CSRDA
Center for Social Research and Data Archives,
Institute of Social Science, The University of Tokyo

SUSTAINABLE DEVELOPMENT GOALS

CSRDA supports the Sustainable Development Goals

# CSRDA Discussion Paper

## Gender gap on the preference of college major choice : Evidence from Conjoint Survey Experiment

No. 13

Date Apr.2021

SDGs

4 QUALITY EDUCATION

5 GENDER EQUALITY

Name Sho Fujihara, Keisuke Kawata, Shoki Okubo

# Gender gap on the preference of college major choice: Evidence from Conjoint Survey Experiment*

Sho Fujihara† Keisuke Kawata‡ Shoki Okubo§

2021-04-13

**Abstract**

The gender gap in college major choice has become an important issue in higher education. In particular, female students are underrepresented in natural science majors in many countries. This paper estimates students' preference for college majors, in addition to other attributes of college education programs. We use a fully randomized conjoint survey experiment, which allows us to estimate preference for each program attributes, including tuition fees, labor market conditions after graduation, gender composition within the university and the department, and college major. We find a large gender gap in the preference for college major even when other attributes are held constant. The female respondents do not prefer natural science majors, on average. The heterogeneity analysis also shows consistent results: a positive preference for natural science majors is not observed even in female respondents with high math grades in junior high school.

Key words: Conjoint survey experiment; Heterogeneous effect; Gender gap.

*We thank Ayako Kondo and Fumio Ohtake, and the participants of the Labor Economics Workshop in Japan for their useful comments and discussions. Of course, we are responsible for any remaining errors. The authors also thank the Management Committee of the Japanese Life Course Panel Surveys for allowing us to use data from the Panel Survey of Junior High School Students and their Mothers' data. This work was supported by the Japan Society for the Promotion of Science (JSPS KAKENHI) [JP15H05397, JP19H01637] and the Director's Funds for Internal Projects ("Social Science Research by Social Surveys and Experiments on the Web).

†Institute of Social Science, University of Tokyo.

‡Corresponding author. Institute of Social Science, University of Tokyo. keisukekawata@iss.u-tokyo.ac.jp

§Institute of Social Science, University of Tokyo.

# 1 Introduction

The gender gap in major college choices has become an important issue. A well-known stylized fact is that women are underrepresented in the STEM (science, technology, engineering, or math) majors in most countries (Stoet and Geary 2018; Mostafa 2019). Previous studies (for instance, Kahn and Ginther 2017; Stoet and Geary 2018; Patnaik, Wiswall, and Zafar 2020) point out potential factors of the gender difference in the college major choice. However, no studies have compared the relative importance of the attributes of education programs.

This paper quantitatively compares the relative importance of these factors using a fully randomized conjoint survey experiment (Hainmueller, Hopkins, and Yamamoto 2014) among young respondents (20 years old). In the experiment, respondents are asked to choose their preferred college program from two hypothetical programs. The programs consist of the following attributes: (1) major (natural sciences or humanities), (2) the gender composition in the university, (3) the gender composition in the program, (4) university location, (5) tuition fee, (6) average income of male graduates at 30 years old, and (7) average income of female graduates at 30 years old. The randomization of these attributes allows us to quantitatively identify causal factors affecting the program choice, which can be interpreted as the educational program preference.

The main interest is the college major, natural science majors vs other majors. The survey results show a significant gender difference in major choices. Even when other attributes are given, female respondents tend to avoid natural science, but no clear effect of college major is observed among male respondents.

We employ heterogeneity analysis with machine learning (Semenova and Chernozhukov 2017) to discover within-gender groups. The results show that female respondents do not prefer natural science majors even if their math and natrual science grades are above the median. In contrast, male respondents prefer the natural science majors if their grades are above the median.

We also find other factors affecting college program choice. The expected income of graduates with the same gender as the respondent increases the probability of program choice, while the expected income of graduates with the opposite gender does not. The other factors, including location, tuition, and gender composition in the program/university, have similar effects for male and female respondents.

The rest of the paper is organized as follows. Section 2 presents the survey design and descriptive statistics of our data. In Section 3, we introduce the empirical strategy. Section 4 shows the estimation results, and Section 5 concludes the paper.

