# Supplementary Information for "Multiple Hypothesis Testing in Conjoint Analysis"*

Guoer Liu† Yuki Shiraito‡

First draft: January 5, 2021
Submitted for publication: October 9, 2022

## Contents



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

†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.

# A False Positive Results (Theoretical)

Figure A.1 below shows the FWER under $\alpha$ being .1, .05, and .01. Because most existing conjoint designs adopt $\alpha = .05$ as the threshold, the panel in the middle corroborates most applied cases. The trend 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.

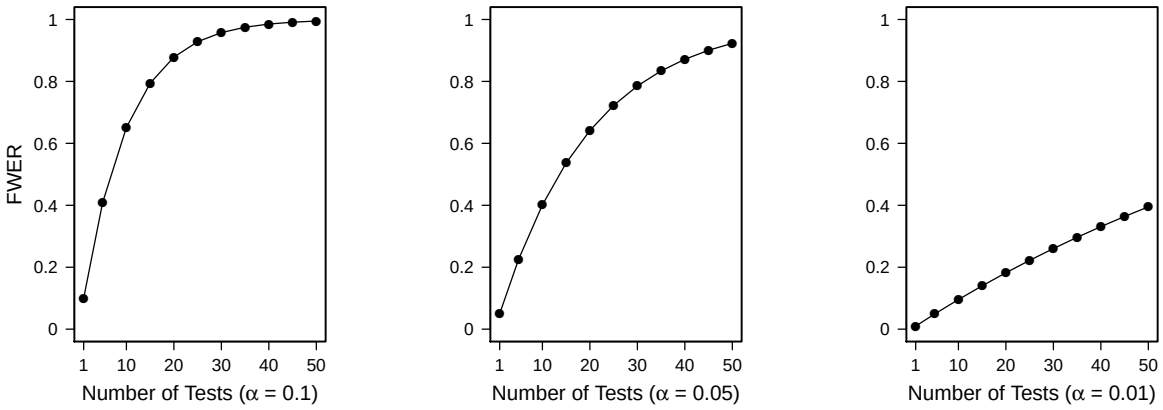



Figure A.1: FWER at Varying Number of Tests Given a Significance Level.

# B Simulations

## B.1 Data Generating Process for Simulations

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. In all simulations, we use the observed profiles in Hainmueller, Hopkins and Yamamoto (2014) to generate the dummy variables, but we simulate respondent $i$'s choice on profile $j$ in $k$th paired-comparison by the following process:

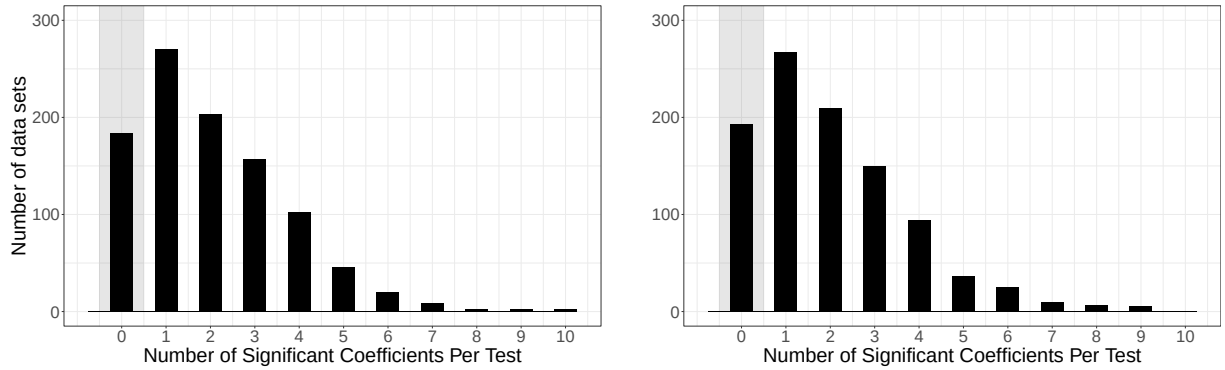
$$Y_{ijk} = \begin{cases} 1, & \underset{j'}{\text{argmax}}(T'_{ij'k}\boldsymbol{\beta} + \epsilon_i) = j \\ 0, & \text{otherwise} \end{cases} \qquad \text{and} \quad \sum_j^J Y_{ijk}(T_{ijk}) = 1 \tag{1}$$

where $T_{ijk}$ is a vector indicates the treatment given to respondent $i$ as the $j$th profile in her $k$th 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.

