PA

Multiple Hypothesis Testing in Conjoint Analysis

Guoer Liu[1] **and Yuki Shiraito**[2]

[1] *Ph.D. Candidate, Department of Political Science, University of Michigan, Ann Arbor, MI, USA. E-mail: guoerliu@umich.edu*
[2] *Assistant Professor, Department of Political Science, University of Michigan, Ann Arbor, MI, USA; Center for Political Studies, 4259 Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104-2321, USA. E-mail: shiraito@umich.edu*
URL: shiraito.github.io

Abstract

Conjoint analysis is widely used for estimating the effects of a large number of treatments on multidimensional decision-making. However, it is this substantive advantage that leads to a statistically undesirable property, multiple hypothesis testing. Existing applications of conjoint analysis except for a few do not correct for the number of hypotheses to be tested, and empirical guidance on the choice of multiple testing correction methods has not been provided. This paper first shows that even when none of the treatments has any effect, the standard analysis pipeline produces at least one statistically significant estimate of average marginal component effects in more than 90% of experimental trials. Then, we conduct a simulation study to compare three well-known methods for multiple testing correction, the Bonferroni correction, the Benjamini–Hochberg procedure, and the adaptive shrinkage (Ash). All three methods are more accurate in recovering the truth than the conventional analysis without correction. Moreover, the Ash method outperforms in avoiding false negatives, while reducing false positives similarly to the other methods. Finally, we show how conclusions drawn from empirical analysis may differ with and without correction by reanalyzing applications on public attitudes toward immigration and partner countries of trade agreements.

Keywords: conjoint analysis, multiple hypothesis testing, false discovery rate, empirical Bayes

1 Introduction

Conjoint analysis has been one of the most widely used survey experimental designs in political science, since Hainmueller, Hopkins, and Yamamoto (2014) defined the average marginal component effect (AMCE) as an estimand in conjoint designs and developed a simple estimator. In a typical conjoint experiment, respondents are asked to assess pairs of profiles and choose a preferred one in each paired comparison. The profiles consist of theoretically relevant attributes that reflect multiple dimensions of respondents' preferences, and the attributes are independently randomized across the profiles. For instance, Hainmueller and Hopkins (2015) examined individual-level attributes of a hypothetical immigrant such as gender, education, occupation, and the country of origin. Using a conjoint experiment, the authors estimated the AMCEs of those attributes on the probability that the immigrant's admission is preferred. After this canonical study, conjoint designs are used to study voting (e.g., Carnes and Lupu 2016; Incerti 2020; Ono and Burden 2019; Teele, Kalla, and Rosenbluth 2018), bureaucratic selection (e.g., Liu 2019; Oliveros and Schuster 2018), and other types of multidimensional decision-making (e.g., Fournier, Soroka, and Nir 2020; Sen 2017; Shafranek 2021).[1]

Conjoint analysis "enables researchers to estimate the causal effect of multiple treatment components and assess several causal hypotheses simultaneously" (Hainmueller *et al.* 2014, 1). This property is extremely valuable substantively. Since a number of factors contribute to decisions, isolating the causal effect of each factor under all combinations of the others would require impractically many experimental conditions. Conjoint analysis overcomes this difficulty by identifying the AMCEs of multiple attributes at once. AMCE is the causal effect of an attribute

1 For a more comprehensive list of conjoint experiment papers, see de la Cuesta, Egami, and Imai (2022).

Political Analysis (2023)

DOI: 10.1017/pan.2022.30

Corresponding author
Yuki Shiraito

Edited by
Jeff Gill

averaged over all profiles of the other attributes, and it has an intuitive interpretation (Bansak *et al.* 2022). The combination of conjoint designs and AMCE enables researchers to estimate the effects of multiple features simultaneously.

Despite this substantive advantage, producing many estimates leads to a statistically undesirable property, multiple hypothesis testing. Testing multiple hypotheses in statistical inference is problematic, because the more null hypotheses are tested, the more likely at least one of them is to be rejected, even if all of them are true. The pre-specified critical value, conventionally set at .05, represents the probability of falsely rejecting the null hypothesis assuming that only one is tested. When several hypotheses are tested simultaneously, the test procedure needs to be modified. In political science, multiple testing has not been considered as a common concern, because studies usually intend to examine only a few hypotheses.[2] However, since conjoint analysis is designed exactly for estimating multiple effects, it cannot avoid multiple statistical tests. The immigration application in Hainmueller *et al.* (2014), for example, involves 41 hypothesis tests in total. Theoretically, even if all 41 AMCEs are zero in truth, estimates of two AMCEs will be statistically distinguishable from zero on average across experimental trials. The promise of conjoint analysis implies many statistical tests, and false-positive conclusions may follow as a result.

To the best of our knowledge, existing studies in political science using conjoint analysis do not correct for multiple testing in their main analysis except for Hainmueller, Hangartner, and Yamamoto (2015), which use the Bonferroni correction (BC). A few others, for example, Clayton, Ferwerda, and Horiuchi (2021), confirm their results with corrections as robustness checks. In fact, researchers are aware that multiple hypothesis testing is an inherent problem with conjoint designs. Bansak *et al.* (2021a, 28) point out that the concerns about multiple comparisons make pre-registration and pre-analysis plans especially valuable. However, no systematic assessments have been done on the severity of the problem in the literature. Moreover, to avoid haphazard selection, applied researchers need guidance on which correction method among several well-known ones is appropriate under their circumstances.

In this paper, we quantify the multiple testing problem in conjoint designs and assess easy-to-implement correction strategies. First, we show that under a classic conjoint setup the standard analysis pipeline produces at least one statistically significant AMCE estimate in more than 90% of experimental trials even when all AMCEs are zero.

Second, we compare the strengths and limitations of two well-known correction methods: the BC (Bland and Altman 1995; Dunn 1961) and the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995). In addition, we introduce a recently developed correction method, adaptive shrinkage (Ash) (Gerard and Stephens 2018; Stephens 2017). While none of the methods completely resolves the problem, all of them are better than the standard practice. Among the three methods, the BC guards against false-positive conclusions, but the cost of false-negative conclusions can be significant. On the other hand, the BH is the least susceptible to false-negative conclusions, but it is most lenient with false positives. The Ash takes a middle ground.

To illustrate how different correction methods perform in real data, we reanalyze two conjoint design applications. The first application using the dataset of Hainmueller *et al.* (2014) demonstrates that results corrected by the Ash are more consistent with the original authors' argument than other methods. Second, reanalysis of an experiment in Vietnam about the selection of trade agreement partners (Spilker, Bernauer, and Umaña 2016) shows that corrected methods remove the statistical significance on an attribute that is hard to interpret given Vietnam's security policy.

