Multiple Hypothesis Testing in Conjoint Analysis*

Guoer Liu† Yuki Shiraito‡

First draft: January 5, 2021
This draft: February 14, 2022

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. All three methods are more accurate in recovering the truth than the conventional analysis without correction. Moreover, the adaptive shrinkage 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.

*The authors thank Nahomi Ichino, Yusaku Horiuchi, Naijia Liu, Tom Pepinsky, Kevin Quinn, 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, and members of the Ichino lab, the Quinn research group, and the Shiraito research group for helpful comments and discussions on earlier drafts.

†Ph.D. Candidate, Department of Political Science, University of Michigan. Email: guoerliu@umich.edu.

‡Assistant Professor, Department of Political Science, University of Michigan. Center for Political Studies, 4259 Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104-2321. Phone: 734-615-5165, Email: shiraito@umich.edu, URL: shiraito.github.io.

1 Introduction

Conjoint analysis is one of the most widely used survey experimental designs in political science. Its popularity is attributed to Hainmueller, Hopkins and Yamamoto (2014), which 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, in the earliest application of AMCE, 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; Teele, Kalla and Rosenbluth, 2018; Ono and Burden, 2019; Incerti, 2020), bureaucratic selection (e.g., Liu, 2019; Oliveros and Schuster, 2018), and other types of multi-dimensional decision making (e.g., Sen, 2017; Fournier, Soroka and Nir, 2020; Shafranek, 2019).[1]

An attractive property of conjoint analysis is that it "enables researchers to estimate the causal effect of multiple treatment components and assess several causal hypotheses simultaneously" (Hainmueller, Hopkins and Yamamoto, 2014, p.1). This is extremely valuable from a substantive point of view. Since a number of factors contribute to decisions, isolating the causal effect of each factor under each possible combination of the other factors would require experimental manipulation of numerous combinations. Logistics and resource challenges in such designs would be insurmountable. Conjoint analysis overcomes this difficulty by identifying the AMCEs of multiple attributes at once. AMCE is the causal effect of an attribute averaged over all profiles of the other attributes, and it has an intuitive interpretation (Bansak et al., 2020). The combination of conjoint designs and AMCE enables researchers to estimate the causal effects of multiple features on preference formation effectively.

[1]For a more comprehensive list of conjoint experiment papers, see de la Cuesta, Egami and Imai (2021).

Although it is a great substantive advantage that researchers can estimate the effects of a number of causes, producing many estimates leads to a statistically undesirable property, multiple hypothesis testing. The multiple hypothesis testing problem arises whenever more than one hypothesis are tested in statistical analysis. It is problematic because the more null hypotheses are tested, the more likely it is that at least one hypothesis is rejected, even if all the null hypotheses are true. The prespecified critical value, conventionally set at .05, represents the probability of falsely rejecting the null hypothesis assuming that only one hypothesis is tested. When statistical analysis involves several hypothesis tests 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 one or 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, Hopkins and Yamamoto (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 our knowledge, existing empirical studies in political science using conjoint analysis do not correct for the number of hypotheses in their main analysis, with the exception of Hainmueller, Hangartner and Yamamoto (2015) where the authors use the Bonferroni correction. A few other studies, an example of which is Clayton, Ferwerda and Horiuchi (2019), confirm their results with corrections in appendices as robustness checks. In fact, researchers are aware that multiple hypothesis testing is an inherent problem with conjoint designs. Bansak et al. (2021b, p.28) point out that the concerns about multiple comparisons in conjoint designs make pre-registration and pre-analysis plans especially valuable. However, no systematic assessments have been done on how serious the problem can be in conjoint analysis, and while several correction methods are used in applied statistics, the consequences of the choice of a correction method have not been empirically evaluated. To avoid haphazard selection, applied researchers need guidance on which correction method is appropriate under their circumstances.

[2]Recently, however, multiple testing correction has become to be used more often as robustness checks than before. We thank Yusaku Horiuchi for pointing this out.

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 experiment setup the standard analysis pipeline produces at least one statistically significant AMCE estimate in more than 90% of simulated data sets even when all AMCEs are zero in truth.

Second, we compare the strengths and limitations of two well-known correction methods, the Bonferroni correction (Bland and Altman, 1995, BC) and the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995, BH). In addition, we introduce a recently developed correction method, adaptive shrinkage (Stephens, 2017; Gerard and Stephens, 2018, Ash). ASh shows great potential in biology applications that involve hypothesis testing on numerous gene expressions. While none of the methods completely resolves the problem, all of them are better than the standard practice. Among the three methods, the BC is most conservative. However, while it guards against false positive conclusions, the cost of false negative conclusions can be significant at times. The BH is the least susceptible to false negative conclusions, but the rank-based procedures do not produce uncertainty measures of estimates. The ASh takes a middle ground: it best detects true positives while avoiding false negatives.

To illustrate how different correction methods perform in real data, we reanalyze two prominent conjoint design applications in American politics and international relations. The first application, which uses the data set of Hainmueller, Hopkins and Yamamoto (2014), demonstrates that ASh corrected results are more consistent with what we would substantively expect than conclusions from other correction methods or no corrections. Second, an application to an experiment in Vietnam about the selection of trade agreement partners (Spilker, Bernauer and Umaña, 2016) shows that the corrected results recover the null result on an attribute that should have been excluded based on substantive knowledge.