To simulate heterogeneous marginal component effects (MCE) across respondents, we generate $\boldsymbol{\beta}$ as random draws. For simulations presented in Section 2 and 4.1, the coefficients are drawn from the normal distribution. In particular, $\boldsymbol{\beta}_{ijl} \overset{iid}{\sim} N(.06, .015^2)$ for a half of respondents and $\boldsymbol{\beta}_{ijl} \overset{iid}{\sim} N(-.06, .015^2)$ for the other half. The standard deviation for the normal distributions is set at the median of the standard errors of $\boldsymbol{\beta}$ in the original paper. The AMCE is zero nonetheless in this setting. The error term, $\epsilon_i$, is generated as $\epsilon_i \overset{iid}{\sim} N(0, .01^2)$.

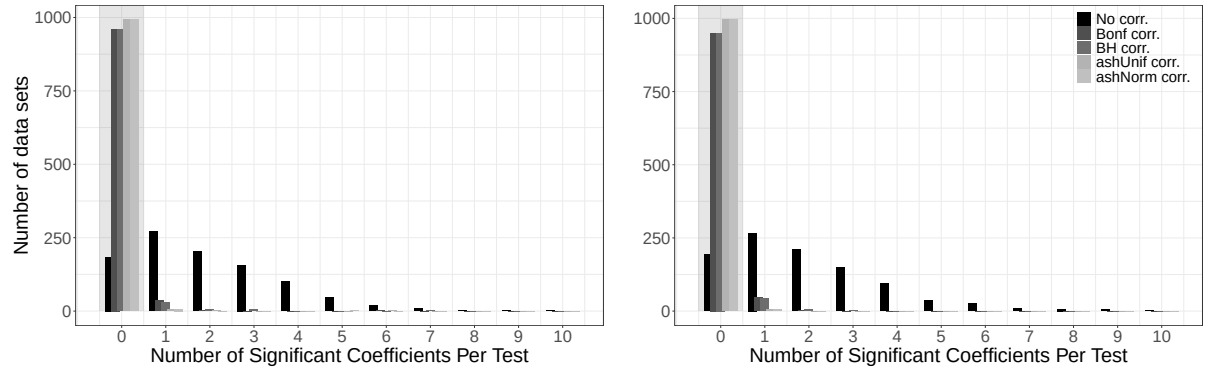
Additional simulation results with less noisy data generating processes are shown in Figures B.1 and B.2. To remove heterogeneity across individuals, we generate the coefficients by $\boldsymbol{\beta}_{ijkl} \overset{iid}{\sim} N(0, .015^2)$ The error term follows the same distribution. Figures B.1a and B.2a summarizes the simulation results. The bars in the figures show the number of simulated data sets for each number of significant findings. The significance level is set at $\alpha$=.05. Although the results without correction are better than Figure 1, in less than 200 out of 1,000 simulated data sets all statistical tests correctly accept the null hypothesis. Improvement by the use of the correction methods is even greater than Figure 3. Only less than ten data sets produce false positive results when a correction method is used.

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 $i$, $\pi_i(\boldsymbol{t}_1, \boldsymbol{t}_0) \equiv$




(a) Parametric Approach (b) Potential Outcome Framework

Figure B.1: **False Positive Results of Estimated AMCEs when the Null Hypothesis is True: Zero Individual MCE.** (a) The true MCE for each individual is indepndently distributed as $\mathcal{N}(0, .015^2)$ for all respondents, all profiles, all attributes at all levels. The individual error term follows the normal distribution $\mathcal{N}(0, .01^2)$. (b) The true marginal component effects are zero under the potential outcome framework.