Compared to other studies that propose improvements on conjoint survey designs, this paper exclusively focuses on statistical inference. Existing studies have examined estimands and

2 Recently, however, multiple testing correction is used more often as robustness checks than before. We thank Yusaku Horiuchi for pointing this out.

interpretation (Abramson *et al.* 2020; Abramson, Koçak, and Magazinnik 2022; Bansak *et al.* 2022; de la Cuesta *et al.* 2022; Egami and Imai 2019; Ganter 2021), implementation (Bansak *et al.* 2018; 2021b), social desirability bias (Horiuchi, Markovich, and Yamamoto 2020), and subgroup analysis (Clayton *et al.* 2021; Leeper, Hobolt, and Tilley 2020). While this paper does not directly engage with any of these, the issue of multiple testing is relevant to any statistical inference with conjoint analysis due to its multiple comparison feature, unless the purpose of the analysis is exclusively exploration of higher-order interaction effects (Egami and Imai 2019).

This paper proceeds in four sections. First, we discuss why multiple testing is a problem in conjoint designs and quantify the problem. Then, we examine three correction methods and compare their performance in a simulation study. Third, we apply the correction methods to two conjoint experiment datasets. Finally, we summarize the paper and discuss suggested analysis pipelines for conjoint designs in the concluding section.

2 False-Positive Findings in Conjoint Analysis

When a large number of hypothesis tests are conducted, some reject null hypotheses purely by chance. With the conventional significance level of .05, a test rejects a true null hypothesis with probability .05. That is, the test tolerates five false positives out of 100 experimental trials on average. However, the probability that *at least one of multiple tests* rejects its null hypothesis can be much larger depending on the number of hypotheses. When 10 hypotheses are tested, this probability, known as the *familywise error rate* (FWER), is 1 – Pr(None of the 10 tests rejects the null) = $1 - (1 - .05)^{10} = .401$. If the number of tests is 20, the FWER increases to .642. (See Section A of the Supplementary Material.) Since the number of hypotheses is greater than 20 in most conjoint experiments, the problem is even more severe—in fact, it is almost guaranteed that at least one AMCE will be deemed statistically distinguishable from zero in any conjoint experiment, even if all AMCEs are zero in truth.

To illustrate how likely conjoint experiments may produce false-positive findings, we conducted a simulation study. Simulated datasets are generated from the conjoint design of Hainmueller *et al.* (2014). The design consists of nine attributes with total 50 levels, and therefore requires 41 comparisons excluding a reference level in each attribute. The forced-choice design is simulated by

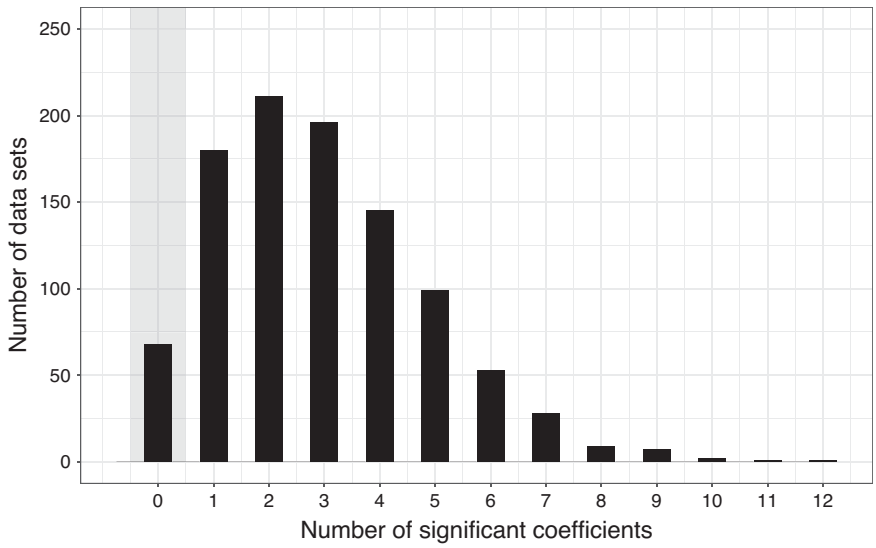

Figure 1. False-positive results of estimated AMCEs when all null hypotheses are true. Each bar presents the number of datasets (*y*-axis) for each number of statistically significant estimates (*x*-axis), with the truth (no significant findings) shaded by gray.

coarsening linear continuous responses into a binary choice in each pair of profiles. A total of 1,000 simulation datasets are generated under the scenario that the true AMCEs of all attributes are zero. In particular, the individual marginal component effect (MCE) is generated from $\mathcal{N}(.06, .015^2)$ for a half of the respondents and from $\mathcal{N}(-.06, .015^2)$ for the other half. We estimate AMCEs for each simulated dataset following the standard analysis pipeline for conjoint analysis and test the null hypothesis that each AMCE is zero.[3]

Figure 1 shows that only less than 75 out of 1,000 experimental trials correctly conclude that none of the attribute levels has any average effect. In other words, more than 90% of experiments may produce false-positive findings. Although we observe that the rate of false-positive findings is a little lower (around 80%) under some other simulation settings (see Section B of the Supplementary Material), the high false-positive rate is concerning for applied research.

3 Multiple Testing Correction Methods

This section briefly introduces two popular methods, BC and BH procedure, and a recently developed method, Ash. Then, the respective advantages and limitations of these methods will be illustrated by Monte Carlo simulations.

3.1 Bonferroni Correction

The BC (Bland and Altman 1995; Dunn 1961) reduces the FWER by using a more stringent threshold as the number of tests increases. To control the FWER below α, the BC tests each hypothesis at the significance level $\alpha/(\text{number of tests})$. For instance, when five hypotheses are tested at the conventional 5% level, each test is conducted at the 1% level. The BC is easiest to implement among the methods to control the FWER, since researchers only need to implement the standard test procedure and construct confidence intervals with a new significance level.

One caveat is that the BC can be overly conservative. In many applications, the BC reduces the FWER substantially lower than the level set by the user. Hence, the BC may suffer low statistical power and false-negative findings. We illustrate this point later in our simulation study.

Another critique of the BC is that the total number of tests in a "family" cannot be unambiguously defined and tracked (Sjölander and Vansteelandt 2019). Hochberg and Tamhane define *family* as "[a]ny collection of inferences for which it is meaningful to take into account some combined measure of errors" (1987, 5). While conjoint designs clearly pre-specify the number of attribute levels, researchers often conduct many tests to ensure survey quality such as balance and attention checks. Moreover, many applications include subgroup comparisons (Leeper *et al.* 2020). It may not be obvious which tests should be included in the "family" when using the BC.

While the decision on the number of tests may increase the researchers' degree of freedom, this problem should be ameliorated by pre-registration, as Bansak *et al.* (2021a) suggest for conjoint experiments in general. What constitutes a family depends on whether the type of research is exploratory or confirmatory. "In purely exploratory research the question of interest (or lines of inquiry) are generated by data-snooping. In purely confirmatory research they are stated in advance. Most empirical studies combine aspects of both types of research" (Hochberg and Tamhane 1987, 5). Discussing this issue in greater detail is beyond the scope of this paper, but pre-registration will ameliorate this ambiguity to some extent. In the conclusion section, we provide a recommendation checklist for conjoint users.