# 2 Data

This paper employs a fully randomized conjoint survey experimental design (Hainmueller, Hopkins, and Yamamoto 2014) to elicit a preference for academic majors. The fully randomized design has become popular due to its robustness (Hainmueller, Hangartner, and Yamamoto 2015; Bansak et al. 2019, 2018).

Our survey was conducted via the Internet in March 2020. The respondents were 19 years old and included both college and noncollege students. There were 399 and 501 male and female respondents, respectively. They first enrolled in the conjoint experiments and then answered the background questions.

An advantage of our data is the panel structure. We can access the survey results of the same respondents in junior high school and can then examine the relationship between respondents' college major preference and their grades in junior high school.

# 3 Estimation strategy

This section formally defines our estimands and the choice problem in our conjoint survey experiments. Our main interest is the gender-specific preference on the academic major; natural science VS Social science and Humanity.

Each respondent was asked her/his preferred program from two education programs $j$ and $-j$. Program $j$ is characterized by a vector of other program attributes $a_j = \{a_1, ..., a_l, ..., a_7\}$, where $a_l$ is the level of attribute $l$.

Our estimands are defined by the potential outcome as $Y_{i,j}(a)$ where $a = [a_j, a_{-j}]$. $Y_{i,j}(a) = 1$ is equal to one if respondent $i$ prefers program $j$ over alternative $-j$.

In the choice experiment, respondents state their preferred program from two programs $j$ and $-j$ after observing realized attributes $A_i = [A_{i,j}, A_{i,-j}]$. The respondent's statement is summarized by choice dummy $Y_{i,j}^{obs}$; $Y_{i,j}^{obs} = 1$ if respondent $i$ chooses program $j$ and $Y_{i,j}^{obs} = 0$ if she/he chooses alternative $-j$.

An important property of the fully randomized design is the independence between the potential outcome and realized program attributes. Formally,

$$Y_{i,j}(d, a) \perp A_i.$$

Independence allows us to identify the conditional average potential outcome $\mu(a|g, x) = E[Y_{i,j}(a)|G_i = g, X_i = x]$ as

$$\mu(a|g, x) = E[Y_{ij}^{obs}|A_i = a, G_i = g, X_i = x] \quad (1)$$

where $G_i$ is respondent $i$'s sex, and $X_i$ is background characteristics.

Above identification results cannot directly apply for estimation with our data because of the over-fitting problem. Even the sample size is limited, $A_i$ and $X_i$ are high-dimensional. The statistical estimation of $E[Y_{ij}^{obs}|A_i = a, G_i = g, X_i = x]$ is then difficult. We introduce marginalized quantities, which are easily estimated.

## 3.1 Average marginal potential outcome

To describe the gender-specific preference on an education program, the average marginal potential outcome is introduced. The average marginal potential outcome is defined as $\mu_l(a_l|g) = \sum_{a_{-l,j}, a_{-j}, x} \mu(a|g, x) \times f(a_{-l,j}, a_{-j}, x)$, which is marginalized the conditional average potential outcome given attribute $l$'s level and gender. Equation (1) directly identifies the average marginal potential outcome as

$$\mu(a|g) = E[Y_{ij}^{obs}|A_{i,j,l} = a, G_i = g].$$

The average marginal potential outcome is easily estimated by the sub-sample mean given gender of respondent and the level of the attribute.

## 3.2 Preference on college major

Next, the preference structure of college major is examined. We then estimate the best linear predictors of the conditional average difference of potential outcome;

$$\mu(d = 1, \hat{a}|g, x) - \mu(d = 0, \hat{a}|g, x)$$

where $d$ be the indicator of academic major of a program $j$, which is equal to one with natural science and zero otherwise. $\hat{a}$ is other attributes.