(a) Parametric Approach (b) Potential Outcome Framework

Figure B.2: **False Positive Results of Estimated AMCEs when the Null Hypothesis is True Using Different Correction Methods: Zero Individual MCE** (a) The true MCE for each individual is indepndently distributed as $\mathcal{N}(0, .015^2)$ for all respondents, all profiles, all attributes at all levels. The individual error term follows the normal distribution $\mathcal{N}(0, .01^2)$. (b) The true marginal component effects are zero under the potential outcome framework. 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`).

$Y_i(\boldsymbol{t}_1) - Y_i(\boldsymbol{t}_0)$. For each paired-comparison, $J = 2$, let $Y_{ijk}(\bar{\boldsymbol{t}})$ indicates whether respondent $i$ chooses the $j$th profile in her $k$th 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 Figures B.1b and B.2b. The results are almost identical to the parametric approach shown above.

## B.2 Three Attributes Have Non-zero AMCEs

In this simulation, we set all levels for `Gender`, `Edutation`, and `English` as significant and all other attributes have zero AMCE with the parameters below. Table 2 presents the results.

| | **Gender** | **Education** | **English** | **Others attributes** |
|---|---|---|---|---|
| *Reference level* | `Female` $\sim \mathcal{N}(0, .015^2)$ | `No formal` $\sim \mathcal{N}(0, .025^2)$ | `Fluent` $\sim \mathcal{N}(0, .015^2)$ | |
| Other levels | Male = $-.06$ | 4th grade = .015 | Broken Eng. = $-.05$ | $\sim \mathcal{N}(0, .015^2)$ |
| | | 8th grade = .02 | Tried but unable = $-.1$ | |
| | | High school = .045 | Use Interpreter = $-.15$ | |
| | | Two-year college = .1 | | |
| | | College = .13 | | |
| | | Graduate = .17 | | |

## B.3 One Level in Each Attribute Has a Non-zero AMCE

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

| Attributes | *Reference level* | Significant level | Other Levels |
|---|---|---|---|
| **Gender** | `Female` $\sim \mathcal{N}(0, .01^2)$ | $-.02$ | 0 |
| **Education** | `No formal` $\sim \mathcal{N}(0, .025^2)$ | .02 | 0 |
| **English** | `Fluent` $\sim \mathcal{N}(0, .03^2)$ | $-.01$ | 0 |
| **Country origin** | `India` $\sim \mathcal{N}(0, .1^2)$ | .05 | 0 |
| **Profession** | `Janitor` $\sim \mathcal{N}(0, .02^2)$ | .02 | 0 |
| **Job experience** | `None` $\sim \mathcal{N}(0, .05^2)$ | .1 | 0 |
| **Job plan** | `Will look for work` $\sim \mathcal{N}(0, .015^2)$ | .01 | 0 |
| **App. reason** | `Family reunion` $\sim \mathcal{N}(0, .01^2)$ | $-.01$ | 0 |
| **Prior trip exp.** | `Never` $\sim \mathcal{N}(0, .05^2)$ | .025 | 0 |

The results are shown in Table B.1. There should be nine significant estimates in each data set. Because there is more noise in the data, the performance difference across correction methods is not clear-cut. Nonetheless, only about 10% of experimental trials have accurately picked out the significant coefficients without correction. In some simulation data sets, more than ten false positive findings are produced. The BC almost doubles the number of accurate tests, but at a significant cost of false negative conclusions. The BH and Ash almost tripled the successful tests, with BH correction risking more false positive conclusions and Ash risking more false negative conclusions.