3.2 Benjamini–Hochberg Procedure

The BH procedure (Benjamini and Hochberg 1995) controls another measure of false-positive findings, the false discovery rate (FDR), which is defined as

3 For greater details of the simulation settings, see Section B of the Supplementary Material.

$$\text{FDR} \equiv \mathbb{E}\left[\frac{\#\text{ false discoveries}}{\#\text{ total discoveries}}\right].$$

The FDR indicates the average proportion of false positives among all statistical findings. Therefore, lowering the FDR implies that researchers can be more confident in their findings. The BH is a method for containing the FDR under a pre-set level α. The value of α is commonly set to .05, that is, 5% of null hypothesis rejections are false positives on average. The key idea of the BH is to remove some findings after conducting standard hypothesis tests. In other words, it prunes significant estimates so that researchers obtain fewer false findings.

The BH is a rank-based method with four steps. (1) For m hypotheses, an m-vector of p-values is produced. (2) Rank the p-values in the ascending order and index by i. (3) Define $k \equiv \max\{i : p_i \leq \alpha \times i/m, 0 \leq i \leq m\}$. (4) Reject null hypotheses H_i for $i = 1, 2, \ldots, k$, whose p-values are smaller than or equal to p_k, or reject none if k does not exist.

Although discussing theoretical properties of the BH (e.g., Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001) is beyond the scope of this paper, the BH is less susceptible to false-negative conclusions than the BC, because it accepts all statistically significant findings in its first step. However, the BH eliminates fewer false-positive findings. Moreover, the BH does not offer confidence intervals because it uses the p-values. We illustrate these limitations below by simulations and applications.

3.3 Adaptive Shrinkage

The Ash is a recently-proposed, empirical Bayes approach to controlling the FDR developed by Stephens (2017) and Gerard and Stephens (2018). Applied researchers can easily incorporate the Ash in conjoint analysis routine using the `ashr` package in **R** (Stephens *et al.* 2020).

The basic idea of the Ash is *post hoc* regularization of estimated coefficients using a spike-and-slab prior (see Figure 2). Regularization, in general, decreases the sampling variance of an estimator by introducing additional information into estimation. For the Ash, the spike-and-slab prior is such auxiliary information. On the one hand, the spike part reflects the fact that some estimates are false positives, inducing estimates to be zero with a certain probability. On the other hand, the slab part allows estimates to be non-zero with the remaining probability. As a result, the Ash shrinks estimated coefficients and produces narrower confidence intervals and smaller mean squared error. As shown in Section 5, the Ash moves point estimates of small absolute

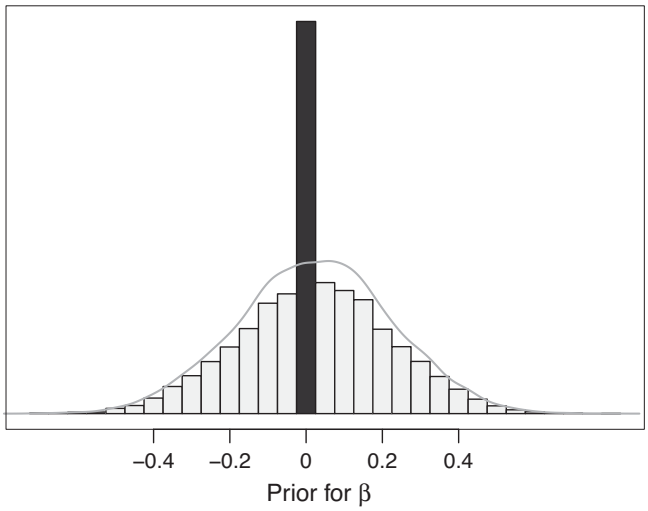

Figure 2. Example of the spike-and-slap prior distribution. The spike (point mass) is at zero, and the slap (gray curve) follows a normal distribution.

values toward zero and removes their statistical significance. By contrast, large point estimates are preserved and their confidence intervals are shortened.

Formally, let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ denote estimates for J attribute levels, let $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)$ denote point estimates of $\boldsymbol{\beta}$, and let $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_J)$ be the standard errors of $\hat{\boldsymbol{\beta}}$. Consider the posterior distribution of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{s}}$:

$$p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}}) \propto p(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \hat{\boldsymbol{s}}) p(\boldsymbol{\beta}|\hat{\boldsymbol{s}}). \tag{1}$$

The likelihood in Equation (1) is the sampling distribution of $\hat{\boldsymbol{\beta}}$ approximated by the normal distribution with mean $\boldsymbol{\beta}$ and variance $\hat{\boldsymbol{s}}^2$. To regularize a large number of estimates, independent spike-and-slab prior distributions are placed. Since the Ash is an empirical Bayes method, the mixture probabilities of the spike-and-slab prior are estimated by maximizing the penalized likelihood and then the posterior parameters are estimated using the prior parameter estimates. The confidence intervals are constructed based on the posterior distribution of $\boldsymbol{\beta}$. Section C.1 of the Supplementary Material provides a greater detail of the model and estimation.

The Ash delivers an additional benefit because of the shrinkage property. Its regularization leads to smaller mean squared errors of the point estimates. This is attractive because in many social science applications, researchers are interested not only in "whether factor X affect respondents' choice," but also in "to what extent." The classic immigrant conjoint experiment, for instance, found a bonus for some education relative to no formal education. When researchers would like to estimate the amount of the education bonus, the other correction methods do not reduce the sampling error of point estimates. The Ash, however, enables us to get more precise estimates in a principled manner. Section C.2 of the Supplementary Material illustrates this point by simulations.

4 Comparing Correction Methods

This section examines the performance of the three methods by a series of simulations. In all simulations, we generate 1,000 samples from simulation experiment using the immigrant profile data of Hainmueller *et al.* (2014), and conduct hypothesis tests at the conventional significance level of .05. The total number of tests for the BC is set to the total number of comparisons of attribute levels and a reference category. First, we apply the correction methods to the case where the true AMCE is zero for all attributes (identical to Section 2). Second, we compare the correction performance in more realistic cases where some attributes have non-zero AMCEs.

4.1 Zero AMCEs

The results are summarized in Figure 3. As in Figure 1, the bars represent the number of datasets for each number of statistically significant estimates. Note that the black bars are identical to Figure 1. Figure 3 also shows the results of the BC, the BH, and the Ash with a mixture of uniform components and with a mixture of normal components.

All three correction methods dramatically reduce the probability of false findings. Because all null hypotheses are true, all simulations should result in zero significant coefficients. As we discussed in Section 2, more than 90% of experimental trials would produce at least one significant estimate without correction. By contrast, both the BC and the BH remove false findings in more than 90% of simulation datasets. The Ash performs even better. It eliminates almost all false-positive findings.

4.2 Non-Zero AMCEs

It is perhaps rare that all AMCEs are zero in applications, because attributes are designed to capture promising theoretical hypotheses. We consider two sets of more realistic simulations where some AMCEs are not zero to see how the correction methods perform in such settings.

