Multiple Hypothesis Testing in Conjoint Analysis*

Guoer Liu† Yuki Shiraito‡

This draft: October 30, 2021
First draft: January 5, 2021

Abstract

Conjoint analysis is widely used for estimating the effects of a large number of treatments on multidimensional decision making such as voting for electoral candidates. However, it is this substantive advantage that leads to a statistically undesirable property, multiple hypothesis testing. With a few exceptions, existing applications of conjoint analysis do not correct for the number of hypotheses to be tested. This paper shows that even when none of the treatments has any effect, the standard analysis pipeline produces at least one statistically significant estimate of an average marginal component effect in more than 80% of experimental trials. Then, we conduct a simulation study to compare three methods for multiple testing correction, the Bonferroni correction, the Benjamini-Hochberg procedure, and the adaptive shrinkage method. 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, slum brokers, and partner countries of trade agreements.

*The authors thank Naijia Liu and participants at the Joint Conference of Asian Political Methodology Meeting VIII and Australian Society for Quantitative Political Science Meeting IX for helpful comments and discussions on an earlier draft.

†Ph.D. student, 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. Undoubtedly, 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 for it. 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. Since 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. Generally, 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 0.05, 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. For example, the canonical immigration application in Hainmueller, Hopkins and Yamamoto (2014) 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 multiple testing problem can be in the context of conjoint analysis. Moreover, while several correction methods are used in applied statistics, the consequences of the choice of a correction methods 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 80% of simulated data sets even when all AMCEs are zero in truth. Moreover, our simulations suggest that this probability is higher in real-world applications, because heterogeneity in the effects among respondents increases the chance of getting false positive results.

Second, we compare the strengths and limitations of two well-known correction methods, the Bonferroni correction (Bland and Altman, 1995) and the Benjamini-Hochberg procedure (BH) (Benjamini and Hochberg, 1995). In addition, we introduce a newly developed correction method, adaptive shrinkage (ASh) (Stephens, 2017; Gerard and Stephens, 2018). ASh shows great potential in biology applications that involve multiple 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 Bonferroni procedure produces the most conservative correction. However, while it guards against false positive conclusions, the cost of false negative conclusions can be significant at times. BH is the least susceptible to false negative conclusions, but the rank-based procedures do not produce uncertainty measures of estimates. 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 three prominent conjoint design applications in American politics, comparative 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 a study on clientelism in India (Auerbach and Thachil, 2018) demonstrates the difference between BH and ASh, although both are based on the idea of controlling false discovery rate. Third, 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 inter-

pretation (Egami and Imai, 2019; de la Cuesta, Egami and Imai, 2021; Abramson, Koçak and Magazinnik, 2019; Abramson et al., 2020; Bansak et al., 2020), 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. Second, 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

The fact that a large number of null hypotheses are tested 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 *family-wise error rate* (FWER), is $\text{FWER} = 1 - \mathbb{P}(\text{None of the ten tests rejects the null}) = 1 - (1 - .05)^{10} = .401$. If the number of tests is 20, FWER increases to .642. (See Figure 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. In this section, we show the results of a simulation study to quantify how serious the problem can be in a typical conjoint analysis setting.

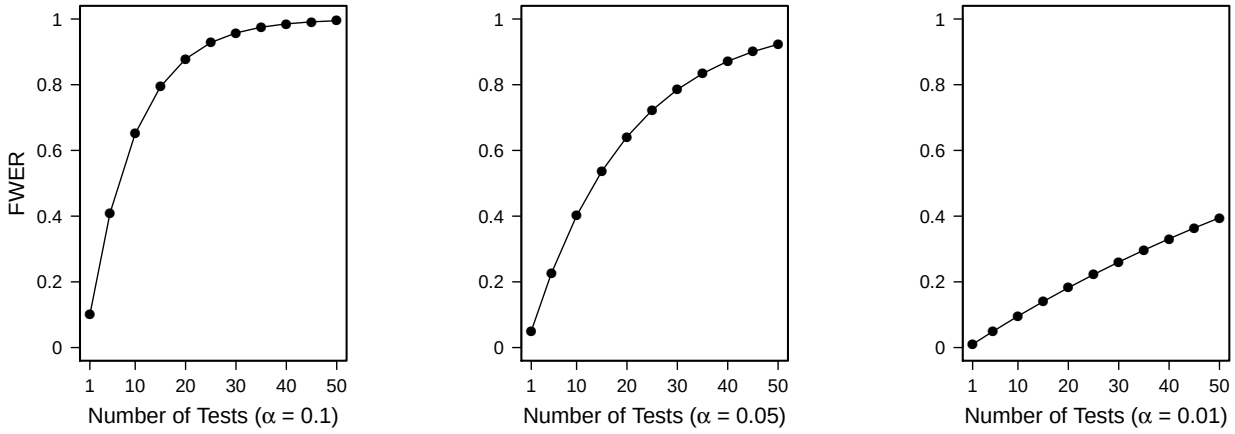


Figure 1: FWER at Varying Number of Tests Given a Significance Level. Panels show the FWER under α being 0.1, 0.05, and 0.01. Because most existing conjoint designs adopt $\alpha = 0.05$ as the threshold, the panel in the middle corroborates most applied cases. The trend clearly suggests that when the number of tests grows, it is almost guaranteed that the conjoint experiment will produce at least one significant AMCE due to chance.

2.1 Simulation Setup

The conjoint design of our simulations follows exactly the experimental design of Hainmueller, Hopkins and Yamamoto (2014). We tested the forced-choice case under the most difficult circumstances where AMCE is zero. The quantities of interests are the effects of immigrant attributes on preference for admission to the United States. There are nine discretely valued attributes: *Gender* is a binary variable; *Education* takes seven levels; *Language* has four levels; *Country of origin* consists of 10 counties; *Profession* include 11 job categories; *Job experience* takes four levels; *Job plans* includes four scenarios; *Application reasons* includes three commonly stated reasons for immigration; and *Prior trips to the U.S.* includes five types of previous travel experience to the U.S.. Each attribute contains a reference category, so the interpretation of AMCE for a given attribute is always relative to the marginal effect of the reference category. For example, "no formal education" is the reference category for *education*. The interpretation of "college degree" is that shifting an immigrant's education level from no formal education to having a college degree increases the probability of being preferred for admission to the United States by some percentage points.

There are two cases where AMCE is zero: 1) zero individual Marginal Component Effect

(MCE); 2) nonzero individual MCE but zero AMCE. We examine the two cases separately.

Zero Individual MCE

We could view the data generating process from two different perspectives. One is that the chosen outcome is a linear combination of a set of dummy variables for attribute values. The coefficient estimate of the respective dummy variable is the AMCE of the comparison category relative to the reference category. Dummy variables come from the immigrant conjoint experiment discussed above, but we simulate respondent i's choice in kth paired-comparison with j profiles using the following scheme:

$$Y_{ijk} = \begin{cases} 1, & \underset{j}{\text{argmax}}(T'_{ijk}\boldsymbol{\beta} + \epsilon_i) \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \quad \sum_{j}^{J} Y_{ijk}(T_{ijk}) = 1 \tag{1}$$

where T_{ijk} is a vector indicates the treatment given to respondent i as the jth profile in her kth paired-choice tasks and $|T_{ijk}| = L$, which is the total number of attributes in a conjoint study. In our case, $L = 9$. Because each attribute l is a categorical or ordinal variable, it can be decomposed as a set of dummy variables.

For zero individual MCE, we generate 1000 datasets by setting $\boldsymbol{\beta}$ and ϵ_i as the following, $\boldsymbol{\beta}_{ijkl} \overset{iid}{\sim} N(0, 0.015^2)$ Standard deviation is set as the same as the median of empirical standard error for $\boldsymbol{\beta}$ in the paper. $\epsilon_i \overset{iid}{\sim} N(0, 0.01^2)$. Figure 2a summarizes the simulation results. The X-axis indicates the number of significant coefficients per test. The significance level is set at α=0.05. Y-axis is the number of data sets. Out of 1000 tests, less than 200 correctly conclude that there are no statistically significant attribute levels. Some tests have identified as many as 10 significant attribute-levels.