| | | | No. of False Positives | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| No. of True Positives | No corr. | 7 | | 2 | | 1 | 1 | | | | | | | |
| | | 8 | 9 | 20 | 18 | 18 | 11 | 7 | 6 | 2 | 3 | 1 | | |
| | | 9 | 99 | 201 | 197 | 143 | 110 | 62 | 40 | 22 | 20 | 3 | 3 | 1 |
| | Bonf corr. | 5 | 6 | | | | | | | | | | | |
| | | 6 | 72 | 5 | | | | | | | | | | |
| | | 7 | 273 | 28 | 4 | | | | | | | | | |
| | | 8 | 366 | 34 | 7 | 1 | | | | | | | | |
| | | 9 | 182 | 17 | 3 | 2 | | | | | | | | |
| | BH corr. | 6 | 2 | 2 | 1 | | | | | | | | | |
| | | 7 | 35 | 13 | 6 | | 1 | 1 | | | | | | |
| | | 8 | 163 | 67 | 37 | 23 | 5 | | | | | | | |
| | | 9 | 322 | 179 | 68 | 39 | 23 | 8 | 2 | 1 | 1 | 1 | | |
| | ashUnif corr. | 5 | | 1 | | | | | | | | | | |
| | | 6 | 17 | 2 | 3 | | | | | | | | | |
| | | 7 | 76 | 23 | 7 | 2 | 1 | | | | | | | |
| | | 8 | 212 | 99 | 45 | 14 | 4 | 1 | | | | | | |
| | | 9 | 271 | 143 | 40 | 21 | 11 | 2 | 4 | | 1 | | | |
| | ashNorm corr. | 5 | 1 | | | | | | | | | | | |
| | | 6 | 18 | 4 | 3 | | | | | | | | | |
| | | 7 | 85 | 26 | 7 | 2 | | | | | | | | |
| | | 8 | 221 | 93 | 41 | 14 | 4 | 1 | | | | | | |
| | | 9 | 268 | 138 | 40 | 19 | 9 | 1 | 4 | | 1 | | | |

Table B.1: **Number of Data Sets for Each Number of True and False Positive Findings when the True AMCE of One Level of Each Attribute is Non-zero.** Empty cells indicate zero data set. The true AMCEs of all other levels are set to be zero. For exact simulation parameters, see Appendix B.3. The gray shaded cell represents the perfect test restuls, where nine true positives and no false findings.

# C Adaptive Shrinkage

## C.1 Model and Estimation

Here we briefly outline the model. Detailed discussion of the method and its properties including the FDR can be found in Stephens (2017) and Gerard and Stephens (2018). Consider the posterior distribution:

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

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:

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

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

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

$\delta_0$ 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 $\pi_0$ to be as large as permitted by the observed data):

$$\hat{g} = \underset{g \in \mathcal{U}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{\beta}}|g, \hat{\boldsymbol{s}}) = \underset{g \in \mathcal{U}}{\operatorname{argmax}} \prod_{j=1}^{J} \int_{\beta_j} g\mathcal{N}(\hat{\beta}_j|\beta_j, \hat{s}_j^2) \mathrm{d}\beta_j \tag{6}$$

$$= \underset{g \in \mathcal{U}}{\operatorname{argmax}} \prod_{j=1}^{J} \sum_{k=0}^{K} \pi_k \mathcal{N}(\hat{\beta}_j; 0, \delta_k^2 + \hat{s}_j^2)$$