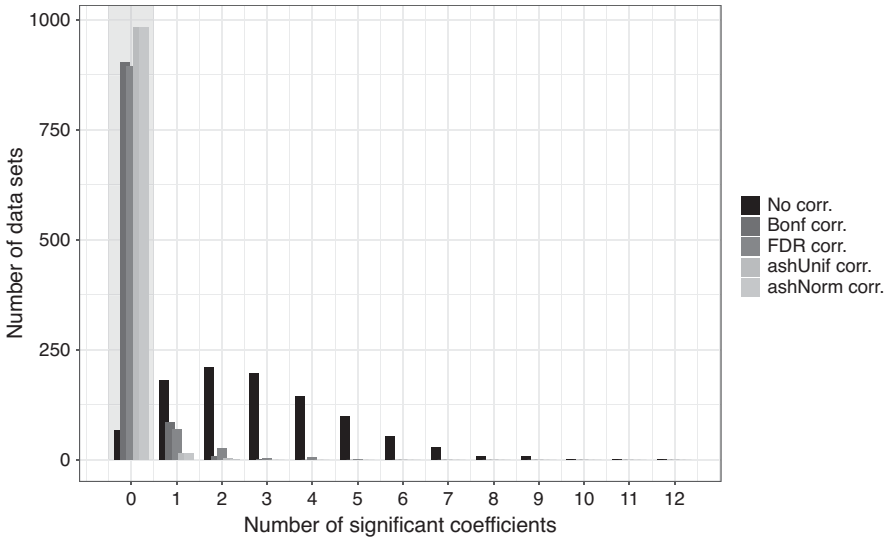

Figure 3. False-positive AMCE estimates when all null hypotheses are true with correction methods. Whereas the standard analysis pipeline correctly accepts all null hypotheses in fewer than 80 datasets, the BC, the BH, and the Ash all correct multiple testing in more than 900 experimental trials with the Ash performing the best.

In the first scenario, one binary attribute has a non-zero AMCE, and the results are shown in Table 1. In the original profile data, this attribute corresponds to *Gender*. We vary the noise in simulations by changing the heterogeneity of AMCEs and the error variance of the regression model for latent responses. Since only *Gender* has an effect, the shaded cells are the target we would like to hit: tests identify only one true-positive finding and no false findings. The pattern is quite consistent with the simulation results shown in Section 4.1. Without correction, about 80% of experimental trials produce at least one false-positive finding. All correction methods improve the situation remarkably, with the Ash has the best performance in all circumstances.

In the second scenario, all levels of the attributes that correspond to *Gender*, *Education*, and *English* in the original data have non-zero AMCEs, whereas the AMCEs of the others are zero.[4] Table 2 presents the results. Because 10 levels have non-zero AMCEs, the shaded cells indicate the number of datasets in which hypothesis tests are perfectly accurate. All cells to the right (above) are the number of samples where some false positives (negatives) are produced. For example, without correction, 248 experimental trials successfully detect exactly the true non-zero AMCEs; 314 detect those AMCEs, plus one false-positive result; three experiments do not yield any false-positive findings, but missed one non-zero effect.

Table 2 shows the trade-off in using correction methods. On the one hand, the use of a correction method dramatically improves the number in the shaded cells. In contrast to 248 without correction, almost all correction methods find the truth in more than 600 samples. On the other hand, as the *Sum* column indicates, all correction methods produce false negatives more often than the standard approach. Reducing the number of false positives comes at a cost of increasing the number of false negatives. Moreover, the trade-off exists among the correction methods, too. As the most conservative correction method, the BC produces false negatives in about 30% of experimental trials. The BH is least likely to miss the true AMCEs, but it produces more false-positive conclusions than the other two. The Ash takes the middle ground: it produces false-negative findings less likely than the BC, and false-positive results less likely than the BH.

Given this trade-off, should researchers use a correction method? In Figure 3 and Table 1, the answer is clear: any correction method dominates non-correction. When only zero or one attribute

4 Section B.2 of the Supplementary Material describes simulation parameters.

Table 1. Number of datasets for each number of true- and false-positive findings when the AMCE of `Gender` is non-zero. (a) The effect of `male` is −.06 and the effects of `female` and all other attributes are drawn independently from $\mathcal{N}(0, .015^2)$. The error variance of the regression model for continuous responses is $.01^2$. (b) AMCEs are identical to Table 1a, but the error variance of the regression model is $.1^2$. (c) The effect of `male` and the other attributes and the error variance are identical to Table 1b, but the effect of `female` is independently drawn from $\mathcal{N}(0, .12^2)$. Empty cells indicate zero.

			No. of False Positives								
			0	1	2	3	4	5	6	7	8
No. of True Positives	No corr.	1	230	290	215	123	69	42	19	9	3
	Bonf. corr.	1	966	32	2						
	BH corr.	1	931	61	7	1					
	ashUnif corr.	1	996	4							
	ashNorm corr.	1	998	2							

(a) Baseline

			No. of False Positives											
			0	1	2	3	4	5	6	7	8	9	10	11	12
No. of True Positives	No corr.	1	237	253	223	134	83	38	17	6	2	6			1
	Bonf. corr.	1	962	37	1										
	BH corr.	1	930	55	7	5	1	1	1						
	ashUnif corr.	1	984	14	2										
	ashNorm corr.	1	987	12	1										

(b) Larger Error Variance

			No. of False Positives										
			0	1	2	3	4	5	6	7	8	9	10
No. of True Positives	No corr.	1	191	288	228	125	79	42	30	9	3	2	3
	Bonf. corr.	1	951	43	6								
	BH corr.	1	902	83	12	3							
	ashUnif corr.	1	982	15	3								
	ashNorm corr.	1	985	13	2								

(c) Larger Heterogeneous AMCE and Error Variance

level has AMCE, the use of correction methods reduces the risk of false-positive findings at no cost since there is nothing to be missed. However, if many levels have AMCEs as in Table 2, correction methods decrease the number of false positives in exchange for an increase of the number of false negatives. Hence, correction methods may not be uniformly better than not correcting.

Figure 4 presents a measure to evaluate this trade-off. It shows the distribution of the true-positive rate (TPR) minus the false-positive rate (FPR) across samples in the same simulations as Table 2. The TPR is the number of true positives divided by the number of true non-zero AMCEs while the FPR is the number of false positives divided by the number of true zero AMCEs. If a test is perfect, its TPR is one and FPR is zero, because the ideal test finds all non-zero AMCEs and does not falsely reject the null on any zero AMCEs. Therefore, the higher density is concentrated on the right in Figure 4, the better. The figure shows that the BH and the Ash achieve a value larger than .85 in almost all simulated samples while the distribution without correction has a longer tail on the left. The figure indicates that researchers are more likely to get the ideal outcome with a correction method than without any.

These simulations demonstrate the promise and pitfalls of multiple testing correction methods. First, researchers should always use some correction method when conducting conjoint survey

Table 2. Number of datasets for each number of true- and false-positive findings when the true AMCEs of all levels in `Gender`, `Education`, and `English` are non-zero. Obtaining 10 true positives and zero false positives (shaded) is the ground truth. Empty cells indicate zero.

