Political Analysis (2024), **32**, 329–344
doi:10.1017/pan.2023.41

PA

ARTICLE

Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis

Dae Woong Ham[1], Kosuke Imai[1,2] and Lucas Janson[1]

[1]Department of Statistics, Harvard University, Cambridge, MA, USA. URL: http://lucasjanson.fas.harvard.edu; [2]Department of Government and Statistics, Harvard University, Cambridge, MA, USA. URL: https://imai.fas.harvard.edu/

Corresponding author: Dae Woong Ham; Email: daewoongham@g.harvard.edu

(Received 8 September 2022; revised 5 April 2023; accepted 7 April 2023; published online 8 February 2024)

Abstract

Conjoint analysis is a popular experimental design used to measure multidimensional preferences. Many researchers focus on estimating the average marginal effects of each factor while averaging over the other factors. Although this allows for straightforward design-based estimation, the results critically depend on the ways in which factors interact with one another. An alternative model-based approach can compute various quantities of interest, but requires correct model specifications, a challenging task for conjoint analysis with many factors. We propose a new hypothesis testing approach based on the conditional randomization test (CRT) to answer the most fundamental question of conjoint analysis: Does a factor of interest matter *in any way* given the other factors? Although it only provides a formal test of these binary questions, the CRT is solely based on the randomization of factors, and hence requires no modeling assumption. This means that the CRT can provide a powerful and assumption-free statistical test by enabling the use of any test statistic, including those based on complex machine learning algorithms. We also show how to test commonly used regularity assumptions. Finally, we apply the proposed methodology to conjoint analysis of immigration preferences. An open-source software package is available for implementing the proposed methodology. The proposed methodology is implemented via an open-source software R package *CRTConjoint*, available through the Comprehensive R Archive Network https://cran.r-project.org/web/packages/CRTConjoint/index.html.

Keywords: factorial design; heterogeneous treatment effect; causal interactions; design-based inference

Edited by: Jeff Gill

1. Introduction

Conjoint analysis, introduced more than half a century ago (Luce and Tukey 1964), is a factorial survey-based experiment designed to measure preferences on a multidimensional scale. It has been extensively used by marketing firms to determine desirable product characteristics (e.g., Bodog and Florian 2012; Green, Krieger, and Wind 2001). Recently, conjoint analysis has gained popularity among social scientists (Hainmueller, Hopkins, and Yamamoto 2014; Raghavarao, Wiley, and Chitturi 2010) who are interested in studying individual preferences concerning elections (e.g., Ono and Burden 2018), immigration (e.g., Hainmueller and Hopkins 2015), employment (e.g., Popovic, Kuzmanovic, and Martic 2012), and other issues.

When analyzing conjoint experiments, the *design-based* approach, pioneered by Hainmueller *et al.* (2014), has been the most popular among social scientists. The main advantage of this nonparametric approach is its simplicity—it uses the difference-in-means estimator or linear regression to infer the average marginal component effect (AMCE) of each factor. However, because the AMCE represents the marginal effect of one factor averaged over all the other factors, it may fail to capture important

© The Author(s), 2024. Published by Cambridge University Press on behalf of The Society for Political Methodology.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

interactions. This is potentially problematic given that practitioners often use a small AMCE to conclude that a factor does not matter (e.g., Hainmueller and Hopkins 2015; Hainmueller *et al.* 2014; Ono and Burden 2018). Although a narrow AMCE-based confidence interval containing zero only implies that the factor has a weak *marginal* effect, it is possible that the same factor has substantial interaction effects.

A possible solution is the *model-based* approach that ranges from traditional parametric regression models (Campbell, Mhlanga, and Lesschaeve 2013; Green and Srinivasan 1990; McFadden 1973) to more recent machine learning (ML) algorithms (Abramson *et al.* 2020; de la Cuesta, Egami, and Imai 2022; Egami and Imai 2019; Goplerud, Imai, and Pashley 2022). In conjoint analysis, however, there exist a large number of potential interaction effects. Thus, the model-based approach often assumes the absence of certain interaction terms or uses regularization, yielding possible misspecification or regularization bias. While subgroup analysis, a common practice to analyze only a subset of the data, is simpler, it suffers from the well-known problem of multiple testing, which is of serious concern in conjoint analysis given the large number of possible causal effects of interest (see also Shiraito and Liu, 2022 for a discussion of multiple testing problems in conjoint analysis). Finally, the use of ML algorithms, which is becoming increasingly common, cannot yield even consistent estimates in high-dimensional settings without strong assumptions.

In this paper, we propose a new approach to analyzing data from conjoint analysis that combines the strengths of the existing design-based and model-based approaches (Section 3). Specifically, we show how to conduct assumption-free hypothesis testing based on the conditional randomization test (CRT; Candès *et al.* 2018). In the causal inference literature, the CRT has been used to test interference between units (Aronow 2012; Athey, Eckles, and Imbens 2018). Instead of estimating a particular causal effect, we ask the most fundamental question of conjoint analysis: Does a factor of interest matter *in any way* given the other factors? In many conjoint analyses, researchers are interested in investigating this binary question regarding a specific factor (e.g., country effects in immigration preferences [Hainmueller and Hopkins 2015] and gender effects in candidate evaluation [Ono and Burden 2018]). The proposed approach answers this question with greater statistical power than the AMCE by utilizing flexible ML algorithms but without making *any* assumption about the underlying causal structure. Despite its flexibility, the CRT has an attractive statistical property that the resulting p-values are *exactly* valid regardless of the sample size or the number of factors.

We also show that the proposed methodology can test the validity of assumptions commonly invoked in conjoint analysis (Hainmueller *et al.* 2014). They include the assumptions of no profile order effect, no carryover effect, and no fatigue effect (Bansak *et al.* 2018, 2019). Thus, the proposed hypothesis testing approach can serve as the first step of analyzing conjoint data without assumptions, complementing existing approaches that estimate causal quantities of interest.

For empirical illustration, we apply the proposed methodology to a conjoint analysis of immigration preferences among U.S. citizens (Section 4). While some researchers contend that U.S. citizens generally prefer high-skilled immigrants regardless of their countries of origin, others have suggested that racially prejudiced respondents discriminate against non-European immigrants (Hainmueller and Hopkins 2015; Newman and Malhotra 2019). By combining ML algorithms with the CRT, we find that respondents do differentiate according to whether immigrants are from Mexico or European countries. In addition, we conduct simulations studies whose results are reported in the Supplementary Material.

2. Motivation

In this section, we briefly describe a motivating empirical application concerning the role of ethnocentrism in immigration preferences. In Appendix G of the Supplementary Material, we present an additional application about the role of gender discrimination in political candidate evaluations. We also discuss the limitations of the commonly used approach based on the AMCE that motivate the proposed methodology. In Section 4, we revisit the application and apply our hypothesis testing approach.

2.1. Role of Country of Origin in Immigration Preference

Immigration is one of the most contentious issues in the United States today. In an influential study, Hainmueller and Hopkins (2015) use a conjoint analysis to empirically examine the immigrant

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

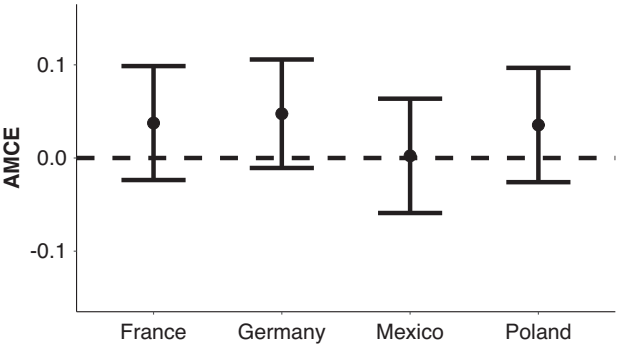


Figure 1. The estimated average marginal component effects (AMCEs) of immigrants' countries of origin in the Hainmueller and Hopkins (2015) study. The plot shows the estimated AMCEs for *France*, *Germany*, *Mexico*, and *Poland*, which represent the average differences in the estimated probability of choosing an immigrant profile with a specific level of the "country of origin" factor, marginalizing other attributes. The baseline factor level is *India*, and the 95% confidence intervals are also shown.

characteristics favored or disfavored by U.S. citizens. The study used the forced-choice design, in which each respondent was presented with a pair of hypothetical immigrant profiles and asked which immigrant they would "personally prefer to see admitted to the United States." Each of 1,396 respondents rated five pairs of profiles.