Another approach is to view the data generating process in the potential outcome framework. For any pair of profile set $\boldsymbol{t}_0$ and $\boldsymbol{t}_1$, the *unit treatment effect* is the difference between the two potential outcomes under the two profile sets for respondent i, $\pi_i(\boldsymbol{t}_1, \boldsymbol{t}_0) \equiv Y_i(\boldsymbol{t}_1) - Y_i(\boldsymbol{t}_0)$. For each paired-comparison, $J = 2$, let $Y_{ijk}(\bar{\boldsymbol{t}})$ indicates whether respondent i chooses the jth profile in her kth comparison when she receives a sequence of profile attributes $\bar{\boldsymbol{t}}$. For zero individual

MCE for all attributes, Y_{ijk} is independent of T_{ijk} and follows

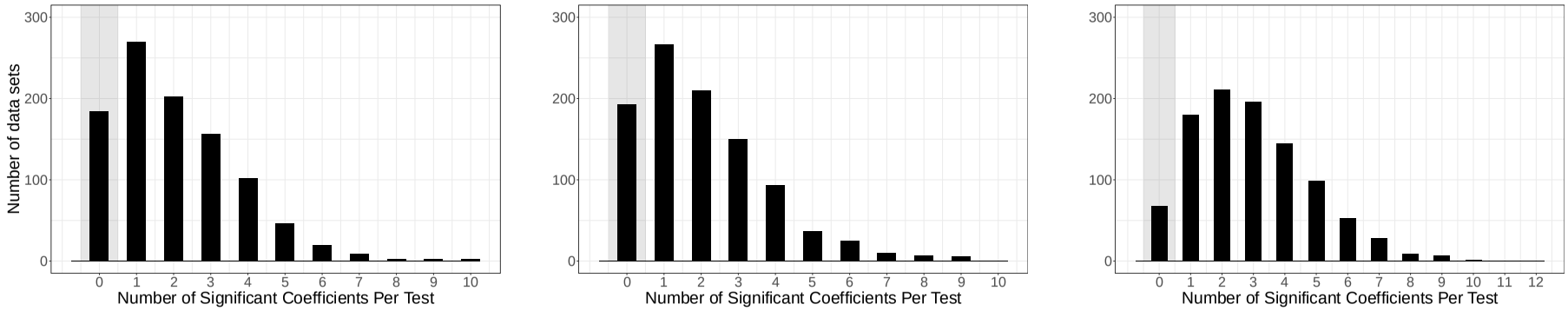
$$\begin{cases} Y_{i,j,k} & \overset{\text{i.i.d.}}{\sim} \mathsf{Bern}(.5) \\ Y_{i,-j,k} & = 1 - Y_{i,j,k} \end{cases}$$

The simulation results is represented in 2b. Similar to the parametric approach, less than 200 tests find no significant attribute. Some tests have identified 10 false-positive results, even when there is none.

Nonzero Individual MCE But Zero AMCE

Following the parametric approach in equation 1. For half of the respondents, we set individual MCE to be positive and significant, where $\boldsymbol{\beta}_{ijkl} \overset{iid}{\sim} N(0.06, 0.015^2)$. For the other half, individual MCE is negative and significant, $\boldsymbol{\beta}_{-ijkl} \overset{iid}{\sim} N(-0.06, 0.015^2)$. AMCE is zero nonetheless. Standard deviation is set to be $\epsilon_i \overset{iid}{\sim} N(0, 0.01^2)$. Figure 2c summarizes the results. Even fewer tests draw the correct conclusion: less than 75 tests identified the null results. Compared with zero individual MCE results, the curve shifts rightward—indicating a tendency for more false-positive conclusions. In extreme cases, 12 significant attribute levels were identified despite there is none. This difference is expected: nonzero individual MCE adds more noise to the data, making false-positive conclusions more likely.

This pattern is particularly concerning because nonzero individual MCE is a more realistic scenario in applied research, whereas zero individual MCE is the rare case. Attributes are included in a conjoint design because researchers expect some MCE based on theoretical reasoning or existing empirical evidence. And MCE may be more pronounced for some subpopulation than others. Design elements like these could induce more false-positive findings.



(a) Parametric Approach: Zero Individual MCE

(b) Potential Outcome Framework: Zero Individual MCE

(c) Parametric Approach: Nonzero Individual MCE but Zero AMCE

Figure 2: **False Positive Results of Estimated AMCEs when the Null Hypothesis is True.** (a) The true effects of the randomly assigned immigrant attributes on the probability of being preferred for admission to the United States is indepndently distributed as $\mathcal{N}(0, 0.015^2)$ for all respondents, all profiles, all attributes at all levels. Individual standard errors follow the normal distribution $\mathcal{N}(0, 0.01^2)$. (b) The true marginal component effects of the randomly assigned immigrant attributes on the probability of being preferred for admission to the United States are zero under the potential outcome framework. (c) For half of the respondents, the true AMCE is independently distributed as $\mathcal{N}(0.06, 0.015^2)$, and for the other half the true AMCE is independently distributed as $\mathcal{N}(-0.06, 0.015^2)$. Individual standard errors follow the normal distribution $\mathcal{N}(0, 0.01^2)$. The results are across 1000 simulated datasets using the profiles from immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014).

3 Multiple Testing Correction Methods

There exists a wealth of correction methods for multiple testing problems. In this section, we discuss the intuition of two relatively well known methods, Bonferroni Correction and Benjamini-Hochberg Procedure, and introduce a recently developed method, adaptive shrinkage. Their respective advantages and limitations will be illustrated in a simulation study. All the methods can be implemented in standard software packages like `R`.

3.1 Bonferroni Correction

Bonferroni correction (Dunn, 1961) tackles the multiple testing problem by adopting a more stringent threshold as the number of tests increases. The new significance level is proportional to the number of tests using the cutoff at $\frac{\alpha}{\#\text{ of tests}}$. For instance, when we test one hypothesis, it is the same as the case without correction. When we test five hypotheses simultaneously, the new significance level is $\alpha^* = \frac{0.05}{5} = 0.01$. We can reject a null hypothesis only when the p-value is smaller than or equal to α^*. With Bonferroni correction, the FWER is well contained at the intended significance level α overall.

While there are other useful methods to control for FWER (Sarkar and Chang, 1997; Ludbrook, 1998; List, Shaikh and Xu, 2019), Bonferroni correction provides perhaps the most intuitive solution to the multiple testing problem. Because α is the only parameter that we need modify, it is easy to implement. The confidence interval construction follows the usual procedure with the new α^*. Nevertheless, the theoretical results assume that all tests are independent of each other. If tests are correlated, which is usually the case, Bonferroni correction can be overly conservative. Hence to avoid false positive findings, it fails to reject false null hypothesis—so we accept that the null hypothesis even when it is truly false. The power becomes very low as a result. We illustrate this point later in the simulation study.

A philosophical critique of Bonferroni correction is that "the total number of tests" cannot be clearly defined and tracked (Sjölander and Vansteelandt, 2019). What tests count towards the collection of tests for multiplicity adjustment? The experiment setup has alleviated the problem somehow, since all relevant items of interests are included in the survey. Even so, there are tests

that researchers do to ensure survey quality, such as filtering responses that do not pass the attention checks. Further, many conjoint analyses often carry out subgroup comparisons (Leeper, Hobolt and Tilley, 2020), should we include these tests in the denominator? The ambiguous standard is another limitation in Bonferroni correction.

3.2 Benjamini-Hochberg Procedure

Whether being conservative is a virtue depends on the contexts. If the goal is to identify a set of candidate attributes that can be used in later lab or field experiments for validation, it may be ideal to risk having some false positive findings as long as the number of findings are controlled and bounded. So controlling FWER is not the focus. Instead, we pivot to false discovery rate (FDR), the ratio of false discoveries out of the total number of discoverties. That is, $\mathbb{E}[\frac{\text{No. false discoverties}}{\text{No. total discoverties}}] \leq \alpha$. Benjamini-Hochberg procedure is perhaps the most well-known tool to control FDR.

BH is a rank-based method that takes four steps. 1) For each of the total m hypotheses of interests, we follow the standard procedure to compute test statistics and obtain a p-value by comparing the test statistic against the appropriate distribution. We get a m-vector of p-values. 2) Rank all the elements in the m-vector from the smallest to the largest. 3) Define a threshold, $k = \max\{i : p_i \leq \alpha\frac{i}{m}, 0 \leq i \leq m\}$, which contains the FDR α that we are willing to tolerate. 4) Reject all 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.

A detailed discussion on the mathmetical properties of BH is beyond the scope of this paper. Interested readers could 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 extentions. Compared with Bonferroni correction, BH is less susceptible to false negative conclusions because we aim at controlling FDR rather than FWER. BH also gives more power than Bonferroni correction, although when all m null hypotheses are true, controlling FDR using BH is identical to controll FWER using Bonferroni, which we will see in the simulation study. However, BH corrected results can be sensitive to the pre-specified FDR by researchers. Another limitation is that because BH focuses on the p-values, it does not offer a straightforward way to construct uncertainty measures.

3.3 Empicial Bayes Shrinkage

Empirical Bayes Shrinkage is a recently proposed method that approaches the traditional FDR controlling from a Bayesian perspective. While there is a vast literature on this topic (Efron, 2010), our discussion draws upon Stephens (2017); Gerard and Stephens (2018)'s adaptive shrinkage (ASh), because ASh is a relatively simple model and it can be implemented with `ash` package in `R`. Applied researchers could easily incorporate the correction procedures in conjoint analysis routine.

Despite its philosophical identity problem, Empirical Bayes provides a useful approach to multiple hypotheis testing (Greenland and Robins, 1991; Efron, 2019). It allows us to regularizes the uncertainty measure with the Bayesian set-up, but the prior parameters are estimated by optimizing the marginal likelihood of the data at hand. Note that this is different from the standard Bayesian, where the priors are chosen by the researchers before any data analysis. So Empirical Bayes can be underunderstood as a frequentist regularization in the sense that the regularization is achived by the MLE of the data subject to some constraints.

For ASh in particular, the constraint is that it decomposes the prior distribution into a *spike-and-slab* prior distribution: the *spike* is the point mass at 0 and the *slab* is a mixture of normal distributions (or other distributions). The likelihood of parameters comes from the data. The posterior distribution of parameters can then be computed with the priors and the likelihood. Therefore, decomposing the prior distribution is a way to regularize the posterior estimates—it can get the point mass as large as possible and remain consistent with the observed data at the same time.

Here we briefly outline the model. Detailed discussion can be found in Stephens (2017) and Gerard and Stephens (2018). Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)$ denote all J attribute levels of interests. Suppose the estimated effect size from available data is $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_J)$ and the standard error is $\hat{\boldsymbol{s}} = (\hat{s}_1, ..., \hat{s}_J)$. The goal is to estimate the posterior distribution of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{s}}$. Both are estimated from standard hypotheses testing. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{s}}$ are two arguments that we need to supply in applications.