			No. of false positives											
			0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	***Sum***
No. of true positives	No corr.	9	3	2	1		2	2						10
		10	248	314	195	116	56	33	14	9	2	2	1	990
	Bonf corr.	8	35	3										38
		9	289	13	1									303
		10	625	33	1									659
	BH corr.	8		1										1
		9	45	17	6	2								70
		10	589	253	57	16	9	4				1		929
	ashUnif corr.	8	18	1										19
		9	151	18	4	3								176
		10	620	151	26	3	4	1						805
	ashNorm corr.	8	15	2										17
		9	178	23	6	1								208
		10	645	106	18	3	2	1						775

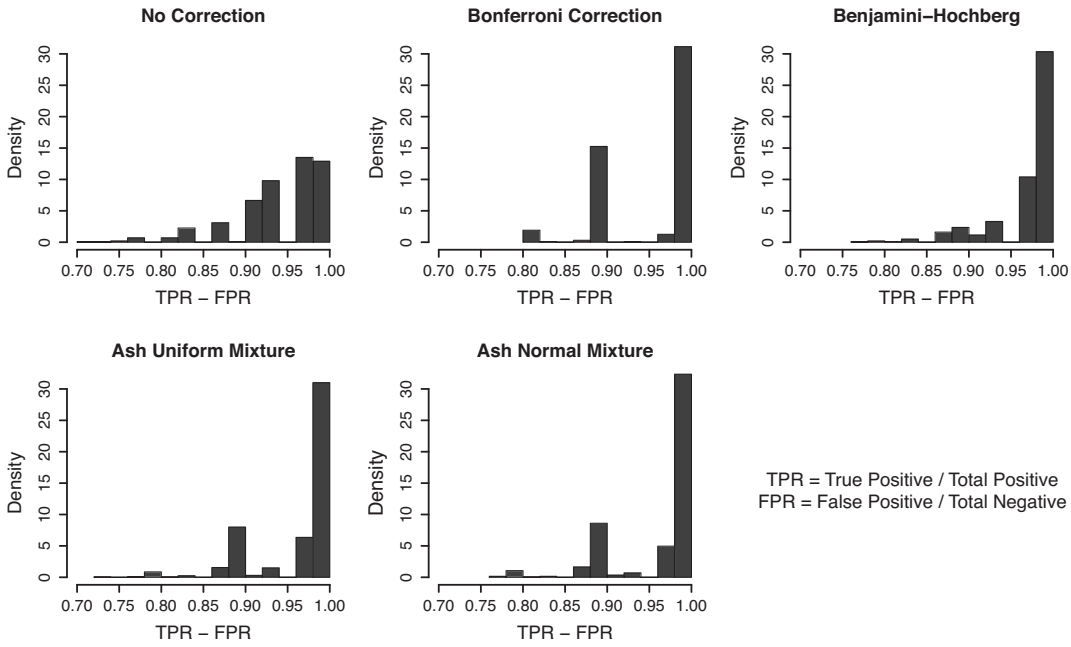

Figure 4. Density Histogram of the Difference between True Positive Rate (TPR) and False Positive Rate (FPR). A larger value on the *x*-axis indicates better performance. The figure is based on the same simulations as Table 2.

experiments. Since conjoint analysis inherently requires a large number of hypothesis tests, some, if not all, statistically significant findings are likely to be false positives. Second, the risk of false-positive findings cannot be entirely eliminated, and correction methods differ across the ability and cost of reducing the number of false positives. The BC is most aggressive in avoiding false positives, but its cost of missing true findings may be substantial. The BH is the opposite, and the Ash is in between the two. Although none provides the perfect solution, researchers should choose a correction method that best suits their needs. In particular, the choice should be based

on a careful assessment on the relative tolerance of false positives and false negatives.[5] We provide a checklist as an additional guidance in concluding remarks.

5 Reanalysis

To illustrate how the use of the correction methods may change empirical conclusions, we apply the correction methods to two published applications of conjoint experiment.[6] Overall, the pattern we observe in the reanalysis is consistent with the simulations. The BC reduces the number of findings the most, and some of the results that are changed to null are substantively questionable. On the other hand, the BH does not eliminate any findings of the original papers. The Ash corrects fewer findings away than the BC, but its results seem to make the most substantive sense.

5.1 Selecting Immigrants in the United States

In the seminal paper on conjoint designs for causal inference, Hainmueller *et al.* (2014) employ the conjoint design to explore the AMCEs of immigrants' attributes on preference for admission to the United States. There are nine attributes: *Gender*, *Education*, *Language*, *Origin*, *Profession*, *Job experience*, *Job plans*, *Application reasons*, and *Prior trips to U.S.* To exclude unrealistic attributes combinations, the randomization for *Education*, *Profession*, *Country of Origin*, and *Application reasons* are conditionally independent given some constraints, and the randomization for the other five attributes are completely independent. The outcome variable is whether a respondent prefers a given profile in a paired comparison.

We focus on two attributes, *Country of origin* and *Profession*, shown in Figure 5.[7] The left panel of Figure 5 shows the estimates of the AMCE of each country of origin relative to India, with no correction, the BC, the BH, and the Ash. The most noticeable pattern is that the BC eliminates the statistical significance of all estimates except for the effect of Iraq. If we believe the BC results, respondents in their survey did not distinguish immigrants from India, Mexico, France, Germany, Sudan, and Somalia. On the other hand, coefficients adjusted by the BH and the Ash largely preserve the original paper's conclusion that immigrants from Sudan, Somalia, and Iraq are less preferred than those from India.

The right panel of Figure 5 presents the results on the *Profession* attribute. *Janitor* is the reference category. The original results suggest that there is a bonus for financial analysts, construction workers, teachers, computer programmers, nurses, research scientists, and doctors. Again, the BC renders more coefficients insignificant: financial analysts and computer programmers are indistinguishable from janitors. While the BH preserves all the original findings, the Ash changes the results for construction workers—the bonus for construction workers is indistinguishable from zero. The Ash result is in fact consistent with the argument of the original paper that high-skilled immigrants are preferred to low-skilled workers.

While we cannot adjudicate on the differences with certainty because the true value is unknown, some correction methods lead to more substantively understandable results over the others. The BC seems overly conservative, and its null findings may require further theoretical justification. The BH results agree with most non-corrected results, including some unexpected significant estimates. The Ash corrects some findings away but not as aggressively as the BC does, and it leads to conclusions that make most substantive sense in this application.

5.2 Selecting Trading Partners in Vietnam

Conjoint experiment is also useful in examining attributes of units other than individuals. Spilker *et al.* (2016) explore what types of countries are preferred partners for Preferential Trade Agree-

5 Additional simulation results with more noisy data are shown in Section B.3 of the Supplementary Material.
6 Section D.3 of the Supplementary Material shows the reanalysis of another paper in comparative politics.
7 For the entire replication results, see Figure D.1 in the Supplementary Material.