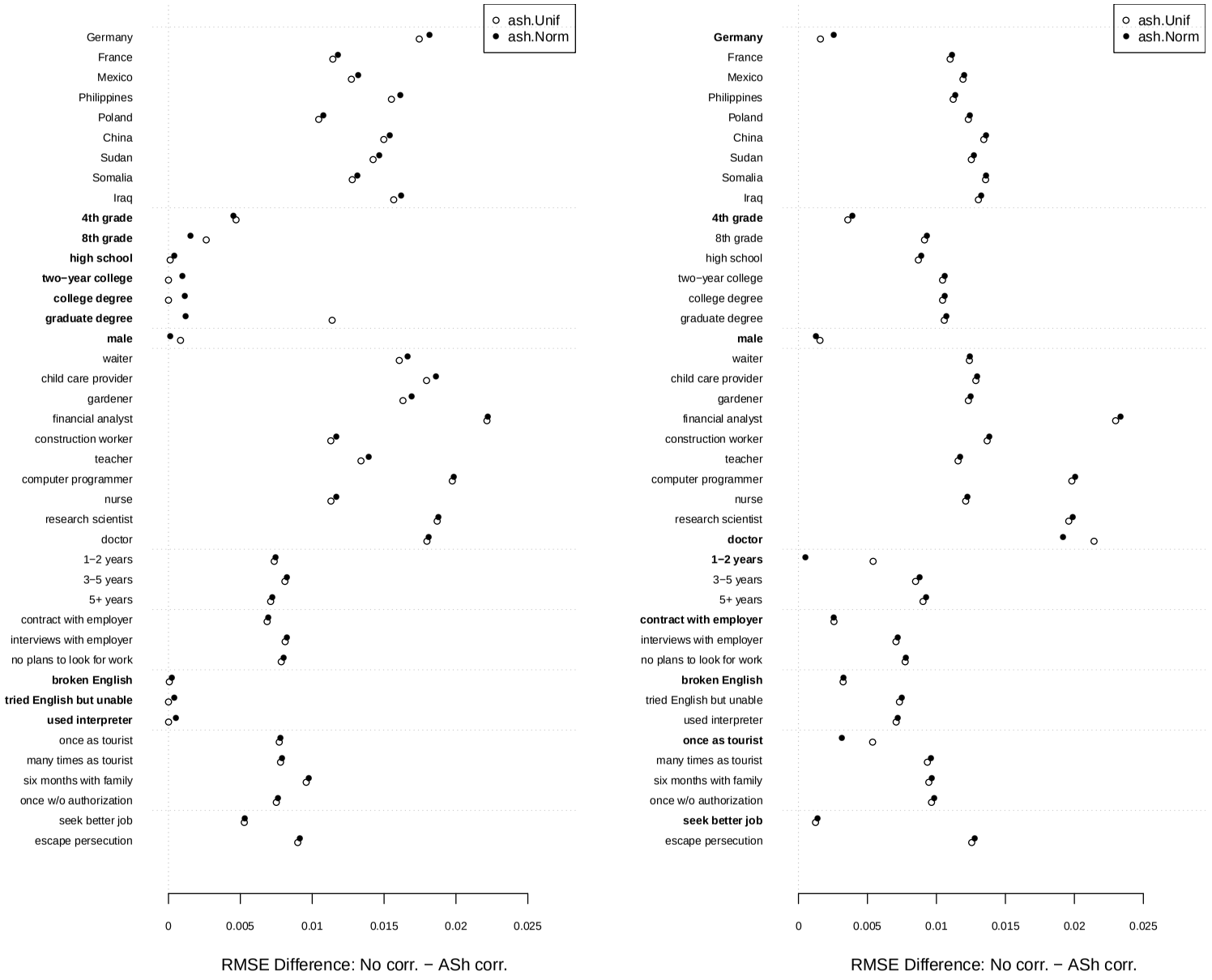
2) Compute the posterior distribution $p(\beta_j|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}, \hat{\boldsymbol{s}})$ and summarize the distributions.

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.

## C.2 Estimation: Smaller RMSE with Ash

As we discussed in Section 3.3, Ash not only regularizes uncertainty measures, but also it 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 C.1a presents RMSE difference when the true AMCEs of all levels of *Gender*, *Education*, and *English* are significant using the same parameters as Appendix B.2. 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 B.2). Figure C.1b summarizes the RMSE difference where the true AMCE of one level of each attribute is non-zero, where the simulation parameters can be found in Appendix B.3.




(a) Non-zero True AMCE for all levels in three attributes (bold)

(b) Non-zero True AMCE for each attribute (bold)

Figure C.1: Comparing Ash corrected RMSE and non-corrected RMSE.