According to Bayes's rule,

$$p(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}}) \propto p(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, \hat{\boldsymbol{s}})p(\boldsymbol{\beta}|\hat{\boldsymbol{s}}) \tag{2}$$

The likelihood for $\boldsymbol{\beta}$ follows a normal approximation:

$$p(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}}) = \prod_{j=1}^{J} \mathcal{N}(\hat{\beta}_j; \beta_j, \hat{s}_j^2) \tag{3}$$

$p(\boldsymbol{\beta}|\hat{\boldsymbol{s}})$ is the prior distribution of $\boldsymbol{\beta}$. Under unimodal assumption, we get

$$\beta_1, ..., \beta_J \overset{iid}{\sim} g \in \mathcal{U} \tag{4}$$

where $\mathcal{U}$ is a space of unimodal distribution with mode at 0. To formalize the idea, we introduce another parameter $\boldsymbol{\pi}$, which denotes the proportion of a point mass at 0 and a mixture of normal distribution centered at zero:

$$p(\boldsymbol{\beta}|\hat{\boldsymbol{s}}, \boldsymbol{\pi}) = \prod_{j=1}^{J} g(\beta_j; \pi) \tag{5}$$

$$\text{where} \qquad g(\cdot; \boldsymbol{\pi}) = \pi_0 \delta_0(\cdot) + \sum_{k=1}^{K} \pi_k \mathcal{N}(\cdot; 0, \delta_k^2)$$

$$\sum_{k=0}^{K} \pi_k = 1 \quad \text{and} \quad \pi_k \geq 0$$

δ_0 denotes a point mass, and $\delta_1, ..., \delta_K$ to be a large and dense grid of fixed positive numbers spanning a wide range. As shown in the supplementary material (Stephens, 2017), g does not have to be symmetric, and the normal mixture is not the only mixture distribution that the model allows.

The estimation takes two steps: 1) Estimate $\hat{g}$, therefore $\boldsymbol{\pi}$, by maximizing a penalized likelihood (to encourage π_0 to be as large as permitted by the observed data):

$$\begin{aligned} \hat{g} = \underset{g \in \mathcal{U}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{\beta}}|g, \hat{\boldsymbol{s}}) &= \underset{g \in \mathcal{U}}{\operatorname{argmax}} \prod_{j=1}^{J} \int_{\beta_j} g\mathcal{N}(\hat{\beta}_j|\beta_j, \hat{s}_j^2) \mathrm{d}\beta_j \\ &= \underset{g \in \mathcal{U}}{\operatorname{argmax}} \prod_{j=1}^{J} \sum_{k=0}^{K} \pi_k \mathcal{N}(\hat{\beta}_j; 0, \delta_k^2 + \hat{s}_j^2) \end{aligned} \tag{6}$$

2) Compute the posterior distribution $p(\beta_j|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{s}})$ and summarize the distributions.

ASh correction relies on the unimodal effects assumption that, as the authors suggest, is both "plausible and beneficial" in many contexts (Stephens, 2017, p.280). It is intuitive to think large effects to be rare, and small effects to be common. Moreover, even if the *detectable* non-zero effect is multimodal, with some being positive value and others being negative, it is nevertheless consistent with the main idea that *all* effects are distributed unimodal.

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. ASh also outperforms other two methods in accurately recovering the true positives results: both false positive and false negative results are well contained. 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 that is unattainable in other correction methods: 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. The *hypotheis testing* approach allow us to answer the question "*whether* applicants who finished high school is preferred to applicants with no formal education." But we cannot answer "*to what extent* applicants who finished high school is preferred to applicants with no formal education." Hypothesis testing is not set up for this task. To answer this question, we need an "estimation" approach. ASh enables us to get more precise estimates in a principled manner. We will illustrate this point in the next section with simulation data, where we know the true point estimates.

3.4 Comparing Correction Methods

To compare across different methods, we adopt the conventional significance level of $\alpha = 0.05$. First, we apply different procedures to cases where the true AMCE is zero for all attributes of interest. Second, we compare the correction performance in more noisy—perhaps more realistic—cases where the true AMCE is non-zero for some attributes. For each case, we generated 1000 datasets. All immigrant profile data comes from Hainmueller, Hopkins and Yamamoto (2014).

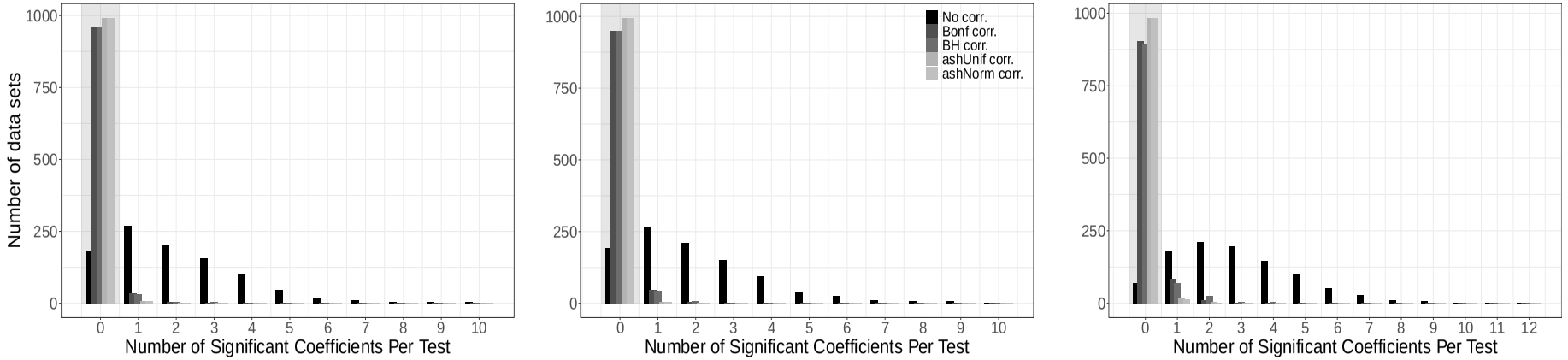
3.5 Simulation Study 1: Zero AMCE

The data generating process for zero AMCE is the same that in Section 2.1. The results are summarized in Figure 3. Note that the black bars are the same as the black bars in Figure 2. The height represents the number of datasets with corresponding significant coefficients per test without any correction. The difference is that we apply Bonferroni correction, BH correction, and ASh with a mixture of uniform components and with a mixture of normal components. Corrected results are represented by bars filled with different shades of gray.

Because the null hypotheses are true for all AMCE, ideally, all tests should result in zero significant coefficients. But as we discussed in Section 2.1, this is not the case with no correction: more than 80% tests produced at least one significant coefficient. Applying any of the correction methods we discussed in the paper improves test results dramatically. Bonferroni and BH correction removed more than 90% false-positive results on average. Considering the philogophical critique of Bonferroni correction, we define the total number of tests as the total number of pair-wise comparisons in the study, which is the most straigtforward and least punishing threshold. ASh with either a uniform mixture or a normal mixture eliminates almost all false-positive findings.

3.6 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 further divide the distribution of significant coefficients into three categories. First, only one binary attribute is significant. In this application, we choose *Gender*. To cover a wide range of cases, we vary the standard deviation of effect size and



(a) Parametric Approach: Zero Individual MCE

(b) Potential Outcome Framework: Zero Individual MCE

(c) Parametric Approach: Nonzero Individual MCE but Zero AMCE

Figure 3: **False Positive Results of Estimated AMCEs when the Null Hypothesis is True Using Different Correction Methods.** (a) The true effects of the randomly assigned immigrant attributes on the probability of being preferred for admission is indepndently distributed as $\mathcal{N}(0, 0.015^2)$ for all respondents, all profiles, all attributes at all levels. Individual standard errors follow the normal distribution $\mathcal{N}(0, 0.01^2)$. (b) The true marginal component effects of the randomly assigned immigrant attributes on the probability of being preferred for admission are zero under the potential outcome framework. (c) For half of the respondents, the true AMCE is independently distributed as $\mathcal{N}(0.06, 0.015^2)$, and for the other half the true AMCE is independently distributed as $\mathcal{N}(-0.06, 0.015^2)$. Individual standard errors follow the normal distribution $\mathcal{N}(0, 0.01^2)$. For each tick on the x-axis, the black bar on the further left represents the number of datasets with corresponding significant coefficient with no correction. The other four bars, from darker to lighter shades, show the number of datasets that use Bonferroni correction (`Bonf corr.`), BH correction (`BH corr.`), ASh with a mixture of uniform components (`ash.Unif`), and ASh with a mixture of normal components (`ash.Norm`). The results are across 1000 simulated datasets using the profiles from immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014).

individual standard errors. Table 1 summarizes the results. Because only *Gender* has an effect, the shaded cell is the target we would like to hit: tests identify only gender significant attribute and nothing else. All the cells directly to the right of the shaded cells are the number of tests that contain some false positive results. We observe a similar pattern to that in zero AMCE. Without correction, about 80% of tests identified at least *one other* coefficient as significant. Different correction methods improved the situation remarkably, with ASh has the best performance in all circumstances. A quick way to interpret the table isBecause only *Gender* has an effect, the shaded cell is the target we would like to hit: tests identify only gender significant attribute and nothing else.

