Supplementary Material

Authors: Dae Woong Ham, Kosuke Imai, and Lucas Janson

A Proof of Theorem 3.1

Proof. We first prove that if $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$, then $\mathbf{Y}(\mathbf{x},\mathbf{z}) \overset{d}{=} \mathbf{Y}(\mathbf{x}',\mathbf{z})$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$.

$$\begin{aligned} P(\mathbf{Y} \mid \mathbf{Z} = \mathbf{z}) &= P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \\ &= P(\mathbf{Y}(\mathbf{x},\mathbf{z}) \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \\ &= P(\mathbf{Y}(\mathbf{x},\mathbf{z}) \mid \mathbf{Z} = \mathbf{z}) \\ &= P(\mathbf{Y}(\mathbf{x},\mathbf{z})), \end{aligned}$$

where the third and fourth equalities follow from the randomization of $\mathbf{X}$ and that of $\mathbf{Z}$, respectively. Similarly, we can show $P(\mathbf{Y} \mid \mathbf{Z} = \mathbf{z}) = P(\mathbf{Y}(\mathbf{x}',\mathbf{z}))$, thus we have shown $P(\mathbf{Y}(\mathbf{x},\mathbf{z})) = P(\mathbf{Y}(\mathbf{x}',\mathbf{z}))$ for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$.

To the prove the other direction, we want to show that $\mathbf{Y}(\mathbf{x},\mathbf{z}) \overset{d}{=} \mathbf{Y}(\mathbf{x}',\mathbf{z})$ implies $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}$, or equivalently $P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}', \mathbf{Z} = \mathbf{z})$ for any value of $\mathbf{z} \in \mathcal{Z}$ and any value of $\mathbf{x}$, $\mathbf{x}' \in \mathcal{X}$.

$$\begin{aligned} P(\mathbf{Y}(\mathbf{x},\mathbf{z})) &= P(\mathbf{Y}(\mathbf{x},\mathbf{z}) \mid \mathbf{Z} = \mathbf{z}) \\ &= P(\mathbf{Y}(\mathbf{x},\mathbf{z}) \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) \\ &= P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}), \end{aligned}$$

where the first two equalities follow from the randomization of $\mathbf{Z}$ and that of $\mathbf{X}$, respectively. The same argument shows $P(\mathbf{Y}(\mathbf{x}',\mathbf{z})) = P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}', \mathbf{Z} = \mathbf{z})$. Finally, because $\mathbf{Y}(\mathbf{x},\mathbf{z}) \overset{d}{=} \mathbf{Y}(\mathbf{x}',\mathbf{z})$ we have that $P(\mathbf{Y}(\mathbf{x},\mathbf{z})) = P(\mathbf{Y}(\mathbf{x}',\mathbf{z}))$, implying $P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = P(\mathbf{Y} \mid \mathbf{X} = \mathbf{x}', \mathbf{Z} = \mathbf{z})$ for any value of $\mathbf{z} \in \mathcal{Z}$ and any value of $\mathbf{x}$, $\mathbf{x}' \in \mathcal{X}$. □

B Relation to Finite-Population Inference

In Section 3, we introduced H_0 under the super-population framework. Here, we consider the relationship between H_0 and Fisher's sharp null of no treatment effect (Fisher, 1935). Under the finite-population framework where the potential outcomes are fixed and the randomness comes only from the randomization of treatment assignment, if we assume no interference between units, H_0 reduces to testing $Y_{ij}(\mathbf{x}_{ij}, \mathbf{z}_{ij}) = Y_{ij}(\mathbf{x}'_{ij}, \mathbf{z}_{ij})$ for all i, j and all possible values of $\mathbf{x}_{ij}, \mathbf{x}'_{ij} \in \mathcal{X}$ and $\mathbf{z}_{ij} \in \mathcal{Z}$. Under the super-population framework, if we make the same no-interference assumption, H_0 reduces to testing $Y_{ij}(\mathbf{x}_{ij}, \mathbf{z}_{ij}) \overset{d}{=} Y_{ij}(\mathbf{x}'_{ij}, \mathbf{z}_{ij})$ for all i, j and all possible values of $\mathbf{x}_{ij}, \mathbf{x}'_{ij}, \mathbf{z}_{ij}$, where the potential outcomes are assumed to be drawn from a population. These two null hypotheses are not equivalent even though the CRT can test H_0 under both the finite-population and super-population frameworks (as proven in Theorem 3.1). This is because the distribution of $\mathbf{Y}(\mathbf{x},\mathbf{z})$ can still be equal to $\mathbf{Y}(\mathbf{x}',\mathbf{z})$ even if $Y_{ij}(\mathbf{x},\mathbf{z})$ is different than $Y_{ij}(\mathbf{x}',\mathbf{z})$ for some i, j.

C Grouping Factor Levels

In this appendix, we further detail how to test H_0^{General} when the analyst is interested in grouping multiple factor levels. For example, in the immigration conjoint application (Section 4), we wish to test whether respondents differentiate immigrants from Mexico and those from Europe where there are three distinct levels for countries from Europe: *Germany*, *France*, and *Poland*. We now formally show how to group these levels up into one category *Europe* and test the hypothesis H_0^{General}.

We introduce a coarsening function c that takes q factors of interest $\mathbf{X}$ as input and transforms them to a new set of grouped factors $\overline{\mathbf{X}}$. Formally, this function is defined as $c : \mathcal{X} \mapsto \overline{\mathcal{X}}$ where $\overline{\mathcal{X}}$ represents the support of the grouped factors $\overline{\mathbf{X}}$ with $|\mathcal{X}| \geq |\overline{\mathcal{X}}|$. For the above example, we define the function c on the "country of origin" factors such that the *Mexico* level takes one value whereas the *France*, *Germany*, and *Poland* levels all take another value: *Europe* $\in \overline{\mathbf{X}}$. All other factor levels are mapped to different values in $\overline{\mathbf{X}}$.

Next, we define the outcome for each level of newly transformed factor levels. Given the coarsening function c defined above, we introduce the marginalized potential outcome variable $\overline{\mathbf{Y}}(\overline{\mathbf{x}}, \mathbf{z})$, which averages over the distribution of original factor levels that are grouped. Formally, this new outcome variable has the following mixture structure,

$$\overline{\mathbf{Y}}(\overline{\mathbf{x}}, \mathbf{z}) = \frac{\sum_{\mathbf{x}' \in \mathcal{X}} \mathbf{1}\{c(\mathbf{x}') = \overline{\mathbf{x}}\} \mathbf{Y}(\mathbf{x}', \mathbf{z}) P(\mathbf{X} = \mathbf{x}' \mid \mathbf{Z} = \mathbf{z})}{\sum_{\mathbf{x}' \in \mathcal{X}} \mathbf{1}\{c(\mathbf{x}') = \overline{\mathbf{x}}\} P(\mathbf{X} = \mathbf{x}' \mid \mathbf{Z} = \mathbf{z})}, \tag{11}$$

where $\mathbf{z} \in \mathcal{Z}$, $\overline{\mathbf{x}} \in \overline{\mathcal{X}}$, and $P(\mathbf{X} = \mathbf{x} \mid \mathbf{Z} = \mathbf{z})$ represents the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ used in the experiment. For example, if we group three European countries—*France*, *Germany*, and *Poland*—and create one new factor level *Europe*, then its marginalized potential outcome will be a mixture distribution of the original potential outcomes for the three countries weighted by their known randomization probabilities conditional on the other factors.

Furthermore, the previously introduced coarsening function h now takes the newly grouped up factor $\overline{\mathbf{X}}$ and maps it to the new coarsened factor, i.e., $h : \overline{\mathcal{X}} \mapsto \widetilde{\mathcal{X}}$. Consequently, our updated generalized null hypothesis is,

$$\overline{H}_0^{\text{General}} : \overline{\mathbf{Y}}(\overline{\mathbf{x}}, \mathbf{z}) \overset{d}{=} \overline{\mathbf{Y}}(\overline{\mathbf{x}}', \mathbf{z}) \text{ for all } \overline{\mathbf{x}}, \overline{\mathbf{x}}' \in \overline{\mathcal{X}}, \text{ such that } h(\overline{\mathbf{x}}) = h(\overline{\mathbf{x}}') \text{ and } \mathbf{z} \in \mathcal{Z}. \tag{12}$$

Finally, it can also be shown by applying the same argument as the one used to prove Theorem 3.1 that $\overline{H}_0^{\text{General}}$ is equivalent to the following conditional independence relation,

$$\overline{\mathbf{Y}} \perp\!\!\!\perp \overline{\mathbf{X}} \mid h(\overline{\mathbf{X}}), \mathbf{Z}. \tag{13}$$

To test this null hypothesis, we keep the original test statistic T_{HierNet} shown in Equation (5) under the same symmetry constraints given in Equation (6) except that we use $\overline{\mathbf{X}}$ in place of $\mathbf{X}$ to account for coarsening based on the function c.

D Simulations

A primary advantage of the CRT is that it can yield powerful statistical tests by incorporating machine learning algorithms to capture complex interactions in high dimensions. The CRT achieves this while maintaining the finite sample validity of the resulting p-values. In this section, we conduct simulation studies to show that the CRT with the HierNet test statistic can be substantially more powerful than the AMCE-based test.

For simplicity, we focus on the setting in which each respondent only has one evaluation, i.e., $J = 1$. Figure 5 in Appendix D.5 presents additional simulations that have multiple evaluations per respondent based on respondent random effects. In the setting where each respondent evaluates only one task, we treat each response as an independent observation and drop subscript j. We also assume that every factor $(\mathbf{X}, \mathbf{Z})$ is uniformly and independently randomized, implying that all treatment combinations are equally likely. As before, $\mathbf{X}$ represents the main factors of interest while $\mathbf{Z}$ denotes the other factors. For simplicity, we assume that all factors $(\mathbf{X}, \mathbf{Z})$ are binary with their success probabilities equal to 0.5, and we have one factor of interest ($q = 1$) with no respondent characteristics $\mathbf{V}$.

