Gender Composition in Classrooms: Influences on Post-Secondary Schooling Choices

(very preliminary. Please do not circulate)

Jaime Polanco-Jiménez[*]
Kristof De Witte [†]
Gloria L. Bernal [‡]

March 1, 2024

Abstract

This study explores the impact of male students' presence in classrooms on the choice of university majors among female students. To unpack this, we employ a fixed-effects methodology and a staggered difference-in-difference (S-DiD) approach, specifically examining gender compositions within classrooms concerning university major choices. Additionally, we use schools transitioning from single-gender to coeducational settings in Colombia as a robustness check. Our findings reveal that, compared to their male counterparts, female students show a significantly higher inclination toward pursuing academic majors associated with human sciences. These results suggest that classrooms with a high composition of male students narrow the gap in STEM majors for women, except in fields such as Law and Medicine. Furthermore, no gender differences were found based on classroom gender composition in majors related to agriculture, zootechnics, and veterinary sciences.

JEL Codes:

Key Words:

[*] Corresponding author: Department of Economics, Pontificia Universidad Javeriana. (email: jaime.polanco@javeriana.edu.co)

[†] KU Leuven & UNU-Merit, Maastricht University. (email: kristof.dewitte@kuleuven.be).

[‡] Department of Economics, Pontificia Universidad Javeriana. (email: gbernal@javeriana.edu.co).

1 Introduction

The decision regarding post-secondary studies carries profound implications, extending beyond mere career preferences to enduring consequences, notably in terms of income differentials. As highlighted by National Center for Science and Engineering Statistics (2023), careers in science, technology, and mathematics provide a substantial 35% income advantage compared to fields such as humanities, social sciences, and educational sciences. It is essential to recognize, however, that considerations beyond financial incentives, such as the cultivation of critical thinking, problem-solving, and communication skills, exert a significant influence, particularly in humanities[1].

In the specific case of Colombia, a developing country in South America, where approximately half a million students complete their secondary education annually, the implications of post-secondary choices become apparent. During the subsequent 7.5 years post-secondary, on average 33% of women and 29% of men pursue and complete a college academic program. Despite a higher rate of women completing university degrees, a mere 16% of them opt for careers in mathematics, engineering, science, architecture, construction, among others. In contrast, 24% of men choose these fields, leading to a noteworthy gender-based disparity in the selection of academic disciplines.

The literature emphasizes the crucial role of the learning environment in shaping students' career choices, particularly in STEM fields. The impact of classroom demography on girls' STEM performance and persistence (Pregaldini et al., 2020; Bottia et al., 2015) underscores the importance of considering the social context in fostering interest and success in STEM subjects.

The influence of self-efficacy and academic performance on STEM degree choices is highlighted (Pregaldini et al., 2020; Bottia et al., 2015). Moreover, the recognition of psychological, social, and cultural mechanisms driving gender differences in STEM choices (Wang & Degol, 2013; Tyler-Wood et al., 2018) emphasizes the need for interventions at multiple levels to address these disparities.

Early STEM experiences and learning environments are identified as critical factors shaping girls' STEM persistence (Buschor et al., 2014; Bottia et al., 2015). This insight suggests that interventions targeted at the early stages of education can have a lasting impact on encouraging girls to pursue STEM careers.

Building on these observations, our research delves into the influence of gender composition in classrooms on post-secondary study decisions across various knowledge domains. Our hypothesis posits that male students, often characterized by higher levels of competitiveness, demonstrate a propensity to pursue university majors in mathematics, engineering, and science, as evidenced in existing literature[2]. Consequently, the proportion of male students in a classroom serves as a potential mechanism for the transfer of competitive aspirations from male to female students during social interactions. By exploring how gender composition contributes to the observed gender gap in post-secondary schooling decisions, we aim to provide nuanced insights into the complex interplay of factors shaping educational trajectories.

Furthermore, the literature underscores that the difference in competitiveness between female and male students manifests as early as the first 4 to 5 years of age (Sutter

[1] See, for example, Alesina, Giuliano, and Nunn (2013); Clifford D. Evans (2006); Evans and Diekman (2009); Lent, Brown, and Hackett (1994)

[2] e.g., Buser, Niederle, and Oosterbeek (2014); Kohen and Nitzan (2022)

and Glätzle-Rützler (2010)). Notably, male students exhibit a greater perception of motor and spatial competitiveness, while female students lean towards a greater perception of verbal competitiveness (Shahar Gindi and Pilpel (2019)). These characteristics become significant as male students are more inclined towards careers in STEM, while female students tend to gravitate towards social and human sciences. Our study is the first of its kind to establish a connection between gender school composition and post-secondary study preferences in a developing country.

In our empirical investigation, we utilize datasets from the Integrated School Enrollment System in Colombia (SIMAT) and the National Higher Education Information System (SNIES). These datasets serve as essential tools, allowing us to track individuals over time and explore the majors students choose after completing secondary school. These rich data also include individual characteristics, contributing to our research methodology and minimizing biases. Additionally, we delve into detailed school attribute information from the Formal Education Survey (EDUC) to gain a holistic understanding of the nuanced interplay between gender, school composition, and post-secondary study decisions. Merging insights from these diverse sources enhances both the depth of our understanding and the robustness of our analytical framework.

On the basis of this foundation, we employ an analytical approach. Specifically, we segment the data into specific proportions of male students within each classroom, utilizing these proportions to estimate the log-odds and the probability of a female student selecting a particular university major. The analytical model includes essential components such as school and year fixed effects, along with control variables. To address potential biases inherent in our investigation, we strategically leverage a real-world intervention—the transition from single-sex schools to co-educational schools. Our methodology include a Staggered Difference-in-Differences design (S-DiD), allowing us to draw robust comparisons between schools that have undergone this transition and those that have not. This meticulous approach provides a comprehensive and reliable examination of the causal impact of the transition on student choices at the university level, specifically in relation to gender composition within classrooms.

2 Context

Educational system in Colombia

- compulsory education
- Higher education
- access to education.
- Constraints for Human Capital Accumulation
- participation in higher education

Single-gender schools

Coeducational schools
~

Why does gender composition matter?
Transition from single-sex to coeducational schools

3 Data and Descriptive Statistics

In our empirical investigation into the intricate relationship between gender, school composition, and post-secondary study decisions, we leverage comprehensive datasets from the Integrated School Enrollment System in Colombia (SIMAT) and the National Higher Education Information System (SNIES). These systems serve as invaluable tools, enabling us to track individuals over time and gain insights into the majors students choose after completing secondary school. Additionally, the richness of this data extends to individual characteristics, which we incorporate into our research methodology to meticulously minimize bias.

To gain a more holistic understanding of the contextual nuances shaping this relationship, we further complement our analysis by delving into detailed information about school attributes. This additional layer of insight is sourced from the Formal Education Survey (EDUC). Our overarching goal is to unravel the complex connections between gender, school composition, and post-secondary study decisions. By merging data from these diverse sources, we not only refine our understanding of the nuanced interplay of factors influencing educational choices but also enhance the robustness of our analytical framework.

The Integrated School Enrollment System in Colombia (SIMAT)
The National Higher Education Information System (SNIES)
Formal Education Survey (EDUC)
Descriptive Statistics.

We present the distribution of students in different gender compositions. It can be observed that X_i is a continuous variable with a normal distribution, where the values at 0 (female schools) and 1 (male schools) represent single-gender schools. These single-gender schools are excluded from the descriptive analysis because when calculating the relationship between studying a major and gender as a conditioning factor, they lack sample variation within the school.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.377	0.0.462	0.457	0.546	1.000