In the second scenario, we set all levels for *Gender*, *Education*, and *English* as significant, whereas other attributes have zero AMCE. The parameters can be found in Appendix 6.1.2. Table 2 presents the results. Because there should be ten significant coefficients in a given test, the shaded cells correspond to the situation where we accurately identified the significant coefficients and nothing else. 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 resuls. For example, with no correction, 258 tests successfully detected only the significant attributes; 270 tests detected significant attributes, but another false-positive result; two tests did not identify any false-positive results, but missed one significant attribute.

Note that now there are false negative conclusions with or without correction methods. Unsurprisingly, false-negative findings are more likely to occur when we apply multiple-testing correction methods. Without correction, about 75% of tests result in at least one false-positive conclusions. In an extreme case, there are 9 false-positive results. Correction methods alleviate the problem, but at a cost. All approaches have almost tripled the number of accurate tests: correctly identified only the ten significant attributes and nothing else. Differences in the correction methods start revealing. As the most conservative correction method, about 30% of tests corrected by the Bonferroni method result in false-negative findings. BH correction is the least conservative, but there are more false-positive conclusions consequently. ASh takes the middle ground. It is not as conservative as the Bonferroni: 20% tests have false-negative findings, but there are also more false-positive results.

Third, we simulate the situation where each attribute has exactly one significant level. The parameters can be found in Appendix 6.1.3. Table 3 summarizes the results. For a given test, there should be nine significant coefficients. Because there is more noise in the data, the performance of different correction methods is not clear-cut.

Nonetheless, only about 10% tests have accurately picked out the significant coefficients without correction. The extreme scenario leads to as many as 12 false-positive conclusions per dataset. Bonferroni correction almost doubles the number of accurate tests, but at a significant cost of false negative conclusions. BH and ASh tripled the successful tests, with BH correction risking false-positive conclusions and ASh risking false negative.

Lastly, we set the standard deviation for *Job Experience* reference category to be four times larger than that in Table 3. The parameters can be found in Appendix 6.1.4. Everything else remains the same as previously. The change would make successful detection much harder. As Table 4 suggests, while the relative strengths and weakness of different correction methods remain, none of them have precisely recovered all significant coefficients.

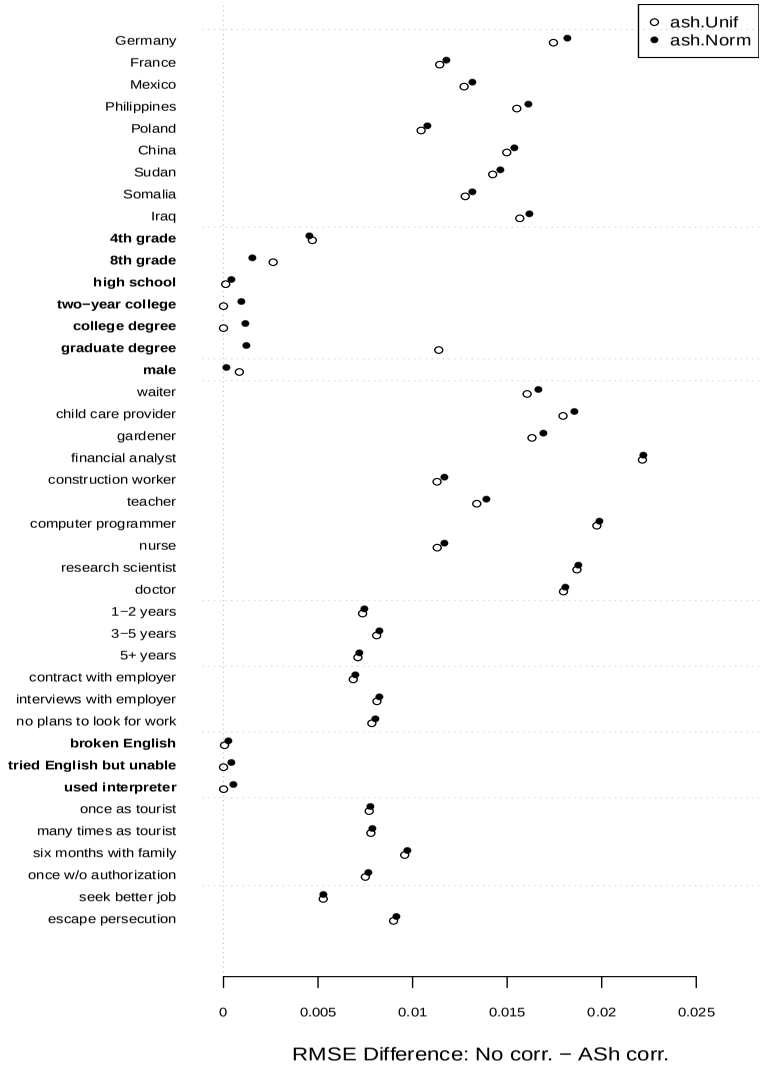
The results demonstrate the limitation of correction methods: they can only accurately recover the true results as far as data permits. If the effects of interests are *not detectable*, we should figure out other ways to capture the variables of interests or reconsider the theories behind. No correction methods could address such concerns.

3.7 Estimation: Smaller RMSE with ASh

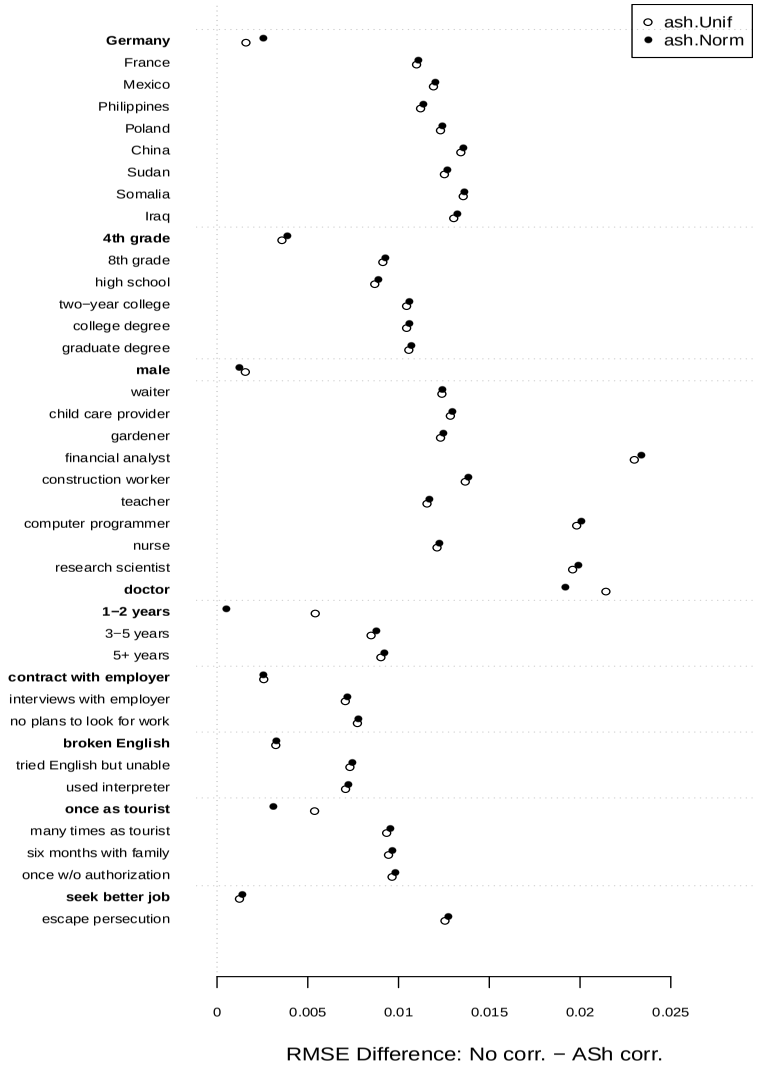
As we discussed earlier, ASh not only regularizes uncertainty measures, it also produces more accurate point estimates. This feature sets ASh apart from Bonferroni and FDR, which exclusively focus on hypothesis testing rather than estimation (Stephens, 2017). With simulation data, we can compare RMSE difference between non-corrected estimates and ASh corrected results.

Figure (4a) presents RMSE difference when the true AMCE for all levels in *Gender*, *Education*, and *English* are significant using the same parameters as the previous section. Note that all the RMSE differences are positive, meaning that non-corrected point estimates has larger RMSE than ASh corrected ones. This confirms the corrected effect size has smaller error. Additionally, ASh with a uniform mixture or with a normal mixture perform similarly in RMSE. So at least in this

application, the improvement in RMSE is not sensitive to the choice of mixture distribution. We may notice that the difference in RMSE is close to zero for significant attribute levels. This is an artifact due parameters we chose to generate the true effect size: the reference category is a random variable but the effect sizes are fixed (see Appendix 6.1.2). Figure (4b) summarizes the RMSE difference where the true AMCE for each attribute has one significant level, which relies on the same data in the previous section. The parameters can be found in Appendix 6.1.3.



(a) True AMCE for all levels in three attributes (bold) are significant



(b) True AMCE for each attribute has one significant level (bold)

Figure 4: Comparing ASh corrected RMSE and non-corrected RMSE.