To avoid over-fitting problem, the double machine learning estimation (Semenova and Chernozhukov 2017) is employed. Equation (1) can be rewritten by the double robust score function (Robins, Rotnitzky, and Zhao 1994);

$$\mu(1, \hat{a}, g, x) - \mu(0, \hat{a}, g, x) = E[S_{ij}(\hat{a}, g, x)|\hat{A}_i = a, G_i = g, X_i = x],$$

where

$$S_{ij}(\hat{a}, g, x) = f_Y(1, \hat{a}, g, x) - f_Y(0, \hat{a}, g, x)$$

$$+\frac{I(D_{ij} = 1) \times (Y_{ij}^{obs} - f_Y(1, \hat{a}, g, x))}{0.5} - \frac{I(D_{ij} = 0) \times (Y_{ij}^{obs} - f_Y(0, \hat{a}, g, x))}{0.5},$$

where $f_Y(d, \hat{a}, g, x) = E[Y_{ij}^{obs} | D_{ij} = d, \hat{A}_i = \hat{a}, G_i = g, X_i = x]$. $f_Y(d, \hat{a}, g, x)$ is estimated by LASSO method (Tibshirani 1996), which is lower out-of-sample MSE among boosting (Friedman 2001) and random forest (Breiman 2001).

Semenova and Chernozhukov (2017) shows the best linear predictor on a variable $z \in [a, x]$ can be estimated by the OLS estimation of $\mu(1, \hat{a}, g, x) - \mu(0, \hat{a}, g, x)$ on $z$. In the following analysis, we estimate the best linear predictors of program attributes and academic performance in the junior high school.

# 4 Result

## 4.1 Average marginal potential outcome

The following figure reports the average marginal potential outcome, which is separately estimated with male and female samples.

Figure 1

Our main interest is the preference of academic major, differs significantly by gender. Female respondents tend to not prefer natural science majors, even when other attributes are held constant. For male respondents, no clear preference is observed because the point estimator of the academic major is close to zero and not statistically significant. Preferences on the expected income are also clearly different. On average, respondents are concerned only about the income of graduates with the same gender.

The estimated average potential outcome of other attributes are almost the same between the female and male respondents. Respondents of both sexes prefer an education program with low tuition fees, located in their home countries, and with a larger share of women in the university and the department.

## 4.2 Natrual science penalty/premium

The best linear predictors of the natural science penalty/premium are reported.

### 4.2.1 Program attributes

The best linear predictors of program attributes are follows. First, the best linear predictors of female respondents are shown in Figure2.

Figure 2

The figure shows the natural science penalty of female respondents regardless of other attributes. Even considering confidence intervals, the natural science major robustly decreases the choice probability.

Figure 3 presents The best linear predictors of male respondents.

Figure 3

Same as female respondents, our data cannot find clear interactions between the college major and other program attributes. The natural science major then dose not have clear effects on the choice probabilities.

#### 4.2.2 Grades in junior high school

Next figures report the best linear predictors of grades in junior high school. Estimation results of female respondents are in Figure 4.

Figure 4

Estimated linear predictors are consisted with the grades of math and natrual science. Respondents with lower grade tend to strongly avoid the natural science major. Surprisingly, the natural science penalty is still observed even among female respondents with high math/natural science grades.

Figure 5 reports the best linear predictors of male respondents.

Figure 5

College major preference of male respondents is also consisted with math and natural science grades in the junior high school. Positive effect of the natural science department is observed among respondents with high grades.