D.1 The Basic Setup

To clearly separate main and interaction effects, we use the sum-to-zero constraint by coding each binary factor as $(-0.5, 0.5)$. Our data generating process uses the following logistic regression model under the forced-choice design,

$$\begin{aligned}\Pr(Y_i = 1 \mid X_i, \mathbf{Z}_i) \ &= \ \operatorname{logit}^{-1}\left[\beta_X(X_i^L - X_i^R) + \beta_Z^\top(\mathbf{Z}_i^L - \mathbf{Z}_i^R)\right. \\ &\quad \left. + 2\gamma^\top\{(X_i^L\mathbf{Z}_i^L) - (X_i^R\mathbf{Z}_i^R)\} + 2\delta^\top\{(X_i^L\mathbf{Z}_i^R) - (X_i^R\mathbf{Z}_i^L)\} + 2\tilde{\gamma}^\top\{(\mathbf{Z}_i^L \times \mathbf{Z}_i^L) - (\mathbf{Z}_i^R \times \mathbf{Z}_i^R)\}\right],\end{aligned}$$

where β_X and β_Z represent the coefficient vectors for the main effects of $\mathbf{X}$ and $\mathbf{Z}$, respectively, and γ and δ denote the coefficient column vectors for the within-profile and between-profile interactions between $\mathbf{X}$ and $\mathbf{Z}$, respectively. For simplicity, we omit between-profile interactions among $\mathbf{Z}$ and consider only within-profile interactions among $\mathbf{Z}$ with effect sizes $\tilde{\gamma}$. To facilitate interpretation, each interaction coefficient is multiplied by 2 because our encoding of interaction effects, e.g., $\{(X_i^L\mathbf{Z}_i^L) - (X_i^R\mathbf{Z}_i^R)\}$, results in three possible values $(-0.5, 0, 0.5)$.

This data generating process also implies the absence of profile order effects. Lastly, we note that the logistic regression has a latent variable, Y_i', representation such that $Y_i = 1$ if $Y_i' > 0$ and 0 otherwise, where we let ϵ_i follow a standard logistic distribution and

$$\begin{aligned}Y_i' &= \beta_X(X_i^L - X_i^R) + \beta_Z^\top(\mathbf{Z}_i^L - \mathbf{Z}_i^R) \\ &\quad + 2\gamma^\top\{(X_i^L\mathbf{Z}_i^L) - (X_i^R\mathbf{Z}_i^R)\} + 2\delta^\top\{(X_i^L\mathbf{Z}_i^R) - (X_i^R\mathbf{Z}_i^L)\} + 2\tilde{\gamma}^\top\{(\mathbf{Z}_i^L \times \mathbf{Z}_i^L) - (\mathbf{Z}_i^R \times \mathbf{Z}_i^R)\} + \epsilon_i.\end{aligned}$$

We consider settings in which $\mathbf{X}$ has one main effect of $\beta_X = 0.1$, which is fixed for all simulations. In addition, $\mathbf{Z}$ consists of ten factors with four non-zero main effects with a magnitude of 0.1 with alternating signs, which is fixed for all simulations, i.e., $\beta_Z = (0.1, -0.1, 0.1, -0.1, 0, \dots, 0)$. Lastly, we fix $\tilde{\gamma}^\top$ to have fifteen non-zero entries of 0.05, where the non-zero interactions are randomly chosen from all possible within-profile interactions among $\mathbf{Z}$. The remaining entries are all zero.

The sample size is fixed to $n = 3{,}000$ throughout the simulations. For each simulation, we generate within-profile and between-profile interactions between $\mathbf{X}$ and $\mathbf{Z}$ by randomly selecting the specified number of interactions from all possible interactions. The total number of non-zero interaction effects between $\mathbf{X}$ and $\mathbf{Z}$ varies from 0 to 18. The number of non-zero within-profile interactions between $\mathbf{X}$ and $\mathbf{Z}$ is kept identical to that of non-zero between-profile interactions between $\mathbf{X}$ and $\mathbf{Z}$. We make all non-zero within-profile interactions positive and all non-zero between-profile interactions negative while fixing all non-zero interaction effects between $\mathbf{X}$ and $\mathbf{Z}$ to be equal in magnitude. We explore additional simulations in Appendix D.3 where there are heterogeneous interaction effects. We set $B = 200$ for all simulations presented in this paper.

To calculate the statistical power of each test, we compute a p-value for each of 1,000 Monte Carlo data sets. For the CRT, we use the HierNet test statistic given in Equation (5) and impose the constraints in Equation (6), whereas we use the t-test based on the estimated regression coefficient of $\mathbf{X}$ for the AMCE-

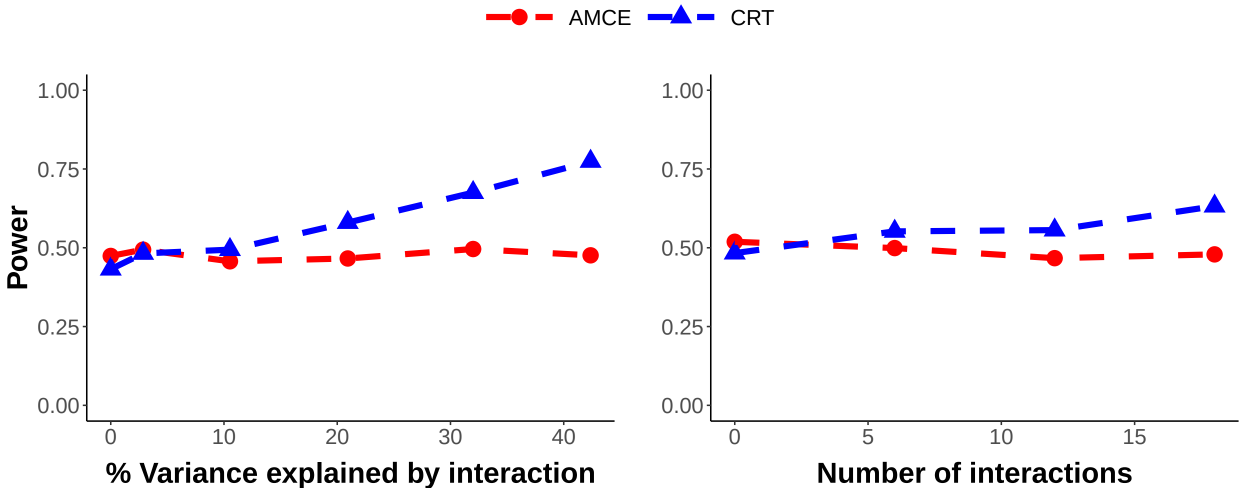

Figure 2: The figure shows how the power of the CRT and AMCE-based tests varies as the size of interaction effects (left plot) or the number of non-zero interaction effects (right) increases. The AMCE-based test (red circles) is based on the t-test from the estimated regression coefficient. The CRT uses the HierNet test statistic given in Equation (5). The sample size is $n = 3,000$. Finally, the standard errors are negligible with a maximum value of 0.016.

based test.[11] The AMCE-based analysis assumes no profile-order effect. Therefore, as suggested by Hainmueller, Hopkins and Yamamoto (2014), we run the linear regression by stacking all the left and right profiles, resulting in $2n$ rows. More formally, we have the new response $\mathbf{Y}^{\text{AMCE}} = [\mathbf{Y}; \mathbf{1} - \mathbf{Y}]$ regressed on $\mathbf{X}^{\text{AMCE}} = [\mathbf{X}^L; \mathbf{X}^R]$, where $\mathbf{X}^L = [X_1^L; X_2^L; \dots; X_n^L]$ and $\mathbf{X}^R$ is defined similarly (note we have dropped the j subscript since $J = 1$). Finally, the standard errors are clustered by respondents as suggested by Hainmueller, Hopkins and Yamamoto. Since $J = 1$ for our simulation, we cluster the standard errors on each evaluation task, i.e., each unique cluster consists of the left and right profile for each task. We then compute the power as the proportion of p-values that are less than $\alpha = 0.05$.

D.2 Main Simulation Results

We now present the results for the simulations described above. The left plot of Figure 2 shows how the statistical power of each test varies as the size of interaction effects between $\mathbf{X}$ and $\mathbf{Z}$ increases. The number of non-zero interaction effects between $\mathbf{X}$ and $\mathbf{Z}$ is fixed at six. For ease of interpretation, we plot the percentage of total outcome variance explained by the interaction effects between $\mathbf{X}$ and $\mathbf{Z}$ on the x-axis.[12] In our setup, the total variance represents the outcome variance explained by all main and interaction effects under the latent representation of the logistic regression model. Consequently, we define the variance explained by the interaction effects between $\mathbf{X}$ and $\mathbf{Z}$ and all other "remaining" effects (main effects of $\mathbf{X}$, $\mathbf{Z}$, and interactions

[11] Although this is a valid procedure to compute the AMCE estimate for $\mathbf{X}$, practitioners typically compute the AMCE estimates of all factors $(\mathbf{X}, \mathbf{Z})$ simultaneously with a single linear regression of $\mathbf{Y}$ on all $(\mathbf{X}, \mathbf{Z})$. Figure 4 of Appendix D.4 shows that the power of the AMCE remains indistinguishable when using all factors $(\mathbf{X}, \mathbf{Z})$ in a single linear regression. Lastly, although our goal is to compare the CRT with the AMCE, we acknowledge that practitioners may also use the omnibus F-test for testing interactions by including all the two-way interactions. We show in Appendix J that such an approach leads to inflated p-values, thus we omit this as a baseline comparison here.