An immigrant profile consists of nine factors—prior trips to the United States, reason for application, country of origin, language skills, profession, job experience, employment plans, education level, and gender, each of which has multiple levels (see Table 5 in Appendix H of the Supplementary Material as well as the original article for details). Most factors are independently and uniformly randomized across their levels with the exception of two restrictions to avoid implausible pairs. First, immigrant profiles that list *escape persecution* as the "reason of immigration" can only have *Iraq*, *Sudan*, or *Somalia* as their "country of origin." Second, a high-skill "profession" such as *financial analyst*, *research scientist*, *doctor*, and *computer programmer* is possible only if the "education level" is at least *2 years of college*. This restricted randomization scheme induces dependencies between these factors. The survey also contains information about respondents' age, education, ethnicity, gender, and ethnocentrism. The study contains a random sample of 14,018 profiles.

In this study, Hainmueller and Hopkins estimate the AMCE, which represents the marginal effect of a factor of interest averaging over the other factors. Based on the statistically insignificant estimates for the AMCEs of the "country of origin" factor for Mexico and European countries (reproduced in Figure 1),[1] they conclude that "despite media frames focusing on low-skilled, unauthorized immigration from Mexico, there is little evidence of penalty specific to Mexicans" (539). The authors obtain these estimates by fitting a linear regression model, where the outcome variable indicates whether the profile is selected and the predictors are the nine randomized factors. To account for the restricted randomization, they also include two sets of interaction terms, one between "country of origin" and "reason of immigration" and the other between "profession" and "education level." To obtain the estimated AMCE of *Germany*, for example, Hainmueller and Hopkins take the main effect of *Germany* (the baseline is *India*) and then add it to the average of all the interaction terms between *Germany* and the "reason of immigration" factor. Clustered standard errors are computed by clustering on each respondent to account for dependency within a respondent.

Despite this overall finding, the AMCE-based approach may mask relevant interactions and heterogeneous treatment effects. Indeed, Hainmueller and Hopkins conduct a subgroup analysis and find that the "country of origin" factor has statistically significant interactions with the respondents'

[1] All data and results throughout this paper are publicly available at https://doi.org/10.7910/DVN/ENI8GF (Ham, Imai, and Janson 2023).

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

ethnocentrism. They measure ethnocentrism using the feeling thermometer score (ranging from 0 to 100) for the respondent's own groups minus the average feeling thermometer across the other groups. In addition, Newman and Malhotra (2019) reanalyze the same dataset and estimate three-way interactions among respondents' ethnocentrism, "country of origin," and "profession." The authors compute the AMCEs of high-skilled immigrants (baseline of *janitor*) separately for each country of origin and respondent's ethnocentric group. They find that these AMCEs are different between Mexican and European immigrants when compared among highly ethnocentric respondents (see Figure 1 in Newman and Malhotra 2019 for further details).

In this paper, we apply the proposed hypothesis testing approach to testing whether or not immigrants from Mexico and those from Europe are viewed differently in *any way* while controlling for all the other experimental factors as well as the respondent characteristics. The rejection of this null hypothesis would mean that the country of origin of an individual plays a statistically significant role in some U.S. citizens' preferences about that individual's immigration to the United States.

2.2. Limitations of Existing Approaches

Although the AMCE is a useful causal quantity of interest and can be easily and reliably estimated, it is not free of limitations. The AMCE is a marginal effect based on two types of averaging: (1) averaging over the distribution of other attributes and (2) averaging over the responses (and hence respondents). Recall that in the standard causal inference setting with a binary treatment, a zero average treatment effect does not necessarily imply zero treatment effect for everyone. The treatment may benefit some and harm others, and these positive and negative effects can cancel out through averaging. The AMCE suffers from a similar problem, potentially masking important causal heterogeneity if there are interactions among attributes and/or between attributes and respondent characteristics.

Additionally, although an AMCE-based confidence interval that does not contain zero represents evidence that the factor matters, a narrow AMCE-based confidence interval that contains zero only implies that a factor has a weak *marginal* effect. Nevertheless, practitioners tend to use narrow AMCE-based confidence intervals that contain zero to conclude that a factor does not matter. For example, Hainmueller *et al.* (2014) conclude that the "candidates' income does not matter much" and that the "candidates' racial and ethnic backgrounds are even less influential," based on the AMCE-based confidence interval for income and ethnicity (19). Ono and Burden (2018) also claim that "the bias against female candidates [...] is limited to presidential rather than congressional elections" (585) after obtaining a statistically insignificant AMCE estimate for gender among congressional candidates. Lastly, Spilker, Bernauer, and Umaña (2016) interpret statistically insignificant country effects to conclude that "in spite of different national contexts, individuals in [Costa Rica, Nicargua, and Vietnam] hold similar preferences regarding potential [Preferential Trade Agreement] partners" (712). In all cases, the lack of a significant AMCE estimate is not sufficient to conclude that the factor of interest does not matter. Instead, it only implies that the factor may have little impact "on average." Furthermore, although the AMCE can be generalized to account for interactions, it requires researchers to choose a specific interaction term out of many possible interactions and may lead to the problems similar to those of subgroup analysis, including multiple testing.

Although the AMCE is popular, there also exist model-based approaches to flexibly estimate potentially any quantity of interest. In particular, logistic regression remains a popular model-based alternative in conjoint analysis (Campbell *et al.* 2013; Green and Srinivasan 1990; McFadden 1973) especially in marketing research. Although a hierarchical modeling approach remains another popular model-based alternative in conjoint studies, we do not consider it here because it is based on a Bayesian framework rather than frequentist approach taken in this paper (Andrews, Ansari, and Currim 2002). Model misspecification, however, remains a significant challenge. Although researchers may add more interactions to account for all possible effects, such an approach can reduce statistical power and more importantly lead to invalid p-values (Candès and Sur 2018). We show in Figure 9 in Appendix J of the Supplementary Material that using logistic regression and accounting for all two-way interactions to reduce model misspecification can easily lead to invalid p-values.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

Consequently, a consensus among researchers has emerged that flexible ML algorithms are necessary for capturing these causal interactions (Abramson *et al.* 2020; Bansak *et al.* 2020; de la Cuesta *et al.* 2022; Goplerud *et al.* 2022). Yet ML algorithms, despite their flexibility, cannot yield consistent estimates in high-dimensional settings without strong assumptions. In addition, statistical inference in small samples remains a challenge (Chernozhukov *et al.* 2017; Dezeure *et al.* 2015; Imai and Li 2021). Our goal is to address these problems through an assumption-free approach based on the CRT.

3. The Proposed Methodology

In this section, we describe the proposed methodology based on the CRT.

3.1. Notation and Setup

For concreteness, we focus on the forced-choice conjoint design, under which a respondent is asked to choose one of two profiles. Our methodology is general and can be extended to other designs. Let n be the total number of respondents. As is often done in practice, suppose that each respondent evaluates J pairs of profiles, yielding a total of nJ responses (for notational simplicity, we assume the same number of evaluations for each respondent).

We use $Y_{ij} \in \{0,1\}$ to represent the binary outcome variable for evaluation j by respondent i, which equals 1 for selecting the left profile and 0 for choosing the right profile. Although for convenience we use "left" and "right" to distinguish two profiles under each evaluation, the profiles do not necessarily have to be placed side by side on the actual survey platform. We use the following $nJ \times 1$ stacked vector representation for this outcome variable $\mathbf{Y} = [\mathbf{Y}_1; \mathbf{Y}_2; \ldots; \mathbf{Y}_n]$, where $\mathbf{Y}_i = [Y_{i1}; Y_{i2}; \ldots; Y_{iJ}]$ of dimension $J \times 1$ denotes the outcome variable for respondent i. We use $[a_1; a_2; \ldots; a_n]$ to denote a vertical stacking of vectors or matrices $a_1, a_2, \ldots, a_n$. We often observe some characteristics of the respondents, and we use $\mathbf{V}_i$ to denote a $J \times r$-dimensional matrix of r pre-treatment covariates for respondent i that are repeated across J rows.

Next, let p represent the total number of attributes or factors[2] used for each conjoint profile. We use a scalar $X^L_{ij\ell} \in \{1, 2, \ldots, K_\ell\}$ to denote the value of the ℓth factor of interest for evaluation j by respondent i, where the superscript distinguishes the factors for the left (L) and right (R) profiles, and $K_\ell \geq 2$ is the total number of factor levels for factor ℓ. We use $\mathbf{X}^L_{ij} = [X^L_{ij1}; \ldots; X^L_{ijq}]$ to denote a q-dimensional column vector, containing all q factors of interest for the left profile for respondent i in the jth evaluation where $q \leq p$. We define $\mathbf{X}^R_{ij}$ similarly for the right profile. In addition, we use $\mathbf{X}_{ij} = [\mathbf{X}^L_{ij}; \mathbf{X}^R_{ij}]$ as a column vector of length $2q$ to represent the main factors of interest from two profiles together. Lastly, the remaining $(p-q)$ factors are denoted by $\mathbf{Z}_{ij} = [\mathbf{Z}^L_{ij}; \mathbf{Z}^R_{ij}]$, where each term is similarly defined. For example, in the immigration conjoint experiment, if the main factor of interest is "country of origin," the other factors include "education" and "profession."