Table 1: Caption

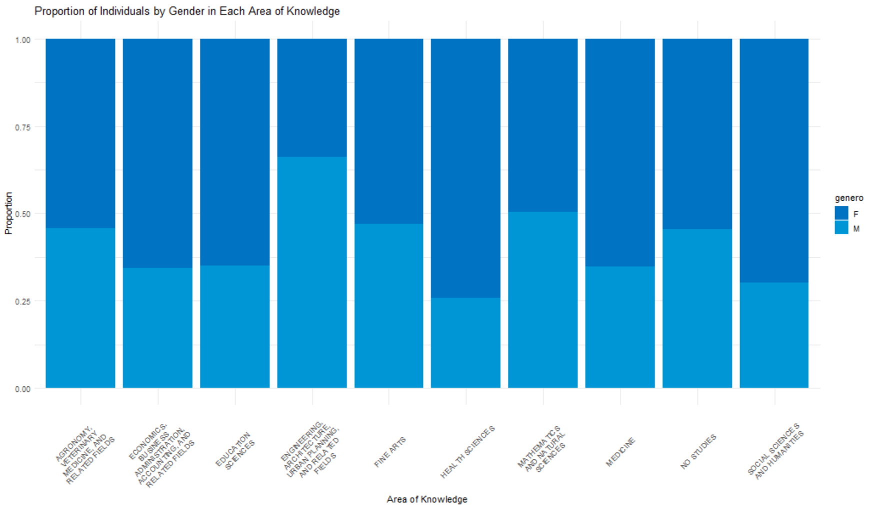

Figure 1: Distribution by Gender according to the Area of Knowledge

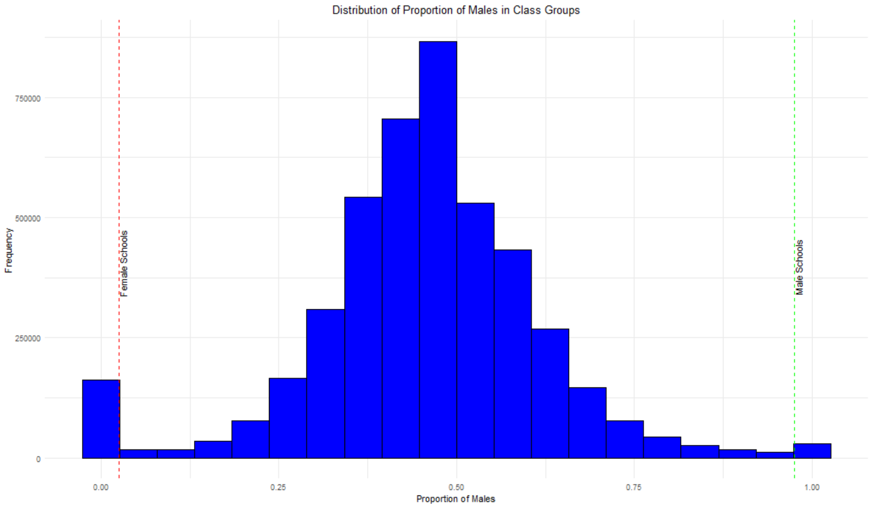

Figure 2: Distribution of Proportion of Males in Class Groups

4 Conceptual framework

—- —— —- —— THINK ABOUT WHAT IS GOING FIRST THE DIF AND DIFF OR THE BIASED ESTIMATION, IF I SAY THAT THE DIF AND DIFF HAS PROBLEMS ESTIMATING JUST THE CORNES COMBINATION THEN IS BETTER TO SHOW FIRST THE DIF AND THEN THE BIASED —- —— —- ——

The accumulation of human capital and post-secondary school decisions, particularly in the selection of a major university, is discussed in studies such as Hanushek (1979) and Todd and Wolpin (2003). In these studies, the decision results from a multifaceted interaction between expected future income, budget constraints when the student makes the decision, family attributes, student attributes, school attributes, and social influences to which students are exposed. Among these factors, student attributes play a critical role, encompassing intelligence, motivation, study habits, competitiveness, and other influential characteristics [3].

Cooper and Liu (2019a) explicates the life cycle of households in three distinct stages: education, early work, and late work. This progression is supported by theoretical models of human capital accumulation (refer, for instance, to Ben-Porath (1967)). Theoretical models of human capital accumulation predominantly hinge on the time available for students to dedicate to their studies. Having more time to study directly correlates with spending less time on early work. However, this also implies a budgetary constraint when deciding whether to pursue further studies and which field to study. Despite the primacy of future income expectations, the costs associated with pursuing higher education annually can range between xxx (looking for cite) and xxx (looking for cite, is it including the cost of materials, accommodation, transport, food, well-being?).

Environmental factors, including social support and media, significantly influence students' interest in STEM careers, as revealed by Wang et al. (2023). The study highlights that male students generally exhibit higher interest in STEM careers compared to female students. The influence mechanisms differ, with social support being more impactful for males and media playing a greater role for females.

Another perspective on gender roles is provided by Alesina et al. (2013), who explore the historical origins of cross-cultural differences in beliefs and values regarding the appropriate role of women in society. The study focuses on the influence of traditional agricultural practices, particularly plough agriculture, on the evolution of gender norms. The findings suggest that societies with a heritage of traditional plough use exhibit less equal gender norms in contemporary beliefs and practices, even among the descendants living in diverse environments. This underlines the lasting impact of historical practices on contemporary societal values and gender roles.

According to the OECD (OECD (2013)), motivated students achieve better performance in mathematics, and in this sense mathematics performance in school is a predictor of choosing a STEM major at university (Kohen and Nitzan (2022)).

Is motivation affected by gender composition? how? Explain in a paragraph. Show data from Colombia

The investigation conducted by Sutter and Glätzle-Rützler (2010) revealed a notable gender-based divergence in competitiveness, wherein males exhibited a heightened level by 15 to 20 percentage points compared to the average competitiveness observed in females. This discernible gap in competitiveness is particularly intriguing as it manifests as early as 4 to 5 years of age, underscoring the early emergence of gender disparities in this trait. Additionally, Shahar Gindi and Pilpel (2019) extended these findings by highlighting that boys tend to manifest greater competitiveness in motor and spatial domains, while girls demonstrate heightened competitiveness in verbal areas.

Furthermore, empirical support for the association between the desire for competitiveness and career choices is provided by Buser et al. (2014). This experimental study establishes a robust connection between individuals' inclination toward competitiveness and their career preferences. Notably, the findings of this study suggest that men

[3] Studies by Lönnqvist, Verkasalo, Walkowitz, and Wichardt (2015) and Buser et al. (2014) demonstrate the impact of risk attitudes and self-confidence on students' academic choices.

tend to display a greater proclivity for competitiveness compared to women. Moreover, it underscores a direct correlation between the level of competitiveness and the chosen academic major at the university level. Intriguingly, highly competitive individuals predominantly opt for disciplines within the Science, Technology, Engineering, and Mathematics (STEM) domain.

Proportion of male students in a classroom ⇒ Female competitiveness ⇒ University major choice

Therefore, it is anticipated that the gender composition within a classroom significantly influences female competitiveness. A higher proportion of male students may suggest a narrowing gender gap in career selection, primarily attributed to the impact of competitiveness emanating from male peers. This influence could exhibit variations across academic domains, with fields associated with humanities, social sciences, economics, and languages—traditionally dominated by female students—experiencing a decline in female participation relative to male involvement when influenced by their male counterparts.

In contrast, within STEM-related disciplines, it is expected that female students will demonstrate a reduction in the gender gap, driven by the substantial influence exerted by male peers in the classroom. Essentially, the presence of male students alters the behavioral dynamics of female students, leading to a diminished participation gap between men and women.

4.1 Analytical Framework

In this study, we examine how the gender composition in secondary schools impacts students' choices of university majors. We utilize a fixed effect approach with a binary dependent variable to analyze this relationship, and we augment this approach with a staggered difference-in-difference (S-DiD) design comparing female schools transitioning to co-educational schools in Colombia.

In this first analysis, we calculate the probability that a secondary school student denoted as i chooses a university major c given the gender composition of their class. For instance, a gender composition of 0.2 implies that a secondary female student i studied in a classroom where 20% of the students were male students, and 80% were female.

To capture the probability of a secondary school student choosing a university major based on their class's gender makeup, we segment the data by specific proportions of male students within each classroom. These proportions range from 0% to 100%, subdividing into intervals like less than 5%, 5-15%, 15-25%, and so forth, up to 85-95%.

—- —— —- —— EXPLAIN IN DETAIL WHY IS RELEVANT NO JUST CALCULATE THE CORNERS POINT AS IN THE DIF AND DIF ESTIMATION BASED ON THE FACT THAT THE CAUSATION IN THE TRANSITION FROM SIGLE SEX SCHOOL TO COEDUCATIONAL IS JUST A DROP IN A GLASS OF WATER —- —— —- ——

This estimate pertains to each student represented as i. Therefore, the relationship is expressed as follows:

$$\log\left(\frac{P(Y_{i,s,t}^{c}=1)}{1-P(Y_{i,s,t}^{c}=1)}\right)=\beta_1\times Gender_{i,s,t}+\beta_2\times X_{i,s,t}^{c}+\gamma_t+\gamma_s+\varepsilon_{i,s,t} \qquad (1)$$

Where $Y_{i,s,t}^{c}$ is the binary response variable for students i who have completed secondary school in the school s. It takes the value of 1 when a student i chooses a university major c and 0 otherwise. $Gender_{i,s,t}$ takes the value of 1 for female students of a secondary school. X_i is a vector of student i characteristics in each secondary school. The model includes school fixed effects (γ_t), year fixed effects (γ_t), and the usual error term $\varepsilon_{i,s,t}$.

β_1 is the coefficient of interest that specifically represents the relationship between being a female student (denoted as $Gender = Female$) and the log-odds of a female student choosing a particular university major c, while holding other variables constant in the model. A positive value for β_1 suggests a positive correlation between being a female student and the likelihood of choosing the academic major 'c' compared to male students. This means that, all else being equal, being a female student is associated with a higher probability of choosing the specified academic major 'c' as compared to being a male student.

An additional interpretation regarding marginal effects and the significance of these coefficients could enhance the understanding of how the likelihood of choosing a specific major changes with varying gender compositions in secondary schools. For example, a positive β_1 might suggest a certain increase in the probability of a female student choosing the specified major compared to male students, specifically considering a change in the gender composition of their class.

Moreover, the incorporation of school-fixed effects enables the consideration of inherent and stable characteristics unique to different educational institutions. School-fixed effects allow us to account for the intrinsic and stable characteristics of different schools. For instance, some schools may emphasize technology, business, or industry-related subjects. The geographical features and physical attributes of a school may remain consistent and unique across different generations. It includes the socio-demographic and economic conditions of the students in a school that remain constant.

Furthermore, time-fixed effects help to capture systematic variations that occur over time. Changes in societal norms, economic conditions, and government policies, as well as other temporal trends, can influence educational choices over different periods. These time-specific effects are crucial for a more comprehensive understanding of the changing landscape of educational decisions made by individuals, particularly women entering university programs.

4.1.1 Robustnesss Check

To address potential biases intrinsic to fixed effect estimation, we employ a distinct approach by leveraging the transition from single-sex schools to co-educational settings. This unanticipated shift serves as an intervention influencing student behavior, enabling a more robust examination of its effects.