Compared to other studies that propose improvements on conjoint survey designs, this paper exclusively focuses on statistical inference. Existing studies have examined estimands and interpretation (Egami and Imai, 2019; de la Cuesta, Egami and Imai, 2021; Abramson, Koçak and Magazinnik, 2019; Abramson et al., 2020; Bansak et al., 2020; Ganter, 2022), implementation (Bansak et al., 2018, 2021a), social desirability bias (Horiuchi, Markovich and Yamamoto, 2020), and subgroup analysis (Leeper, Hobolt and Tilley, 2020; Clayton, Ferwerda and Horiuchi, 2019).

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).

The 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 three conjoint experiment data sets. Finally, we summarize the paper and discuss suggested analysis pipelines for conjoint designs in the concluding section.

2 False-Positive Findings in Conjoint Analysis

A large number of hypothesis tests in conjoint analysis make it more likely that some statistically significant findings are false positives. The conventional significance level of $\alpha = .05$ means that a test rejects a true null hypothesis with probability .05. In other words, the test tolerates five false positives (a.k.a. Type I error) out of 100 experimental trials on average. On the other hand, the probability that *at least one of multiple tests* rejects the null hypothesis can be much larger than .05 depending on the number of hypotheses to be tested. When ten hypotheses are tested, this probability, known as the *familywise error rate* (FWER), is $1 - \Pr(\text{None of the ten tests rejects the null}) = 1 - (1 - .05)^{10} = .401$. If the number of tests is 20, the FWER increases to .642. (See Figure A.1.) Since the number of null hypotheses is greater than 20 in most conjoint experiments, the problem is even severer – 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.

A simulation study presented in Figure 1 shows how likely a conjoint experiment may produce false-positive findings. The conjoint design of our simulations follows exactly the experimental design of Hainmueller, Hopkins and Yamamoto (2014). The design consists of nine attributes, one being binary, one with three levels, three with four levels, one with five levels, one with seven levels, one with ten levels, and the other with eleven levels. To simulate the forced-choice design, we generate a linear continuous response variable for each respondent and profile and coarse it

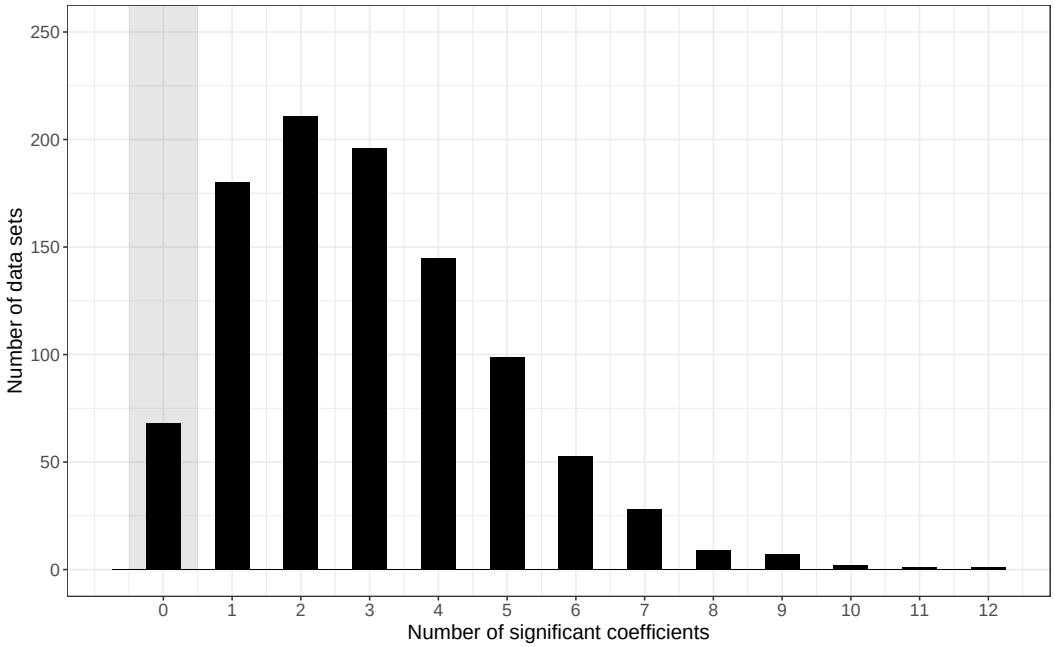

Figure 1: **False-positive Results of Estimated AMCEs when All Null Hypotheses are True.** The x-axis indicates the number of statistically significant AMCE estimates at the 5% level and the y-axis indicates the number of data sets for each number of sigfinicant estimates. Gray shaded area is the correct number, meaning that no effect is distinguishable from zero.

into a choice between an observed pair of profiles. 1,000 simulation data sets 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 data set 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. That is, researchers may obtain false-positive findings in more than 90% of experiments. Although we observe that the rate of false-positive findings is a little lower (around 80%) under some other simulation settings shown in Appendix B, the high false positive rate is concerning for applied research.

[3]Appendix B describes the simulation settings and parameters in greater detail.