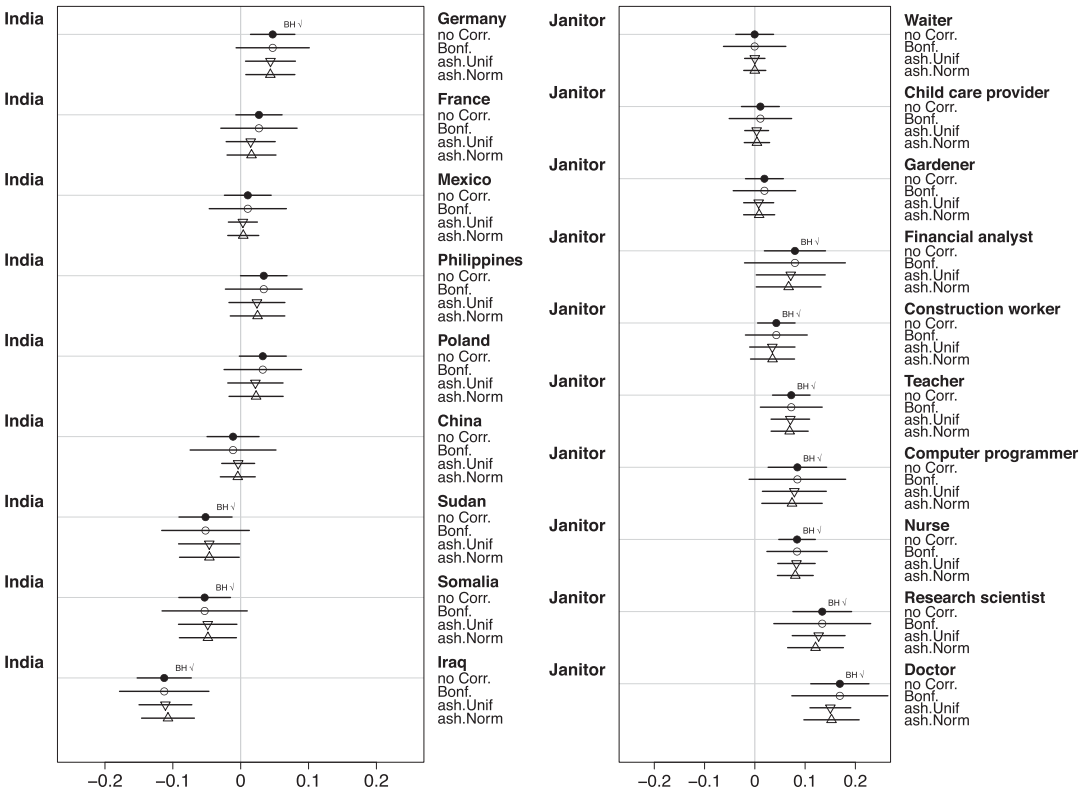

Figure 5. Effects of the immigrant's country of origin (left) and profession (right) on the probability of being preferred for admission to the United States. For `country of origin`, the reference category is India; for `profession`, the reference category is janitor. The plot shows estimates with no correction, the BC (`Bonf`), the Ash with a mixture of normal components (`ash.Norm`), and the Ash with a mixture of uniform components (`ash.Unif`) for each pair of comparison. `BH`✓ next to a point estimate indicates the BH corrected coefficient is significant for the corresponding attribute level. The estimates are based on regression estimators with clustered standard errors at the respondent level; the bars represent 95% confidence intervals. The estimates with no correction replicate the results for the corresponding attributes in Figure 3 in Hainmueller *et al.* (2014, 21).

ments (PTAs) by conducting conjoint surveys in Costa Rica, Nicaragua, and Vietnam. They include eight attributes in their design: *Distance* from the partner country's capital with three levels; *Size of the economy* with three levels; *Culture*, a binary variable indicating similarity in tradition and language of the partner country; *Religion*, which contains three religions for Costa Rica and Nicaragua and four religions for Vietnam; *Political system*, three levels of the extent to which citizens democratically elect political leaders; *Military ally*, a binary variable indicates whether the partner country has a security alliance with respondents' home country; *Environmental protection standards* and *Worker rights protection standards*, each takes three levels. All these attributes are completely randomized, and no profile is excluded in the original surveys. The outcome is binary, whether respondents choose a country profile in a paired comparison.

Figure 6 focuses on the effect of two attributes *Military ally* and *Environmental protection standards* on the respondents in Vietnam.[8] Among the three countries, Vietnam is the only one where non-military allies are punished relative to military allies. The original paper justifies the finding by its geopolitical location and military-security rivalries in the region (Spilker *et al.* 2016, 710, 714). However, Vietnam has a "Three Nos" defense policy since 1998: no military alliance, no aligning with one country against another, and no foreign military bases on Vietnamese soil.[9] The

8 The complete replication results can be found in Figure D.2 in the Supplementary Material.
9 Socialist Republic of Vietnam Ministry of National Defence, 2009.

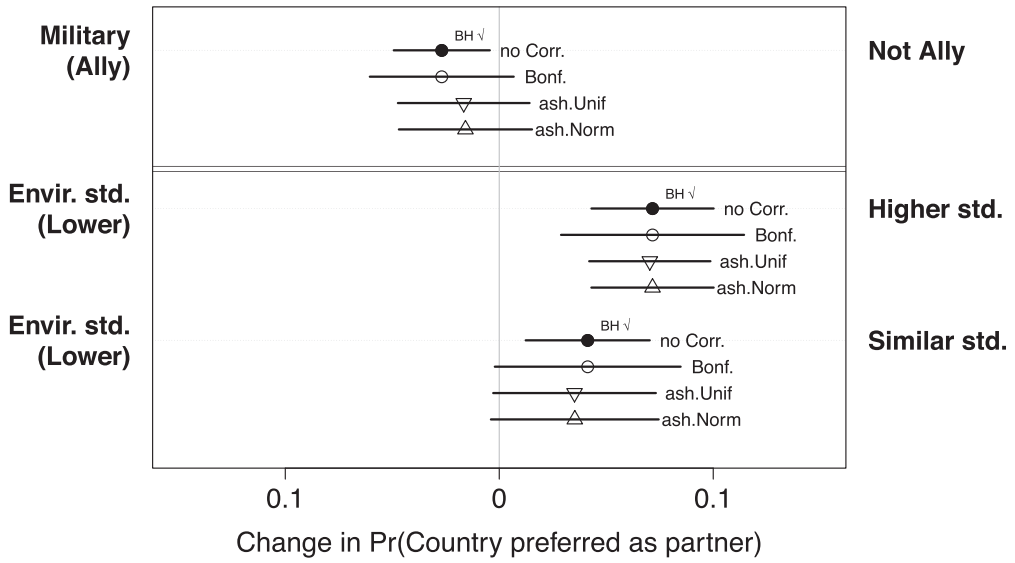

Figure 6. Effects of Military ally (top) and Environmental protection standards (bottom) on the probability of being preferred as trading partners in Vietnam. For `Military ally`, the reference category is allied; for `Environmental Protection Standards`, the reference category is lower standards. The plot shows estimates with no correction, the BC (`Bonf`), the Ash with a mixture of normal components (`ash.Norm`), and the Ash with a mixture of uniform components (`ash.Unif`) for each pair of comparison. `BH✓` next to a point estimate indicates the BH corrected coefficient is significant for the corresponding attribute level. The estimates are based on regression estimators with clustered standard errors at the respondent level; the bars represent 95% confidence intervals. The estimates with no correction replicate the results for the corresponding attributes in Figure 1.3 in Spilker *et al.* (2016, 715).

context makes it difficult to interpret the significant result, because it is unclear what military allies mean to Vietnam given this defense policy. While the BH preserves the original finding, the BC and the Ash correct it away.