Our study encompasses all public schools in Colombia that have undergone this transition, allowing us to track student enrollments across diverse academic fields at the university level. Employing a Staggered Difference-in-Differences design (S-DiD), we compare schools that have undergone the transition with those that, as of 2020, had not yet experienced the shift (treated schools versus no treated yet schools). This methodological strategy helps reveal the causal impact of the transition on student choice at university, offering insight into the implications arising from the change in school structure on students' university enrollment patterns.

$$\hat{\tau} = \bar{Participation}^{P}_{\text{after transition}} - \bar{Participation}^{P}_{\text{before transition}} \tag{2}$$

Where $\hat{\tau}$ represents the change in the average participation of students in the academic major P [4] before and after the transition from single school sex to co-educational school.

Given your focus on the participation of students in major P before and after a transition, the adapted formula might look something like this:

$$\text{Participation}^{P}_{c,t,j} = \beta_0 + \sum_{\varphi=-S}^{-2} \mu_{\varphi} \cdot D_{c,\varphi} + \sum_{\varphi=0}^{M} \mu_{\varphi} \cdot D_{c,\varphi} + \sigma_t + \gamma_c + \varepsilon_{c,t} \tag{3}$$

Here $\text{Participation}^{P}_{c,t,j}$ represents the level of participation of students in major P at a particular school c and time t. β_0 is the intercept or baseline level of participation in major P. μ_{φ} are the parameters associated with the different time periods or treatment phases (φ). $D_{c,\varphi}$ are dummy variables denoting the treatment status (e.g., before and after the transition) for school c at time φ. σ_t captures time-specific effects. γ_c captures school-specific effects. $\varepsilon_{c,t}$ is the error term.

By including both time-specific and school-specific effects in the estimation, the analysis can better account for and control various unobserved factors that might influence students' choices of university majors. This helps in providing a more accurate understanding of the specific influence of transitioning from single-sex to co-educational schooling on the participation of students in the specified academic major, resulting in more robust and reliable estimations.

The time-specific fixed effect is crucial as it captures broader trends or fluctuations that might affect student participation in academic majors, regardless of the transition being studied. Societal changes, economic shifts, or educational reforms occurring independently of the transition could impact students' major choices. By including these effects, the model more effectively isolates the transition's specific impact on student decisions.

Conversely, the school-specific fixed effect addresses persistent differences between schools, unrelated to the transition itself. Each school possesses unique attributes, teaching methods, or cultural distinctions that could influence students' major choices. Incorporating these school-specific effects helps the model accommodate these differences, effectively separating the transition's impact from inherent school-specific variations.

Identification Strategy

Our study aims to identify the causal effect of Colombian schools transitioning from single-sex to coeducational on students' choice of university major after graduating. We employ a staggered difference-in-differences (S-DiD) design, comparing schools that

[4] Refer to the breakdown of university major choices in Section A.1

have already transitioned to coeducational to those that have not yet transitioned as of 2020.

—- —— —- — WRITE AS A TALE, AVOID BULLETS POINT—- —— —- —

The S-DiD design relies on several key assumptions:

1. **Parallel trends:** In the absence of the transition, we assume the trends in students choosing major P would have evolved similarly between schools that transitioned and those that have not yet transitioned. This assumes common unobserved factors influencing students' choice of major across schools.
2. **No concurrent shocks:** There are no other policy changes or shocks differentially affecting treated and not-yet-treated schools over the study period, apart from the transition itself. The time fixed effects help adjust for broader secular trends.
3. **No anticipation:** We assume the transition does not impact student behavior until it actually occurs. We test for anticipation effects by examining leads of the treatment indicator.
4. **Irreversibility:** Once a school transitions, it remains coeducational over the study period. No school switches back to single-sex education in our sample.
5. **Overlapping cohorts**: Each school has students from multiple cohorts overlapping at any given time. This allows us to observe treated and not-yet-treated cohorts to identify the effect.
6. **Composition stability**: The composition of students and schools remains stable over time. The school fixed effects help adjust for time-invariant compositional differences.

The staggered timing of schools' transitions provides variation in treatment status over time. By comparing outcomes between treated schools and not-yet-treated schools before and after the transitions, and conditioning on school and time fixed effects, we can isolate the causal effect of the transition under the assumptions above.

We assess the plausibility of these assumptions by examining pretrends in outcomes, checking for anticipatory effects, and evaluating the overlap in covariates between treated and not-yet-treated schools over time. Violations of the parallel trends assumption could bias the estimates, so we pay close attention to the pretreatment fit.

5 Results

Results intro
General results
Detailed result
Heterogeneities
robustness
Discussion

The proportion of males within a classroom seems to have mixed behavior between them, on one hand, a higher proportion of male students increases the probability of any student choosing careers in Engineering, Architecture and related Careers. Similarly, a higher proportion of male students is also related to the fact that students are not following additional studies. A general correlation can be seen in Table A.1.

The gender composition within a classroom has a profound impact on students' career choices. When female students find themselves in classrooms predominantly composed of male students, it turns out that their career choices may differ significantly from those studying in more gender-balanced settings. In this subsection, we delve into the intriguing dynamics of classroom heterogeneity and its pivotal role in influencing the career decisions of female students.

The first scenario we investigate pertains to the probability of a female student opting not to pursue higher education in any major (Refer to Figure 3)[5]. The decision-making process behind women's choices to continue their university education post-secondary school has been extensively examined in the literature. These studies have underscored various influential factors, such as familial caregiving responsibilities, financial constraints, personal interests, pursuit of independence, and the absence of adequate career guidance[6].

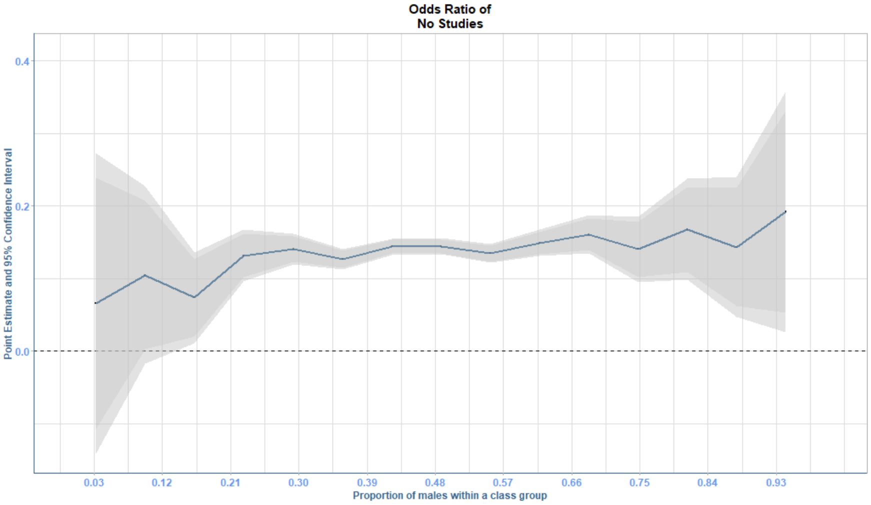

Figure 3: Likelihood of a female student does not select any university major

[5] In order to obtain gender variation We exclude from the sample all single-sex schools.

[6] See for example: Heenan (2002), Capilla Navarro Guzmán and Antonio Casero Martínez (2012), McDermott (2012) and Manisha Joshi (2016)

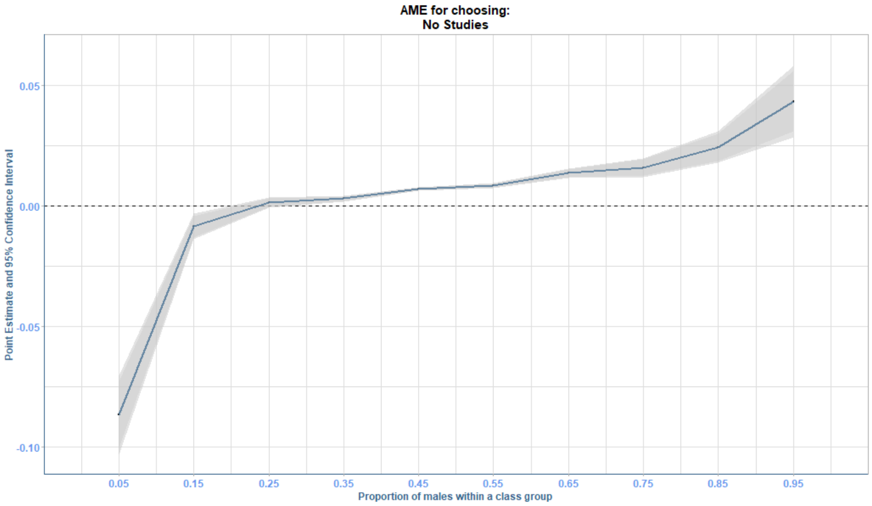

Figure 4: Average Marginal Effect of a female student does not select any university major

The correlation in figure 3 provides the relationship between the proportion of male students in a classroom and the probability that a female student will not continue with any study after finishing school compared with male students. Notably, the figure illustrates that a higher proportion of male students in a classroom is associated with an increased probability of female students not pursuing further education after finishing school, as opposed to their male counterparts. Furthermore, the analysis reveals that when the proportion of male students is less than 25%, there is no significant difference in the educational choices between genders.