3 Multiple Testing Correction Methods

There exists a wealth of correction methods for multiple testing problems. In this section, we briefly introduce two widely known methods, Bonferroni Correction and Benjamini-Hochberg Procedure, and a recently developed method, adaptive shrinkage. Their respective advantages and limitations will be illustrated in a simulation study.

3.1 Bonferroni Correction

The Bonferroni correction (Dunn, 1961, henceforth BC) tackles the multiple testing problem by adopting a more stringent threshold as the number of tests increases. The method sets the significance level at $\alpha/(\#\text{ of tests})$ instead of α. For instance, when one tests five hypotheses at the 5% level simultaneously, the significance level to be used is $\alpha^* = .05/5 = .01$. With the BC, the probability of rejecting each null hypothesis is reduced so that the FWER is contained at the intended significance level α. While there are other useful methods to control for FWER (Sarkar and Chang, 1997; Ludbrook, 1998; List, Shaikh and Xu, 2019), the BC is perhaps the easiest to implement. α is the only parameter that one needs modify, and the confidence interval construction follows the usual procedure with the new α^*.

A caveat is that the BC can be overly conservative in many applications. Since its theoretical property assumes that all tests are independent, null hypotheses are more likely to be accepted than should be when tests are correlated. Hence, the BC may suffer low statistical power and false-negative findings because tests using the same data set tend to be correlated. We illustrate this point later in our simulation study.

A philosophical critique of the BC is that "the total number of tests" cannot be clearly defined and tracked (Sjölander and Vansteelandt, 2019). While conjoint designs clearly pre-specify the number of attribute levels, researchers often conduct tests to ensure survey quality such as balance and attention checks. Further, many conjoint analyses often carry out subgroup comparisons (Leeper, Hobolt and Tilley, 2020). Discussing this issue in greater detail is beyond the scope of this paper, but we note that pre-registration (Bansak et al., 2021b) will ameliorate this ambiguity in the use of the BC.

3.2 Benjamini-Hochberg Procedure

The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995, henceforth BH) addresses the multiple testing problem by controlling another measure, the false discovery rate (FDR). The FDR is defined as $\mathbb{E}[\#\text{ false discoverties}/\#\text{ total discoverties}]$. The BH is a widely known method to contain the FDR under a pre-set level α. In other words, the BH prunes statistically significant estimates so that researchers obtain fewer false findings.

The BH is a rank-based method that takes four steps. 1) For each of the total m hypotheses of interests, a m-vector of usual p-values is produced. 2) Rank the p-values in the ascending order and index by i. 3) Define $k \equiv \max\{i : p_i \leq \alpha \times i/m, 0 \leq i \leq m\}$, where α is the FDR tolerance level. 4) Reject null hypotheses H_i for $i = 1, 2, ..., k$, whose p-values are smaller than or equal to p_k. No hypothesis is rejected if the maximum does not exist.

Discussion on the mathematical properties of BH is beyond the scope of this paper,[4] but compared to the BC, the BH is less susceptible to false-negative conclusions because it accepts all statistically significant findings in its first step. On the other hand, it is less aggressive in eliminating false-positive findings. Another limitation is that because BH focuses on the p-values, it does not offer a straightforward way to construct uncertainty measures. We illustrate this trade-off later by simulations and applications.

3.3 Adaptive Shrinkage

The adaptive shrinkage (Ash) is a recently proposed method that is based on an empirical Bayes approach to controlling the FDR. While empirical Bayes approaches to multiple testing correction is not new (e.g., Greenland and Robins, 1991), the Ash in particular is developed by Stephens (2017); 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. Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)$ denote all J attribute levels of interests, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_J)$ denote point estimates of $\boldsymbol{\beta}$, and $\hat{\boldsymbol{s}} = (\hat{s}_1, ..., \hat{s}_J)$ be the standard errors of $\hat{\boldsymbol{\beta}}$. Consider the posterior

[4] Interested readers may refer to Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001); Storey (2002); Romano and Shaikh (2006a,b); Romano, Shaikh and Wolf (2008); Romano and Wolf (2010); Romano, Shaikh and Wolf (2008) for more detail and related extensions.

distribution of $\boldsymbol{\beta}$ given the estimates and standard errors:

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

The likelihood in Equation 1 is the sampling distribution of $\hat{\boldsymbol{\beta}}$, and hence approximated by the normal distribution with mean $\boldsymbol{\beta}$ and variance $\hat{\boldsymbol{s}}^2$. To regularize a large number of estimates, the prior distribution is assumed to be a mixture of a point mass at 0 and normal (or other) distributions with mean 0. Since the Ash is an empirical Bayes method, the mixture probabilities of this spike-and-slab prior are estimated by maximizing the penalized likelihood and then the posterior parameters are estimated based on the prior parameter estimates. Appendix C.1 provides a greater detail of the model and estimation.

By making the modeling assumptions transparent, ASh accommodates different correction strategies as researchers see fit. While expert knowledge would certainly inform the choice appropriate mixture distributions (e.g., uniform or normal mixture), as we will see in the simulation and application, the corrected results under ASh is relatively consistent. Additionally, because the Bayesian approach gives us the entire posterior distribution of the coefficients, their uncertainty measures are readily available.