As done for the outcome variable, we stack all evaluation-specific factors to define respondent-level factor matrices, which are further combined to yield the factor matrix $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2; \ldots; \mathbf{X}_n]$ and $\mathbf{Z} = [\mathbf{Z}_1; \mathbf{Z}_2; \ldots; \mathbf{Z}_n]$ of dimension $nJ \times 2q$ and $nJ \times 2(p-q)$, respectively, where $\mathbf{X}_i = [\mathbf{X}^\top_{i1}; \mathbf{X}^\top_{i2}; \ldots; \mathbf{X}^\top_{iJ}]$ and $\mathbf{Z}_i = [\mathbf{Z}^\top_{i1}; \mathbf{Z}^\top_{i2}; \ldots; \mathbf{Z}^\top_{iJ}]$ are matrices of dimension $J \times 2q$ and $J \times 2(p-q)$, respectively. Lastly, we also stack all respondent characteristics $\mathbf{V} = [\mathbf{V}_1; \mathbf{V}_2; \ldots; \mathbf{V}_n]$ of dimension $nJ \times r$.

Finally, we use $\mathbf{Y}(\mathbf{x}, \mathbf{z})$ to denote the nJ-dimensional vector of the potential outcomes when $\mathbf{X} = \mathbf{x}$ and $\mathbf{Z} = \mathbf{z}$. We avoid the assumption of no interference effect since our vector of potential outcomes is a function of the entire set of treatments $\mathbf{X}$ and $\mathbf{Z}$ (Rubin 1990). We assume a super-population framework, where the potential outcomes $\mathbf{Y}(\mathbf{x}, \mathbf{z})$ are assumed to be drawn from a population of infinite size. In Appendix B of the Supplementary Material, we discuss how our framework is related to a finite-population framework, which is the basis of Fisher's randomization test. In conjoint analysis, the profile attributes are randomized according to a known distribution, $P(\mathbf{X}, \mathbf{Z})$, and we allow for

[2] Throughout the paper, we use "factors" and "attributes" interchangeably.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

any randomization distribution. In general, the randomization of the factors implies the following independence relation:

$$\mathbf{Y}(\mathbf{x},\mathbf{z}) \perp\!\!\!\perp (\mathbf{X},\mathbf{Z}) \quad \text{for all } \mathbf{x} \in \mathcal{X}, \text{ and } \mathbf{z} \in \mathcal{Z}, \tag{1}$$

where we use $\mathcal{X}$ and $\mathcal{Z}$ to represent the support of $\mathbf{X}$ and that of $\mathbf{Z}$, respectively, and $\mathbf{x}$ and $\mathbf{z}$ to denote the value of $\mathbf{X}$ and that of $\mathbf{Z}$, respectively (see Chapter 3.6 of Imbens and Rubin 2015).

3.2. The Conditional Randomization Test

The CRT is an assumption-free approach that can combine design-based inference with flexible ML algorithms. For ease of presentation, we first introduce the CRT without incorporating the respondent characteristics $\mathbf{V}$ and then return in Section 3.6 to show how $\mathbf{V}$ can be incorporated in all the proposed methods. The CRT allows us to examine whether the factors of interest $\mathbf{X}$ change the response $\mathbf{Y}$ while holding the other factors $\mathbf{Z}$ constant. Specifically, we test the following null hypothesis:

$$H_0 : \mathbf{Y}(\mathbf{x},\mathbf{z}) \overset{d}{=} \mathbf{Y}(\mathbf{x}',\mathbf{z}) \quad \text{for all } \mathbf{x},\mathbf{x}' \in \mathcal{X}, \text{ and } \mathbf{z} \in \mathcal{Z}, \tag{2}$$

where we use $\overset{d}{=}$ to denote distributional equality. As a reminder, H_0 states that our entire *vector* of potential outcomes are equal in distribution for any values of $\mathbf{X}$. Our alternative hypothesis states that $\mathbf{X}$ affects $\mathbf{Y}$ in some way while keeping $\mathbf{Z}$ unchanged. This is formalized as

$$H_1 : \mathbf{Y}(\mathbf{x},\mathbf{z}) \overset{d}{\neq} \mathbf{Y}(\mathbf{x}',\mathbf{z}) \quad \text{for some } \mathbf{x},\mathbf{x}' \in \mathcal{X}, \text{ and } \mathbf{z} \in \mathcal{Z}. \tag{3}$$

We emphasize that the null hypothesis defined in Equation (2) implies the absence of any causal effects involving the main factor(s) of interest. For example, the null hypothesis is false if $\mathbf{X}$ affects $\mathbf{Y}$ for *any* individual respondent or subgroup of respondents. Similarly, the null hypothesis does not hold if $\mathbf{X}$ influences $\mathbf{Y}$ only when $\mathbf{Z}$ takes a certain set of values. Thus, the null hypothesis precludes any heterogeneous or interaction effects between the selected factors of interest X and other factors Z included in the experiment.

Contrast this hypothesis test formulation with that of the standard AMCE-based analysis, which asks whether each factor of interest $\mathbf{X}_i$ matters *on average*. More specifically, Hainmueller, Hopkins, and Yamamoto assume that each individual's potential outcome is only a function of its own profile task, that is, $Y_{ij}(\mathbf{X},\mathbf{Z}) = Y_{ij}(\mathbf{X}_{ij},\mathbf{Z}_{ij})$, and computes the marginal importance of $\mathbf{X}_i$ by averaging each individual potential outcome over $\mathbf{Z}_i$ as well as the respondents, which are assumed to be exchangeable, leading to the following null hypothesis:

$$H_0^{\text{AMCE}} : \mathbb{E}\{Y_{ij}(\mathbf{x},\mathbf{Z}_{ij})\} \;=\; \mathbb{E}\{Y_{ij}(\tilde{\mathbf{x}},\mathbf{Z}_{ij})\}, \tag{4}$$

where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the specified values of the main factors and the expectation is taken over $\mathbf{Z}_{ij}$ (other factors) and the respondents. As briefly explained in Section 2.2, the limitation of the AMCE-based approach is that averaging over other factors can mask important causal interaction and heterogeneity.

We now establish the equivalence between the null hypothesis about the potential outcomes defined in Equation (2) and the conditional independence relation among observed variables. This result allows us to use the CRT, which is a general assumption-free methodology for testing conditional independence relations in designed experiments (Candès *et al.* 2018). We state this result as the following theorem whose proof is given in Appendix A of the Supplementary Material.

Theorem 3.1 (Equivalence). *The null hypothesis defined in Equation (2) is equivalent to the following conditional independence hypothesis under the randomization assumption of Equation (1):*

$$H_0^{CRT} : \mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}.$$

The CRT produces exact p-values without asymptotic approximation while enabling the use of any test statistic, including ones based on complex ML algorithms, without making any modeling assumptions. In the conjoint analysis literature, researchers have used traditional regression modeling

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

Algorithm 1: Conditional Randomization Test (CRT)

Input: Data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, test statistic $T(\mathbf{x}, \mathbf{y}, \mathbf{z})$, total number of re-samples B, conditional distribution $\mathbf{X} \mid \mathbf{Z}$;

for $b = 1, 2, \ldots, B$ **do**

  Sample $\mathbf{X}^{(b)}$ from the distribution of $\mathbf{X} \mid \mathbf{Z}$ conditionally independently of $\mathbf{X}$ and $\mathbf{Y}$;

Output: p-value $:= \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{X}^{(b)}, \mathbf{Y}, \mathbf{Z}) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\}\right]$[3]

Table 1. The p-values based on the conditional randomization test (CRT) and the average marginal component effect (AMCE) estimation. The first p-values are from the HierNet-based CRT and AMCE-based test statistics, testing whether the immigrant's "country of origin" (*Mexico* or *Europe*) matters for immigration preferences. The other three p-values are from the HierNet-based CRT test statistics, testing the regularity conditions commonly used for conjoint analysis.