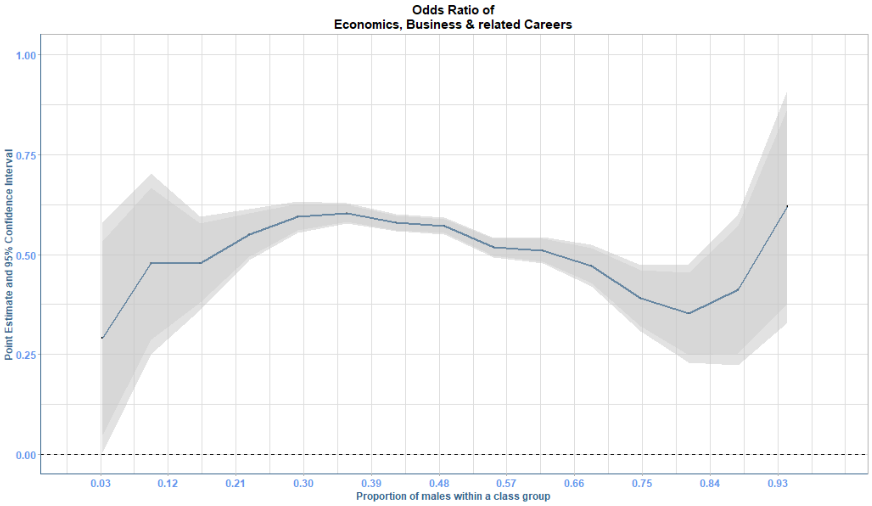

Figure 5: Likelihood of a female student choosing careers related to economics and business (breakdown of 10%)

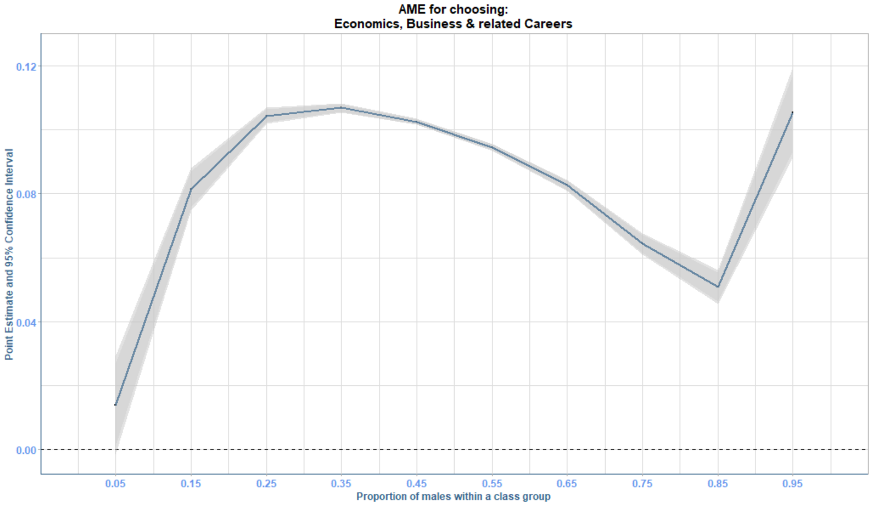

Figure 6: Average Marginal Effect of a female student choosing careers related to economics and business (breakdown of 10%)

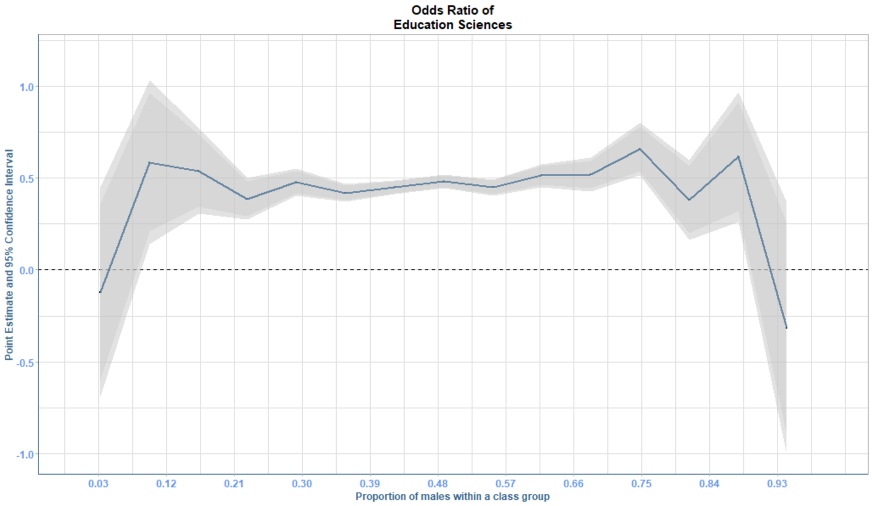

Figure 7: Likelihood of a female student choosing careers related to education sciences

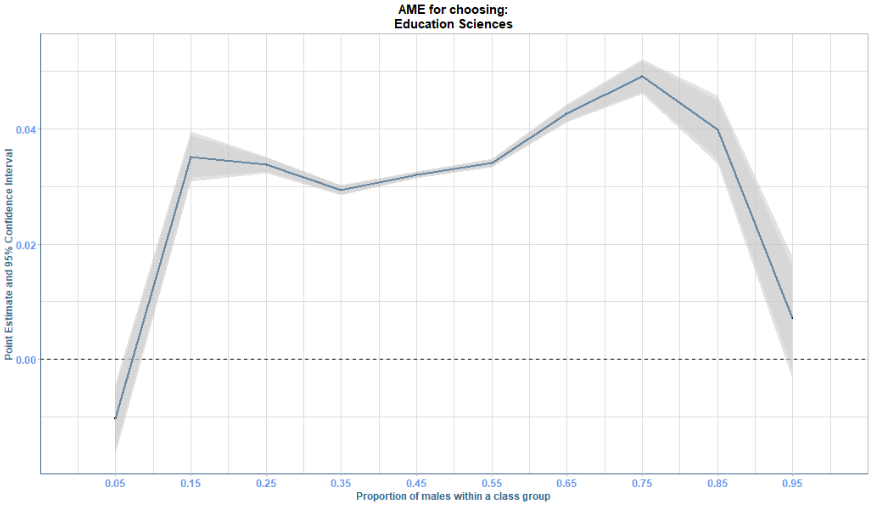

Figure 8: Average Marginal Effect of a female student choosing careers related to education sciences

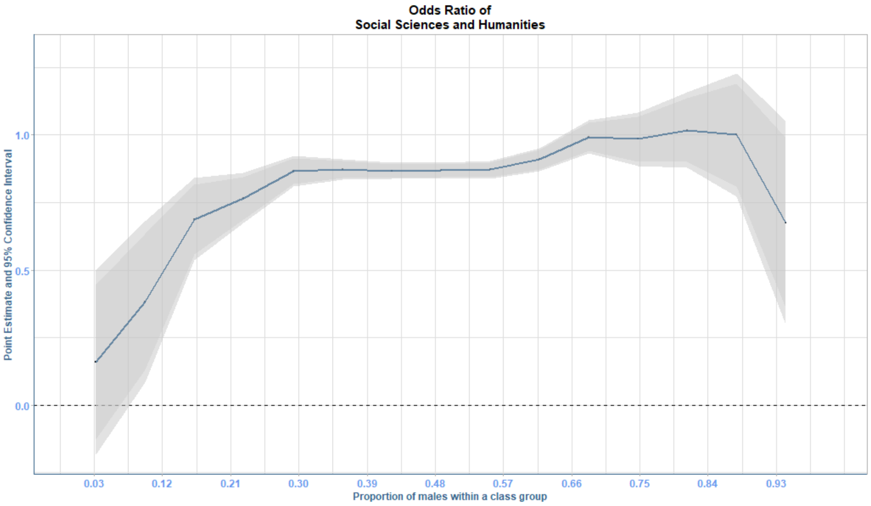

Figure 9: Likelihood of a female student choosing careers related to social science and humanities (except law)

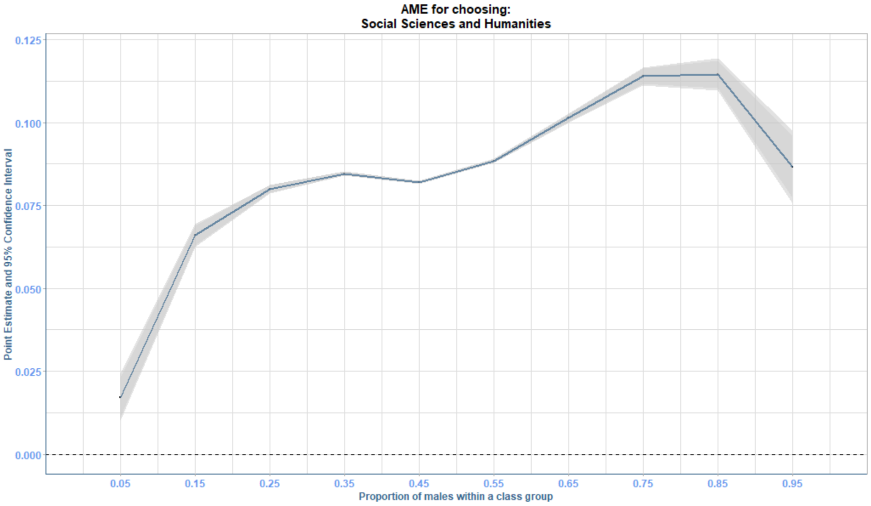

Figure 10: Average Marginal Effect of a female student choosing careers related to social sciences and humanities (except law)