No. of True Positives			No. of False Positives								
			0	1	2	3	4	5	6	7	8
	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: True AMCE for `Gender` is significant (I)

No. of True Positives			No. of False Positives											
			0	1	2	3	4	5	6	7	8	9	10	11	12
	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: True AMCE for `Gender` is significant (II)

No. of True Positives			No. of False Positives										
			0	1	2	3	4	5	6	7	8	9	10
	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: True AMCE for `Gender` is significant (III)

Table 1: **Number of Data Sets with Corresponding True Positives and False Positives Using Different Correction Methods When True AMCE for `Gender` Is Significant.**
(a) The true effect of `male`$= -0.06$ and the reference category `female` is independently distributed as $\mathcal{N}(0, 0.015^2)$. The true AMCE for all other attributes is independently distributed as $\mathcal{N}(0, 0.015^2)$. Individual standard errors follow the normal distribution $\mathcal{N}(0, 0.01^2)$.
(b) The distribution as well as true effect of `male` follow that in panel A, but individual standard errors follow the normal distribution $\mathcal{N}(0, 0.1^2)$, which has larger variance than panel A.
(c) The true effect of `male`$= -0.06$ and the reference category `female` is independently distributed as $\mathcal{N}(0, 0.12^2)$, which has larger variance than panel (a). The true AMCE for all other attributes is the same as panel A. Individual standard errors follow the data generating process in panel B.
Empty cells indicate zero data set. Shaded cells are the ones we want to hit. The results are across 1000 simulated datasets using the profiles from the immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014). For instance, with no correction in (a), 230 tests successfully detected `male` as the only significant attribute; 290 tests detected `male` as a significant attribute but there is another false positive result.

			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 with Corresponding True Positives and False Positives Using Different Correction Methods When True AMCE for All levels in `Gender`, `Education`, and `English` Are Significant.** True AMCE for all other attributes are set to be zero. There should be ten significant coefficients in a given test. Empty cells indicate zero data set. Shaded cells are the ones we want to hit. The results are across 1000 simulated datasets using the profiles from the immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014).

			No. of False Positives												
			0	1	2	3	4	5	6	7	8	9	10	11	12
No. of True Positives	No corr.	7		2				1							
		8	10	22	27	16	22	8	2	3	1				
		9	118	194	179	169	86	58	39	19	13	7	2	1	1
	Bonf corr.	5	7	3											
		6	77	5	2										
		7	244	15	7										
		8	396	37	5										
		9	180	20	2										
	BH corr.	6	5	2											
		7	37	15	5	1	1								
		8	147	89	36	11	4	1		3					
		9	321	187	75	35	12	8	1	3	1				
	ashUnif corr.	6	12	3	1	1									
		7	84	25	4	1	1								
		8	220	99	23	12	1	1							
		9	294	130	46	29	8	2	2		1				
	ashNorm corr.	5		1											
		6	11	5	2	1									
		7	98	21	5	2									
		8	224	100	24	10	1	1							
		9	295	124	42	21	7	2	2	1					

Table 3: **Number of Data Sets with Corresponding True Positives and False Positives Using Different Correction Methods When True AMCE for Each Attribute Has One Significant Level (I).** True AMCE for all other levels are set to be zero. There should be nine significant coefficients in a given test. Empty cells indicate zero data set. Shaded cells are the ones we want to hit. The results are across 1000 simulated datasets using the profiles from the immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014). For instance, with no correction, 118 tests successfully detected only the significant levels; 194 tests detected the true significant levels, but there is another false-positive result.

			No. of False Positives													
			0	1	2	3	4	5	6	7	8	9	10	11	12	13
No. of True Positives	No corr.	6			5	7	4	4	1	1						
		7			41	46	34	17	8	9	3					
		8			115	100	88	52	22	16	12	6	1		2	1
		9			100	116	82	49	31	17	5	1	4			
	Bonf corr.	4		1	37											
		5		2	247	14	1									
		6		4	365	15	1									
		7		4	224	7	3	1								
		8		2	63	2										
		9			7											
	BH corr.	4			3											
		5			32	4	2									
		6			106	28	7	4	2							
		7			212	70	17	8	1	1						
		8			229	82	38	9	7	2	1	1				
		9			77	34	13	5	3	2						
	ashUnif corr.	4			2	1		1								
		5		1	52	13	4									
		6		1	176	50	13	5								
		7			233	72	14	11	1	1						
		8			180	62	23	6	1	1	2					
		9			40	20	10	2	1	1						
	ashNorm corr.	4			4			1								
		5		1	47	13	4									
		6		1	174	49	11	3								
		7		234	71	17	8		1							
		8			187	63	23	7	1	2	1					
		9			43	20	11	1	2							

Table 4: **Number of Data Sets with Corresponding True Positives and False Positives Using Different Correction Methods When True AMCE for Each Attribute Has One Significant Level (II).** The standard deviation for the reference category for `Job Experience` is four times larger than Table 3. True AMCE for all other levels are set to be zero. There should be nine significant coefficients in a given test. Empty cells indicate zero data set. Shaded cells are the ones we want to hit. The results are across 1000 simulated datasets using the profiles from the immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014). For instance, with no correction, 0 test successfully detected only the significant levels; 0 test detected only the significant attributes and one false positive coefficient; 100 tests detected the true significant levels, but there are two false-positive results.

4 Reanalysis

Using the correction methods discussed above, we reanalyze the data from three published conjoint experiments. To compare across correction methods directly, for each replication, we reproduce the findings in the original paper, and then apply correction methods. Because BH does not give us uncertainty measures directly, BH corrected results are denoted by a significance indicator in each pairwise comparison.

4.1 Selecting Immigrants in the US

In the seminal paper on using conjoint design 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. The outcome variable is binary, indicating whether respondents prefer a given profile in a paired comparison. There are nine attributes in total: *Gender*, *Education*, *Language*, *Origin*, *Profession*, *Job experience*, *Job plans*, *Application reasons*, and *Prior trips to U.S.*. A description of all attributes can be found in Section 2.1. 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.

To compare the corrected results with the findings in the paper, we focus on two attributes *Country of origin* and *Profession*. For the entire replication results, see Figure 8 in Appendix. India is the reference category for the country of origin attribute. The interpretation of AMCE for this attribute is relative to the marginal effect of an applicant being Indian. The left panel in Figure 5 compares the results. The most noticeable pattern is that other than Iraq, none of the Bonferroni corrected coefficients are significant. Nevertheless, coefficients adjusted by BH and ASh largely preserves the original paper's conclusion that immigrants from Sudan, Somalia, and Iraq were less preferred than those from India. If we believe the Bonferroni corrected results, it

means that respondents did not distinguish immigrants from India, Mexico, France, Germany, Sudan, and Somalia conditioning on other attributes.

On *Profession*, janitor is the reference category. Results in the paper suggest a bonus for financial analysts, construction workers, teachers, computer programmers, nurses, research scientists, and doctors. Again, Bonferroni correction renders more coefficients insignificant: the probability of admission for financial analysts and computer programmers are indistinguishable from janitors. BH preserves all significant findings in the paper. ASh preserves all the results other than construction workers—the probability of admission for construction workers is the same as baseline janitors. This result is consistent with what we would substantively expect.

Because the true value is unknown, we cannot adjudicate the differences with certainty. The idea is that overly conservative correction, such as the Bonferroni method, may lead to counter-intuitive findings that require further theoretical justification. BH corrected results agree completely with the non-corrected results, including some unexpected significant attribute levels. ASh takes a middle ground. As illustrated in the simulation study, it can correct false-positive results and leave true positive results intact as long as it is consistent with the data.

4.2 Selecting Brokers in India

The conjoint design has been widely used in subfields other than American politics. Auerbach and Thachil (2018) conducted an ethnographically informed conjoint experiment in slums in urban India. Focusing on how clients shape the broker-client relationship, they examine factors that affect client preference for brokers in the context where multiple brokers compete for a following. Using a forced-choice design, they ask 2,199 slum residents to choose the preferred candidate for Development Council Presidency in a given hypothetical candidate pair. The attributes include *Broker Caste*, a binary variable indicating whether the candidate is from the same caste as the respondent; *Broker religion*, a binary variable indicating whether the candidate has the same religion as the respondent; *Broker State*, a binary variable indicating if the candidate comes from the same state; *Ethnic Rank* takes three different categories; *Broker Partisanship*, a binary variable indicating co-partisanship; *Broker Incumbent Status* contains incumbent, opposition, and independent; *Broker Connectivity* is a three-level attribute and *Broker Capability* an ordinal variable