ASh also delivers an additional benefit: it regularizes the point estimates in addition to uncertainty measures—so that the estimated effect size should have smaller error. This is an attractive property especially in social sciences applications. Because in most cases, what interests researchers is not simply "whether factor X affect respondents' choice," but also "to what extent." In the classic immigrant conjoint experiment, for instance, researchers found an education bonus for immigrant applicants with some education relative to those with no formal education. Although researchers would like to estimate the amount of the education bonus precisely, the other correction methods do not affect the sampling error of point estimates. The Ash, however, enables us to get more precise estimates in a principled manner. Appendix C.2 illustrates this point by simulations.

4 Comparing Correction Methods

To compare the performance across different methods, we conduct a series of simulations. For each case, we generated 1000 simulated data sets. In all simulations, we use immigrant profile data of Hainmueller, Hopkins and Yamamoto (2014) and adopt the conventional significance level of $\alpha = .05$. 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 described above to the case where the true AMCE is zero for all attributes (identical to the simulations in Section 2). Second, we compare the correction performance in more noisy—perhaps more realistic—cases where the true AMCE is non-zero for some attributes.

4.1 Simulation Study 1: Zero AMCEs

The results are summarized in Figure 2. As in Figure 1, the height of bars represents the number of data sets with the corresponding number of statistically significant findings. Note that the black bars show the identical results as in Figure 1. However, Figure 2 also shows the results of the BC, BH, and ASh with a mixture of uniform components and with a mixture of normal components. Those are represented by bars filled with different shades of gray.

All three correction methods we discussed above dramatically reduce the probability of false findings. Because the null hypotheses are true for all AMCE, ideally, all simulations should result in zero significant coefficients. As we discussed in Section 2, more than 90% experimental trials would produce at least one significant estimate without correction. However, both the BC and BH remove false findings in more than 90% of simulation data sets. ASh with either a uniform mixture or a normal mixture eliminates almost all false-positive findings. The results provides a clear suggestion for applied researchers: in any conjoint study, at least one correction method should be used.

4.2 Simulation Study 2: Non-zero AMCE

Zero AMCE for all attribute-levels is an extreme case in applied research, because attributes are carefully chosen to capture theoretically and empirically relevant concepts. To see how correction methods perform in more realistic settings, we conduct simulations under scenarios where some

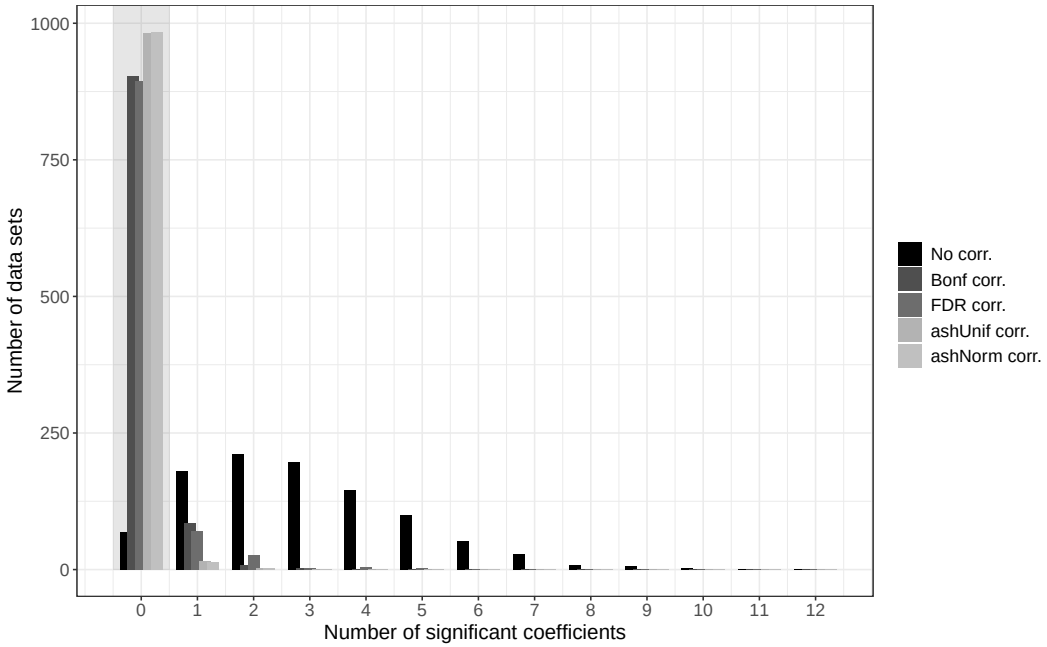

Figure 2: **False-positive Results of Estimated AMCEs when All Null Hypotheses are True with Correction Methods.** The light gray shaded area is the correct number of findings, i.e., no effects are found. While the standard analysis pipeline correctly accepts all null hypotheses in fewer than 80 data sets, the BC, BH, and Ash all correct multiple testing in more than 900 experimental trials. Among those, the Ash performs the best.

AMCEs are not zero in truth.