[12] The x-axis ticks correspond to interaction sizes of $0, 0.025, 0.05, 0.075, 0.1$, and 0.125, respectively. For example, the 20% point on the x-axis refers to an interaction size of 0.05.

among $\mathbf{Z}$) as,

$$\begin{aligned}\sigma^2_{\text{Interaction}} &:= \mathbb{V}(2\gamma^\top\{(X_i^L\mathbf{Z}_i^L) - (X_i^R\mathbf{Z}_i^R)\} + 2\delta^\top\{(X_i^L\mathbf{Z}_i^R) - (X_i^R\mathbf{Z}_i^L)\})\\ \sigma^2_{\text{Remaining}} &:= \mathbb{V}(\beta_X(X_i^L - X_i^R) + \beta_Z^\top(\mathbf{Z}_i^L - \mathbf{Z}_i^R) + 2\tilde{\gamma}^\top\{(\mathbf{Z}_i^L \times \mathbf{Z}_i^L) - (\mathbf{Z}_i^R \times \mathbf{Z}_i^R)\}),\end{aligned}$$

where $\mathbb{V}(\cdot)$ denotes the variance of the respective random variable. The total variance is defined as $\sigma^2_{\text{Interaction}} + \sigma^2_{\text{Remaining}}$. Since $\beta_X, \beta_Z, \tilde{\gamma}$ are fixed for all simulations, $\sigma^2_{\text{Remaining}}$ is fixed with a value of:

$$\sigma^2_{\text{Remaining}} = 9 \times 0.1^2 \times 2\mathbb{V}(X_i^L) + 15 \times 4 \times 0.05^2 2\mathbb{V}(X_i^L X_i^R) = 9 \times 0.1^2 \times \frac{1}{2} + 15 \times 4 \times 0.05^2 \times \frac{1}{8} = 0.06375,$$

where the first equality holds because all random variables $(X_i^L, X_i^R, \mathbf{Z}_i^L, \mathbf{Z}_i^R)$ are independent, centered at zero, and identically distributed (element-wise identically distributed for the multivariate $\mathbf{Z}_i^L$ and $\mathbf{Z}_i^R$).

Furthermore, given a signal size I and the number of interactions n_I (n_I denotes the total number of within-profile and between-profile interactions), we have that:

$$\sigma^2_{\text{Interaction}} = 8I^2\mathbb{V}(X_i^L X_i^R)n_I = \frac{I^2 n_I}{2}.$$

Finally, since the left plot of Figure 2 contains six interactions ($n_I = 6$), the x-axis is computed by $\frac{3I^2}{3I^2+0.06375}$, where I takes the following values $(0, 0.025, 0.05, 0.075, 0.1, 0.125)$.

Consistent with our theoretical expectation, the CRT (blue triangles) becomes more powerful than the AMCE-based test (red circles) as the interaction size increases. For example, when the interaction size is strong enough to account for about 30% of the total variance, the CRT is approximately 20 percentage points more powerful than the AMCE-based test. When there is no interaction effect, the CRT is only slightly less powerful (by about 3 percentage points) than the AMCE-based test.

The right plot of Figure 2 shows how the power of the tests change as one varies the number of non-zero interaction effects. The size of interaction effects is fixed to 0.06, around half the size of the main effect. We find that as expected, the CRT becomes more powerful than the AMCE-based test as the number of interaction increases. For example, when there are twelve interactions the CRT is approximately 10 percentage points more powerful than the AMCE-based test. Even when there is no interaction effect at all, the loss of statistical power is minimal. Appendix E presents additional simulation results, showing that the use of no profile order constraints given in Equation (6) increases the power of test.

D.3 Heterogeneous Interaction Size

As shown in the above simulations, we fix all interaction sizes to be equal for every simulation. Here, we examine if the simulation results presented in Figure 2 change if there are heterogeneous interaction effects.

We compare the statistical power under two different data generating processes using the CRT with test statistic in Equation (5). We denote the original data generating process as the "homogeneous" scenario since all interaction sizes are equal under this scenario. We create an additional "heterogeneous" scenario that contains two unique varying interaction effects - one that is strong, I_s, and one that is weak I_w. To facilitate a fair comparison between the "heterogeneous" and "homogeneous" scenario, we force the total variance explained by the interactions to be equal under both scenarios. We assign all strong interaction effects to the within-profile interaction and all weak interaction effects to the between-profile interaction. Suppose there are only two non-zero interactions between $\mathbf{X}$ and $\mathbf{Z}$. Since we impose $\sigma^2_{\text{Interaction}}$ to be equal under both the "homogeneous" and "heterogeneous" scenario, we have the following equation:

$$I_w^2 + I_s^2 = 2I^2,$$

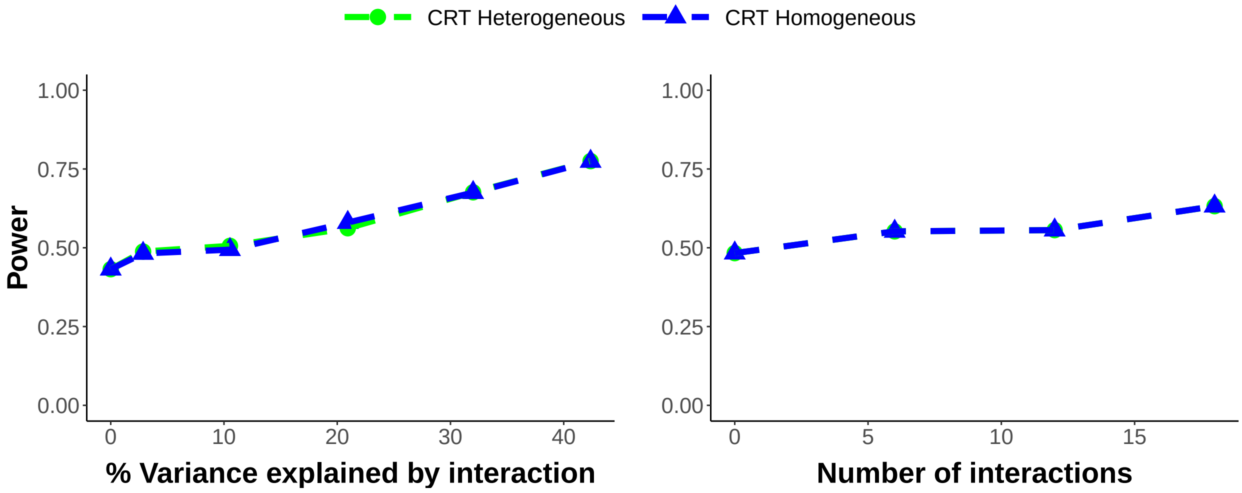

Figure 3: The figure shows the power of the "heterogeneous" (light green circles) and "homogeneous" (blue triangles) scenario in the same simulation setting as in Figure 2.

where I is the interaction size for the "homogeneous" scenario.

The possible values of strong and weak effects lie on the circle of radius $\sqrt{2}I$ centered at the origin. We pick $I_s(I) = \sqrt{\frac{4}{3}}I$ and $I_w(I) = \sqrt{\frac{2}{3}}I$, i.e., the point corresponding to the thirty degree angle of the circle. The variance explained by the interaction is equal for both the "heterogeneous" and "homogeneous" scenario when there are two non-zero interactions. To create a power curve for the "heterogeneous" scenario analogous to the left plot in Figure 2, we create three interactions of size $I_s(I)$ for the within-profile interaction and three interactions of size $I_w(I)$ for the between-profile interaction. The two scenarios still maintain equal variance explained by the interaction effects because all random variables are independent and centered at zero. For the analogue of the right plot of Figure 2, we similarly keep the relative proportion of $I_w(I)$ and $I_s(I)$ fixed and increase the number of non-zero interactions to match the "homogeneous" scenario. For example, if there are twelve non-zero interaction effects, then there are six within-profile interactions of size $I_s(I)$ and six between-profile interactions of size $I_w(I)$ for the "heterogeneous" scenario.

Figure 3 shows that the power of the "heterogeneous" and "homogeneous" scenario is indistinguishable under the same simulation setting in Figure 2. This shows that we lose no generality by considering only the simple "homogeneous" scenario in the main simulations in Figure 2.

D.4 Additional AMCE Simulations

The AMCE computed in Figure 2 was based on a linear regression of $\mathbf{Y}$ on $\mathbf{X}$, without $\mathbf{Z}$ included among the predictors. Although this is valid and sufficient to compute the AMCE of $\mathbf{X}$, practitioners often compute the AMCE of all factors $(\mathbf{X}, \mathbf{Z})$ simultaneously with a single linear regression of $\mathbf{Y}$ on all $(\mathbf{X}, \mathbf{Z})$. We compute the power of the "long AMCE" that is based on the t-test for the estimated regression coefficient of $\mathbf{X}$ obtained by regressing the response $\mathbf{Y}$ on all $(\mathbf{X}, \mathbf{Z})$. Figure 4 shows that the power of the "long AMCE" (orange squares) is indistinguishable from that of the original AMCE presented in Figure 2 (red circles).