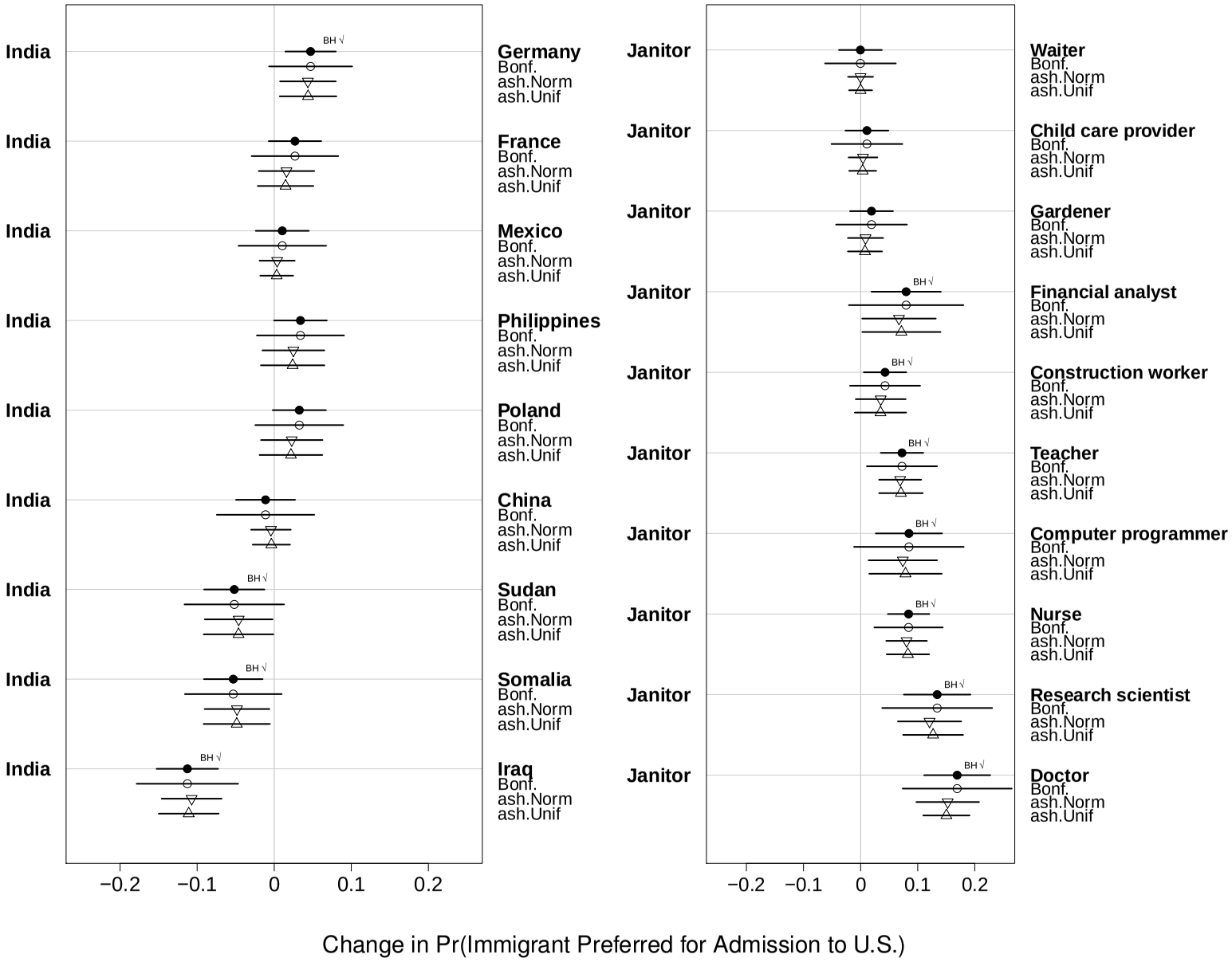


Figure 5: Effects of Immigrants Country of Origin (left) And Profession (right) on the Probability of Being Preferred for Admission to the United States. For `country of origin`, the reference category is India; for `profession`, the reference category is janitor. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), ASh with a mixture of normal components (`ash.Norm`), and 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. `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 corretion, which replicate corresponding attributes in Figure 3 in Hainmueller, Hopkins and Yamamoto (2014, p.21).

proxied by the education level. The randomization for all attributes is completely independent of each other.

We focus on the attribute *Broker Connectivity* here. For the entire replication results, see Figure 9 in Appendix. To avoid social desirability bias, broker's connectedness to urban bureaucracies is proxied by candidates' occupations. Occupations entirely contained inside the slum are considered as "low connectivity," which is the baseline. Occupations located outside the slum but not explicitly connected to municipal authorities are "medium connectivity." "High connectivity" occupations refer to those that are directly connected to municipal authorities.

As Figure 7 shows, the original analysis suggests a positive and significant AMCE of highly connected candidates relative to those who work inside the slum. The AMCE for moderately connected candidates is positive, but not significant. As a key finding in the paper, the result implies that clients prefer candidates with higher connectivity conditional all other relevant attributes. It adds to the conventional wisdom of co-ethnic and co-partisans preference in clientelistic relationships.

BH correction gives us exactly the same results as the original paper. This is guranteed by the property of BH: because there are only eight significant discoveries in the paper, the idea of controlling FDR at $\alpha = 0.05$ would remove less than one significant finding. However, both Bonferroni correction and ASh suggest otherwise. The probability of being selected as slum president is not higher for well-connected or moderately connected candidates relative to the baseline. The null result is certainly not definite. Nonetheless, it calls for more evidence to support the argument.

4.3 Selecting Trading Partners in Vietnam

Conjoint design 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) with a conjoint design using national surveys in Costa Rica, Nicaragua, and Vietnam. The outcome variable for the main analysis is binary, whether respondents choose a country profile in a paired comparison. The attributes of interests are: *Distance* from the partner country's capital, which takes three levels; *Size of the economy*, which takes three levels; *Culture*,

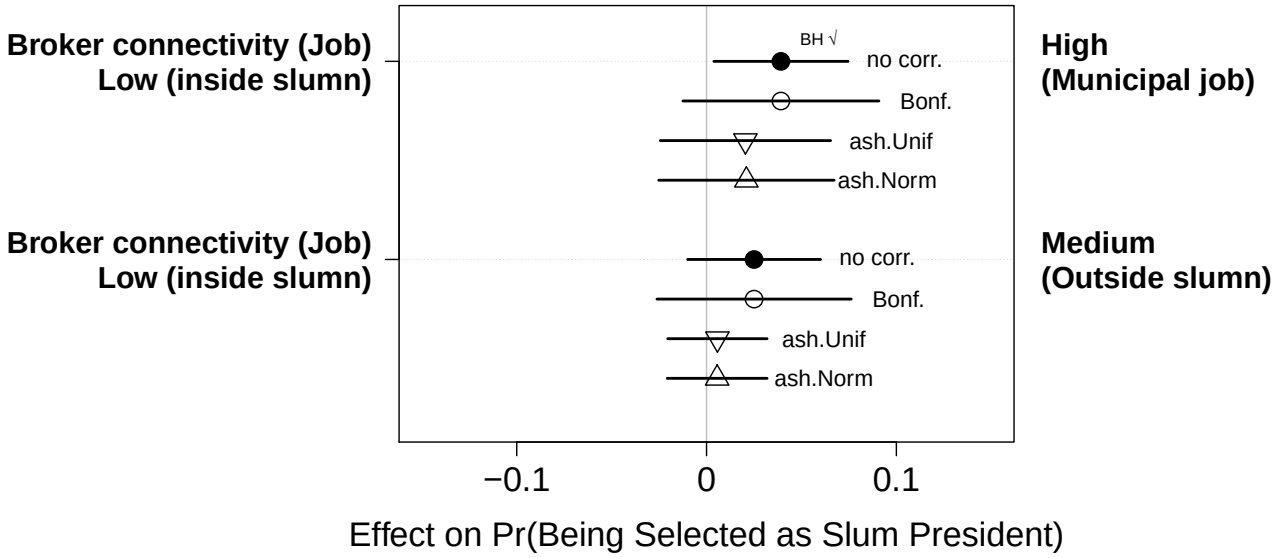


Figure 6: Effects of the Randomly Assigned Slum Leader's Connectivity on the Probability of Being Preferred for President of the Slum Development Council. The reference category is low connectivity jobs: occupations entirely contained within the slum. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), ASh with a mixture of normal components (`ash.Norm`), and 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 the corresponding attribute in Figure 1 in Auerbach and Thachil (2018, p.784)

.

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* means the extent to which citizens democratically elect political leaders, and it takes three levels; *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. The realized value for each of the eight attributes is completely independent for a given country profile.

We discuss two attributes *Military ally* and *Environmental protection standards* for Vietnam in this paper. The complete replication results can be found in Figure 10 in Appendix. Vietnam is the only country where non-military allies are punished relative to military allies, the baseline. The 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: no military alliance, no aligning with one country against another, and no foreign military bases on Vietnamese soil.[3] The context makes it difficult to interpret the significant result, because AMCE of military ally should not be significant. With the Bonferroni method or ASh, the significant finding will be corrected 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. Both correction methods render it a false positive conclusion. BH agrees completely with the original conclusion, but we cannot rule out the possibility that this is guranteed by the design of BH: there are not enough significant discoveries to begin with to control for FDR. A higher FDR may be needed to accomodate 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.