The first scenario is that only one binary attribute has a non-zero AMCE, and the results are shown in Table 1. In the profile data of Hainmueller, Hopkins and Yamamoto (2014), this attribute corresponds to *Gender*. We vary the noise in simulation data, by changing the heterogeneity of AMCEs and the error variance of the regression that generates latent linear responses. Since only *Gender* has an effect, the shaded cells is 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 significant estimate in addition to the true finding. All correction methods improve the situation remarkably, with ASh has the best performance in all circumstances.

In the second scenario, we set all levels of attributes *Gender*, *Education*, and *English* as having non-zero AMCEs, whereas all the other attributes have zero AMCEs. The parameters can be found in Appendix B.2 and table 2 presents the results. Because ten significant coefficients should be found in each simulated data set, the shaded cells correspond to the situation where the

			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) Panel A: Baseline (a)

			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) Panel B: Larger Error Variance (b)

			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	903	82	12	3							
	ashUnif corr.	1	982	15	3								
	ashNorm corr.	1	985	13	2								

(c) Panel C: Larger Heterogeneous AMCE and Error Variance (c)

Table 1: **Number of Data Sets for Each Number of True- and False-positive Findings when the True AMCE of `Gender` is Non-zero.** Empty cells indicate zero data set. (a) The true effect of `male` is set to $-.06$ and the effect of the reference category `female` is independently distributed as $\mathcal{N}(0, .015^2)$. The true AMCE for all other attributes is independently distributed as $\mathcal{N}(0, .015^2)$. The error term of the regression equation for the continuous response variable follows the normal distribution $\mathcal{N}(0, .01^2)$. (b) The true AMCEs are identical to Panel A, but the error term of the regression follows the normal distribution $\mathcal{N}(0, .1^2)$. (c) The true effect of `male` is set to $-.06$ and the effect of the reference category `female` is independently distributed as $\mathcal{N}(0, .12^2)$. The true AMCEs of all other attributes are identical to Panel A and regression errors follow the same data generating process as Panel B.

			No. of False Positives									
			0	1	2	3	4	5	6	7	8	9
No. of True Positives	No corr.	9	2	8	3	1	4	1				
		10	258	270	196	133	54	42	13	10	4	1
	Bonf corr.	8	38									
		9	305	6	2							
		10	623	25	1							
	BH corr.	8	4									
		9	47	25	4		1					
		10	607	208	66	23	7	6	2			
	ashUnif corr.	8	17	2								
		9	160	26	4	1		1				
		10	620	127	30	6	5	1				
	ashNorm corr.	8	21	2								
		9	172	29	3	1	1					
		10	647	99	14	7	4					

Table 2: **Number of Data Sets for Each Number of True- and False-positive Findings when the True AMCEs of All levels in `Gender`, `Education`, and `English` are Non-zero.** Empty cells indicate zero data set. True AMCE for all other attributes are set to be zero. There should be ten significant coefficients in a given test.

hypothesis testing accurately identifies true-positive conclusions and does not produce any false-positive findings. All the cells directly to the right of the shaded ones are the number of tests that contain some false-positive results, and all those directly above are the number of tests that contain some false-negative results. For example, with no correction, 258 experimental trials successfully detect exactly the true non-zero AMCEs; 270 detect those AMCEs, plus one false-positive result; two experiments do not yield any false-positive findings, but missed one non-zero effect.

Table 2 shows the trade-off that researchers need to consider in choosing a multiple testing correction method. There are false-negative conclusions with or without correction methods. Unsurprisingly, false-negative findings are more likely to occur when one applies multiple testing correction. While all correction methods increase the number of experimental trials with accurate test results remarkably, reducing the number of false positives comes at a cost of increasing the number of false negatives. Without correction, the conventional hypothesis testing misses the

AMCEs in few data sets. However, as the most conservative correction method, the BC produces false-negative conclusions in about 30% of experimental trials. The BH is the least conservative and hence least likely to miss the true AMCEs, but it produces more false-positive conclusions than the other two correction methods. The ASh takes the middle ground. It is not as conservative as the BC while not as lenient as the BH: it produces false-negative findings less likely than the BC, and false-positive results less likely than the BH.

These simulation results demonstrate the promise and pitfalls of multiple testing correction methods. First, researchers should always use some multiple testing correction method when conducting conjoint survey experiments. Since conjoint analysis inherently requires a large number of hypothesis tests, it is very likely that some, if not many, statistically significant findings are 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 reducing the possibility of false positives, but its cost is missing true findings. The BH is the opposite, and the Ash is in between the two. Since none achieves the perfect correction, 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]

5 Reanalysis

To illustrate how the use of the correction methods discussed above may change empirical conclusions, we present reanalysis of data from two published papers that use conjoint analysis. Also, Appendix D.3 shows reanalysis of an additional paper. In the reanalysis, we reproduce the findings in the original paper and apply the three correction methods.

Overall, the pattern we observe in the reanalysis is consistent with the simulations results in the previous section. In some cases, the BC seems to over-correct statistical significance. It reduces the number of findings the most, and some of the results that are changed to null by the BC are substantively questionable. On the other hand, the BH does not exclude any statistically significant findings of the original paper. The Ash correct less number of findings away than the

[5] Additional simulation results with more noisy data are shown in Appendix B.3

BC, but its results seem to make the most substantive sense.