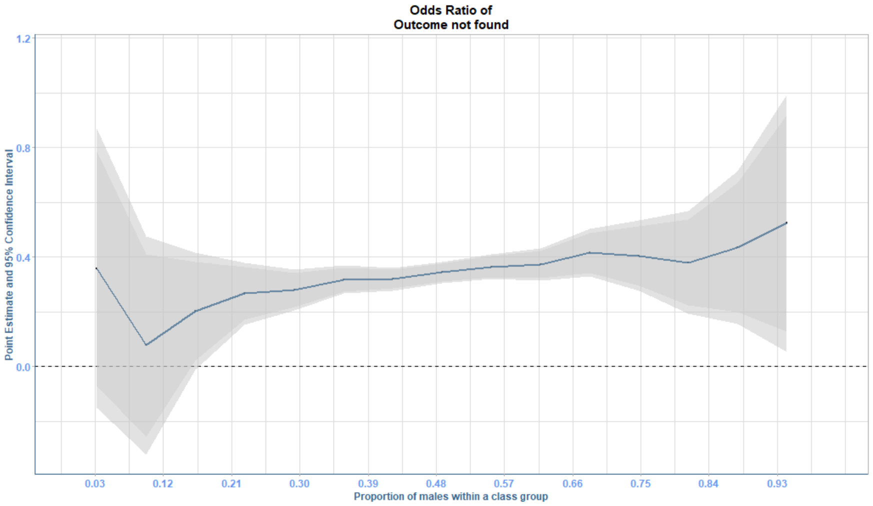

Figure 11: Likelihood of a female student choosing careers related to law

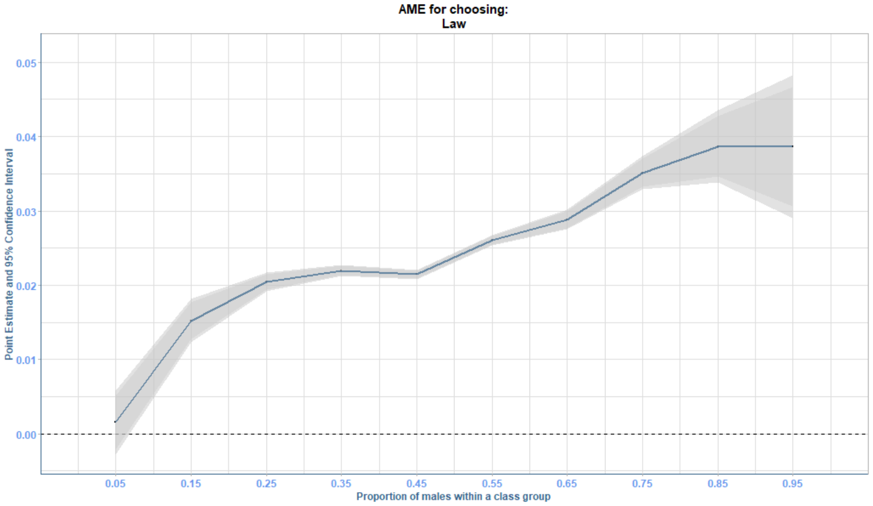

Figure 12: Average Marginal Effect of a female student choosing careers related to law

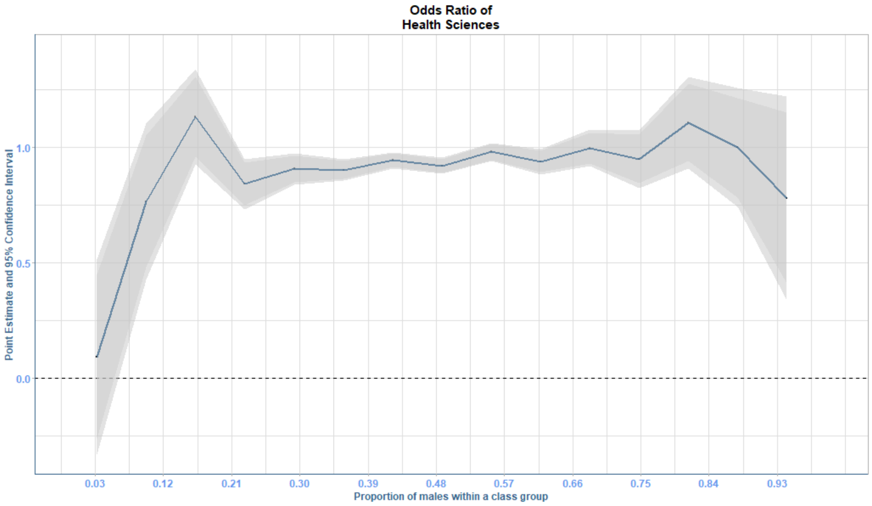

Figure 13: Likelihood of a female student choosing careers related to health sciences (except medicine)

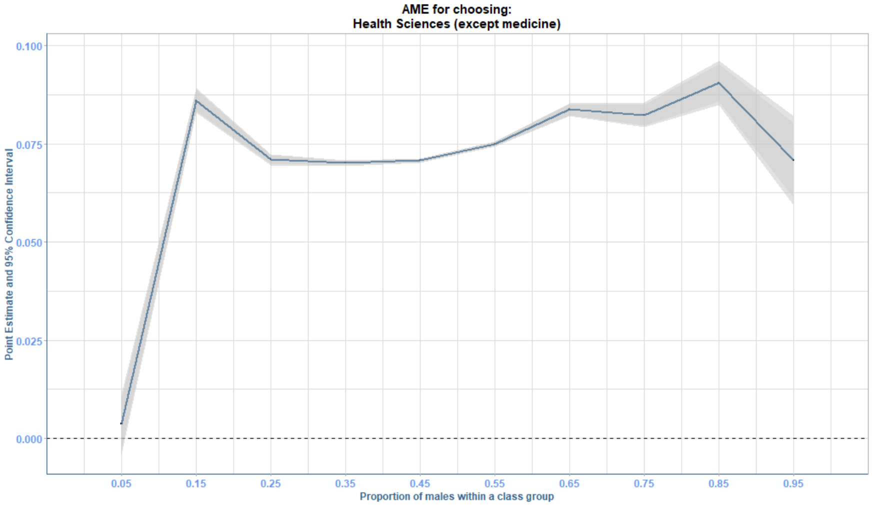

Figure 14: Average Marginal Effect of a female student choosing careers related to health sciences (except medicine)

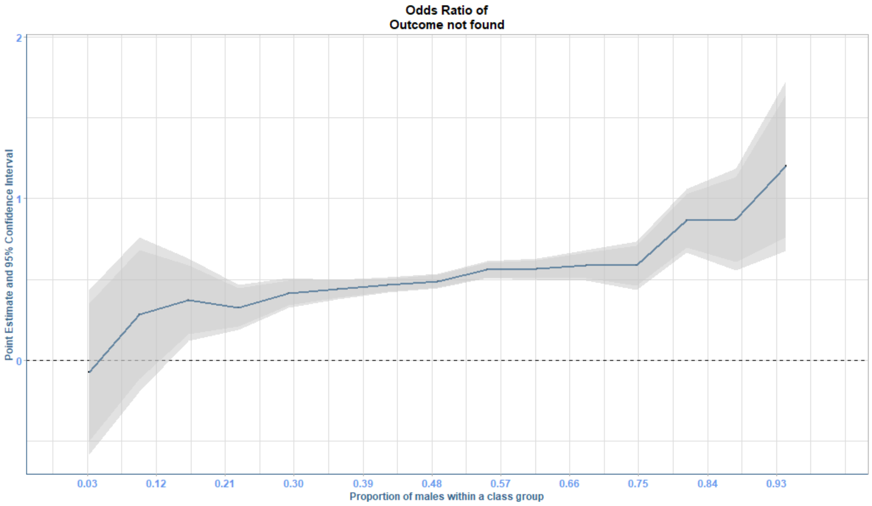

Figure 15: Likelihood of a female student choosing careers related to medicine

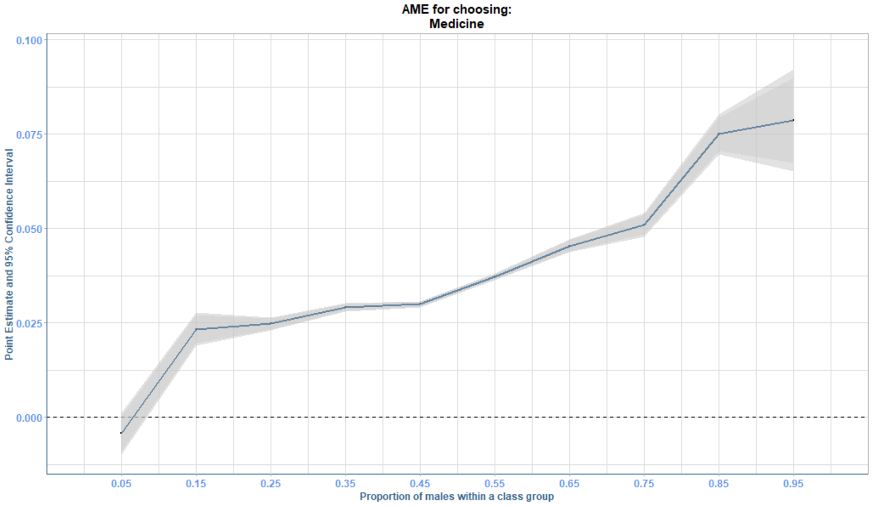

Figure 16: Average Marginal Effect of a female student choosing careers related to medicine