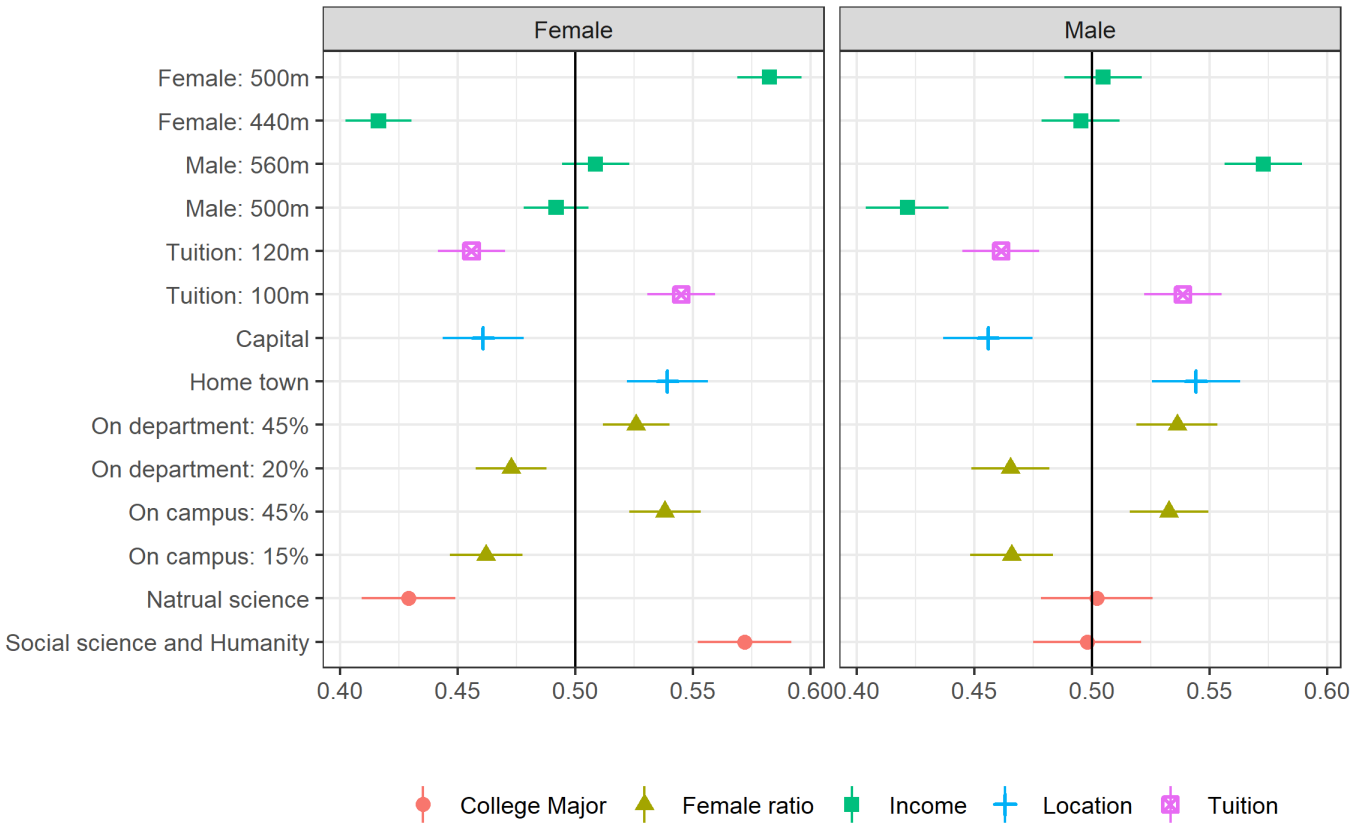
## 5 Conclusion

This paper examines the gender gap in education program preferences using a fully randomized conjoint survey experiment. We find significant gender heterogeneity in college major choice even when other program attributes are held constant. Female respondents tend to avoid natural science majors, while male respondents do not.

Additionally, both male and female respondents prefer a diverse environment; a larger share of female students on campus increases program choice probability. The expected income of graduates with the same gender as the respondents significantly increases the program-choice probability, but the income of graduates with the opposite gender has no clear effects. This result implies that attracting more female students requires improving the female labor market.