	CRT	AMCE	Profile order effect	Carryover effect	Fatigue effect
p-values	0.042	0.27	0.80	0.12	0.45

(e.g., Barone, Lombardo, and Tarantino 2007; Hauber *et al.* 2016; McFadden 1973) and more recently modern ML algorithms (e.g., Abramson *et al.* 2020; Egami and Imai 2019). However, the validity of these analyses critically depends on modeling assumptions, parameter tuning, and/or asymptotic approximation. In contrast, the CRT assumes nothing about the conditional distribution of the outcome $\mathbf{Y}$ given $(\mathbf{X}, \mathbf{Z})$. Indeed, it does not even require the data to be independently or identically distributed, a property which we use later to test carryover and profile order effects. The only requirement is the specification of the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$, which is readily available from the experimental design of conjoint analysis. Although the power of the CRT critically depends on the test statistic, the CRT always controls type 1 error no matter what the true model is. This contrasts with other model-based approaches that require modeling assumptions to be valid (see Appendix J of the Supplementary Material for more details).

Algorithm 1 summarizes the general procedure used to compute the exact p-value for the CRT. Note that if $\mathbf{X}$ and $\mathbf{Z}$ are independently randomized, as is often the case, one can simply sample $\mathbf{X}^{(b)}$ from the marginal distribution of $\mathbf{X}$. If, on the other hand, certain combinations of $\mathbf{X}$ and $\mathbf{Z}$ values (e.g., doctor without a college degree in the immigration conjoint experiment) are excluded, then we must use the appropriate conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$. Critically, Algorithm 1 is valid for complicated experimental designs, so long as one can sample from the conditional distribution $\mathbf{X}$ given $\mathbf{Z}$.

The CRT can be computationally intensive since it requires computing the test statistic T a total of $B + 1$ times. However, these computations can easily be parallelized. Furthermore, recent works (Liu *et al.* 2020; Tansey *et al.* 2018) have shown that certain test statistic constructions also alleviate the need for these computations. In Appendix I of the Supplementary Material, we detail several tricks that can be used to dramatically reduce the computation time when implementing the CRT. For the main application results presented in Section 4 (first column of Table 1), we note that the parallelized computational time was approximately 6 minutes with 50 cores to calculate each p-value with $B = 2{,}000$. Our software package makes it easy for practitioners to use multiple cores and provides a step-by-step instruction for using many cores on Amazon Web Services.[4]

The p-value of the CRT is valid[5] regardless of sample size and test statistic (Candès *et al.* 2018). This is especially attractive because the number of second-order interactions are comparable or even larger than

[3] We add one to the numerator and denominator so that the distribution of the p-value is stochastically dominated by the uniform distribution as suggested by Candès *et al.* (2018).

[4] The detailed instructions and example use cases can be found in a vignette of our open-source software package at https://cran.r-project.org/web/packages/CRTConjoint/vignettes/CRTConjoint.html.

[5] That is, under H_0, $P(p\text{-value} \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

the sample size in typical conjoint studies. To see the validity of the p-values, it suffices to recognize that under the null hypothesis, all $B+1$ test statistics, $T(\mathbf{X},\mathbf{Y},\mathbf{Z})$, $T(\mathbf{X}^{(1)},\mathbf{Y},\mathbf{Z}),\ldots,T(\mathbf{X}^{(B)},\mathbf{Y},\mathbf{Z})$, are exchangeable given $(\mathbf{Y},\mathbf{Z})$. While any test statistic produces a valid p-value under the CRT, the choice of test statistic determines the statistical power. We now turn to this practically important consideration.

3.3. Test Statistics

Although H_0 is logically false, if there is any difference in the potential outcome distribution, the potential outcome distribution for binary random variables is completely characterized by its conditional mean and marginal mean. Therefore, in this setting, both the CRT and AMCE test hypotheses related to the marginal or conditional means. To obtain a powerful test statistic that does not mask important interactions, we consider a test statistic based on the Lasso logistic regression with hierarchical interactions, or HierNet (Bien, Taylor, and Tibshirani 2013). Specifically, HierNet constrains the two-way interaction effects to be smaller in magnitude than their corresponding main effects. For example, this implies that a two-way interaction effect will be set to zero if its relevant main effects are all zero. Although this constraint may not actually hold in practice, a stricter regularization on the interactions is desirable. This is because the space of possible two-way interactions is large and grows quadratically, and many of them are expected to be indistinguishable from zero. Thus, we view the hierarchical sparsity constraint as an important tool that may lead to greater power when interactions are weak or do not exist. Although the practical implementation of the HierNet leverages the sparsity constraint, we emphasize that the validity of the CRT does not depend on the appropriateness of this constraint. When fitting HierNet, we use the dummy variable encoding (i.e., each factor level is represented by its own dummy variable) but do not omit the baseline level. We can fit this overparameterized model because of the regularization of HierNet. The primary advantage of this approach is that the results are no longer dependent on the choice of baseline levels (Egami and Imai 2019). Finally, we use the publicly available *HierNet* package in R, where the response is the nJ-dimensional $\mathbf{Y}$ and the design matrix is the $nJ \times 2p$ dimensional features $(\mathbf{X},\mathbf{Z})$ for both the left and right profiles.

We begin by considering the simplest case where we have a single main factor of interest $\mathbf{X}$ ($q = 1$). Without loss of generality, we assume that this is the first factor among the total of p factors. There are two types of interaction effects to consider (de la Cuesta *et al.* 2022). First, a within-profile interaction effect represents the interaction between one level of the main factor and another level of a different factor within the same profile. Second, a between-profile interaction effect represents the factor interaction between two profiles (left vs. right) that are being compared under the forced choice design.

$$T_{\text{HierNet}} = \underbrace{\sum_{k=1}^{K_1}(\hat{\beta}_k - \bar{\beta})^2}_{\text{main effects}} + \underbrace{\sum_{\ell=2}^{p}\sum_{k=1}^{K_1}\sum_{k'=1}^{K_\ell}(\hat{\gamma}_{1\ell kk'} - \bar{\gamma}_{1\ell k'})^2}_{\text{within-profile interaction effects}} + \underbrace{\sum_{\ell=1}^{p}\sum_{k=1}^{K_1}\sum_{k'=1}^{K_\ell}(\hat{\delta}_{1\ell kk'} - \bar{\delta}_{1\ell k'})^2}_{\text{between-profile interaction effects}}, \tag{5}$$

where $\hat{\beta}_k$ is the estimated main effect coefficient for the kth level of our factor of interest $\mathbf{X}$ with $\bar{\beta}$ denoting the average of these estimated main effect coefficients, and $\hat{\gamma}_{1\ell kk'}$ and $\hat{\delta}_{1\ell kk'}$ represent the estimated within-profile and between-profile interaction effect coefficients between the kth level of the factor of interest $\mathbf{X}$ and the k'th level of the ℓth factor, respectively. Similar to the main effects, $\bar{\gamma}_{1\ell k'}$ and $\bar{\delta}_{1\ell k'}$ denote the averages of their corresponding estimated interaction effect coefficients. We do not consider third- or higher-order interactions because of the typical sample size in a conjoint experiment and a lack of powerful methods to detect such interactions. However, in Appendix G of the Supplementary Material, we illustrate how to incorporate third-order interactions when prior substantive knowledge is available.

This test statistic can be easily generalized to the setting where there is more than one factor of interest ($q > 1$). In such a case, we simply compute Equation (5) for each factor of interest, and then sum the resulting values to arrive at the final test statistic. T_{HierNet} aims to capture any differential effects the levels of $\mathbf{X}$ have on the response through their main effects and relevant interaction effects.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

For example, suppose that $\mathbf{X}$ is "country of origin" in the immigration conjoint experiment (Section 2.1). Under H_0, we would expect all main effects and any interaction effects of *Mexico* and *Germany* to be roughly equivalent, thus making T_{HierNet} close to zero. However, suppose that immigrants from *Germany* with a certain "education level" were favored more than those from *Mexico* with a certain "education level." Then, we would expect these interactions to differ, making T_{HierNet} further from zero.

We use cross-validation[6] to choose the value of HierNet's tuning parameter, which controls the degree of regularization. In addition, when the sample size and the number of factors are large, fitting HierNet can be computationally demanding. To alleviate this issue, we propose computational speedups of the HierNet test statistics, which are detailed in Appendix I of the Supplementary Material. In particular, we drop $\mathbf{X}$ when fitting the HierNet tuning parameter via cross-validation. Since this computationally expensive step does not depend on $\mathbf{X}$, we do not need to re-run it for each $\mathbf{X}^b$.