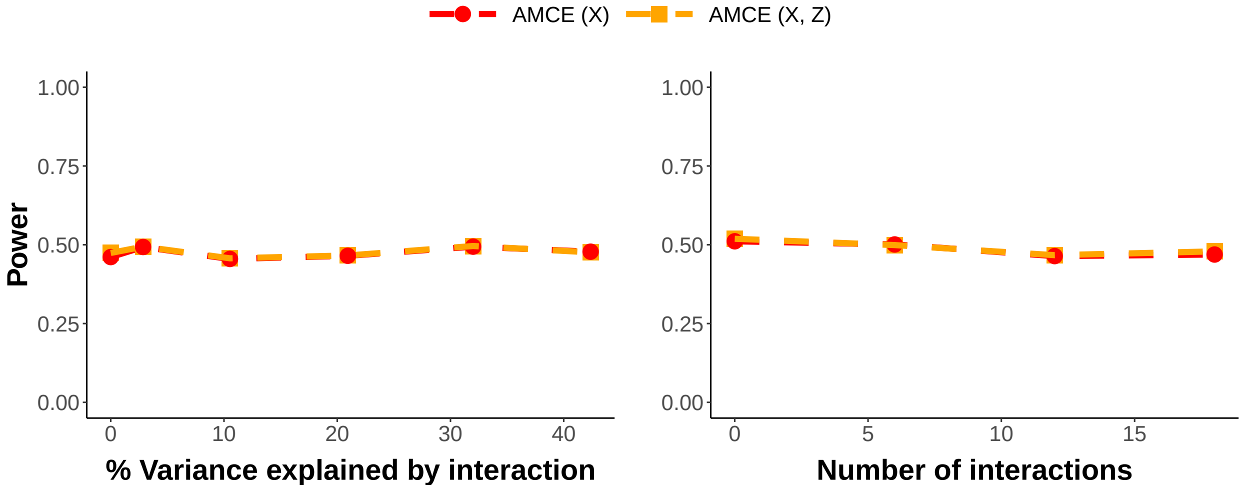

Figure 4: The figure shows the power of the "long AMCE" that uses all $(\mathbf{X}, \mathbf{Z})$ in the linear regression fit (orange squares) and the original AMCE that uses only $\mathbf{X}$ in the linear regression fit (red circles) in the same simulation setting as in Figure 2.

D.5 Simulations with Multiple Tasks per Respondent

Appendix D.1 presents the simulation results where each respondent evaluates only one task ($J = 1$). Here, we consider a simulation setup where each respondent evaluates $J = 5$ tasks while still fixing the total sample size $nJ = 3,000$. We keep the same simulation setup as the one described in Appendix D.1 except that we allow each respondent to have a random effect $U_j \sim N(0, \sigma^2_{RE})$. More formally, our new data generating process is,

$$\begin{aligned}\Pr(Y_{ij} = 1 \mid X_{ij}, \mathbf{Z}_{ij}) \;=\; & \text{logit}^{-1}\Big[\beta_X(X^L_{ij} - X^R_{ij}) + \beta_Z^\top(\mathbf{Z}^L_{ij} - \mathbf{Z}^R_{ij}) \\ & + 2\gamma^\top\{(X^L_{ij}\mathbf{Z}^L_{ij}) - (X^R_{ij}\mathbf{Z}^R_{ij})\} + 2\delta^\top\{(X^L_{ij}\mathbf{Z}^R_{ij}) - (X^R_{ij}\mathbf{Z}^L_{ij})\} + 2\tilde{\gamma}^\top\{(\mathbf{Z}^L_{ij} \times \mathbf{Z}^L_{ij}) - (\mathbf{Z}^R_{ij} \times \mathbf{Z}^R_{ij})\} + U_j\Big],\end{aligned}$$

where U_j is the random effect for each respondent j. We keep all simulation parameters the same as that in Figure 2 except we use the above data generating process with $J = 5$ evaluation tasks and random effects with $\sigma^2_{RE} = 0.1$ to produce Figure 5.

Although the CRT HierNet test statistic does not change with the addition of multiple respondent evaluations, the AMCE estimate must properly account for the respondent effect. As suggested by Hainmueller, Hopkins and Yamamoto (2014), we use the robust clustered standard errors clustered on respondents for the linear regression of $\mathbf{Y}$ on $\mathbf{X}$ and use the t-test based on the estimated regression coefficient of $\mathbf{X}$ to produce the power curve (red). Figure 5 shows the results are similar to those shown in Figure 2, suggesting that our results are not sensitive to the number of evaluations per respondent.

E Enforcing The No Profile Order Effect

We detail here a way to enforce the no profile order effect constraints in Equation (6) for general test statistics. We show through simulations that these constraints, when they hold, can substantially improve statistical power.

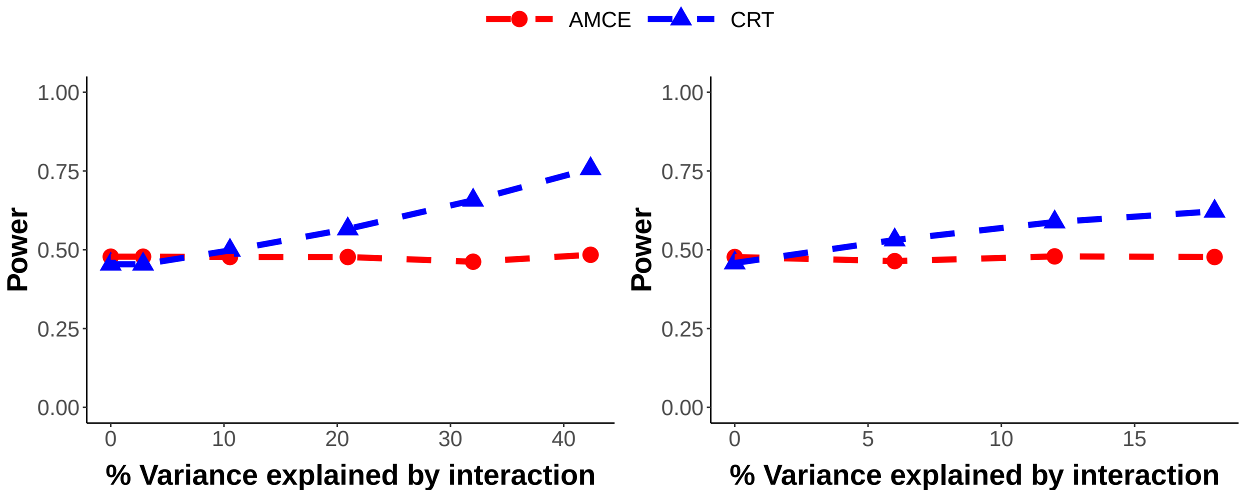

Figure 5: The figure shows the power of the AMCE (red circles) and the CRT (blue triangles) using the HierNet test statistic in Equation (5). We modify the simulation setting in Figure 2 by having each respondent evaluate $J = 5$ tasks with a total of $nJ = 3,000$ responses. Otherwise, the simulation setup remains identical to that in Figure 2. Each respondent has a random effect of $\sigma^2_{RE} = 0.1$.

Row Number	**Country^L**	**Country^R**	**Gender^L**	**Gender^R**	**Respondent Age**	**Y**
1	Mexico	Germany	Male	Female	27	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$nJ + 1$	Germany	Mexico	Female	Male	27	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 2: Visual example of D^c data matrix with $\mathbf{X}$ as country of origin, $\mathbf{Z}$ as gender, and $\mathbf{V}$ as the respondent's age. With a slight abuse of notation Country^L denotes the left profile's country values. Country^R, Gender^L, and Gender^R are defined similarly.

Under these constraints, switching the "left" and "right" profile order does not change the value of the test statistic. We now formalize this intuition. First denote $D \in \mathbb{R}^{nJ\times(2p+r+1)}$ as the regression data matrix composed of $(\mathbf{X}, \mathbf{Z}, \mathbf{V}, \mathbf{Y})$. Let $s_E : \mathbb{R}^{nJ\times(2p+r+1)} \rightarrow \mathbb{R}^{nJ\times(2p+r+1)}$ denote a function that takes a data matrix as an input and swaps the "left" and "right" profile order for rows E of the data matrix, where $E \subset \{1, 2, \ldots, nJ\}$. For example, suppose we swap just the first row and we denote $(\tilde{\mathbf{X}}, \tilde{\mathbf{Z}}, \tilde{\mathbf{V}}, \tilde{\mathbf{Y}})$ as the columns for the output $s_{\{1\}}(D)$. Then $\tilde{\mathbf{X}}^L_{11} = \mathbf{X}^R_{11}$, $\tilde{\mathbf{X}}^R_{11} = \mathbf{X}^L_{11}$, $\tilde{\mathbf{Z}}^L_{11} = \mathbf{Z}^R_{11}$, $\tilde{\mathbf{Z}}^R_{11} = \mathbf{Z}^L_{11}$, and $\tilde{Y}_{11} = 1 - Y_{11}$ and all remaining rows remain identical as the original data D. The function applies no swap to the respondent characteristic $\mathbf{V}$ $(\tilde{\mathbf{V}} = \mathbf{V})$.

We introduce a new data matrix D^c that appends the original data matrix with a data matrix that swaps the profile order for all rows. Formally, $D^c = [D; s_{\{1,2,\ldots,nJ\}}(D)]$. Table 2 shows an example of D^c with $\mathbf{X}$ as country, $\mathbf{Z}$ as gender, and $\mathbf{V}$ as the respondent's age. Conceptually, D^c aims to destroy all information about the profile order since the "left" and "right" profile are now indistinguishable, thus ensuring any test statistic that uses D^c will respect the no profile order effect constraints. The following lemma formally states this result.