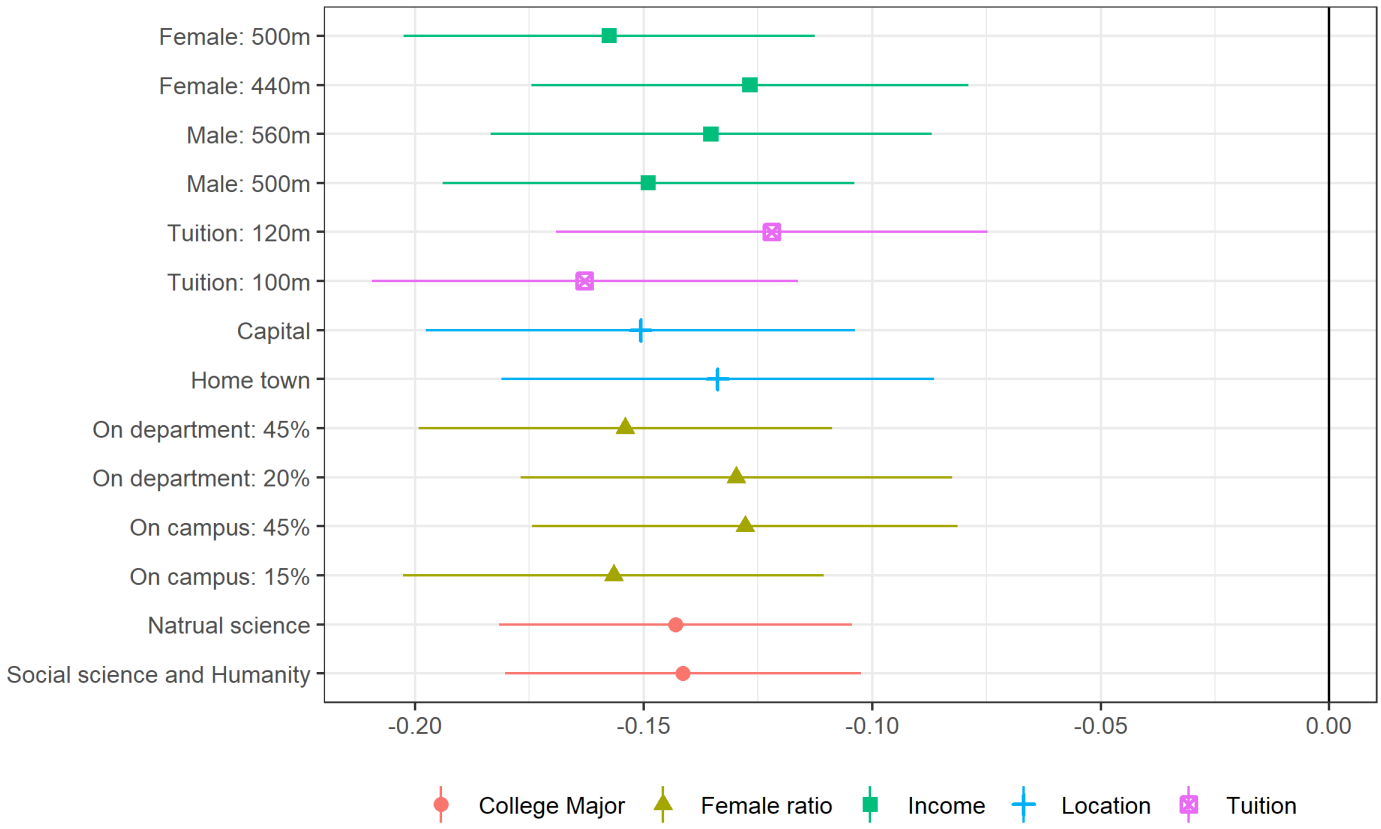
Finally, the paper shows negative results, which means that our data cannot identify a female group with a positive STEM preference. Our results suggest that the conjoint survey experiment should include other attributes, for instance, the share of female faculty, to detect modifiers that may effectively change women's preferences.