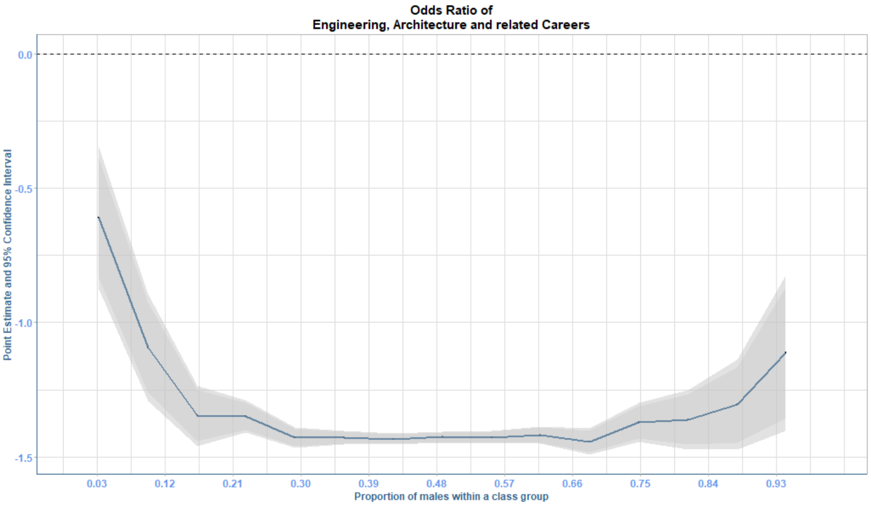

Figure 17: Likelihood of a female student choosing careers related to engineering and architecture

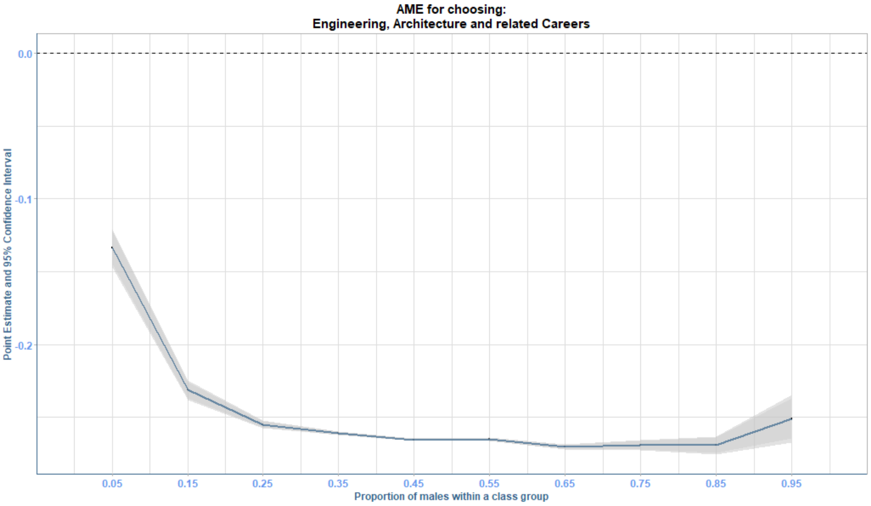

Figure 18: Average Marginal Effect of a female student choosing careers related to engineering and architecture

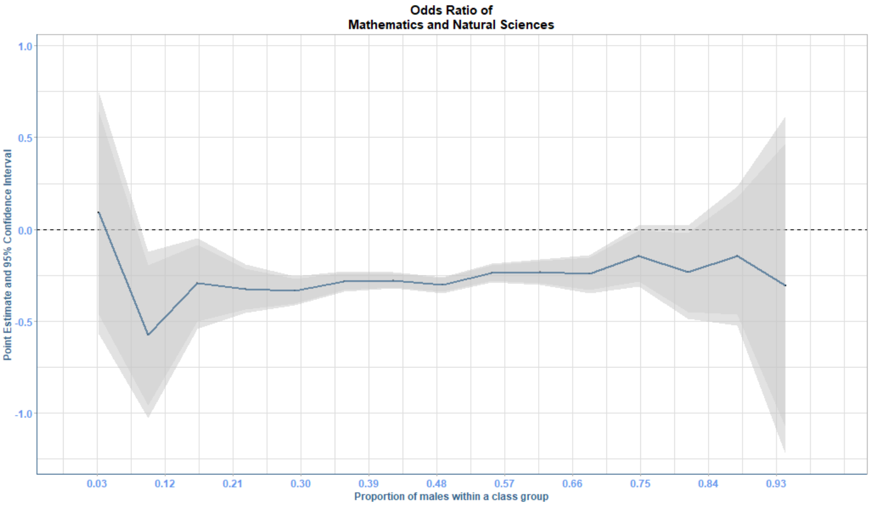

Figure 19: Likelihood of a female student choosing careers related to mathematics and natural science

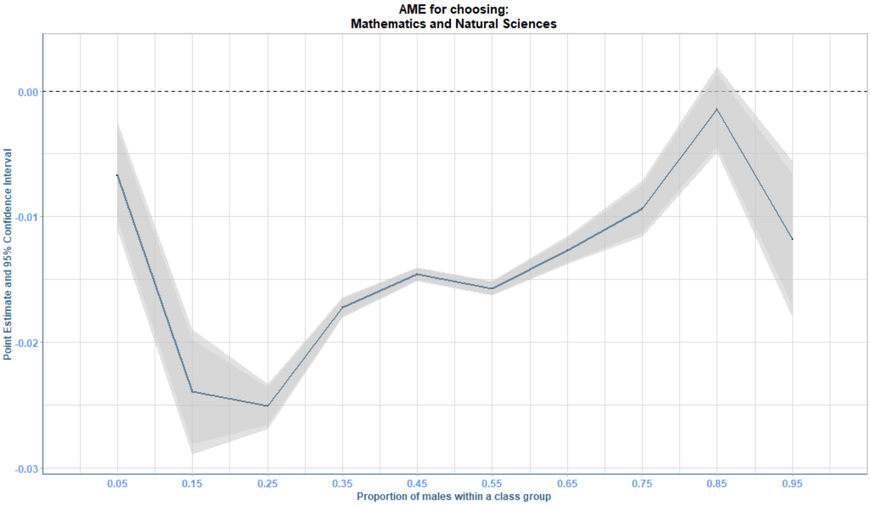

Figure 20: Average Marginal Effect of a female student choosing careers related to mathematics and natural science

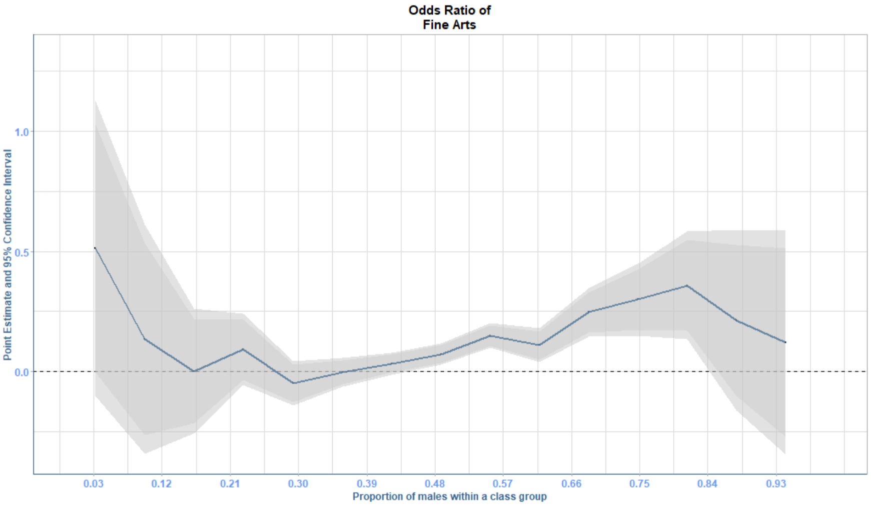

Figure 21: Likelihood of a female student choosing careers related to fine arts

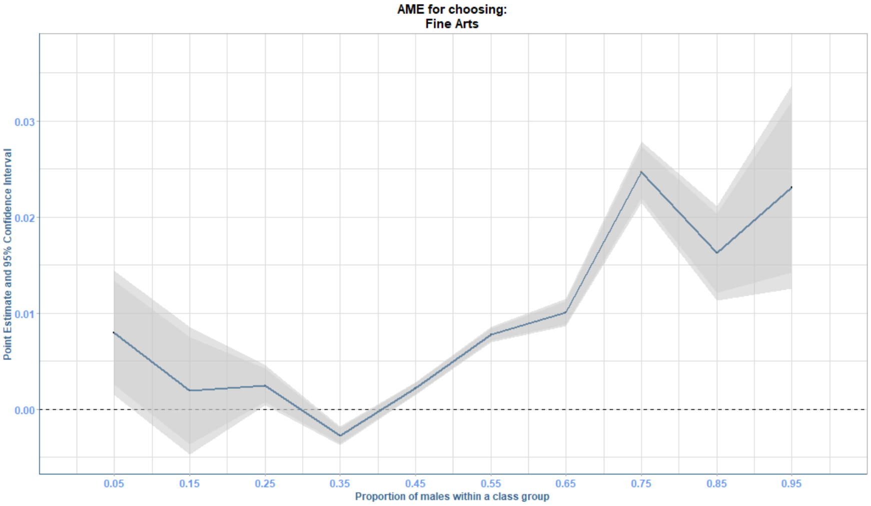

Figure 22: Average Marginal Effect of a female student choosing careers related to fine arts