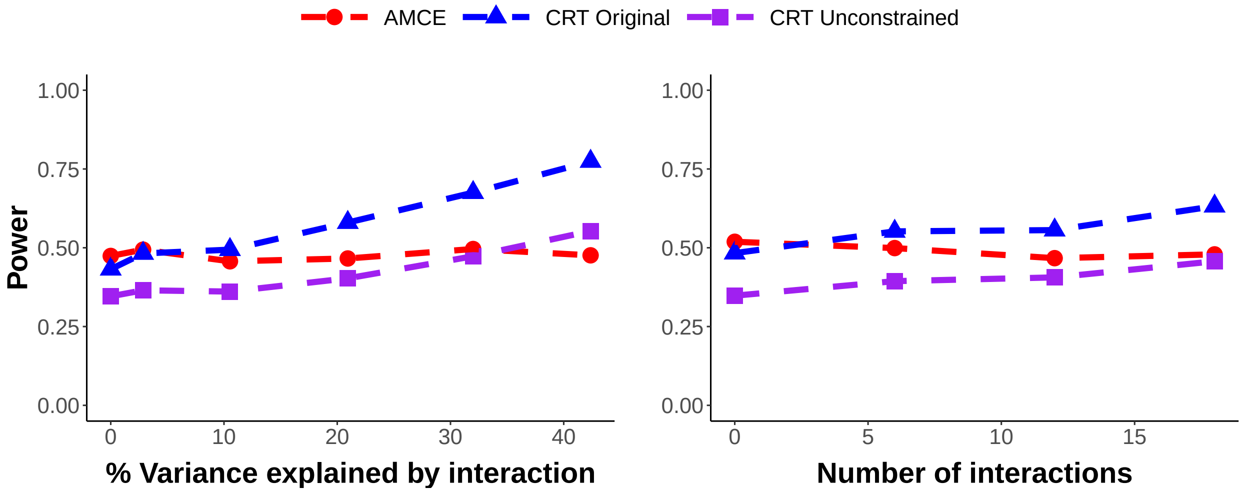

Figure 6: This figure represents the power gain from imposing the no "Profile Order Effect" constraints in Equation (6) under the same simulation setup in Figure 2. We keep the AMCE and the original HierNet statistical power curves (red circles and blue triangles respectively) and add the new unconstrained HierNet test statistic (purple squares).

Lemma E.1. Let $T(\cdot)$ be a row invariant test statistic,[13] then for any $E \subset \{1, 2, \dots, nJ\}$ we have that $T(D^c) = T(s_E(D)^c)$, where $s_E(D)^c = [s_E(D); s_{1,2,\dots,nJ}(s_E(D))]$.

Proof. The lemma holds because $D^c = s_E(D)^c$ up to row permutations. Using the assumption that $T(\cdot)$ is a row invariant test statistic we obtain the equality. Many algorithms like Random Forest have random initializations in the process so $T(D^c) = T(s_E(D)^c)$ may only hold up to distributional equality. □

Lemma E.1 states that the test statistic will remain invariant to any relabelling of "left" and "right" profiles as long the test statistic is a function of the D^c data matrix. All regression based algorithms will respect exact row invariance. Lemma E.1 allows practitioners to build any test statistic, even ones that do not have natural "left" and "right" coefficients, while still enforcing the no profile order effect. In particular, we enforce the constraints in Equation (6) by using D^c as the input in HierNet.

Simulations. We now show that imposing the no profile order effect constraints in Equation (6) can substantially increase statistical power under the same simulation setup in Figure 2 when the no profile order effect assumption is satisfied. To evaluate the power gain, we also fit HierNet without imposing the constraints in Equation (6) by using the original data D. We use $T_{\text{HierNet}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})^L + T_{\text{HierNet}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})^R$ as the test statistic, where $T_{\text{HierNet}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})^L$ is the same as $T_{\text{HierNet}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ in Equation (5) but all coefficients correspond to their respective estimates for the left profile. We similarly define $T_{\text{HierNet}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})^R$.

Figure 6 shows that imposing the no profile order effect constraints can significantly increase power. For example, the power of HierNet without the constraints (purple squares) is roughly equal or smaller than that of even the AMCE (red circles) when the interaction effect accounts for 20% of the variance or when there are as many interactions as twelve. Furthermore, we see the power of using HierNet that imposes the constraints (blue triangles) is consistently higher than that of the HierNet without the constraints.

[13] More formally we say $T(D)$ is row invariant if $T(D) = T(\pi(D))$ for any possible π, where π denotes a possible permutation of the rows of D.

```
Algorithm 2: Generalized $H_0$: Testing $H_0^{\text{General}}$
Input: Data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{V})$, test statistic $T(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{v})$, $h(\mathbf{x})$, total number of re-samples $B$;
for $b = 1, 2, \dots, B$ do
    Sample $\mathbf{X}^{(b)}$ from the distribution of $\mathbf{X} \mid (h(\mathbf{X}), \mathbf{Z}, \mathbf{V})$ conditionally independently of $\mathbf{X}$ and $\mathbf{Y}$;
Output: $p$-value $:= \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{X}^{(b)}, \mathbf{Y}, \mathbf{Z}, \mathbf{V}) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{V})\}\right]$
```

Enforcing No Profile Order Effect in the Carryover Effect Test Statistic. As mentioned in Section 3.5, we enforce the no profile order effect constraints when conducting the test of no carryover effect. We detail here how to implement this.

Since $\mathbf{X}^*$ represents the lag-1 profile values, we do not expect $\mathbf{X}^*$ alone (without interactions with $\mathbf{Z}^*$) to influence respondents' choice of the left versus right profile. Therefore, we force all main effects and interactions among $\mathbf{X}^*$ to be zero in the following way. Let $\mathbf{X}_i^{*L} \in \mathbb{R}^{\frac{J}{2} \times p}$ denote the columns of $\mathbf{X}_i^*$ that correspond to the left profile in $\mathbf{X}_i^*$. More formally, $\mathbf{X}_i^{*L} = [[(\mathbf{X}_{i1}^L); (\mathbf{Z}_{i1}^L)]^\top; [(\mathbf{X}_{i3}^L); (\mathbf{Z}_{i3}^L)]^\top; \dots; [(\mathbf{X}_{i,J-1}^L); (\mathbf{Z}_{i,J-1}^L)]^\top]$. If J is not even, consider up to $J-1$ instead. We define $\mathbf{X}_i^{*R}, \mathbf{Z}_i^{*L}, \mathbf{Z}_i^{*R}$ similarly. We also define $\mathbf{X}^{*L} = [\mathbf{X}_1^{*L}; \dots; \mathbf{X}_n^{*L}] \in \mathbb{R}^{\frac{nJ}{2} \times p}$ with $\mathbf{X}^{*R}, \mathbf{Z}^{*,L}, \mathbf{Z}^{*,R}$ defined similarly. Then, we append copies of the following: $[[(\mathbf{X}^{*R})^\top; (\mathbf{X}^{*L})^\top; (\mathbf{Z}^{*R})^\top; (\mathbf{Z}^{*L})^\top]^\top; [(\mathbf{X}^{*L})^\top; (\mathbf{X}^{*R})^\top; (\mathbf{Z}^{*R})^\top; (\mathbf{Z}^{*L})^\top]^\top; [(\mathbf{X}^{*R})^\top; (\mathbf{X}^{*L})^\top; (\mathbf{Z}^{*L})^\top; (\mathbf{Z}^{*R})^\top]^\top$ to the original data matrix $[(\mathbf{X}^{*L})^\top; (\mathbf{X}^{*R})^\top; (\mathbf{Z}^{*L})^\top; (\mathbf{Z}^{*R})^\top]^\top$, resulting in a total of $2nJ$ rows. Lastly, we also append copies of $[\mathbf{1} - \mathbf{Y}_i^*; \mathbf{1} - \mathbf{Y}_i^*; \mathbf{Y}_i^*]$ to the original response $\mathbf{Y}_i^*$. Appending the first copy of $[(\mathbf{X}^{*R})^\top; (\mathbf{X}^{*L})^\top; (\mathbf{Z}^{*R})^\top; (\mathbf{Z}^{*L})^\top]^\top$ to the original data matrix enforces the familiar constraint in Equation (6) using Lemma E.1. The remaining copies force all the main effects and interactions among $\mathbf{X}^*$ to be zero by appealing to the same reasoning in Lemma E.1.

F CRT Procedure for Testing Extensions

In this appendix, we describe in further detail how to carry out all the resampling procedures for the tests introduced in Sections 3.4 and 3.5.

F.1 Testing H_0^{General}

When testing H_0^{General}, the resampling procedure is different than Algorithm 1 because Equation (8) forces us to hold $(h(\mathbf{X}), \mathbf{Z})$ constant rather than holding only $\mathbf{Z}$ constant. The conditional distribution of $\mathbf{X} \mid (h(\mathbf{X}), \mathbf{Z})$ constrains $\mathbf{X}$ to be randomized only within the factor levels of interest. For concreteness, consider the immigration example in Section 4. In this case, $h(\mathbf{X})$ groups levels *Mexico* and *Europe* to one output while keeping all the other countries of origin the same. Therefore, when obtaining the resamples for the "country of origin" factor, we keep all countries except *Mexico* and *Europe* constant, i.e., countries such as *China* do not get re-randomized. For the entries corresponding to *Mexico* and *Europe*, we resample values *Mexico* and *Europe* with probabilities $(0.25, 0.75)$ respectively (since 3 countries make up Europe). We present this resampling procedure in Algorithm 2. Lastly, we remark that Algorithm 2 remains the same when testing $\overline{H}_0^{\text{General}}$ except we repalce $\mathbf{X}$ with $\overline{\mathbf{X}}$.

```
Algorithm 3: CRT: Profile Order Effect
  Input: Data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$, test statistic $T(\mathbf{x}, \mathbf{y}, \mathbf{z})$, total number of re-samples $B$;
  for $b = 1, 2, \ldots, B$ do
      Sample $nJ$ independent Bernoulli(0.5) independently of $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. $E^b$ is the index set corresponding to values of 1 in the $nJ$ Bernoulli's;
  Output: $p$-value $:= \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{X}^{(E^b)}, \mathbf{Y}^{(E^b)}, \mathbf{Z}^{(E^b)}) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})\}\right]$
```

```
Algorithm 4: CRT: Carryover Effects
  Input: Data $(\mathbf{X}^*, \mathbf{Y}^*, \mathbf{Z}^*)$, test statistic $T(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$, total number of re-samples $B$;
  for $b = 1, 2, \ldots, B$ do
      Sample $(\mathbf{X}_{ij}^b, \mathbf{Z}_{ij}^b)$ from the distribution of $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ independently of $\mathbf{Y}$ for $i = 1, 2, \ldots, n$ and $j = 1, 3, \ldots J-1$. Let $(\mathbf{X}_i^b)^* = [[\mathbf{X}_{i1}^b; \mathbf{Z}_{i1}^b]^\top; [\mathbf{X}_{i3}^b; \mathbf{Z}_{i3}^b]^\top; \ldots; [\mathbf{X}_{i,J-1}^b; \mathbf{Z}_{i,J-1}^b]^\top]$ and $(\mathbf{X}^b)^* = [(\mathbf{X}_1^*)^b; (\mathbf{X}_2^*)^b; \ldots; (\mathbf{X}_n^*)^b]$;
  Output: $p$-value $:= \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbb{1}\{T((\mathbf{X}^b)^*, \mathbf{Y}^*, \mathbf{Z}^*) \geq T(\mathbf{X}^*, \mathbf{Y}^*, \mathbf{Z}^*)\}\right]$
```