# D Replication

## D.1 Selecting Immigrants in the US

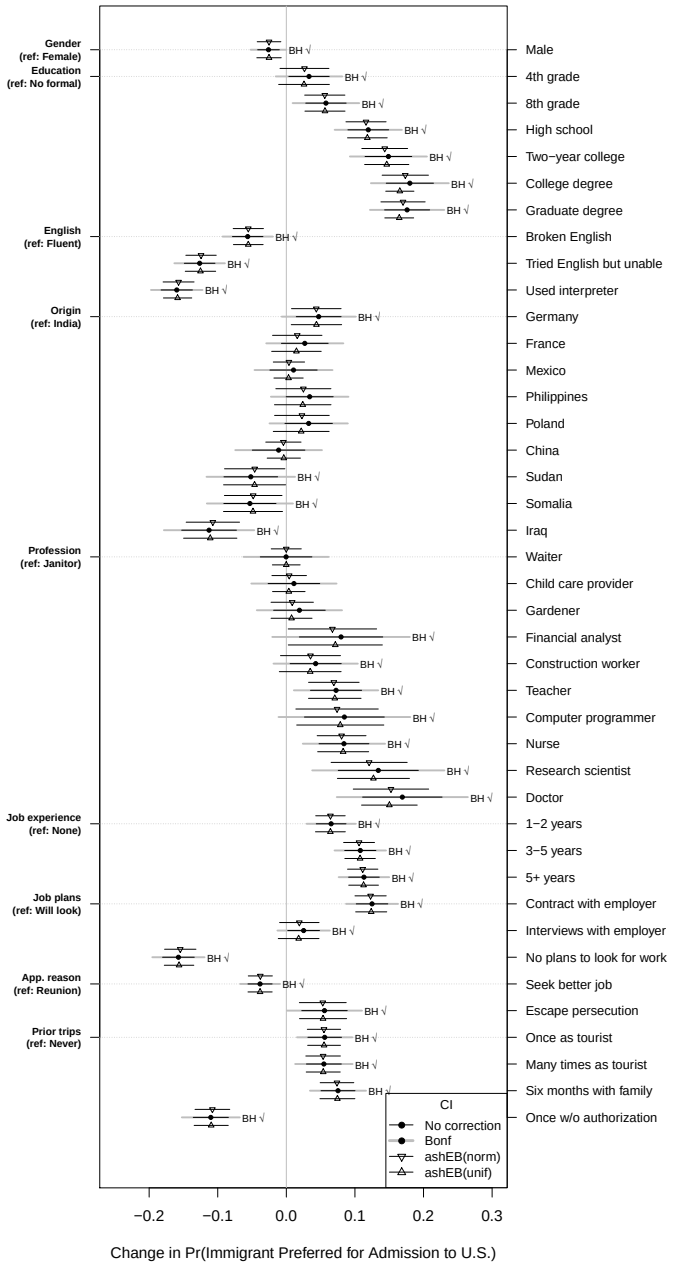



Figure D.1: **Effects of immigrant attributes on the probability of being preferred for admission to the United States. The reference category for each attribute is in parentheses on the left side of the y-axis.** The plot shows estimates with no correction, Bonferroni correction (`Bonf`), empirical 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).