So far, we have constructed our test statistic as if there is no profile order effect. This implies that the effects of each factor do not depend on whether it belongs to the left or right profile. Formally, we have imposed the following symmetry constraints in our HierNet test statistic:

$$\hat{\beta}_k = \hat{\beta}_k^L = -\hat{\beta}_k^R, \quad \hat{\gamma}_{1\ell kk'} = \hat{\gamma}_{\ell\ell' kk'}^L = -\hat{\gamma}_{\ell\ell' kk'}^R, \quad \hat{\delta}_{\ell\ell' kk'} = -\hat{\delta}_{\ell'\ell k'k}, \tag{6}$$

where the superscripts L and R denote the left and right profile effects, respectively. $\hat{\delta}_{\ell\ell' kk'}$ denotes the between profile interaction between the kth level of factor ℓ in the left profile with the k'th level of factor ℓ' in the right profile.[7] The signs of the estimated coefficients reflect the fact that the response variable $\mathbf{Y}$ is recorded as 1 if the left profile is chosen and as 0 if the right profile is selected. These constraints reduce the dimension of parameters to be estimated by half.

Importantly, the validity of the proposed tests does not depend on whether the assumption of no profile order effect holds. Through simulations, Figure 6 in Appendix E of the Supplementary Material shows that these constraints can significantly increase statistical power when there is no profile order effect. Appendix D.2 of the Supplementary Material also presents simulations that show a substantial power gain from using the HierNet-based CRT test statistic compared to using the AMCE-based test statistic. To incorporate this symmetry constraint, we append another copy of the dataset below the original dataset, where the appended copy is identical to the original dataset except that the order of left and right profiles is flipped and the response variable is transformed as $\mathbf{1} - \mathbf{Y}$ before fitting HierNet (see Appendix E of the Supplementary Material for details). In Section 3.5, we show how to use the CRT for testing the validity of the assumption of no profile order effect.

Because the validity of the CRT does not depend on modeling assumptions, one can incorporate a variety of assumptions into test statistics. In general, test statistics have a greater statistical power if the assumptions hold in the true (unknown) data generating process. Therefore, as much as possible the choice of test statistic should reflect researchers' substantive knowledge. Appendix G of the Supplementary Material presents an empirical example of leveraging substantive knowledge in the test statistic.

3.4. Generalization of the Null Hypothesis and Test Statistic

Researchers are often interested in testing only a few levels of interest as opposed to testing the whole factor. Yet, simply dropping the observations that correspond to those irrelevant factor levels can lead to a loss of statistical power. An advantage of the formulation described below is that we can retain all observations including those whose factor levels are irrelevant, which can improve statistical power.

[6] The CRT remains valid if used with cross-validation so long as the resampled test statistics based on $\mathbf{X}^b$ similarly uses cross-validation. This ensures exchangeability.

[7] Equation (6) also implies that the between-profile interactions in Equation (5) for the same factor obey $\hat{\delta}_{\ell\ell kk'} = -\hat{\delta}_{\ell\ell k'k}$ for any factor ℓ and levels k, k'. In particular, this implies that $\hat{\delta}_{\ell\ell kk} = 0$, i.e., between-profile interactions of the same factor and same level are zero, while $\hat{\delta}_{\ell\ell kk'}$ are counted twice in Equation (5), i.e., between-profile interactions of the same factor and levels k and k' are counted twice in Equation (5).

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

For example, suppose we are interested in how respondents differentiate immigrants from *Mexico* and *Germany*. If the way in which respondents differentiate between immigrants from *Mexico* and those from *China* is different from how they distinguish between immigrants from *Germany* and those from *China*, then this implies that the respondents are viewing immigrants from *Mexico* differently than those from *Germany*. Therefore, detecting any differences for even the irrelevant levels may help improve the statistical power.

Here, we generalize the null hypothesis and test statistic, given in Equations (2) and (5), so that the methodology can accommodate any combinations of factor levels. We introduce a coarsening function h that groups factor levels of interest while assigning other factor levels to themselves. Formally, this coarsening function is defined as $h : \mathcal{X} \mapsto \widetilde{\mathcal{X}}$, where $|\mathcal{X}| \geq |\widetilde{\mathcal{X}}|$. Thus, for our aforementioned immigration example, h will assign the same value to immigrants from *Mexico* and *Germany* while leaving all other combinations mapped to different values.

Under this setup, we can test the null hypothesis that specific levels within $\mathbf{X}$ do not affect the potential outcome in any way. Formally,

$$H_0^{\text{General}} : \mathbf{Y}(\mathbf{x}, \mathbf{z}) \overset{d}{=} \mathbf{Y}(\mathbf{x}', \mathbf{z}) \text{ for all } \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \text{ such that } h(\mathbf{x}) = h(\mathbf{x}') \text{ and } \mathbf{z} \in \mathcal{Z}. \tag{7}$$

The condition $h(\mathbf{x}) = h(\mathbf{x}')$ enables the comparison of the factor levels of interest alone. Additionally, H_0 is a special case of H_0^{General} when the coarsening function h is the identity function. Finally, applying the same argument as the one used to prove Theorem 3.1, it can be shown that H_0^{General} is equivalent to the following conditional independence relation:

$$\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid h(\mathbf{X}), \mathbf{Z}. \tag{8}$$

To test this null hypothesis, we first fit the same HierNet with the main and two-way interaction effects. To incorporate the coarsening function h, our test statistic takes the same form as the one given in Equation (5) but is based only on the estimated coefficients that correspond to the factor levels of the group induced by the h function, that is, *Mexico* and *Germany* in the above example. Appendix F.1 of the Supplementary Material contains further details of testing the general null hypothesis and the corresponding CRT algorithm. In Section 4, we also provide an example of applying H_0^{General} that contains more details on this test statistic.

Finally, Appendix C of the Supplementary Material details how to further generalize H_0^{General} when a researcher is interested in grouping factor levels, that is, combining levels *France*, *Germany*, and *Poland* into one level *Europe* when testing H_0^{General}. Although coarsening via h also involved "grouping" levels, the grouping described in Appendix C of the Supplementary Material aggregates factor levels to allow comparison between higher-level categories, while the coarsening function h allows us to focus our hypothesis test only on differences between a subset of factor levels of interest.

3.5. Testing the Regularity Assumptions of Conjoint Analysis

To further demonstrate the flexibility of the CRT, we also show how to use the CRT for testing the validity of several commonly made assumptions of conjoint analysis. Since we are interested in not rejecting the null hypothesis for the hypotheses presented in this section, we propose test statistics that are designed to be reasonably powerful for general settings in conjoint analysis.

3.5.1. Profile Order Effect

The assumption of no profile order effect states that changing the order of profiles, that is, left versus right, does not affect the actual profile chosen (since the value of $\mathbf{Y}$ corresponds to whether the left or right profile is chosen, $\mathbf{Y}$ should be recoded as $\mathbf{1} - \mathbf{Y}$ when the profile order is changed). We denote the potential outcome $Y_{ij}(\mathbf{x}_{ij}^L, \mathbf{x}_{ij}^R, \mathbf{z}_{ij}^L, \mathbf{z}_{ij}^R)$, which is now a function of left and right profiles. Although not necessary, we assume here no interference between responses for notational clarity (see Appendix F.2 of the Supplementary Material for the general case). Lastly, we use $\mathcal{X}_{\text{ind}}$ and $\mathcal{Z}_{\text{ind}}$ to denote the support of

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

$[\mathbf{x}_{ij}^L; \mathbf{x}_{ij}^R]$ and that of $[\mathbf{z}_{ij}^L; \mathbf{z}_{ij}^R]$, respectively, representing the support of factors used in each individual's evaluation (hence "ind" in the subscript).

We formally state the assumption of no profile order effect as the following null hypothesis that reordering of the left and right profiles has no effect on the adjusted response:

$$H_0^{\text{Order}} : Y_{ij}(\mathbf{x}_{ij}^L, \mathbf{x}_{ij}^R, \mathbf{z}_{ij}^L, \mathbf{z}_{ij}^R) \overset{d}{=} 1 - Y_{ij}(\mathbf{x}_{ij}^R, \mathbf{x}_{ij}^L, \mathbf{z}_{ij}^R, \mathbf{z}_{ij}^L), \text{ for all } i, j, [\mathbf{x}_{ij}^L; \mathbf{x}_{ij}^R] \in \mathcal{X}_{\text{ind}}, \ [\mathbf{z}_{ij}^L; \mathbf{z}_{ij}^R] \in \mathcal{Z}_{\text{ind}}.$$

We modify the HierNet test statistic in Equation (5) with the same HierNet fit on $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ but without enforcing the constraints in Equation (6) as

$$T_{\text{HierNet}}^{\text{Order}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \sum_{\ell=1}^{p} \sum_{k=1}^{K_\ell} \left(\hat{\beta}_{\ell k}^L + \hat{\beta}_{\ell k}^R \right)^2 + \sum_{\ell=1}^{p} \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^{p} \sum_{k=1}^{K_\ell} \sum_{k'=1}^{K_{\ell'}} \left(\hat{\gamma}_{\ell\ell' kk'}^L + \hat{\gamma}_{\ell\ell' kk'}^R \right)^2$$
$$+ \sum_{\ell=1}^{p} \sum_{\ell'=1}^{p} \sum_{k=1}^{K_\ell} \sum_{k'=1}^{K_{\ell'}} \left(\hat{\delta}_{\ell\ell' kk'} + \hat{\delta}_{\ell'\ell k'k} \right)^2.$$

Since the symmetry constraints given in Equation (6) must hold under H_0^{Order}, a large value of this test statistic indicates a potential violation of the null hypothesis. To conduct the CRT for testing H_0^{Order}, we resample and recompute our test statistics. Appendix F.2 of the Supplementary Material provides details about the testing procedure.