F.2 Testing the Regularity Assumptions

Profile Order Effect. We first describe testing the assumption of no profile order effect without imposing SUTVA. For any $E \subset \{1, \ldots, nJ\}$, let $\mathbf{x}^{(E)}$ swap the left and right profile values in rows E of $\mathbf{x}$ while leaving the remaining rows unchanged. We similarly define $\mathbf{z}^{(E)}$. Then, let $\mathbf{Y}^{(E)}(\mathbf{x}, \mathbf{z})$ flip the bits of (replace 1's with 0's and vice versa) the entries of $\mathbf{Y}(\mathbf{x}, \mathbf{z})$ corresponding to indices in E while leaving the remaining entries unchanged. For example, for simplicity, assume no $\mathbf{Z}$ while $nJ = 3$, $E = \{1\}$, $\mathbf{x} = [(G, F); (P, G); (M, F)]$ (the left profile values come first), and $\mathbf{Y}(\mathbf{x}) = (1, 0, 1)$. Then $\mathbf{x}^{(E)} = [(F, G); (P, G); (M, F)]$ and $\mathbf{Y}^{(E)}(\mathbf{x}) = (0, 0, 1)$. The observed $\mathbf{Y}^{(E)}$ is defined similarly. We can formally state the null hypothesis of no profile order effect as follows:

$$H_0^{\text{Order}} : \mathbf{Y}(\mathbf{x}, \mathbf{z}) \overset{d}{=} \mathbf{Y}^{(E)}(\mathbf{x}^{(E)}, \mathbf{z}^{(E)}) \text{ for all } E \subset \{1, \ldots, nJ\}, \mathbf{x} \in \mathcal{X}, \text{ and } \mathbf{z} \in \mathcal{Z}.$$

This null hypothesis states that for all possible reorderings of the left and right profiles there is no causal impact on which profile is chosen. For the resampling procedure, we hold the realized values of all $(\mathbf{X}, \mathbf{Z})$ constant while only resampling the subset E, i.e., drawing nJ independent Bernoulli coin flips to determine which of the nJ rows to include as part of E as described above. Algorithm 3 details the procedure to calculate the CRT p-value for testing H_0^{Order}. Lastly, to not enforce Equation (6) in $T_{\text{HierNet}}^{\text{Order}}$, we fit HierNet on the original data matrix $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ rather than D^c.

Carryover Effect. When testing the carryover effect, we need to hold the even numbered tasks $\mathbf{Z}^*$ fixed while resampling all odd numbered tasks $\mathbf{X}^*$. Therefore, we resample all factors $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ for $j = 1, 3, \ldots, J-1$ from the experimental distribution for all the factors while holding $(\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ for $j = 2, 4, \ldots, J$ constant. Algorithm 4 details the procedure to calculate the CRT p-value for testing $H_0^{\text{Carryover}}$.

Fatigue Effect. To carry out the CRT to test the fatigue effect, we re-sample only the task evaluation order index $\mathbf{F}$ for each respondent uniformly from the set of all permutations on $\{1, \ldots, J\}$, denoted as Π_J, while holding all the experimental factor values fixed. Algorithm 5 details the procedure to calculate the CRT p-value for testing H_0^{Fatigue}.

Algorithm 5: CRT: Fatigue Effect

```
Input: Data $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{F})$, test statistic $T(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{f})$, total number of re-samples $B$;
for $b = 1, 2, \dots, B$ do
    Sample $\mathbf{F}^b$ uniformly from $\Pi_J$ the set of all permutations on $\{1, \dots, J\}$;
Output: $p$-value $:= \frac{1}{B+1}\left[1 + \sum_{b=1}^{B} \mathbb{1}\{T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{F}^b) \geq T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{F})\}\right]$
```

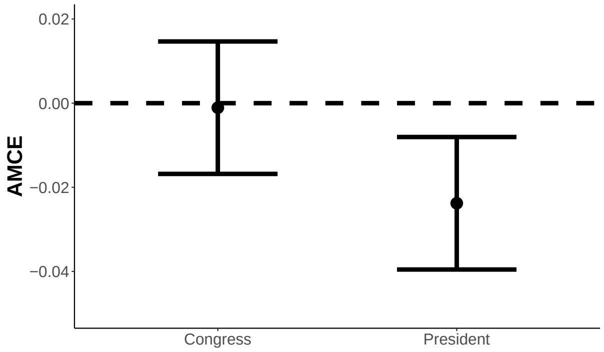

Figure 7: The estimated Average Marginal Component Effect (AMCE) of candidate's gender in the Ono and Burden (2018) study. We present the estimates for congressional candidates (left) and presidential candidates (right). The 95% confidence intervals are also shown.

G Additional Conjoint Application - Role of Gender in Political Candidate Evaluation

We present here an additional empirical application concerning the role of gender discrimination in political candidate evaluations. Recently, several scholars have used conjoint analysis to study the role of gender discrimination in candidate evaluation (e.g., Ono and Burden, 2018; Teele, Kalla and Rosenbluth, 2018). We revisit the study by Ono and Burden (2018) which examines whether voters prefer candidates of one gender over those of another after controlling for other candidate characteristics.[14] The study is based on a sample of voting-eligible adults in the U.S. collected in March 2016 and also uses the forced-choice conjoint design. The following 13 factors are independently and uniformly randomized across their levels: gender, age, race, family, experience in public office, salient personal characteristics, party affiliation, policy area of expertise, position on national security, position on immigrants, position on abortion, position on government deficit, and favorability among the public (see Table 6 in Appendix H and the original article for details). The survey also contains information about the respondents' educational background, gender, age, region, social class, partisanship, political interest, and ethnocentrism. There were 1,583 respondents each given 10 tasks, resulting in 15,830 observations, half of which were for congressional candidates and the other half for presidential candidates.

The original study yields a negative estimate of the AMCE of female candidates, relative to male counterparts, for presidential candidates. However, the estimated AMCE of female candidates is not statistically distinguishable from zero for congressional candidates, based on a simple t-test for the coefficient of *Male* from the linear regression with cluster standard errors. This finding led to the authors' conclusion that gender

[14] This study treats gender as a binary factor with levels *Male* and *Female*.

	CRT	AMCE	Profile order effect	Carryover effect	Fatigue effect
p-values	0.026	0.93, 0.40	0.15	0.97	0.66

Table 3: 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 whether candidate's "gender" matters for voters' preferences of Congressional candidates. The second AMCE-based p-value for "gender" corresponds to a fair comparison with the CRT-based p-value by additionally testing the "gender" interaction with the candidate's "party affiliation" (*Democratic* or *Republican*). The remaining columns report the p-values for testing no profile order effect, no carryover effect, and no fatigue effect for the respective application.

discrimination "is limited to presidential rather than congressional elections" (p. 583). Figure 7 reproduces these AMCE estimates. Like the immigration example, the authors fit a linear regression with all fourteen factors as predictors to obtain these estimates.

In this section, we use the CRT to formally test whether the gender of congressional candidates matters in *any way* for voters' preferences while controlling for the other candidate characteristics. The rejection of this null hypothesis would indicate that gender does matter even for congressional candidates.

G.1 Application Results

To test whether or not the gender of Congressional candidates matters in voter preferences, we use the same data as the one used in Ono and Burden (2018). The Congressional dataset consists of 7,915 observations with 5 tasks performed by each of $n = 1,583$ respondents. Our main factor of interest $\mathbf{X}$ is a binary variable representing *male* or *female*. In addition, we use the remaining 12 randomized factors $\mathbf{Z}$ and all the respondent characteristics $\mathbf{V}$. We test the main null hypothesis H_0 introduced in Section 3.2.

As mentioned in Section 3.3, the use of substantive knowledge can improve the power of the test. To demonstrate this, we leverage the Presidential candidate dataset from the same conjoint experiment to find the strongest interaction with the gender of candidates. We then include this interaction term as an additional main effect in HierNet when computing the test statistic given in Equation (10).[15] By including it as a main effect, HierNet applies less shrinkage on this interaction term. In addition, HierNet will consider potential three-way interactions involving this interaction term and other variables in $\mathbf{Z}$. The power will be greater if strong interactions in the Presidential candidate data are also present in the Congressional candidate data.