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

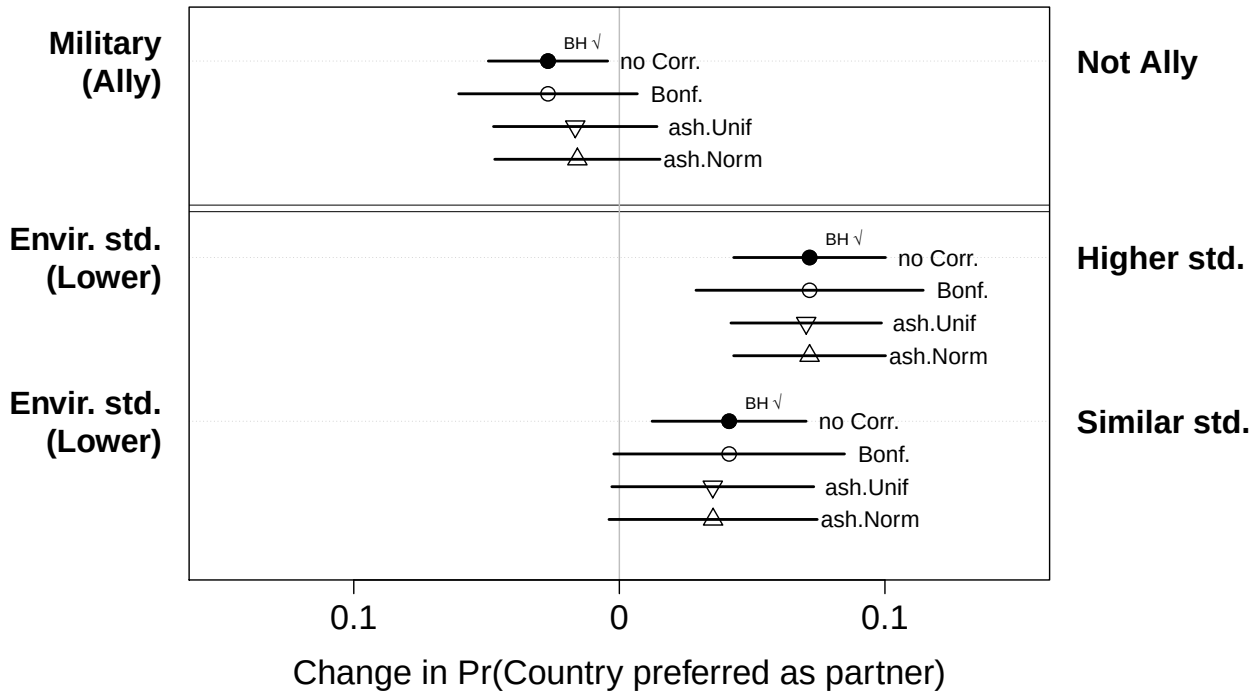


Figure 7: Effects of Military Ally (Top) and Environmental Protection Standards (bottom) on the Probability of Being Preferred as Trading Partners in Vietnam. For `Military ally`, the reference category is allied; for `Environmental Protection Standards`, the reference category is lower standards. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), ASh with a mixture of normal components (`ash.Norm`), and 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 corretion, which replicate corresponding attributes in Figure 1.3 in Spilker, Bernauer and Umaña (2016, p.715).

5 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, no systematic assessments on the severity of the problem nor empirical guidance on the choice of correction methods has 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 tests that the probability of false findings is set. We focused on conjoint analysis in this paper because the number of hypotheses to be tested is unambiguous with this method. 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 due to reviewers' comments, is much harder than to conjoint designs. More research on multiple testing correction in the other contexts than conjoint analysis 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." *https://www.korhankocak.com/publication/akms/AKMS.pdf*.

Auerbach, Adam Michael and Tariq Thachil. 2018. "How Clients Select Brokers: Competition and Choice in India's Slums." *American Political Science Review* 112:775–791.

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. Forthcoming.

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.

Efron, Bradley. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs Cambridge University Press.

Efron, Bradley. 2019. "Bayes, Oracle Bayes and Empirical Bayes." *Statistical Science* 34:177–201.

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.

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.

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.

6 Appendix

6.1 Simulations

6.1.1 Potential Outcome Framework Results: Only *Gender* is Significant

			No. of False Positives											
			0	1	2	3	4	5	6	7	8	9	10	11
No. of True Positives	No corr.	1	196	266	208	150	83	45	27	15	8		1	1
	Bonf corr.	1	958	39	3									
	BH corr.	1	907	68	19	3	3							
	ashUnif corr.	0	2											
		1	971	24	3									
	ashNorm corr.	0	3											
		1	978	17	2									

Table 5: Number of datasets with corresponding true positive and false positives using different correction methods. The true effect of `male`$= -0.06$. The true marginal component effects of other attributes on the probability of being preferred for admission to the United States are zero under the potential outcome framework. Empty cells indicate zero data set. The results are across 1000 simulated datasets using the profiles from immigrant conjoint experiment in Hainmueller, Hopkins and Yamamoto (2014). For instance, with no correction, 196 tests successfully detected `male` as the only significant attribute; 266 tests detected `male` as a significant attribute but there is another false positive result.

6.1.2 Three Attributes are Significant

In this simulation, we set all levels for *Gender*, *Edutation*, and *English* as significant and all other attributes have zero AMCE.

	Gender	**Education**	**English**	**Others attributes**
Reference level	*Female* $\sim \mathcal{N}(0, 0.015^2)$	*No formal* $\sim \mathcal{N}(0, 0.025^2)$	*Fluent* $\sim \mathcal{N}(0, 0.015^2)$	
Other levels	Male = -0.06	4th grade = 0.015	Broken Eng. = -0.05	$\sim \mathcal{N}(0, 0.015^2)$
		8th grade = 0.02	Tried but unable = -0.1	
		High school = 0.045	Use Interpreter = -0.15	
		Two-year college = 0.1		
		College = 0.13		
		Graduate = 0.17		

6.1.3 All Attributes Has One Significant Level (I)

In this simulation, each of the nine attributes has one significant level.

Attributes	*Reference level*	Significant level	Other Levels
Gender	*Female* $\sim \mathcal{N}(0, 0.01^2)$	-0.02	0
Education	*No formal* $\sim \mathcal{N}(0, 0.025^2)$	0.02	0
English	*Fluent* $\sim \mathcal{N}(0, 0.03^2)$	-0.01	0
Country origin	*India* $\sim \mathcal{N}(0, 0.1^2)$	0.05	0
Profession	*Janitor* $\sim \mathcal{N}(0, 0.02^2)$	0.02	0
Job experience	*None* $\sim \mathcal{N}(0, 0.05^2)$	0.1	0
Job plan	*Will look for work* $\sim \mathcal{N}(0, 0.015^2)$	0.01	0
App. reason	*Family reunion* $\sim \mathcal{N}(0, 0.01^2)$	-0.01	0
Prior trip exp.	*Never* $\sim \mathcal{N}(0, 0.05^2)$	0.025	0

6.1.4 All Attributes Has One Significant Level (II)

The parameters in this simulation are identical to those in 6.1.3 but for *Job experience*. We increase the standard deviation for the reference category to be four times larger than previously.

Attributes	*Reference level*	Significant level	Other Levels
Gender	*Female* $\sim \mathcal{N}(0, 0.01^2)$	-0.02	0
Education	*No formal* $\sim \mathcal{N}(0, 0.025^2)$	0.02	0
English	*Fluent* $\sim \mathcal{N}(0, 0.03^2)$	-0.01	0
Country origin	*India* $\sim \mathcal{N}(0, 0.1^2)$	0.05	0
Profession	*Janitor* $\sim \mathcal{N}(0, 0.02^2)$	0.02	0
Job experience	*None* $\sim \mathcal{N}(0, \mathbf{0.2^2})$	0.1	0
Job plan	*Will look for work* $\sim \mathcal{N}(0, 0.015^2)$	0.01	0
App. reason	*Family reunion* $\sim \mathcal{N}(0, 0.01^2)$	-0.01	0
Prior trip exp.	*Never* $\sim \mathcal{N}(0, 0.05^2)$	0.025	0

6.2 Replication

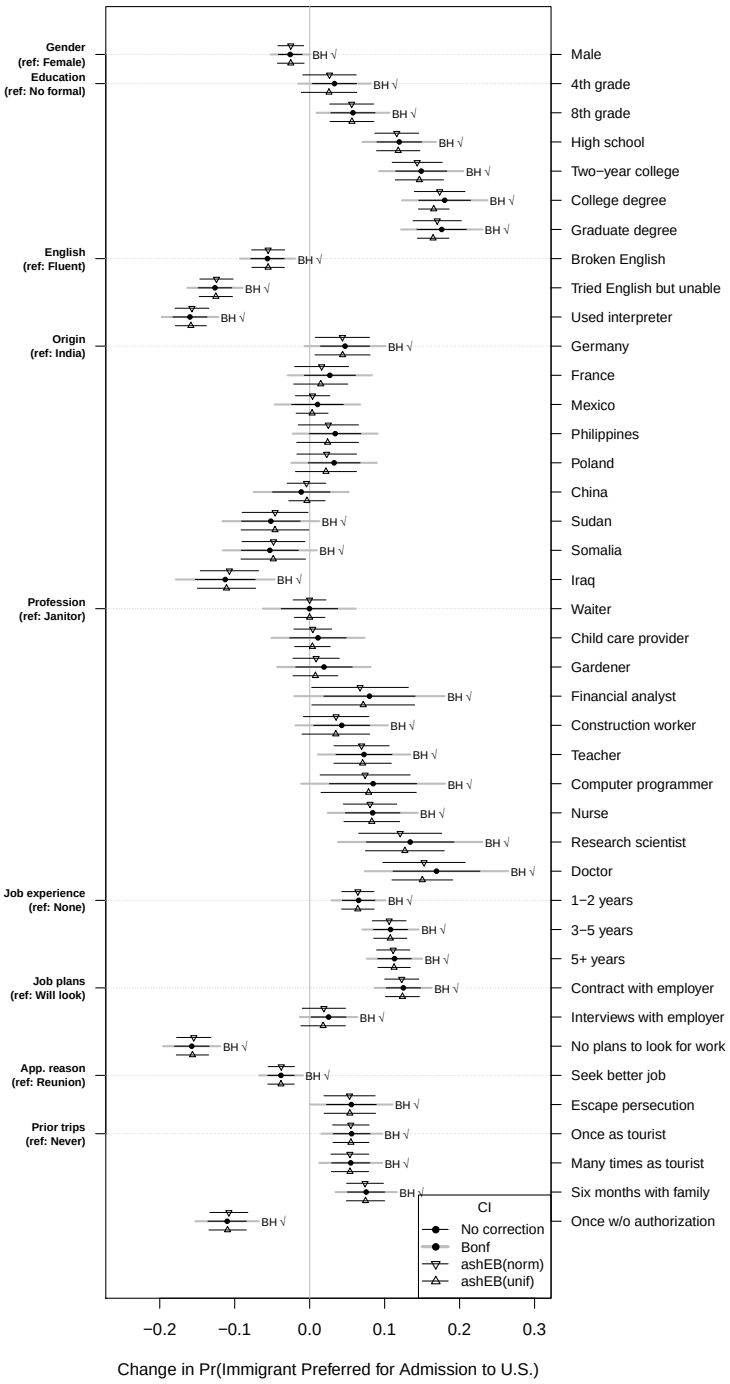


Figure 8: Effects of the randomly assigned immigrant attributes on the probability of being preferred for admission to the United States. The reference category for each attribute is in parentheses on the left side of the y-axis. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), empirical bayes shrinkage with a mixture of normal components (`ash.Norm`), and empirical bayes shrinkage with a mixture of uniform components (`ash.Unif`) for each pair of comparison. 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 corretion, which replicate results in Figure 3 in Hainmueller, Hopkins and Yamamoto (2014, p.21).