3.5.2. Carryover Effect

Researchers also often rely on the assumption of no carryover effect (Hainmueller *et al.* 2014). The assumption states that the order of the J evaluations each respondent performs has no effect on the outcomes. This assumption is violated, for example, if respondents use information from their previous evaluations when assessing a given pair of profiles. To test this carryover effect, we assume no interference across respondents but consider potential interference across evaluations within each respondent.

Let $\mathbf{x}_{i,1:(j-1)}$ represent all the profile attributes that were presented to respondent i from the first evaluation to the $(j-1)$th evaluation. Then, the potential outcome can be written as a function of both current and previous profiles, that is, $Y_{ij}\left(\mathbf{x}_{i,1:(j-1)}, \mathbf{z}_{i,1:(j-1)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}\right)$ for $j \geq 2$, where we assume no interference between respondents but allow interference within a respondent. Our null hypothesis is that, for a given evaluation $j \geq 2$, the response Y_{ij} is independent of all the previous profiles conditional on the current profiles:

$$H_0^{\text{Carryover}} : Y_{ij}\left(\mathbf{x}_{i,1:(j-1)}, \mathbf{z}_{i,1:(j-1)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}\right) \overset{d}{=} Y_{ij}\left(\mathbf{x}'_{i,1:(j-1)}, \mathbf{z}'_{i,1:(j-1)}, \mathbf{x}_{ij}, \mathbf{z}_{ij}\right),$$

where for all $i \geq 1$, $j \geq 2$, $\mathbf{x}_{i,1:(j-1)}, \mathbf{x}'_{i,1:(j-1)} \in \mathcal{X}_{\text{ind}}^{j-1}$, $\mathbf{z}_{i,1:(j-1)}, \mathbf{z}'_{i,1:(j-1)} \in \mathcal{Z}_{\text{ind}}^{j-1}$, $\mathbf{x}_{ij} \in \mathcal{X}_{\text{ind}}$, $\mathbf{z}_{ij} \in \mathcal{Z}_{\text{ind}}$ with $\mathcal{X}_{\text{ind}}^{j-1}$ and $\mathcal{Z}_{\text{ind}}^{j-1}$ denoting the support of $\mathbf{x}_{i,1:(j-1)}$ and that of $\mathbf{z}_{i,1:(j-1)}$, respectively.

We test this null hypothesis by using a test statistic that targets whether the immediately preceding evaluation affects the current evaluation. We believe targeting the lag-1 effect in the test statistic is reasonable because if a carryover effect exists, respondents are likely to be affected most by the immediately preceding evaluation. For example, if respondents believe that they have placed too much weight on profiles' professions in the previous evaluation, they might decide to rely on the current profiles' "country of origin" factor more than its "profession" factor in order to balance across evaluations. Under this scenario, we would expect a significant interaction between previous profiles' "profession" factor and current profiles' "country of origin" factor.

We modify the test statistic given in Equation (5) in the following way. Suppose that J is even (if J is odd, simply consider $J-1$ evaluations). We first define a new response vector $\mathbf{Y}_i^* = [Y_{i2}; Y_{i4}; \ldots; Y_{iJ}]$ by taking every other evaluation. Similarly, we can define new factors of interest $\mathbf{X}_i^* = [[\mathbf{X}_{i1}; \mathbf{Z}_{i1}]^\top; [\mathbf{X}_{i3}; \mathbf{Z}_{i3}]^\top, \ldots; [\mathbf{X}_{i,J-1}; \mathbf{Z}_{i,J-1}]^\top]$ and a new set of conditioning variables

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

$\mathbf{Z}_i^* = [[\mathbf{X}_{i2};\mathbf{Z}_{i2}]^\top;[\mathbf{X}_{i4};\mathbf{Z}_{i4}]^\top,\ldots;[\mathbf{X}_{iJ};\mathbf{Z}_{iJ}]^\top]$. We then fit HierNet with the new response $\mathbf{Y}^* = [\mathbf{Y}_1^*;\mathbf{Y}_2^*;\ldots;\mathbf{Y}_n^*]$ on $(\mathbf{X}^*,\mathbf{Z}^*)$, where $\mathbf{X}^* = [\mathbf{X}_1^*;\mathbf{X}_2^*;\ldots;\mathbf{X}_n^*]$ and $\mathbf{Z}^*$ is defined similarly.

For this particular scenario, HierNet will estimate all main effects and interaction effects of $\mathbf{Z}^*$, and the interaction effects between $\mathbf{X}^*$ and $\mathbf{Z}^*$, which are of primary interest. To increase the power of the test, we set all main and interaction effects of $\mathbf{X}^*$ to zero since we do not expect the previous profile alone to impact the respondent's choice. Furthermore, we also do not expect the interaction effects between $\mathbf{X}^*$ and $\mathbf{Z}^*$ to differ, depending on the ordering (left vs. right) of the relevant factors. Therefore, we enforce all these interaction effects to have equal magnitude as done in Equation (6) (see Appendix E of the Supplementary Material for further details). This leads to the following test statistic:

$$T_{\text{HierNet}}^{\text{Carryover}}(\mathbf{X}^*,\mathbf{Y}^*,\mathbf{Z}^*) = \sum_{\ell=1}^{p}\sum_{\ell'=1}^{p}\sum_{k=1}^{K_\ell}\sum_{k'=1}^{K_{\ell'}}\hat{\gamma}^2_{\ell\ell' kk'},$$

where $\gamma_{\ell\ell' kk'}$ represents the coefficient of an interaction term between the kth level of the ℓth factor of the profile used in the previous evaluation and the k'th level of the ℓ'th factor of the profile used in the current evaluation. Appendix F.2 of the Supplementary Material explains how to resample the test statistic in this setting.

3.5.3. Fatigue Effect

Researchers may be concerned that a respondent performing a large number of conjoint evaluations may experience the "fatigue effect," resulting in a declining quality of responses. Recently, Bansak *et al.* (2018) conducted an empirical study to examine how the pattern of responses depends on the number of evaluations each respondent performs. Here, we show how to use the CRT to formally test the presence of the fatigue effect.

Similar to the carryover effect, we investigate whether there is a fatigue effect within each respondent's potential outcome $Y_{ij}(\mathbf{x}_{ij},\mathbf{z}_{ij})$, where we again assume no interference effect as done when testing no profile order effect. We test the following null hypothesis that the potential outcome is unaffected if respondent i evaluated the same pair of profiles $(\mathbf{x}_{ij},\mathbf{z}_{ij})$ but at a later or earlier evaluation $j' \neq j$:

$$H_0^{\text{Fatigue}} : Y_{ij}(\mathbf{x}_{ij},\mathbf{z}_{ij}) \overset{d}{=} Y_{ij'}(\mathbf{x}_{ij},\mathbf{z}_{ij}) \text{ for all } i,j,j', \mathbf{x}_{ij} \in \mathcal{X}_{\text{ind}}, \text{ and } \mathbf{z}_{ij} \in \mathcal{Z}_{\text{ind}}.$$

We propose a similar HierNet test statistic that reflects a scenario where respondents will only pay attention to a shrinking number of factors as they rate more profiles. In this case, we would expect interactions between the factors and the evaluation order index $\mathbf{F} = (F_1,F_2,\ldots,F_n)$, which represents an nJ-dimensional integer vector with $\mathbf{F}_i = (1,2,\ldots,J)$ for all $i = 1,2,\ldots,n$. Again, for the sake of statistical power, we impose the absence of profile order effects on HierNet as done in Equation (5). Our proposed test statistic is the following from a HierNet fit of $\mathbf{Y}$ on $(\mathbf{X},\mathbf{Z},\mathbf{F})$:

$$T_{\text{HierNet}}^{\text{Fatigue}}(\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{F}) = \sum_{\ell=1}^{p}\sum_{k=1}^{K_\ell}\hat{\gamma}^2_{\ell k},$$

where $\hat{\gamma}_{\ell k}$ represents the coefficient of an interaction term between $\mathbf{F}$ and level k of factor ℓ. Appendix F.2 of the Supplementary Material shows how to resample and recompute the test statistics to test H_0^{Fatigue}.