To find the strongest interaction in the Presidential candidate dataset, we obtain a CRT p-value for each variable in $(\mathbf{Z}, \mathbf{V})$ with a test statistic that focuses on the interaction strength for the corresponding variable under consideration. Specifically, the test statistic uses Lasso logistic regression with all main effects of $(\mathbf{X}, \mathbf{Z}, \mathbf{V})$ and an additional interaction between $\mathbf{X}$ and one variable from $(\mathbf{Z}, \mathbf{V})$. We choose the variable with the lowest p-value as the strongest interaction. The Presidential data shows that the candidate's "party affiliation" (*Democratic* or *Republican*) had the most significant interaction with their "gender". Appendix G.2 contains further details and a robustness check by repeating the same analysis but choosing the variable with the second lowest p-value as the additional main effect.

As shown in the second row of Table 3, the CRT p-value using the HierNet test statistic is 0.026, showing that gender may matter even for Congressional candidates. We find the largest two interactions in the observed test statistic were two three-way interactions: one between "gender", "party affiliation", and "respondent's political interest" and the other between "gender", "party affiliation", and the "respondent's party

[15] Our test statistic $T_{\text{HierNet}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{V})$ is not only a function of the Congressional dataset $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{V})$ but also the entire presidential dataset $(\mathbf{P})$. Therefore, we must now hold all $(\mathbf{Z}, \mathbf{V}, \mathbf{P})$ fixed in the resampling procedure. However, this does not change Algorithm 1 and the resulting p-value remains valid because $\mathbf{X} \mid (\mathbf{Z}, \mathbf{V}, \mathbf{P})$ is still an independent fair coin flip between the levels of *male* and *female*.

affiliation”. This result is consistent with the findings in de la Cuesta, Egami and Imai (2022), which suggests the existence of higher order interactions involving party affiliation. We assess the role of interaction effects using the same procedure described in the immigration example in Section 4. The resulting CRT p-value from a Lasso logistic regression with only the main effects of $(\mathbf{X}, \mathbf{Z}, \mathbf{V})$ and the additional interaction with “gender” and “party affiliation” is 0.15, suggesting that the other interactions were also helpful in detecting significance.

For comparison, we compute the p-value based on the estimated AMCE of “gender” for Congressional candidates as presented in Figure 7. Similar to the common strategy used for the immigration conjoint experiment, we fit a linear regression model using “gender” as the sole predictor while clustering standard errors by respondents. We find that the p-value is 0.89. However, since the CRT leveraged the Presidential candidate data to up-weight the interaction with “party affiliation”, we also create a fair comparison by obtaining analogous p-values for the AMCE. To do this, we again fit a linear regression using “gender” but with an additional main effect of “party affiliation” and the interaction of “gender” and “party affiliation”. We then report the p-value from a F-test for both the main effect of “gender” and the interaction with “party affiliation”. The resulting p-value is 0.40, showing that the AMCE-based result remains statistically insignificant. Finally, the last three columns in the second row show no evidence that the regularity assumptions are violated for this conjoint experiment.[16]

G.2 Leveraging Presidential Data in the Gender Application

We detail here how we leverage the Presidential candidate data to find the strongest interaction, as mentioned above. Because we are only interested in whether the additional interactions are significant, we run the Lasso logistic regression with main effects of $(\mathbf{X}, \mathbf{Z}, \mathbf{V})$ and one interaction between $\mathbf{X}$ (“gender”) and one variable in $(\mathbf{Z}, \mathbf{V})$, producing one CRT p-value for each variable in $(\mathbf{Z}, \mathbf{V})$. For example, when including an interaction between “gender” and “profession”, we include both within-profile and between-profile interactions between all levels of “gender” and “profession”. Consequently, the test statistic for testing the interaction between $\mathbf{X}$ (“gender”) and factor ℓ of $\mathbf{Z}$ is:

$$T_{\text{Presidential}}^{\text{Candidate Factor, } \ell} = \sum_{k=1}^{K} \sum_{k'=1}^{K_\ell} (\hat{\gamma}_{1\ell kk'} - \bar{\gamma}_{1\ell k'})^2 + \sum_{k=1}^{K} \sum_{k'=1}^{K_\ell} (\hat{\delta}_{1\ell kk'} - \bar{\delta}_{1\ell k'})^2. \tag{14}$$

The test statistic for testing the interaction between $\mathbf{X}$ and factor m of the respondent characteristic $\mathbf{V}$ is:

$$T_{\text{Presidential}}^{\text{Respondent Characteristic, } m} = \sum_{k=1}^{K} \sum_{w=1}^{L_m} (\hat{\xi}_{1mkw} - \bar{\xi}_{1mw})^2, \tag{15}$$

where all coefficient estimates refer to the same corresponding estimates in Equation (10) and $K = 2$ refers to levels *male* and *female*. Lastly, when running the CRT we similarly enforce the constraints in Equation (6) (along with the constraints on the respondent characteristics) by fitting the Lasso logistic regression on the appended D^c data matrix.

Table 4 shows the resulting p-value for each variable in $(\mathbf{Z}, \mathbf{V})$. Many of the variables have p-values lower than 0.1. The variables such as “position on immigrants”, “position on abortion”, “position on government deficit”, and “position on national security”, are all related to the disparate views Democratic and Republican candidates may have, in line with “party affiliation” being the most significant. Even among the respondent

[16] We only test the no profile order effect assumption for Congressional candidates because this is the data relevant to the research question. However, for testing the carryover effect and fatigue effect, we use the full dataset including the Presidential candidates in order to increase power.

Variable	***p*-value**
Age	0.060
Race	0.31
Family	0.22
Experience in public office	0.30
Salient personal characteristic	0.45
Party affiliation	0.0049
Policy area of expertise	0.25
Position on national security	0.067
Position on immigrants	0.067
Position on abortion	0.022
Position on government deficit	0.032
Favorability rating among public	0.63
Respondent gender	1.00
Respondent education	1.00
Respondent age	1.00
Respondent class	0.41
Respondent region	1.00
Respondent race	1.00
Respondent partisanship	0.11
Respondent thought on Hillary Clinton	1.00
Respondent interest in politics	0.43
Respondent political ideology	0.042

Table 4: Resulting *p*-values using the CRT Lasso logistic regression with main effects of $(\mathbf{X}, \mathbf{Z}, \mathbf{V})$ and an additional interaction between $\mathbf{X}$ and one variable in $(\mathbf{Z}, \mathbf{V})$ from the Presidential candidate data. The test statistic captures the interaction terms with each variable in $(\mathbf{Z}, \mathbf{V})$ as shown in Equation (14) and Equation (15), respectively. All respondent characteristic variables are labelled with "respondent".

characteristics, the respondent's political ideology, a measure of how conservative or liberal the respondent is, is the most significant variable. We conduct a robustness analysis by repeating the above analysis in Appendix G.1 but using the second most significant variable, the "position on abortion" factor, to interact and include as the additional main effect in HierNet, though we note that "position on abortion" is quite a bit less significant than "party affiliation" with more than four times as large a *p*-value. The resulting *p*-value is 0.078. Although this is not as significant as the main analysis, it still provides suggestive evidence that gender plays a role in voting for Congressional candidates.

H Data Description

Tables 5 and 6 present all the factors and their respective levels used in the immigration conjoint experiment (Section 2.1) and the gender candidate conjoint experiment (Appendix G), respectively.

Factor for Immigration Conjoint Experiment	**Factor Levels**
Education level	No Formal Education, Fourth Grade, Eight Grade, High School, Two Years College, College Degree, Graduate Degree
Gender	Female, Male
Country of origin	Germany, France, Mexico, Philippines, Poland, India, China, Sudan, Somalia, Iraq
Language	Fluent English, Broken English, Tried to speak English but unable to, Spoke through an interpreter
Reason for application	Reunite with family members, Seek better job, Escape political/religious persecution
Profession	Gardener, Waiter, Nurse, Teacher, Child Care Provider, Janitor, Construction Worker, Financial Analyst, Research Scientist, Doctor, Computer Programmer
Job experience	No job training or prior experience, One to two years, Three to five years, More than five years
Employment plan	Has a contract with a U.S. employer, Does not have a contract with a U.S. employer, but has done job interviews, Will look for work after arriving in the U.S., Has no plans to look for work at this time
Prior trips to the U.S.	Never been to the U.S., Entered the U.S. once before on a tourist visa, Entered the U.S. once before without legal authorization, Has visited the U.S. many times before on tourist visas, Spent six months with family members in the U.S.

Table 5: All nine randomized factors and their respective levels in the conjoint experiment used in Hainmueller and Hopkins (2015).

I Computational Details

The HierNet test statistic introduced in Equation (5) is powerful but can be computationally expensive because the CRT requires a total of $B + 1$ cross-validated HierNet fits. To address this problem, we speed up HierNet in three ways. First, we reduce the default convergence tolerance for the optimization algorithm in the HierNet package from 10^{-6} to 10^{-3}. Second, following the "distillation" idea introduced in (Liu et al., 2020), we cross-validate the sparsity parameter lambda only through a HierNet fit of $\mathbf{Y}$ on $\mathbf{Z}$ without involving any $\mathbf{X}$. Because $(\mathbf{Y}, \mathbf{Z})$ remains constant for all $B + 1$ fits, we only need one cross-validation fit. Lastly, we initialize the starting parameters in the optimization algorithm with one HierNet fit that is uniformly and randomly chosen from the $B + 1$ HierNet fits. Since we uniformly choose one out of $B + 1$ HierNet fits as the initialization, this procedure still satisfies the exchangeability needed for the CRT's validity. Because many of the parameters estimated from the $B + 1$ different HierNet fits will likely be similar to each other, the initialization likely saves computation time.