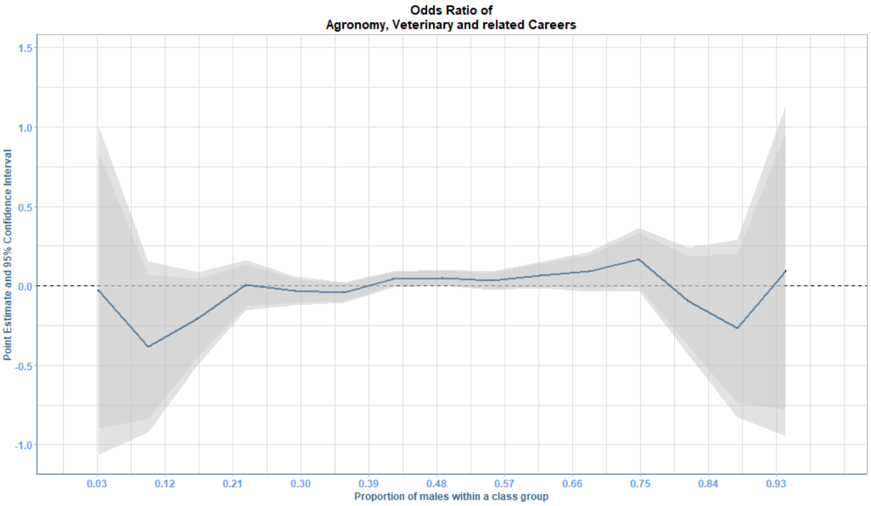

Figure 23: Likelihood of a female student choosing careers related to agronomy and agronomy and veterinary

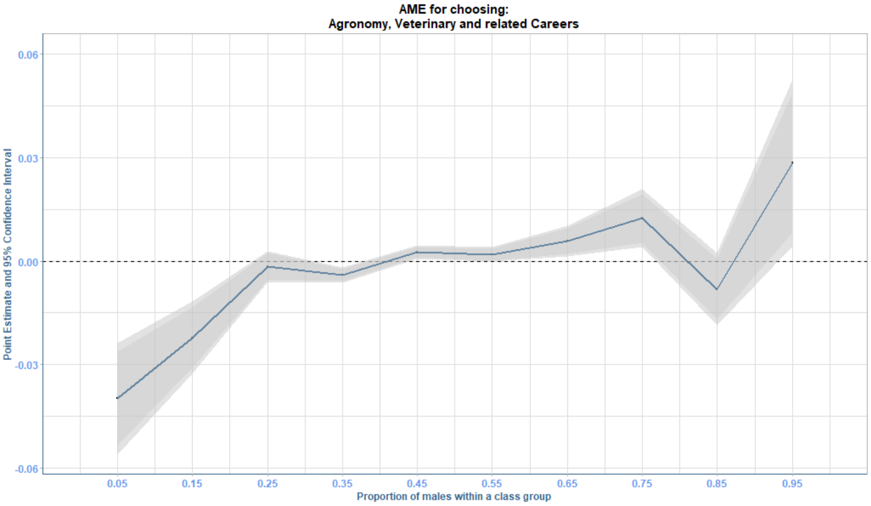

Figure 24: Average Marginal Effect of a female student choosing careers related to agronomy and agronomy and veterinary

Dependent Variables:	Medicine	Econ. and Business	Eng. and Archit.	Fine Arts	Math. and Nat. Scs	Scocial Scs	Agronomy and rel.	Education Scs	Health	No studies	Laws
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Variables											
-5	-0.1842	1.368	0.1344	-0.5156*	-0.1423	-0.2190	-0.1769	0.4220	-0.1065	0.7519	1.139
	(0.1176)	(1.120)	(1.303)	(0.2792)	(0.0983)	(0.3675)	(0.1672)	(0.2895)	(0.1716)	(3.750)	(1.091)
-4	-0.1291	0.3675	-0.5015	-0.2210	-0.0606	-0.1243	-0.0681	0.2610	0.0175	-1.726	-0.2278
	(0.0968)	(0.3524)	(0.7629)	(0.1737)	(0.0555)	(0.2716)	(0.1422)	(0.2801)	(0.1605)	(2.176)	(0.3122)
-3	-0.0445	1.592	0.2667	1.385	0.0789	-0.2409	-0.0452	-0.1471	0.1622	0.3912	-0.0632
	(0.0735)	(0.9749)	(0.7071)	(0.9317)	(0.0727)	(0.4065)	(0.1066)	(0.1977)	(0.1594)	(1.615)	(0.2508)
-2	-0.0800	1.403*	0.7868*	0.0642	0.0278	-0.2223	-0.0739	0.6275	0.0151	0.9828	-0.1498
	(0.0736)	(0.6974)	(0.4337)	(0.1649)	(0.0895)	(0.2779)	(0.0997)	(0.5791)	(0.0818)	(1.408)	(0.2076)
-1	-0.0306	0.2533	0.2070	-0.1812	-0.0036	-0.2743	-0.0767	0.2746	0.0838*	1.216	-0.0304
	(0.0544)	(0.1965)	(0.1803)	(0.1387)	(0.0293)	(0.2242)	(0.0710)	(0.1770)	(0.0452)	(1.065)	(0.1393)
1	-0.0113	0.1299	0.0669	-0.0865	0.0899	-0.1709	-0.0951	0.0060	-0.0428	1.216*	-0.0153
	(0.0501)	(0.2640)	(0.1064)	(0.0778)	(0.0847)	(0.3404)	(0.0747)	(0.1067)	(0.0338)	(0.7062)	(0.0608)
2	-0.0179	-0.1643	0.2392	-0.1414	-0.0164	-0.1244	-0.1251	-0.0350	-0.0359	0.9289	-0.1039
	(0.0459)	(0.1521)	(0.2284)	(0.1072)	(0.0400)	(0.2798)	(0.1015)	(0.0729)	(0.0477)	(1.032)	(0.0855)
3	-0.0903	-0.4023***	-0.0753	-0.1432	0.0736	-0.1864	-0.0325	0.0057	0.0137	-0.0054	-0.0746
	(0.0817)	(0.1473)	(0.2222)	(0.0882)	(0.0610)	(0.2275)	(0.0685)	(0.0956)	(0.0544)	(0.9853)	(0.0882)
4	0.0112	-0.1927	0.0239	-0.0359	0.0342	-0.3216	-0.0437	-0.0982	-0.1084	-0.1944	-0.0746
	(0.0828)	(0.1872)	(0.2690)	(0.0763)	(0.0370)	(0.2426)	(0.0879)	(0.1390)	(0.0922)	(1.165)	(0.1191)
5	0.2898	0.0278	0.1530	-0.2015	0.0369	-0.2842	-0.0634	-0.0121	-0.1249	0.9469	0.0330
	(0.1943)	(0.2613)	(0.4577)	(0.1675)	(0.0450)	(0.2718)	(0.1119)	(0.2045)	(0.1239)	(1.480)	(0.1537)
6	-0.0032	-0.1696	-0.0004	-0.1409	0.0191	-0.2763	-0.1291	-0.0631	0.0817	0.6730	-0.0619
	(0.1356)	(0.2721)	(0.3875)	(0.0945)	(0.0540)	(0.3219)	(0.1181)	(0.2185)	(0.1925)	(1.563)	(0.1468)
7	0.0384	-0.5466*	-0.2923	-0.1405	0.0590	-0.4063	-0.0451	-0.0930	-0.1201	0.2221	0.0039
	(0.1252)	(0.2960)	(0.4391)	(0.1746)	(0.0578)	(0.3227)	(0.1126)	(0.1741)	(0.1450)	(1.801)	(0.1716)
8	0.1721	-0.4737	-0.0768	-0.2292	0.0734	-0.1966	-0.1257	-0.1033	-0.2028	2.356	0.0115
	(0.1434)	(0.3219)	(0.4937)	(0.2413)	(0.0645)	(0.3088)	(0.1472)	(0.2301)	(0.1511)	(2.443)	(0.1762)
Fixed-effects											
id_name	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics											
Observations	265	265	265	265	265	265	265	265	265	265	265
R^2	0.57737	0.43546	0.48926	0.43218	0.29394	0.28442	0.35812	0.39278	0.35315	0.72197	0.35189
Within R^2	0.11627	0.14579	0.03632	0.19577	0.04087	0.02071	0.03342	0.06286	0.06904	0.04221	0.09511