5.1 Selecting Immigrants in the US

In the seminal paper on conjoint designs for causal inference, Hainmueller, Hopkins and Yamamoto (2014) employs the conjoint design to explore the ACME of immigrant 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 scheme is designed such that highly skilled occupations can only be taken by applicants with some college education. Similarly, "escaping persecution" as an application reason only applies to immigrants from countries where the justification is plausible. Therefore, the randomization for *Education*, *Profession*, *Country of Origin*, and *Application reasons* are conditionally independent, and the randomization for the other five attributes are completely independent. The outcome variable is binary, indicating whether respondents prefer a given profile in a paired comparison.

We focus on two attributes, *Country of origin* and *Profession*, for the comparison shown in Figure 3.[6] In the *Country of origin* attribute, India is the reference category. The left panel of Figure 3 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 3 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

[6]For the entire replication results, see Figure D.1 in Supplementary Appendices.

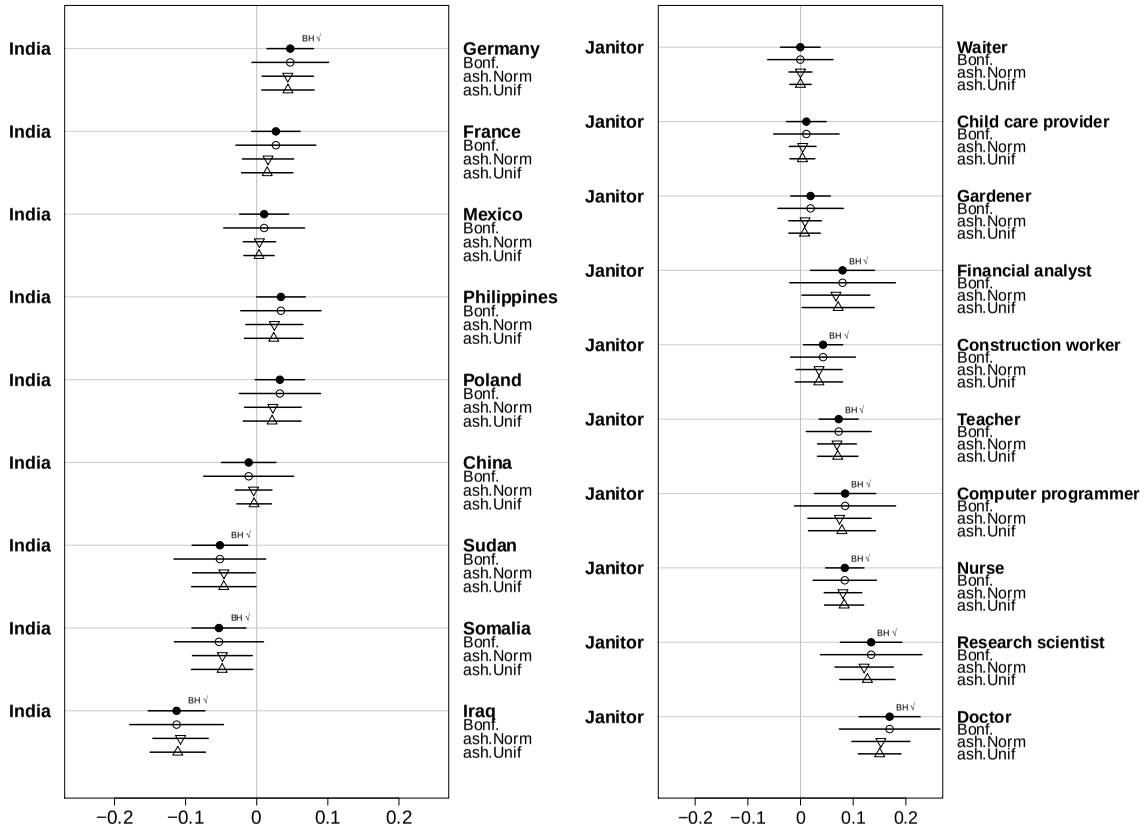

Figure 3: Effects of the Immigrant's Country of Origin (left) And Profession (right) on the Probability of Being Preferred for Admission to the United States. For `country of origin`, the reference category is India; for `profession`, the reference category is janitor. The plot shows estimates with no 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 point estimates indicates BH corrected coefficient is significant for that specific attribute level. Estimates are based on regression estimators with clustered standard errors at respondent level; bars represent 95% confidence intervals. Bars with solid circles are estimates with no correction, which replicate corresponding attributes in Figure 3 of Hainmueller, Hopkins and Yamamoto (2014, p.21).

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, Bernauer and Umaña (2016) explores what types of countries are preferred partners for Preferential Trade Agreements (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 4 focuses on the effect of two attributes *Military ally* and *Environmental protection standards* on the respondents in Vietnam.[7] 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, Bernauer and Umaña, 2016, p.710,714). However, Vietnam has a "Three Nos" defense policy since 1998:

[7]The complete replication results can be found in Figure D.2 of Supplementary Appendices.