Although the above procedure significantly reduces computational complexity, practitioners may worry if there is a significant loss of power from this simplification. Consequently, we plot in Figure 8 the original HierNet power curve shown in Figure 2 that leverages the aforementioned three speed-ups (in blue) and the

Factor For Gender Conjoint Experiment	**Factor Levels**
Gender	Male, Female
Age	36, 44, 52, 60, 68, 76
Race/ethnicity	White, Black, Hispanic, Asian American
Family	Single (never married), Single (divorced), Married (no child), Married (two children)
Experience in public office	12 years, 8 years, 4 years, No experience
Salient personal characteristics	Strong leadership, Really cares about people like you, Honest, Knowledgeable, Compassionate, Intelligent
Party affiliation	Republican, Democrat
Policy area of expertise	Foreign policy, Public safety (crime), Economic policy, Health care, Education, Environmental issues
Position on national security	Cut military budget and keep U.S. out of war, Maintain strong defense and increase U.S. influence
Position on immigrants	Favors guest worker program, Opposes guest worker program
Position on abortion	Pro-choice, Pro-life, Neutral
Position on government deficit	Reduce through tax increase, Reduce through spending cuts, Does not want to reduce
Favorability rating among public	34%, 43%, 52%, 61%, 70%

Table 6: All thirteen randomized factors and their respective levels in the conjoint experiment used in Ono and Burden (2018).

computationally slower HierNet power curve without the three speed-ups (in black). Figure 8 shows that the computational modifications have no significant impact on power.

J Inflated p-values for Logistic Regression

Although the AMCE is popular in conjoint analysis, especially among political scientists, there also exist model-based approaches. Logistic regression remains a popular model-based approach to conjoint analysis (McFadden, 1973; Green and Srinivasan, 1990; Campbell, Mhlanga and Lesschaeve, 2013). We explore in this section how this modeling approach can lead to invalid inference in conjoint analysis. When testing H_0, researchers may want to account for not only the main effects of $(\mathbf{X}, \mathbf{Z})$ but also all two-way interactions, as done similarly in HierNet, to reduce model misspecification. Under this scenario, we show through simulations in Figure 9 that even reasonable sample sizes and dimensions of $(\mathbf{X}, \mathbf{Z})$ can lead to invalid p-values, i.e., the type 1 error is greater than the desired $\alpha \in [0, 1]$.

We use a similar simulation setting as the one in Appendix D.1 but simplify it further. Since we are interested in showing how the p-values obtained from a logistic regression may be invalid in general, we do not have "left" or "right" profiles but only one profile, leading to the following data generating process,

$$\Pr(Y_i = 1 \mid X_i, \mathbf{Z}_i) \;=\; \text{logit}^{-1}\left[\beta_X X_i + \beta_Z^\top \mathbf{Z}_i + \gamma^\top (X_i \mathbf{Z}_i) + \tilde{\gamma}^\top (\mathbf{Z}_i \times \mathbf{Z}_i)\right].$$

All factors have four levels, and we similarly assume one factor of interest $q = 1$ while varying the number of other factors. Since we are interested in the behavior of the p-values under the null H_0, we force all effects of X on the response to be zero, i.e., $\beta_X = \gamma = 0$. For simplicity we also make all effects of $\mathbf{Z}$ zero, i.e., $\beta_Z = \tilde{\gamma} = 0$ and fix the sample size to $n = 5{,}000$. To reflect the researcher's desire to reduce model

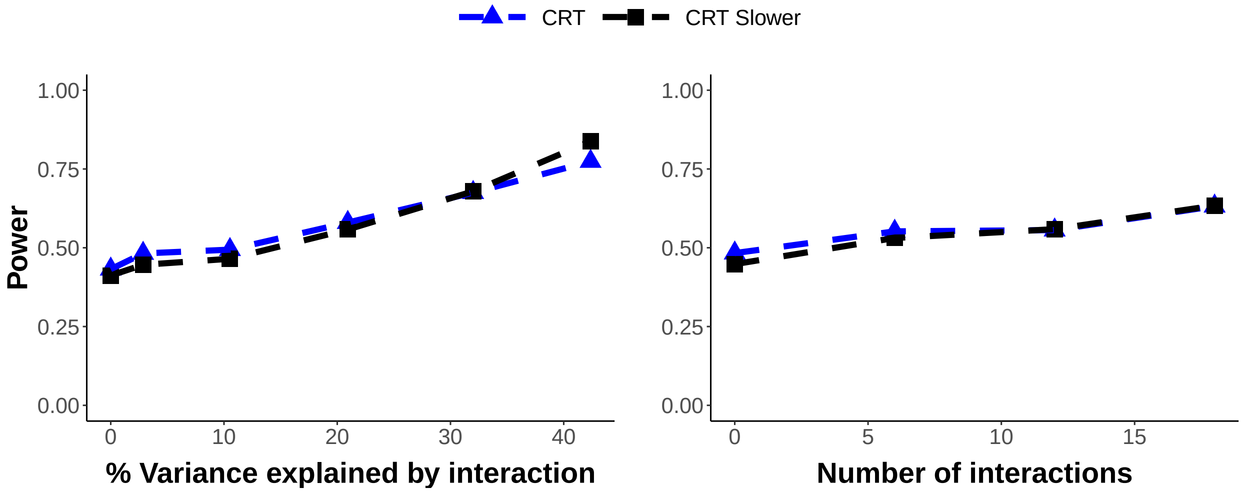

Figure 8: This figure represents the power of the original faster HierNet test statistic (blue triangles) with the three computational speedups and the slower HierNet test statistic (black squares) without the computational speedups in the same simulation setting as in Figure 2.

misspecification by accounting for all two-way interactions as HierNet does, we fit a logistic regression of $\mathbf{Y}$ on all main effects and two-way interactions of $(\mathbf{X}, \mathbf{Z})$. We then obtain a p-value for testing H_0 by an F-test that tests $\beta_X = \gamma = 0$. We obtain 1,000 Monte-Carlo p-values and plot the proportion of p-values less than $\alpha = 0.05$ in the left plot of Figure 9. We also vary the number of factors of $\mathbf{Z}$, which is shown in the x-axis of the left plot. On the right plot of Figure 9, we plot the histogram of the 1,000 p-values obtained when the number of factors of $\mathbf{Z}$ is 12.

Under the null hypothesis, we expect any valid p-value to have type 1 error control, i.e., $P(p\text{-value} \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$. The left plot of Figure 9 shows that only five other factors of $\mathbf{Z}$ is enough to cause the proportion of p-values less than $\alpha = 0.05$ to be noticeably inflated at 7%. The inflation becomes particularly apparent when there are twelve other factors of $\mathbf{Z}$, which causes the proportion of p-values less than 0.05 to be as high as 34%. The histogram on the right plot of Figure 9 visually shows how the p-values are clearly far from the expected uniform distribution and have an undesirable peak at zero, resulting in poor type 1 error control. This phenomenon is studied in (Candès and Sur, 2018) and arises because the p-values' validity in a logistic regression depends on a low-dimensional asymptotic result. We note that a conjoint analysis has typically more than ten factors, where each factor usually has more than three levels. Therefore, Figure 9 shows the potential dangers of using a model-based approach like the logistic regression to flexibly capture all interactions.

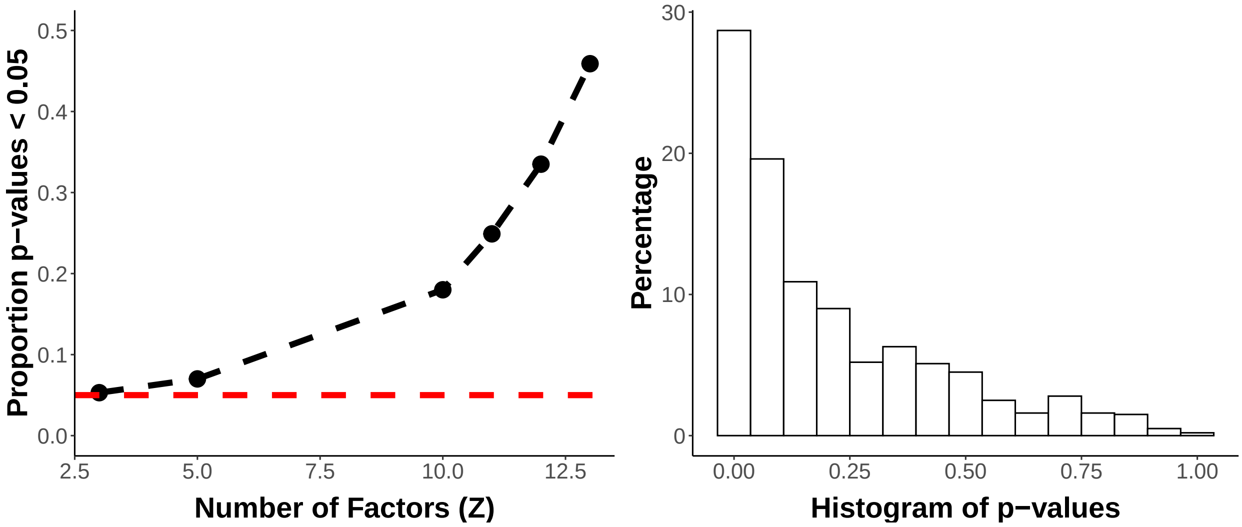

Figure 9: Inflated *p*-values from logistic regression. The left figure shows the proportion of *p*-values, obtained through a F-test from a logistic regression, less than $\alpha = 0.05$ when the number of other factors in $\mathbf{Z}$ is $(3, 5, 10, 11, 12, 13)$ and H_0 is true. The red dotted line at $\alpha = 0.05$ represents the expected proportion of *p*-values less than 0.05 if the *p*-values are valid. The right figure shows the histogram of 1,000 Monte-Carlo *p*-values when the number of other factors of $\mathbf{Z}$ is 12. The sample size is $n = 5,000$ and each factor has four factor levels. All Monte Carlo standard errors are below 0.016.