3.6. Incorporating Respondent Characteristics

In conjoint experiments, researchers often expect factors of interest to interact strongly with respondent characteristics (Hainmueller and Hopkins 2015; Newman and Malhotra 2019; Ono and Burden 2018). It is possible to exploit this fact when applying the CRT by directly incorporating respondent characteristics, $\mathbf{V}$, into the test statistic. Doing so can substantially increase the statistical power.

We incorporate respondent characteristics in the CRT procedure by appending $\mathbf{V}$ to $\mathbf{Z}$ and holding both $(\mathbf{Z},\mathbf{V})$ constant. Since respondent characteristics are not randomized factors, unlike $\mathbf{Z}$, $\mathbf{V}$ is not guaranteed to be independent of the potential outcomes. We can test the following causal null hypothesis

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

that conditions upon $\mathbf{V}$:

$$H_0 : \mathbf{Y}(\mathbf{x},\mathbf{z}) \stackrel{d}{=} \mathbf{Y}(\mathbf{x}',\mathbf{z}) \mid \mathbf{V} \quad \text{for all } \mathbf{x},\mathbf{x}' \in \mathcal{X}, \text{ and } \mathbf{z} \in \mathcal{Z}. \tag{9}$$

Theorem 3.1 can be easily extended to show that this null hypothesis is equivalent to the conditional independence relation $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z},\mathbf{V}$. Algorithm 1 also stays the same except we sample $\mathbf{X}^b$ from the distribution of $\mathbf{X} \mid (\mathbf{Z},\mathbf{V})$ and our test statistic is now a function of $(\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{V})$.

A major benefit of incorporating respondent characteristics is the ability to capture respondent characteristic interactions into the test statistic. Consequently, we incorporate an additional predictor $\mathbf{V}$ when fitting HierNet and modify the HierNet test statistic in Equation (5) as

$$\begin{aligned} T_{\text{HierNet}} = & \underbrace{\sum_{k=1}^{K_1} (\hat{\beta}_k - \bar{\beta})^2}_{\text{main effects}} + \underbrace{\sum_{\ell=2}^{p} \sum_{k=1}^{K_1} \sum_{k'=1}^{K_\ell} (\hat{\gamma}_{1\ell kk'} - \bar{\gamma}_{1\ell k'})^2}_{\text{within-profile interaction effects}} \\ & + \underbrace{\sum_{\ell=1}^{p} \sum_{k=1}^{K_1} \sum_{k'=1}^{K_\ell} (\hat{\delta}_{1\ell kk'} - \bar{\delta}_{1\ell k'})^2}_{\text{between-profile interaction effects}} + \sum_{m=1}^{r} \sum_{k=1}^{K_1} \sum_{w=1}^{L_m} \underbrace{(\hat{\xi}_{1mkw} - \bar{\xi}_{1mw})^2}_{\text{respondent interaction effects}}, \end{aligned} \tag{10}$$

where $\hat{\xi}_{1mkw}$ represents the interaction between the kth level of our factor of interest $\mathbf{X}$ and the wth level of the mth factor of a respondent characteristic. If a respondent characteristic is a numeric variable, we could either coarsen it into a factor variable or directly include it in the model as a numeric variable.[8] Again, $\bar{\xi}_{1mw}$ denotes the average of these estimated interaction coefficients. Similar to Equation (10), we add the additional constraint that $\hat{\xi}_{\ell mkw} = \hat{\xi}^L_{\ell mkw} = \hat{\xi}^R_{\ell mkw}$, where the superscripts L and R similarly denote the left and right profile effects. Finally, we modify $T^{\text{Order}}_{\text{HierNet}}$ similarly by adding $\sum_{\ell=1}^{p} \sum_{m=1}^{r} \sum_{k=1}^{K_\ell} \sum_{w=1}^{L_m} (\hat{\xi}^L_{\ell mkw} + \hat{\xi}^R_{\ell mkw})^2$ to the original test statistic in $T^{\text{Order}}_{\text{HierNet}}$ to account for respondent characteristic interactions. For the empirical applications in Section 4 and Appendix G of the Supplementary Material, we incorporate $\mathbf{V}$ and use the modified test statistics to test H_0 and the no profile order effect.

4. Empirical Application: Immigration Preferences and Ethnocentrism

In this section, we apply the proposed CRT to the conjoint study introduced in Section 2.1. We begin our analysis of the immigration conjoint experiment by testing whether respondents differentiate between immigrants from Mexico and those from European countries. We use the same dataset as the one used in Hainmueller and Hopkins (2015). This gives us a total sample of 6,980 observations with $n = 1{,}396$ respondents each rating $J = 5$ tasks. Our main factor of interest $\mathbf{X}$ is the "country of origin" variable.

Since we are interested in testing how respondents differentiate Mexican and European candidates, we use the generalized hypothesis H_0^{General} defined in Equation (7) that compares the main and interaction effects only between the factor levels of interest and coarsen the three levels—*Germany*, *France*, and *Poland*—into one level called *Europe* (see Appendix C of the Supplementary Material for a formal treatment). Furthermore, the h function in H_0^{General} takes the "country of origin" variable and maps the relevant levels of *Mexico* and *Europe* to one output and the remaining levels to other unique outputs. We include all the other randomized factors and respondent characteristics as $\mathbf{Z}$ and $\mathbf{V}$, respectively, except the ethnocentrism variable, which is only measured for a subset of respondents. We incorporate this variable at the end of this section.

We fit HierNet using $\mathbf{Y}$ as the response and our main factor $\mathbf{X}$, other randomized factors $\mathbf{Z}$, and respondent characteristics $\mathbf{V}$ as the predictors. We then compute the test statistic given in Equation (10)

[8] HierNet standardizes all variables when performing the fit. Therefore, even if a numeric variable is on a different scale, all estimated coefficients remain comparable.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

while imposing the implied no profile order effect constraints given in Equation (6) (with the constraints applied to the respondent characteristic interactions too). As mentioned briefly in Section 3.4, we slightly modify this test statistic by only using the estimated coefficients for *Mexico* and *Europe* while ignoring the other coefficients.

As shown in Table 1, the CRT p-value of this test statistic is 0.042, providing evidence that respondents differentiate immigrants from *Mexico* and *Europe*. For comparison, we also compute the p-value based on the estimated AMCE of being from *Mexico* compared to being from *Europe*. We apply a commonly used linear regression approach described in Section 2.1 to compute this p-value. Specifically, we first fit a linear regression model using "country of origin," "reason of immigration," and their interaction as predictors to account for the restricted randomization (Hainmueller *et al.* 2014). The standard errors are clustered by respondent. We then use the F-test of the linear equality constraint that implies the null hypothesis under the linear model. The resulting p-value for the difference between *Mexico* and *Europe* is 0.27, which is statistically insignificant.

The above result suggests that the CRT may be able to capture complex interactions and yield greater statistical power than the AMCE-based test. The two largest interactions in the observed test statistic are within-profile interactions between "country of origin" and "education" and between "country of origin" and "prior trips to U.S." factors, which included whether or not the immigrant entered the United States illegally. Thus, we next assess the degree to which the interaction effects account for this difference in statistical power. To do this, we use a Lasso logistic regression[9] without interaction terms where we only include the main effects of $(\mathbf{X}, \mathbf{Z}, \mathbf{V})$. The CRT p-value of using only the relevant levels of the main effects of $\mathbf{X}$ as the test statistic is 0.082, which is somewhat larger than the p-value based on HierNet test statistic. This suggests that interactions play some role in yielding a more powerful test than the AMCE-based approach.

Hainmueller and Hopkins suggest that respondents do not differentiate between immigrants from *Mexico* and those from *Europe* based on the results of their main analysis (see Figure 1, which replicates this analysis). However, they also conduct a subgroup analysis and find that "country of origin" has statistically significant interaction(s) with the respondent's ethnocentrism through a subgroup analysis (see also Newman and Malhotra 2019 for related findings). Thus, we now repeat the same analysis as above except that we include this ethnocentrism variable as an additional respondent characteristic in $\mathbf{V}$. Note that unlike the original analysis, we do not dichotomize this variable and use the original continuous scale. Since ethnocentrism is only measured for white and black respondents, the number of total respondents is reduced to $n = 1{,}135$. Despite this reduction in sample size, the inclusion of the ethnocentrism variable produces the p-value of 0.019, which is smaller than the p-value of the analysis without this variable. As expected, the largest interaction in the observed test statistic involves the ethnocentrism variable. All together, our analysis provides evidence that respondents differentiate immigrants from *Mexico* and *Europe*.