## Figure 1



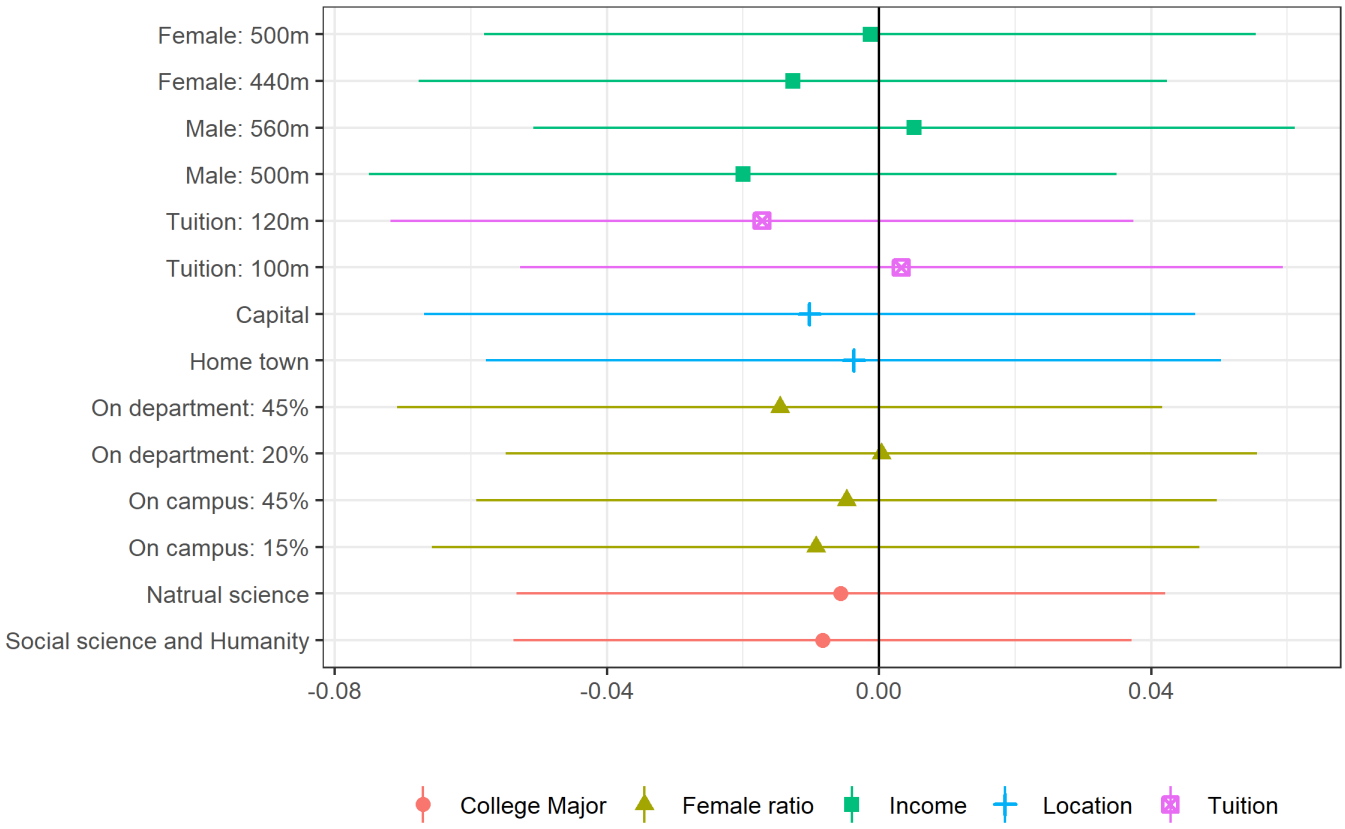

Notes: Each dot shows the point estimators, and bars are the 95th confidence intervals with the respondent-level clustering standard error.