For environmental standards, while the preference for higher standards relative to lower standards is robust to different correction results, the bonus for countries with similar standards is not. Again, the BC and the Ash render it a false-positive conclusion. The BH agrees completely with the original conclusion, but we cannot rule out the possibility that this is guaranteed by the design of BH: there are not enough significant discoveries to begin with to control for FDR. A lower FDR may be needed to accommodate the smaller number of significant findings in social science researches.

The replication exercise demonstrates the usefulness of applying correction methods in conjoint design from a substantive perspective. Correction methods could raise alarms of potential limitations in the profile design. Such warnings would be valuable especially in the phase of pilot research or pretesting. Moreover, results that stand the test of correction would help authors make more convincing arguments. In this application, the authors of the original paper needed to justify their finding on the preference for military allies only in Vietnam, but it is difficult to interpret this finding given the fact that Vietnam has not have military allies for a while and will not for the foreseeable future. The authors could have avoided interpreting this result by using the BC or the Ash, even though they included the *Military ally* attribute, which should have been excluded from the design.

6 Concluding Remarks

Conjoint analysis is widely used in political science because it allows researchers to estimate the effects of many variables on preference formation. Unfortunately, exactly because it is designed for estimating multiple effects, statistical inference on estimates in conjoint designs suffers from the multiple testing problem. However, few systematic assessments on the severity of the problem

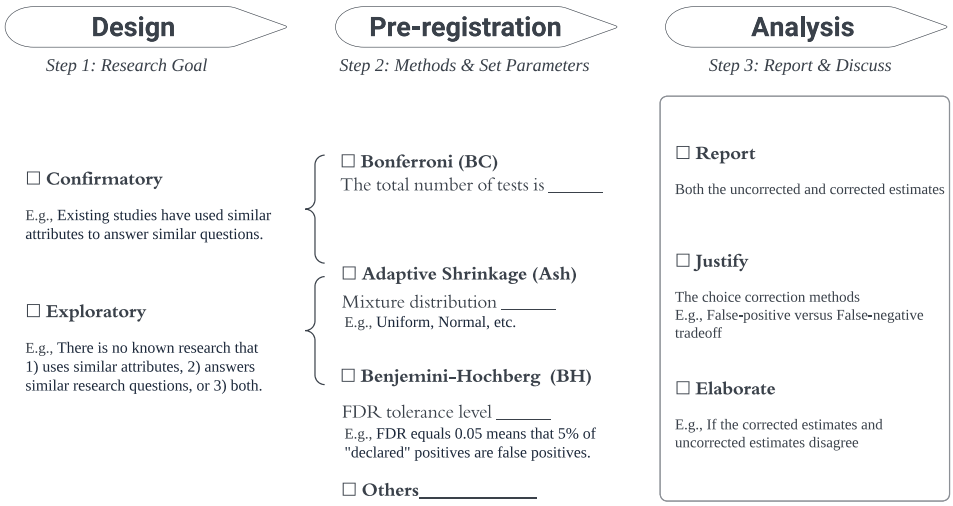

Figure 7. Checklist for multiple hypothesis testing in conjoint analysis.

and little empirical guidance on the choice of correction methods have been provided. In a series of simulations and applications to published data, we examined the probability of getting false-positive conclusions from a typical conjoint survey experiment, and compared the performance of three off-the-shelf multiple testing correction methods.

Although some correction is always better than no corrections, none of the methods provides the perfect solution to the problem. The BC is most conservative. Therefore, it is least likely to mislead researchers to false-positive conclusions, whereas it is most likely to mislead researchers to false-negative conclusions. The BH procedure is the opposite. We even found that the BH procedure does not change the statistical significance of any estimates in some applications. The Ash method takes a middle ground between the two. While it reduces the probability of false positives than the BH, it avoids false negatives better than the BC.

Whether being conservative (or lenient) is a virtue rather than a vice depends on the purpose of researchers. We believe that the Ash method should be recommended when researchers do not have much prior knowledge on the existence of AMCEs, because it is unclear which of false positives or false negatives the researchers need to avoid more. However, the BH procedure might be preferred if previous studies strongly suggest the existence of AMCEs, whereas the BC should be recommended for AMCEs whose existence is considered unlikely. In the former, although the rejection of the null is not surprising, researchers can cast more doubt on the prior knowledge if the null is accepted. In the latter case, passing a more conservative test is valuable information because it is more likely to be a new finding. The comparison in our paper provides a guide in selecting the correction method that suits a particular application.

Figure 7 summarizes our recommendations on the use of multiple testing correction methods in conjoint analysis. It helps researchers reduce missing steps and ensures consistency and completeness. Importantly, for our purpose, it guides researchers to contemplate a series of questions related to multiple hypothesis testing at different stages of the study. The checklist is divided into three sections: *design*, *pre-registration*, and *analysis*.

During the design phase, scholars determine their research objective, whether the conjoint experiment is to confirm findings in existing studies or it is exploratory in nature. The distinction between the two types of research is fuzzy in many empirical studies. This item is *not* designed to force researchers to choose one or the other, but rather it reminds them to be more precise, and their inclination will provide a direction for the pre-registration stage.

As discussed above, if the research objective is confirmatory, we recommend that researchers use the more stringent BC and specify the number of tests they plan to conduct in the

pre-registration. The number of tests is the collection of meaningful inferences from a substantive perspective, defined by the researchers. Usually, the bare minimum includes all the possible attribute-level combinations[10]. It should also include all the subgroup analysis, balancing checks, and other quality check tests that researchers usually perform.[11] On the contrary, we recommend the more lenient BH method if the research is primarily exploratory. Here researchers need to specify the FDR. For instance, the default FDR in many R packages is set at .05, meaning that 5% of the "declared" positive findings will be purged as false positives. For the remaining types, we recommend Ash. Researchers need to specify the mixture distribution they are going to use.[12] Setting the mixture distribution requires some prior knowledge of the subject matter. However, because the number of hypothesis testing in most social science applications is not so large, the corrected results do not diverge drastically when different mixture distributions are used, as our simulation studies demonstrate.

In the analysis and write-up stage, uncorrected and corrected data should be included in the paper regardless of the chosen method. Researchers should consider the false-positive and the false-negative trade-off in this particular application and justify the method of choice. If any of the corrected and uncorrected results differ, the discrepancy should be described and discussed explicitly. In summary, the steps in the checklist are intended to reduce the researchers' degrees of freedom when selecting different methods. Furthermore, it aids researchers in incorporating multiple testing correction into the conjoint analysis routine in a principled and transparent manner.

Multiple hypothesis testing may also be a problem with empirical studies using other methods than conjoint designs. In fact, one of the major sources of publication bias is the property of the frequentist hypothesis testing that the probability of false findings is controlled. We focused on conjoint analysis in this paper because the number of hypotheses to be tested is relatively unambiguous. Applying the correction methods we discussed to studies where the number of statistical hypotheses varies over the stages of research, for example, adding robustness checks to address reviewers' comments, is much harder than to conjoint designs. More research on multiple testing correction in the other contexts is warranted.