Lastly, we use the CRT to test the three commonly made regularity assumptions of conjoint analysis: no profile order effect, no carryover effect, and no fatigue effect. The last three columns in Table 1 present the p-values from the various tests described in Section 3.5. We find no evidence that these assumptions are violated in the immigration conjoint experiment (the first row). In particular, the fact that we do not detect profile order effects suggests that imposing the symmetry constraint as done in Equation (6) likely improves power.

5. Concluding Remarks

Conjoint analysis is a popular methodology for analyzing multidimensional preferences and decision-making. In this paper, we propose an assumption-free approach for conjoint analysis based on the CRT.

[9]Throughout this paper (including the Supplementary Material), we fit all Lasso logistic regressions using the *glmnet* package in R and fit all logistic regressions using the standard *glm()* function, where both the response and design matrices are equivalent to the original HierNet fit unless otherwise specified.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

The proposed methodology allows researchers to test whether a set of factors of interest matter at all without assuming a statistical model. We also extend the proposed methodology to test for differential effects for any combination of factor levels and other regularity assumptions commonly invoked in conjoint analysis like the profile order effect. Although we acknowledge the CRT only provides a formal test of whether a factor matters or not, such an analysis is of substantive interest. In particular, by incorporating ML algorithms and/or domain knowledge, the CRT can leverage complex interactions to improve its power without making modeling assumptions. As a result, the CRT provides a more powerful test than the AMCE that may mask important interactions. The CRT is easy to implement and provides exact (i.e., non-asymptotic) p-values that are valid even in high dimensions. We believe that this flexibility combined with its assumption-free nature makes the CRT a powerful tool for conjoint analysis. The CRT can complement existing methods like the AMCE analysis by providing a useful way to examine whether a factor of interest matters at all.

Acknowledgments. We thank Naoki Egami for advice.

Funding Statement. Imai thanks the Alfred P. Sloan Foundation for partial support (Grant No. 2020-13946). Ham and Janson were partially supported by a CAREER grant from the National Science Foundation (Grant No. DMS-2045981).

Data Availability Statement. Replication code for this article has been published in Ham *et al.* (2023) at https://doi.org/10.7910/DVN/ENI8GF.

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2023.41.

References

Abramson, S., K. Kocak, A. Magazinnik, and A. Strezhnev. 2020. "Improving Preference Elicitation in Conjoint Designs Using Machine Learning for Heterogeneous Effects." Technical report, The Annual Summer Meeting of the Society for Political Methodology.

Andrews, R. L., A. Ansari, and I. S. Currim. 2002. "Hierarchical Bayes versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery." *Journal of Marketing Research* 39: 87–98. *https://doi.org/10.1509/jmkr.39.1.87.18936*

Aronow, P. M. 2012. "A General Method for Detecting Interference between Units in Randomized Experiments." *Sociological Methods & Research* 41: 3–16.

Athey, S., D. Eckles, and G. W. Imbens. 2018. "Exact *P*-Values for Network Interference." *Journal of the American Statistical Association* 113: 230–240.

Bansak, K., J. Hainmueller, D. Hopkins, and T. Yamamoto. 2020. "Using Conjoint Experiments to Analyze Elections: The Essential Role of the Average Marginal Component Effect (AMCE)." *SSRN Electronic Journal*.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2018. "The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments." *Political Analysis* 26: 112–119.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2019. "Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments." *Political Science Research and Methods* 9: 53–71.

Barone, S., A. Lombardo, and P. Tarantino. 2007. "A Weighted Logistic Regression for Conjoint Analysis and Kansei Engineering." *Quality and Reliability Engineering International* 23: 689–706.

Bien, J., J. Taylor, and R. Tibshirani. 2013. "A Lasso for Hierarchical Interactions." *Annals of Statistics* 41: 1111–1141.

Bodog, S., and G. L. Florian. 2012. "Conjoint Analysis in Marketing Research." *Journal of Electrical and Electronics Engineering* 5: 19–22.

Campbell, B. L., S. Mhlanga, and I. Lesschaeve. 2013. "Consumer Preferences for Peach Attributes: Market Segmentation Analysis and Implications for New Marketing Strategies." *Agricultural and Resource Economics Review* 42: 518–541.

Candès, E., Y. Fan, L. Janson, and J. Lv. 2018. "Panning for Gold: Model-X Knockoffs for High-Dimensional Controlled Variable Selection." *Journal of the Royal Statistical Society: Series B* 80: 551–577.

Candès, E. J., and P. Sur. 2018. "The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression." *The Annals of Statistics*. https://api.semanticscholar.org/CorpusID:13804651.

Chernozhukov, V., M. Demirer, E. Duflo, and I. Fern'andez-Val. 2017. "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments." Paper No. 1712.04802. https://ideas.repec.org/p/arx/papers/1712.04802.html.

de la Cuesta, B., N. Egami, and K. Imai. 2022. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis* 30: 19–45.

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press

Dezeure, R., P. Bühlmann, L. Meier, and N. Meinshausen. 2015. "High-Dimensional Inference: Confidence Intervals, *p*-Values and R-Software hdi." *Statistical Science* 30: 533–558. *https://doi.org/10.1214/15-STS527*

Egami, N., and K. Imai. 2019. "Causal Interaction in Factorial Experiments: Application to Conjoint Analysis." *Journal of the American Statistical Association* 114: 529–540.

Goplerud, M., K. Imai, and N. E. Pashley. 2022. "Estimating Heterogeneous Causal Effects of High-Dimensional Treatments: Application to Conjoint Analysis." Technical report. Preprint. arXiv:2201.01357.

Green, P., A. Krieger, and Y. Wind. 2001. "Thirty Years of Conjoint Analysis: Reflections and Prospects." *Interfaces* 31: S56–S73.

Green, P. E., and V. Srinivasan. 1990. "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice." *Journal of Marketing* 54: 3–19.

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: 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: 1–30.

Ham, D. W., K. Imai, and L. Janson. 2023. "Replication Data for: Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis." https://doi.org/10.7910/DVN/ENI8GF

Hauber, A. B., et al. 2016. "Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force." *Value in Health* 19: 300–315.

Imai, K., and M. L. Li. 2021. "Experimental Evaluation of Individualized Treatment Rules." *Journal of the American Statistical Association* 118: 242–256.

Imbens, G. W., and D. B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

Liu, M., E. Katsevich, L. Janson, and A. Ramdas. 2021. "Fast and powerful conditional randomization testing via distillation." *Biometrika* 109 (2): 277–293. https://doi.org/10.1093/biomet/asab039.

Luce, R. D., and J. W. Tukey. 1964. "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology* 1:1–27.

McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers of Econometrics*, edited by P. Zarembka, 105–142. New York: Academic Press.

Newman, B. J., and N. Malhotra. 2019. "Economic Reasoning with a Racial Hue: Is the Immigration Consensus Purely Race Neutral?" *Journal of Politics* 81: 153–166.

Ono, Y., and B. C. Burden. 2018. "The Contingent Effects of Candidate Sex on Voter Choice." *Political Behavior* 41: 583–607.

Popovic, M., M. Kuzmanovic, and M. Martic. 2012. "Using Conjoint Analysis to Elicit Employers' Preferences toward Key Competencies for a Business Manager Position." *Management—Journal for Theory and Practice of Management* 17: 17–26.

Raghavarao, D., J. B. Wiley, and P. Chitturi. 2010. *Choice-Based Conjoint Analysis: Models and Designs*. Boca Raton: Chapman and Hall/CRC.

Rubin, D. B. 1990. "Comments on 'on the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9' by J. Splawa-Neyman. Translated from the Polish and Edited by D. M. Dabrowska and T. P. Speed." *Statistical Science* 5: 472–480.

Shiraito, Y., and G. Liu. 2023. "Multiple Hypothesis Testing in Conjoint Analysis." *Political Analysis* 31: 380–395.

Spilker, G., T. Bernauer, and V. Umaña. 2016. "Selecting Partner Countries for Preferential Trade Agreements: Experimental Evidence from Costa Rica, Nicaragua, and Vietnam." *International Studies Quarterly* 60: 706–718. *https://doi.org/10.1093/isq/sqv024*

Tansey, W., V. Veitch, H. Zhang, R. Rabadan, and D. M. Blei. 2022. "The Holdout Randomization Test for Feature Selection in Black Box Models." *Journal of Computational and Graphical Statistics*. Taylor & Francis, 31 (1): 151–162. https://doi.org/10.1080/10618600.2021.1923520.

Cite this article: Ham, D. W., Imai, K. and Janson, L. (2024). Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis. *Political Analysis*, 329–344. https://doi.org/10.1017/pan.2023.41

https://doi.org/10.1017/pan.2023.41 Published online by Cambridge University Press