## Figure 2



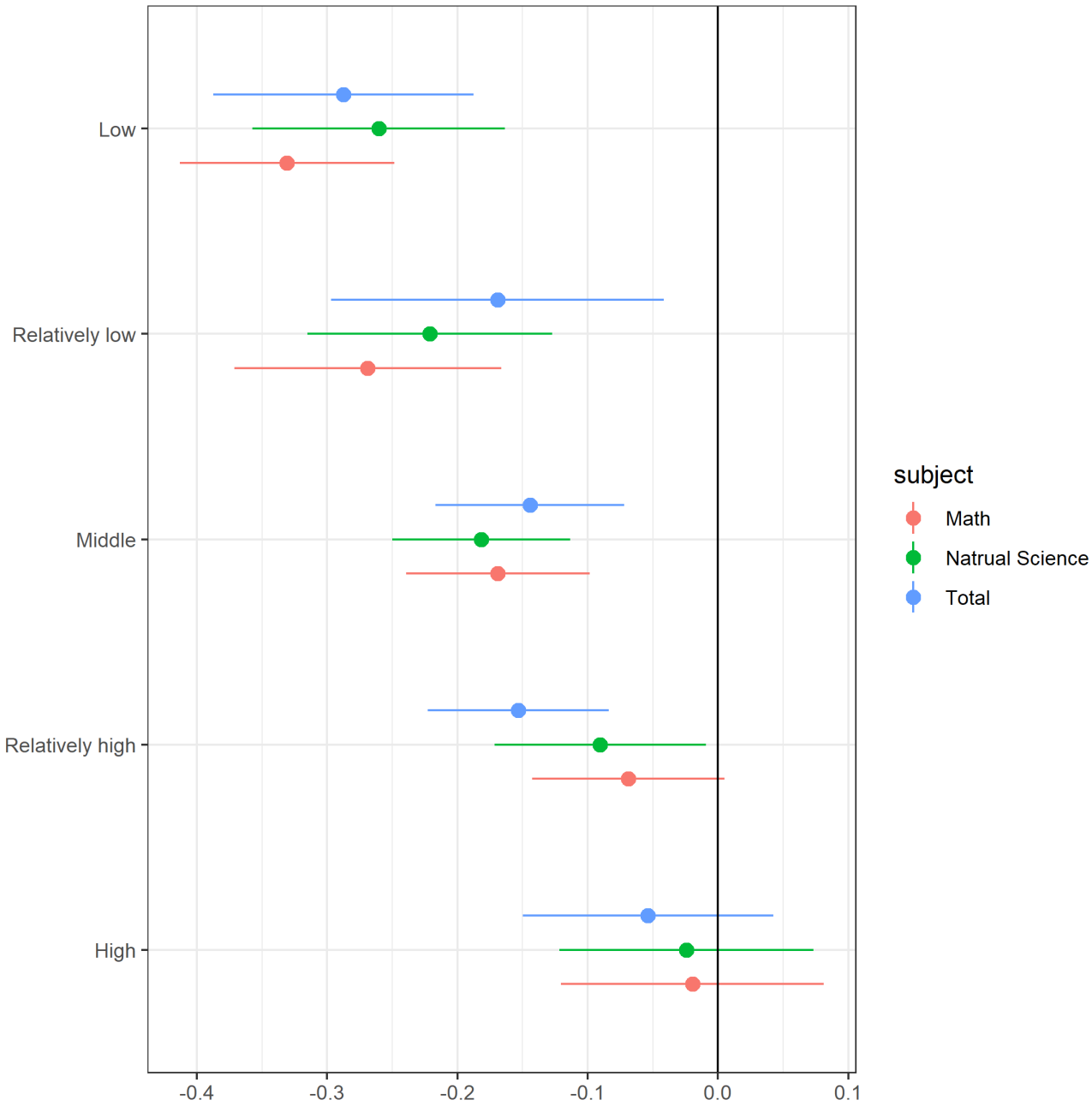

Notes: Each dot shows the point estimators, and bars are the 95th confidence intervals with the respondent-level clustering standard error.