## D.2 Selecting Trading Partners in Vietnam

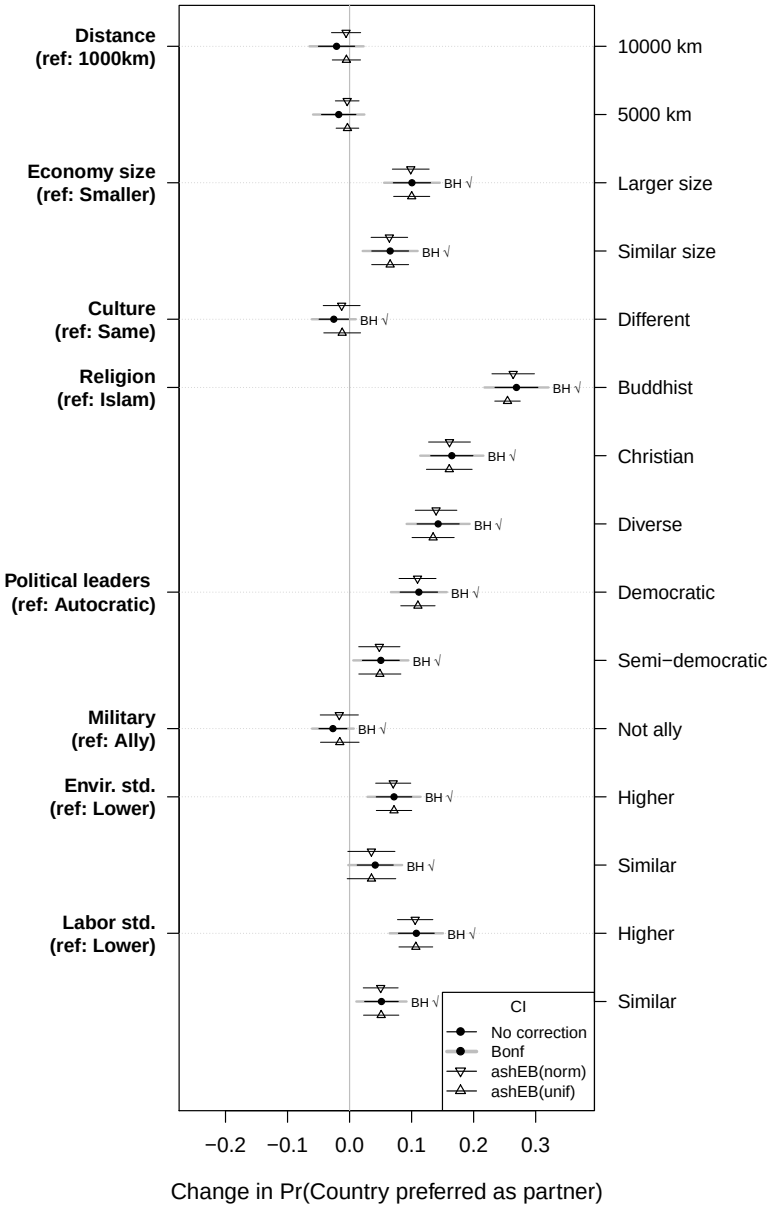



Figure D.2: **Effects of country attributes on the probability of being preferred as trading partners in Vietnam.** The reference category for each attribute is in parentheses on the left side of the y-axis. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), empirical bayes 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).

## D.3 Selecting Brokers in India

We present an additional replication study to demonstrate the difference between BH and Ash, although both are based on the idea of controlling false discovery rate. **?** 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 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 D.4. 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.

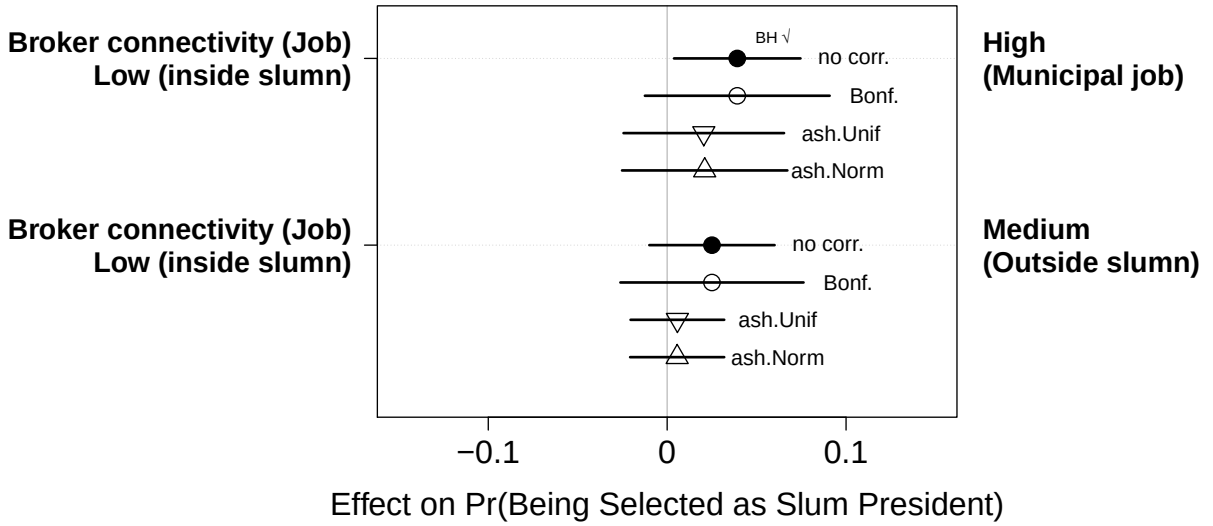



Figure D.3: **Effects of Slum Leader's Connectivity on the Probability of Being Preferred for President of the Slum Development Council.** The reference category is low connectivity jobs: occupations entirely contained within the slum. The plot shows estimates with no correction, Bonferroni correction (`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 **?**, p.784.