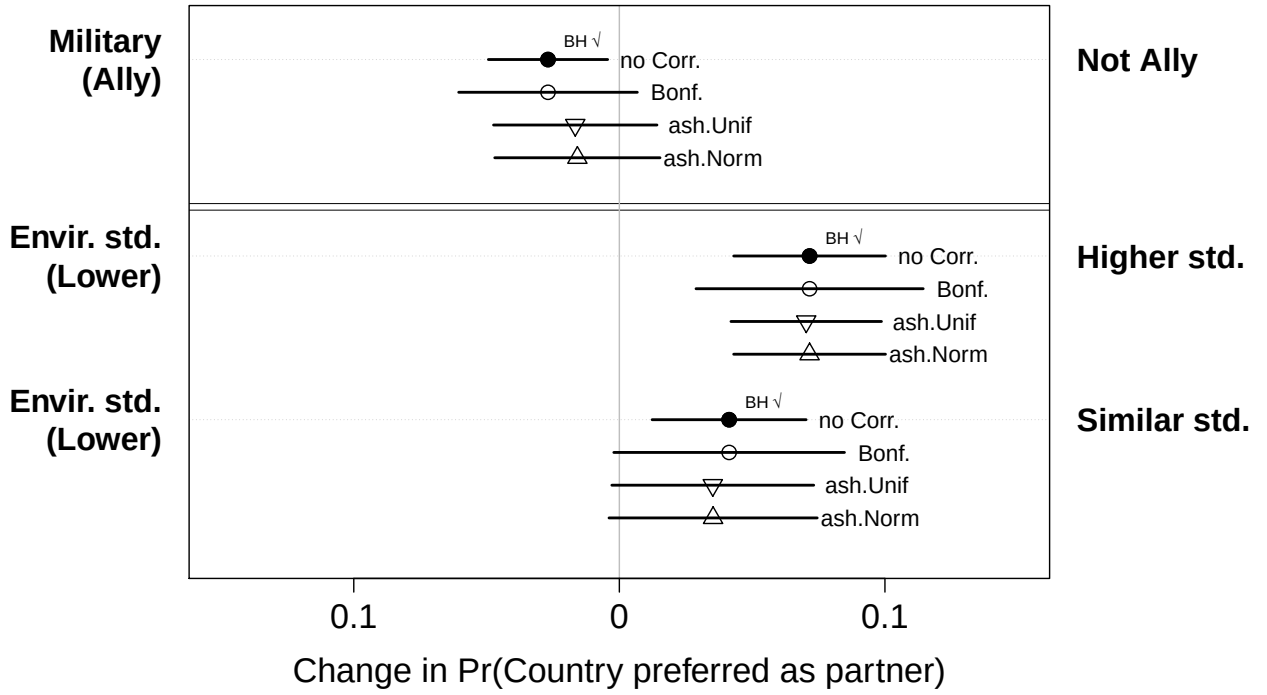

Figure 4: **Effects of Military Ally (Top) and Environmental Protection Standards (bottom) on the Probability of Being Preferred as Trading Partners in Vietnam.** For `Military ally`, the reference category is allied; for `Environmental Protection Standards`, the reference category is lower standards. 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 point estimates indicates the BH corrected coefficient is significant for that specific attribute level. Estimates are based on regression estimators with clustered standard errors at respondent level; bars represent 95% confidence intervals. Bars with solid circles are estimates with no correction, which replicate corresponding attributes in Figure 1.3 in Spilker, Bernauer and Umaña (2016, p.715).

no military alliance, no aligning with one country against another, and no foreign military bases on Vietnamese soil.[8] The 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

[8]Socialist Republic of Vietnam Ministry of National Defence, 2009.

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.

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 and empirical guidance on the choice of correction methods have been provided. In a series of simulations and applications using 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 Bonferroni correction is most conservative, and thus it is least likely to mislead researchers to false positive conclusions. However, it is most likely to mislead researchers to false negative conclusions. The Benjamini-Hochberg procedure is the opposite. We even found that the Benjamini-Hochberg procedure does not change the statistical significance of any estimates in some applications. The adaptive shrinkage method takes a middle ground between the two. While it reduces the probability of false positives than the Benjamini-Hochberg, it avoids false negatives better than the Bonferroni correction.

Whether being conservative (or lenient) is a virtue rather than a vice depends on the purpose of researchers. We believe that the adaptive shrinkage 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 Benjamini-Hochberg procedure might be preferred if previous studies strongly suggest the existence of AMCEs, whereas the Bonferroni correction 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.

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, e.g., 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.

References

Abramson, Scott F, Korhan Koçak and Asya Magazinnik. 2019. "What Do We Learn About Voter Preferences From Conjoint Experiments?" *https://scholar.princeton.edu/sites/default/files/kkocak/files/conjoint_draft.pdf*.

Abramson, Scott, Korhan Kocak, Asya Magazinnik and Anton Strezhnev. 2020. "Improving Preference Elicitation in Conjoint Designs using Machine Learning for Heterogeneous Effects.".

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

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2020. "Using Conjoint Experiments to Analyze Elections: The Essential Role of the Average Marginal Component Effect (AMCE)." *Available at SSRN*.

Bansak, Kirk, Jens Hainmueller, Daniel J. Hopkins and Teppei Yamamoto. 2021a. "Beyond the breaking point? Survey satisficing in conjoint experiments." *Political Science Research and Methods* 9:53–71.