## Figure 3



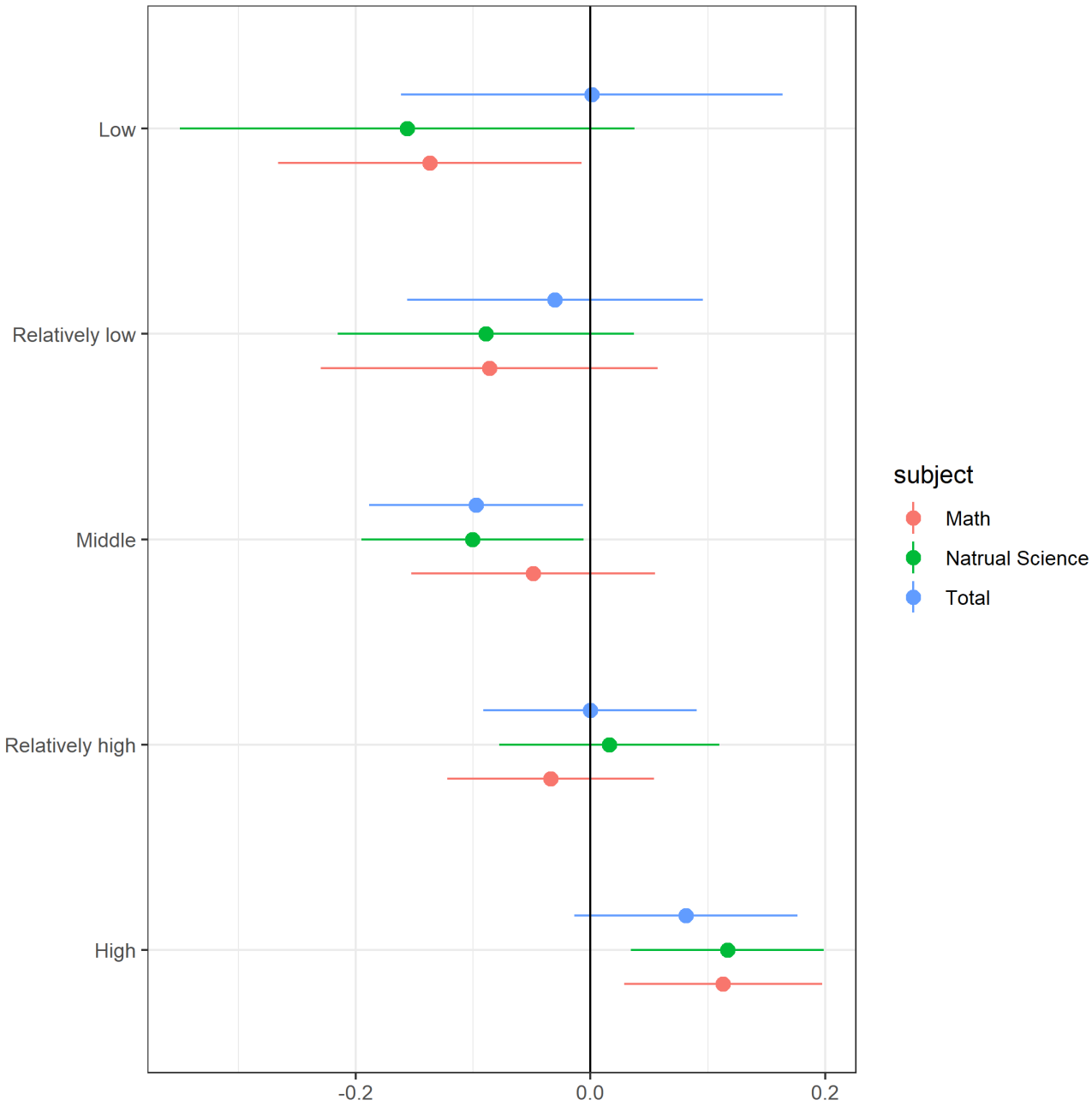

Notes: Each dot shows the point estimators, and bars are the 95th confidence intervals with the respondent-level clustering standard error.

## Figure 4




Notes: Each dot shows the point estimators, and bars are the 95th confidence intervals with the respondent-level clustering standard error.

## Figure 5




Notes: Each dot shows the point estimators, and bars are the 95th confidence intervals with the respondent-level clustering standard error.

# Reference


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 (1): 112–19.

———. 2019. "Beyond the Breaking Point? Survey Satisficing in Conjoint Experiments." *Political Science Research and Methods*, 1–19.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Friedman, Jerome H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, 1189–1232.

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 (8): 2395–2400.

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): 1–30.

Kahn, Shulamit, and Donna Ginther. 2017. "Women and STEM." National Bureau of Economic Research.

Mostafa, Tarek. 2019. "Why Don't More Girls Choose to Pursue a Science Career?" no. 93. https://doi.org/https://doi.org/https://doi.org/10.1787/02bd2b68-en.

Patnaik, Arpita, Matthew J Wiswall, and Basit Zafar. 2020. "College Majors." *National Bureau of Economic Research Working Paper Series*, no. w27645.

Robins, James M, Andrea Rotnitzky, and Lue Ping Zhao. 1994. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed." *Journal of the American Statistical Association* 89 (427): 846–66.

Semenova, Vira, and Victor Chernozhukov. 2017. "Estimation and Inference about Conditional Average Treatment Effect and Other Structural Functions." *arXiv Preprint arXiv:1702.06240*.

Stoet, Gijsbert, and David C Geary. 2018. "The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education." *Psychological Science* 29 (4): 581–93.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.