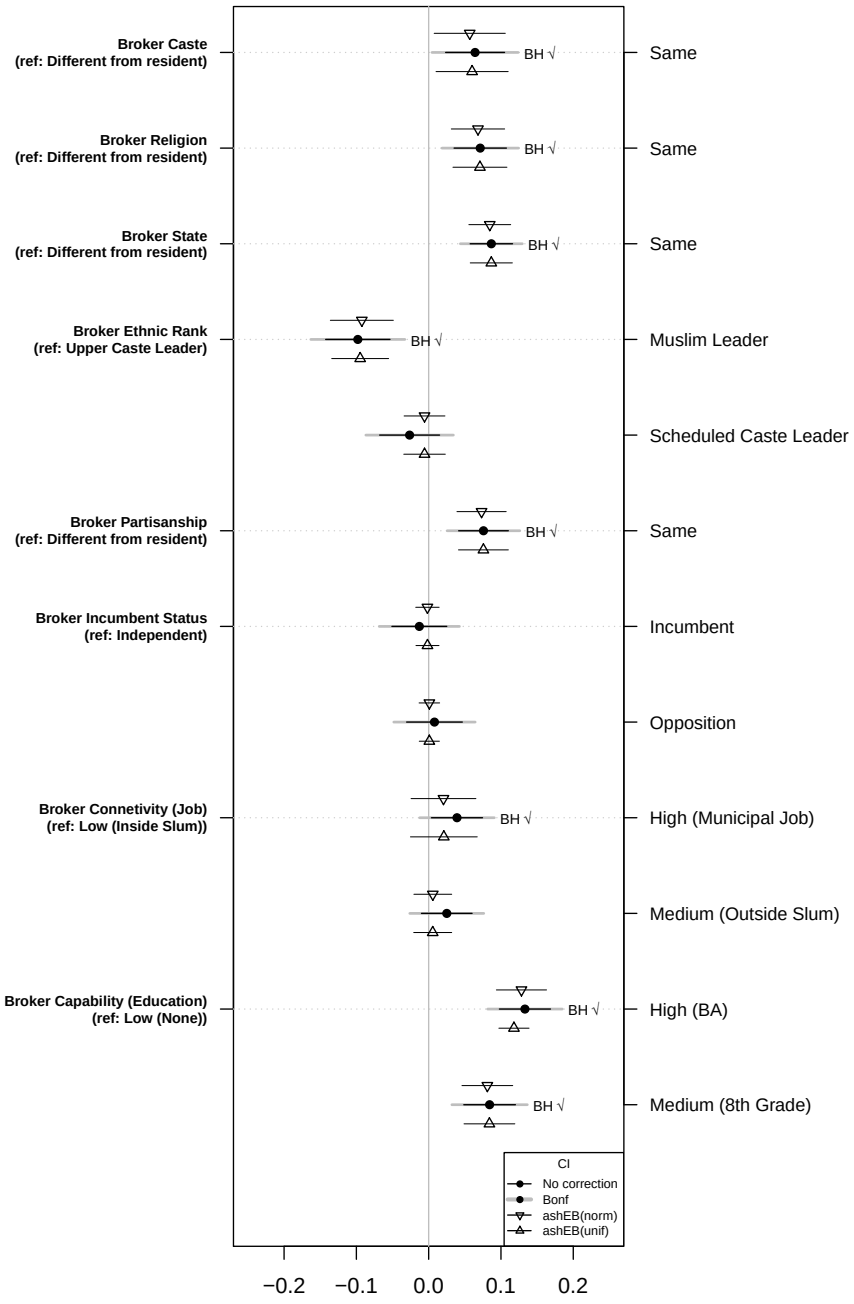


Figure 9: Effects of the randomly assigned slum leader attributes on the probability of being preferred for president of the slum development council. The reference category for each attribute is in parentheses on the left side of the y-axis. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), empirical bayes shrinkage with a mixture of normal components (`ash.Norm`), and empirical bayes shrinkage with a mixture of uniform components (`ash.Unif`) for each pair of comparison. 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 corretion, which replicate corresponding attributes in Figure 1 in Auerbach and Thachil (2018, p.784).

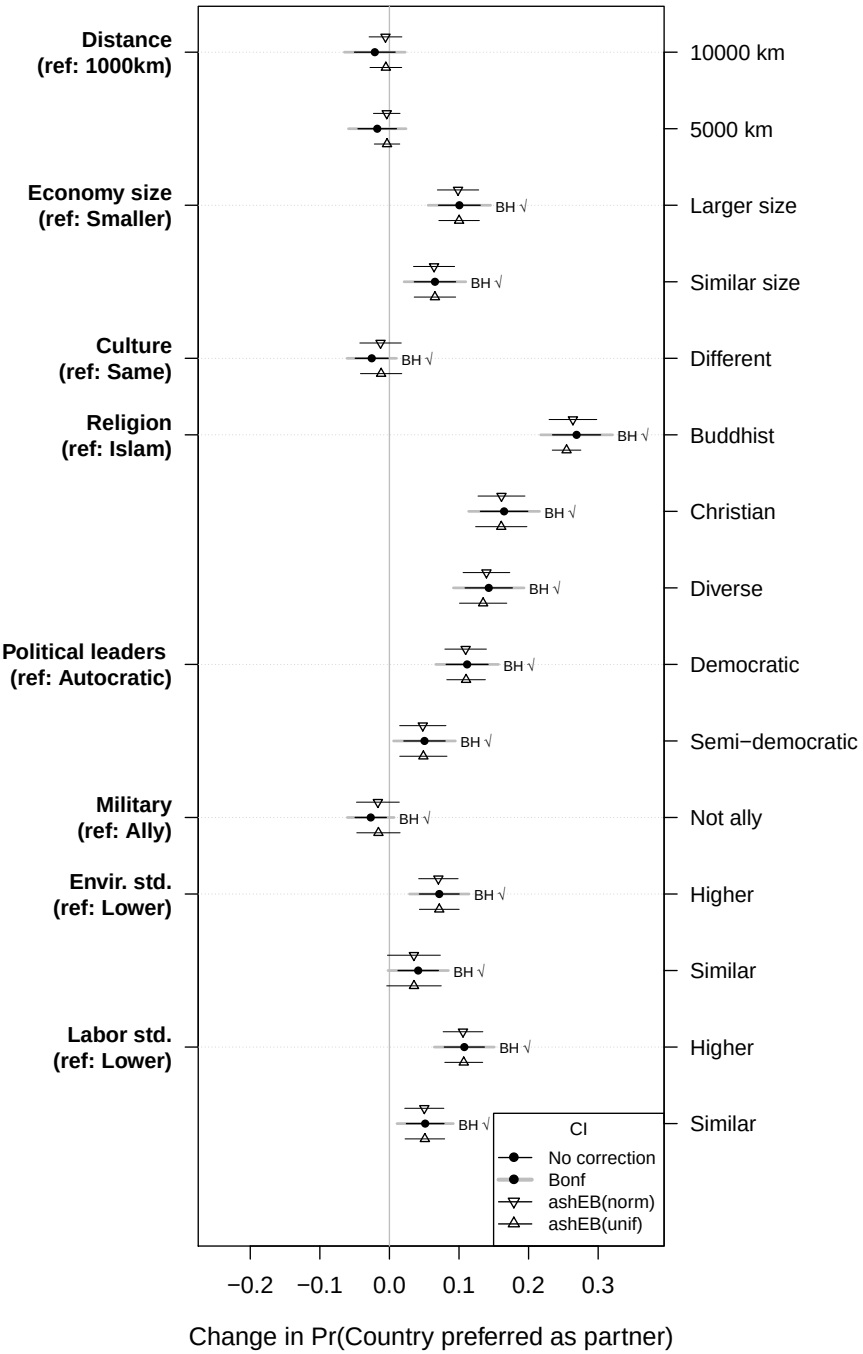


Figure 10: Effects of the randomly assigned country attributes on the probability of being preferred as trading partners in Vietnam. The reference category for each attribute is in parentheses on the left side of the y-axis. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), empirical bayes shrinkage with a mixture of normal components (`ash.Norm`), and empirical bayes shrinkage with a mixture of uniform components (`ash.Unif`) for each pair of comparison. 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 corretion, which replicate results in Figure 1.3 in Spilker, Bernauer and Umaña (2016, p.715).