Bansak, Kirk, Jens Hainmueller, Daniel J Hopkins and Teppei Yamamoto. 2021b. "Conjoint Survey Experiments." In *Advances in Experimental Political Science*, ed. James Druckman and Donald P. Green. Cambridge University Press.

Benjamini, Yoav and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics* 29:1165–1188.

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

Bland, J. Martin and Douglas G. Altman. 1995. "Multiple Significance Tests: The Bonferroni Method." *BMJ* 310:170.

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

Clayton, Katherine, Jeremy Ferwerda and Yusaku Horiuchi. 2019. "Exposure to Immigration and Admission Preferences: Evidence from France." *Political Behavior*.

de la Cuesta, Brandon, Naoki Egami and Kosuke Imai. 2021. "Improving the External Validity of Conjoint Analysis: The Essential Role of Profile Distribution." *Political Analysis*.

Dunn, Olive Jean. 1961. "Multiple Comparisons among Means." *Journal of the American Statistical Association* 56:52 – 64.

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

Fournier, Patrick, Stuart Soroka and Lilach Nir. 2020. "Negativity Biases and Political Ideology: A Comparative Test across 17 Countries." *American Political Science Review* 114:775–791.

Ganter, Flavien. 2022. "Identification of Preferences in Forced-Choice Conjoint Experiments: Reassessing the Quantity of Interest." *Political Analysis*.

Gerard, David and Matthew Stephens. 2018. "Empirical Bayes Shrinkage and False Discovery Rate Estimation, Allowing For Unwanted Variation." *Biostatistics*.

Greenland, Sander and James M. Robins. 1991. "Empirical-Bayes Adjustments for Multiple Comparisons Are Sometimes Useful." *Epidemiology* 2:244–251.

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

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

Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments Against Real-world Behavior." *Proceedings of the National Academy of Sciences* 112:2395–2400.

Horiuchi, Yusaku, Zachary D. Markovich and Teppei Yamamoto. 2020. "Does Conjoint Analysis Mitigate Social Desirability Bias?" *Available at SSRN*.

Incerti, Trevor. 2020. "Corruption Information and Vote Share: A Meta-Analysis and Lessons for Experimental Design." *American Political Science Review* 114:761–774.

Leeper, Thomas J., Sara B. Hobolt and James Tilley. 2020. "Measuring Subgroup Preferences in Conjoint Experiments." *Political Analysis* 28:207–221.

List, John A., Azeem M. Shaikh and Yang Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Economics* 22:773–793.

Liu, Hanzhang. 2019. "The Logic of Authoritarian Political Selection: Evidence from a Conjoint Experiment in China." *Political Science Research and Methods* 7:853–870.

Ludbrook, John. 1998. "Multiple Comparison Procedures Updated." *Clinical and Experimental Pharmacology and Physiology* 25:1032–1037.

Oliveros, Virginia and Christian Schuster. 2018. "Merit, Tenure, and Bureaucratic Behavior: Evidence From a Conjoint Experiment in the Dominican Republic." *Comparative Political Studies* 51:759–792.

Ono, Yoshikuni and Barry C. Burden. 2019. "The Contingent Effects of Candidate Sex on Voter Choice." *Political Behavior* 41:583–607.

Romano, Joseph P. and Azeem M. Shaikh. 2006a. "On stepdown control of the false discovery proportion." *Optimality* 49:33–50.

Romano, Joseph P. and Azeem M. Shaikh. 2006b. "Stepup Procedures for Control of Generalizations of The Familywise Error Rate." *Annals of Statistics* 34:1850–1873.

Romano, Joseph P., Azeem M. Shaikh and Michael Wolf. 2008. "Formalized Data Snooping Based on Generalized Error Rates." *Econometric Theory* 24:404–447.

Romano, Joseph P. and Michael Wolf. 2010. "Balanced Control of Generalized Error Rates." *Annals of Statistics* 38:598–633.

Sarkar, Sanat K. and Chung-Kuei Chang. 1997. "The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics." *Journal of the American Statistical Association* 92:1601–1608.

Sen, Maya. 2017. "How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment." *Political Research Quarterly* 70:374–393.

Shafranek, Richard M. 2019. "Political Considerations in Nonpolitical Decisions: A Conjoint Analysis of Roommate Choice." *Political Behavior*.

Sjölander, Arvid and Stijn Vansteelandt. 2019. "Frequentist versus Bayesian Approaches to Multiple Testing." *European Journal of Epidemiology* 34:809–821.

Spilker, Gabriele, Thomas Bernauer and Víctor Umaña. 2016. "Selecting Partner Countries for Preferential Trade Agreements: Experimental Evidence From Costa Rica, Nicaragua, and Vietnam." *International Studies Quarterly* 60:706–718.

Stephens, Matthew. 2017. "False Discovery Rates: A New Deal." *Biostatistics* 18:275–294.

Stephens, Matthew, Peter Carbonetto, David Gerard, Mengyin Lu, Lei Sun, Jason Willwerscheid and Nan Xiao. 2020. *ashr: Methods for Adaptive Shrinkage, using Empirical Bayes*.

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64:479–498.

Teele, Dawn Langan, Joshua Kalla and Frances Rosenbluth. 2018. "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." *American Political Science Review* 112:525–541.