As Figure D.3 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 guaranteed by the property of BH: because there are only eight significant discoveries in the paper, the idea of controlling FDR at $\alpha = .05$ 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.

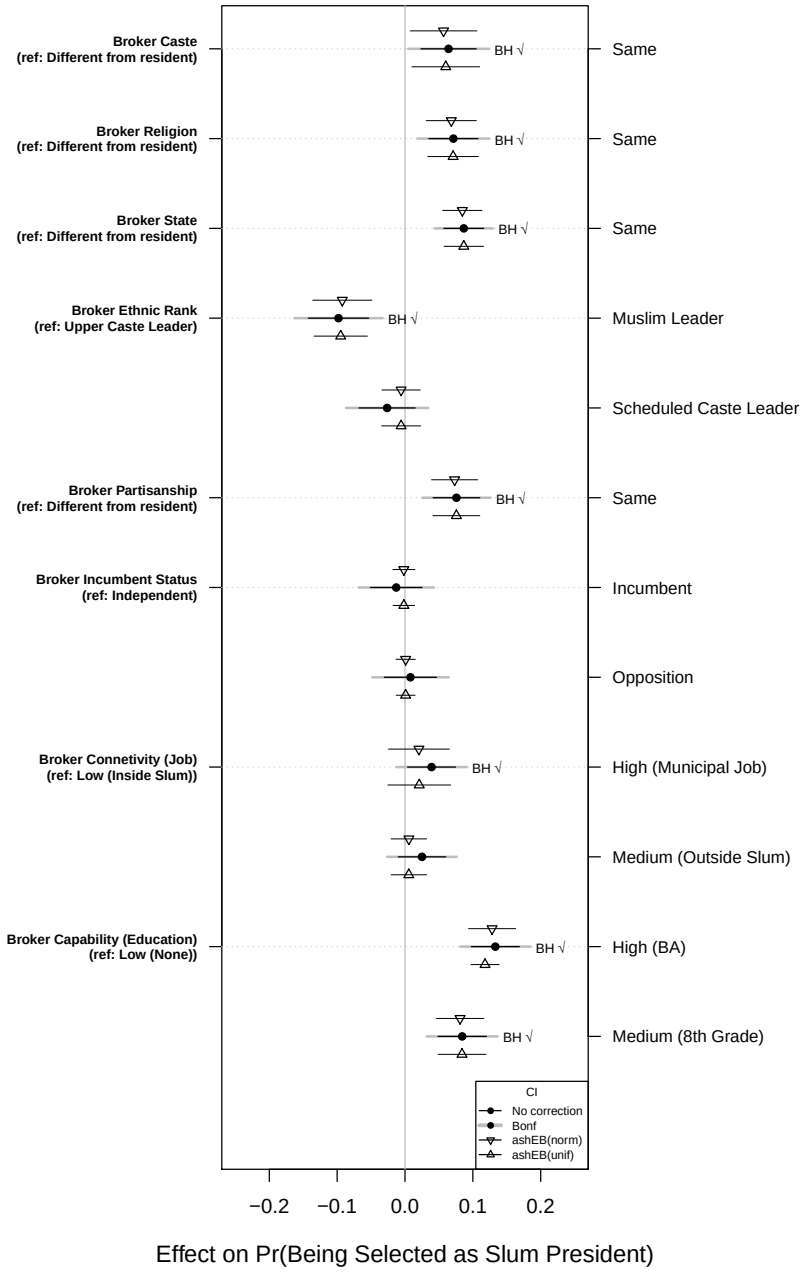



Figure D.4: **Effects of slum leader attributes on the probability of being preferred for president of the slum development council.** The reference category for each attribute is in parentheses on the left side of the y-axis. The plot shows estimates with no correction, Bonferroni correction (`Bonf`), 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 **?**, p.784.