Acknowledgments

The authors thank Scott Abramson, Nahomi Ichino, Yusaku Horiuchi, Naijia Liu, Tom Pepinsky, Kevin Quinn, Teppei Yamamoto, Arthur Yu, Jerry Yu, participants at the Joint Conference of Asian Political Methodology Meeting VIII and Australian Society for Quantitative Political Science Meeting IX, attendees at the "Politics, Sandwiches, and Comments" workshop of the Cornell Department of Government and the University of Michigan Interdisciplinary Seminar in Social Science Methodology, members of the Ichino lab, the Quinn research group, and the Shiraito research group, and two anonymous reviewers for helpful comments and discussions on earlier drafts.

10 Cross-attribute constraints will reduce the total number of tests. For example, the impossible combination of someone whose `occupation` is a doctor and `education level` is no formal education school should not be included in the total number when determining the new significance level $\tilde{\alpha}$ in BC.

11 A reviewer pointed out that some attributes or levels may be included in a conjoint design only to make its profiles look real and therefore it may be more appropriate that estimates for such attributes/levels are not counted as hypotheses to be tested. On the one hand, we agree that, if an attribute is used only for that purpose and not in the study's interest, the number of levels of the attribute can be excluded from the number of hypotheses to be tested. On the other hand, we note that estimates for all levels of all attributes are reported in most previous studies using conjoint surveys. If one excludes some attributes or levels from hypotheses, statistical estimation of the marginal means or AMCEs should ignore those attributes or levels entirely and the analysis is preregistered as such. Researchers also need to be cautious when including attributes or levels that are not interested, because the estimates of the researchers' interest depend on the distribution of those non-interested attributes or levels (de la Cuesta *et al.* 2022).

12 The `ash` function in the `ashr` package (version 2.2-47) supports uniform, normal, halfuniform, +uniform, -uniform, and halfnormal distributions. For more information, see Stephens *et al.* (2020).

Conflict of Interest Statement

The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript and there is no financial interest to report.

Data Availability Statement

Replication materials are available in Liu and Shiraito (2022).

Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2022.30.

References

Abramson, S., K. Kocak, A. Magazinnik, and A. Strezhnev. 2020. "Improving Preference Elicitation in Conjoint Designs using Machine Learning for Heterogeneous Effects." Working Paper. https://polmeth.theopenscholar.com/files/polmeth2020/files/polmeth_magazinnik.pdf

Abramson, S. F., K. Koçak, and A. Magazinnik. 2022. "What Do We Learn about Voter Preferences from Conjoint Experiments?" *American Journal of Political Science* 66: 1008–1020.

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 (1): 112–119.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2021a. "Conjoint Survey Experiments." In *Advances in Experimental Political Science*, edited by J. Druckman, and D. P. Green. Cambridge: Cambridge University Press.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2021b. "Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments." *Political Science Research and Methods* 9 (1): 53–71.

Bansak, K., J. Hainmueller, D. J. Hopkins, and T. Yamamoto. 2022. "Using Conjoint Experiments to Analyze Elections: The Essential Role of the Average Marginal Component Effect (AMCE)." *Political Analysis*: 1–19 (First View).

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Benjamini, Y., and D. Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics* 29 (4): 1165–1188.

Bland, J. M., and D. G. Altman. 1995. "Multiple Significance Tests: The Bonferroni Method." *BMJ* 310 (6973): 170.

Carnes, N., and N. Lupu. 2016. "Do Voters Dislike Working-Class Candidates? Voter Biases and the Descriptive Underrepresentation of the Working Class." *American Political Science Review* 110 (4): 832–844.

Clayton, K., J. Ferwerda, and Y. Horiuchi. 2021. "Exposure to Immigration and Admission Preferences: Evidence from France." *Political Behavior* 43: 175–200.

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 (1): 19–45.

Dunn, O. J. 1961. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56 (293): 52–64.

Egami, N., and K. Imai. 2019. "Causal Interaction in Factorial Experiments: Application to Conjoint Analysis." *Journal of the American Statistical Association* 114 (526): 529–540.

Fournier, P., S. Soroka, and L. Nir. 2020. "Negativity Biases and Political Ideology: A Comparative Test across 17 Countries." *American Political Science Review* 114 (3): 775–791.

Ganter, F. 2021. "Identification of Preferences in Forced-Choice Conjoint Experiments: Reassessing the Quantity of Interest." *Political Analysis*: 1–15 (First View).

Gerard, D., and M. Stephens. 2018. "*Empirical Bayes Shrinkage and False Discovery Rate Estimation, Allowing for Unwanted Variation*." *Biostatistics* 21: 15–32.

Hainmueller, J., D. Hangartner, and T. Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112 (8): 2395–2400.

Hainmueller, J., and D. J. Hopkins. 2015. "The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes toward Immigrants." *American Journal of Political Science* 59 (3): 529–548.

Hainmueller, J., D. J. Hopkins, and T. Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22 (1): 1–30.

Hochberg, Y., and A. C. Tamhane. 1987. *Multiple Comparison Procedures*. New York: John Wiley & Sons, Inc.

Horiuchi, Y., Z. D. Markovich, and T. Yamamoto. 2020. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Political Analysis* 30 (4): 535–549.

Incerti, T. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design." *American Political Science Review* 114 (3): 761–774.
Leeper, T. J., S. B. Hobolt, and J. Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28 (2): 207–221.
Liu, G., and Y. Shiraito. 2022. "Replication Data for: Multiple Hypothesis Testing in Conjoint Analysis." https://doi.org/10.7910/DVN/HIPDOP.
Liu, H. 2019. "The Logic of Authoritarian Political Selection: Evidence from a Conjoint Experiment in China." *Political Science Research and Methods* 7 (4): 853–870.
Oliveros, V., and C. Schuster. 2018. "Merit, Tenure, and Bureaucratic Behavior: Evidence from a Conjoint Experiment in the Dominican Republic." *Comparative Political Studies* 51 (6): 759–792.
Ono, Y., and B. C. Burden. 2019. "The Contingent Effects of Candidate Sex on Voter Choice." *Political Behavior* 41 (3): 583–607.
Sen, M. 2017. "How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment." *Political Research Quarterly* 70 (2): 374–393.
Shafranek, R. M. 2021. "Political Considerations in Nonpolitical Decisions: A Conjoint Analysis of Roommate Choice." *Political Behavior* 43: 271–300.
Sjölander, A., and S. Vansteelandt. 2019. "Frequentist versus Bayesian Approaches to Multiple Testing." *European Journal of Epidemiology* 34 (9): 809–821.
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 (4): 706–718.
Stephens, M. 2017. "False Discovery Rates: A New Deal." *Biostatistics* 18 (2): 275–294.
Stephens, M., et al. 2020. *ashr: Methods for Adaptive Shrinkage, Using Empirical Bayes.* https://CRAN.R-project.org/package=ashr.
Teele, D. L., J. Kalla, and F. Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112 (3): 525–541.