Clustered (id_name) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Dependent Variables:	Medicine	Econ. and Business	Eng. and Archit.	Fine Arts	Math. and Nat. Scs	Scocial Scs	Agronomy and rel.	Education Scs	Health	No studies	Laws
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Variables											
-6	-0.0920***	-0.0117	0.0631	-0.0661	0.1198	-0.0123	0.1036	0.0181	0.0078	0.3264	0.0102
	(0.0349)	(0.1294)	(0.1162)	(0.0441)	(0.1081)	(0.0779)	(0.0767)	(0.0496)	(0.0703)	(0.4050)	(0.0383)
-5	-0.0626**	0.0723	0.0482	0.0950	0.1434	0.4448	0.1020*	-0.0152	-0.0777**	0.2523	0.1231
	(0.0280)	(0.1681)	(0.0941)	(0.1135)	(0.1362)	(0.3914)	(0.0564)	(0.0300)	(0.0388)	(0.5080)	(0.1357)
-4	-0.0461*	-0.1132	0.0047	-0.0023	0.0265	-0.0267	0.1000*	-0.0383	-0.0495	0.2699	-0.0135
	(0.0249)	(0.0692)	(0.0820)	(0.0357)	(0.0217)	(0.0520)	(0.0565)	(0.0243)	(0.0463)	(0.2939)	(0.0406)
-3	-0.0291*	-0.0027	0.0087	-0.0124	0.0245	-0.0057	0.0701	0.0085	-0.0055	-0.0934	0.0227
	(0.0176)	(0.0532)	(0.0663)	(0.0272)	(0.0167)	(0.0484)	(0.0518)	(0.0198)	(0.0380)	(0.2174)	(0.0342)
-2	-0.0383**	0.0371	-0.0789	-0.0142	0.0140	-0.0093	0.0439	0.0087	-0.0344	-0.0776	-0.0081
	(0.0180)	(0.0511)	(0.0542)	(0.0238)	(0.0114)	(0.0391)	(0.0272)	(0.0228)	(0.0432)	(0.2081)	(0.0375)
-1	-0.0228	0.0075	0.0175	-0.0037	0.0134	0.0033	0.2063	0.0196	-0.0714**	0.2501	0.0260
	(0.0211)	(0.0425)	(0.0579)	(0.0193)	(0.0108)	(0.0394)	(0.1591)	(0.0220)	(0.0333)	(0.2433)	(0.0334)
1	-0.0282	0.1025*	0.0054	0.0037	0.0196	0.0140	0.0079	0.0082	-0.0046	0.0823	0.0062
	(0.0259)	(0.0600)	(0.0415)	(0.0250)	(0.0127)	(0.0351)	(0.0287)	(0.0169)	(0.0363)	(0.2100)	(0.0341)
2	-0.0432	0.0700	0.0226	0.0111	0.0098	-0.0553	0.0759*	-0.0099	-0.0017	-0.0971	0.0079
	(0.0343)	(0.0452)	(0.0554)	(0.0286)	(0.0189)	(0.0376)	(0.0397)	(0.0179)	(0.0445)	(0.2186)	(0.0354)
3	-0.0692**	0.0981**	-0.0061	0.0206	0.0269*	-0.0504	0.0096	0.0240	-0.0176	-0.2125	-0.0030
	(0.0300)	(0.0470)	(0.0615)	(0.0355)	(0.0160)	(0.0421)	(0.0155)	(0.0210)	(0.0379)	(0.2444)	(0.0362)
4	-0.0286	0.0162	0.0069	0.0032	0.0062	-0.0277	0.0515*	-0.0033	-0.0068	0.1588	-0.0073
	(0.0292)	(0.0595)	(0.0759)	(0.0218)	(0.0152)	(0.0529)	(0.0279)	(0.0242)	(0.0340)	(0.3869)	(0.0381)
5	-0.0270	0.0809	0.0183	-0.0115	-0.0022	-0.0129	0.0571	-0.0071	0.0236	-0.1657	0.0111
	(0.0349)	(0.0644)	(0.0766)	(0.0295)	(0.0189)	(0.0562)	(0.0604)	(0.0266)	(0.0523)	(0.3338)	(0.0389)
6	-0.0097	0.1154	0.0160	-0.0163	0.0057	0.0075	0.0756	-0.0329	-0.0271	0.2345	0.0144
	(0.0435)	(0.0800)	(0.0905)	(0.0346)	(0.0233)	(0.0682)	(0.0619)	(0.0271)	(0.0497)	(0.4067)	(0.0533)
7	0.0202	0.1386	-0.0794	0.0023	-0.0089	0.0735	0.0040	0.0055	-0.0110	-0.5129	0.0174
	(0.0496)	(0.0887)	(0.1090)	(0.0446)	(0.0271)	(0.0731)	(0.0760)	(0.0358)	(0.0562)	(0.4469)	(0.0536)
8	0.1084**	0.0640	0.0456	0.0944**	0.0264	0.1360	0.0128	-0.0234	0.0366	-0.3876	0.0958*
	(0.0483)	(0.1003)	(0.1429)	(0.0426)	(0.0229)	(0.0849)	(0.0879)	(0.0323)	(0.0543)	(0.5322)	(0.0529)
Fixed-effects											
id_name	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics											
Observations	1,052	1,052	1,052	1,052	1,052	1,052	1,052	1,052	1,052	1,052	1,052
R^2	0.52284	0.45970	0.45576	0.37554	0.29604	0.47764	0.23328	0.32506	0.42700	0.76881	0.39351
Within R^2	0.01291	0.01205	0.00433	0.00738	0.01403	0.03696	0.01784	0.01233	0.00877	0.01312	0.00900

Clustered (id_name) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

6 Conclusion

References

Alesina, A., Giuliano, P., & Nunn, N. (2013). On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics*, *128*(2), 469-530.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, *75*(4), 352–365. Retrieved 2023-10-23, from http://www.jstor.org/stable/1828596

Buser, T., Niederle, M., & Oosterbeek, H. (2014, 05). Gender, Competitiveness, and Career Choices *. *The Quarterly Journal of Economics*, *129*(3), 1409-1447. Retrieved from https://doi.org/10.1093/qje/qju009 doi: 10.1093/qje/qju009

Capilla Navarro Guzmán, & Antonio Casero Martínez. (2012). Análisis de las diferencias de género en la elección de estudios universitarios.

Clifford D. Evans. (2006). Life GOALS: Antecedents IN GENDER BELIEFS AND EFFECTS ON GENDER-STEREOTYPICAL CAREER INTEREST.

Cooper, R., & Liu, H. (2019a). *Mismatch in human capital accumulation* (Vol. 60; Tech. Rep. No. 3). Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/iere.12386 doi: https://doi.org/10.1111/iere.12386

Evans, C. D., & Diekman, A. B. (2009, 6). On Motivated Role Selection: Gender Beliefs, Distant Goals, and Career Interest. *Psychology of Women Quarterly*, *33*(2), 235–249.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources*, *14*(3), 351–388. Retrieved 2023-11-08, from http://www.jstor.org/stable/145575

Heenan, D. (2002, 5). Women, Access and Progression: An examination of women's reasons for not continuing in higher education following the completion of the Certificate in Women's Studies. *Studies in Continuing Education*, *24*(1), 39–55.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*(3), 933–959. Retrieved 2023-11-27, from http://www.jstor.org/stable/23261375

Kohen, Z., & Nitzan, O. (2022). Excellence in mathematics in secondary school and choosing and excelling in stem professions over significant periods in life. *International Journal of Science and Mathematics Education*, *20*(1), 169–191. Retrieved from https://doi.org/10.1007/s10763-020-10138-x doi: 10.1007/s10763-020-10138-x

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, *45*(1), 79-122. Retrieved from https://www.sciencedirect.com/science/article/pii/S000187918471027X doi: https://doi.org/10.1006/jvbe.1994.1027

Lönnqvist, J.-E., Verkasalo, M., Walkowitz, G., & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? an empirical comparison. *Journal of Economic Behavior Organization*, *119*, 254-266. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167268115002115 doi: https://doi.org/10.1016/j.jebo.2015.08.003

Manisha Joshi. (2016). Status of Women in Higher Education.

McDermott, C. L. (2012). To Stand on Their Own Legs: Independence and Other Motivations for Women's Pursuit of Post-Graduate Studies, and Their Parents' Influence on Them, in Vishakhapatnam India. *SSRN Electronic Journal*.

National Center for Science and Engineering Statistics. (2023). *Diversity and stem: Women, minorities, and persons with disabilities* (Report No. NSF 23-315). Retrieved from https://ncses.nsf.gov/pubs/nsf23315/report

OECD. (2013). *Pisa 2012 results: Ready to learn (volume iii).* Retrieved from https://www.oecd-ilibrary.org/content/publication/9789264201170-en doi: https://doi.org/https://doi.org/10.1787/9789264201170-en

Shahar Gindi, J. K.-M., & Pilpel, A. (2019). Gender differences in competition among gifted students: The role of single-sex versus co-ed classrooms. *Roeper Review*, *41*(3), 199-211. Retrieved from https://doi.org/10.1080/02783193.2019.1622163 doi: 10.1080/02783193.2019.1622163

Sutter, M., & Glätzle-Rützler, D. (2010, 06). *Gender differences in competition emerge early in life* (IZA Discussion Paper No. 5015). IZA (Institute of Labor Economics). Retrieved from https://www.iza.org/publications/dp/5015/gender-differences-in-competition-emerge-early-in-life (Largely extended version published in: Management Science, 2015, 61 (10), 2339-2354)

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33. Retrieved 2023-11-08, from http://www.jstor.org/stable/3590137

Wang, N., Tan, A.-L., Zhou, X., Liu, K., Zeng, F., & Xiang, J. (2023). Gender differences in high school students' interest in stem careers: a multi-group comparison based on structural equation model. *International Journal of STEM Education*, *10*(59), 1–14. Retrieved from https://doi.org/10.1186/s40594-023-00351-2 (1843 Accesses, 4 Altmetric) doi: 10.1186/s40594-023-00351-2

A Appendix

A.1 Description of University Major Choices by Knowledge Areas

This section provides an overview of university major choices categorized into distinct knowledge areas. The aggregated classification is structured as follows:

- 1. Economics, Business & related Careers (e.g., Economics, Business Administration, Finance, Accounting, Marketing, Management, Entrepreneurship, International Business, Human Resources)
- 2. Engineering, Architecture and related Careers (e.g., Civil Engineering, Mechanical Engineering, Electrical Engineering, Architecture, Computer Science, Information Technology, Software Engineering, Industrial Design, Environmental Engineering, Biomedical Engineering)
- 3. Fine Arts (Visual Arts, Performing Arts (e.g., Theater, Dance, Music), Graphic Design, Interior Design, Animation)
- 4. Mathematics and Natural Sciences (e.g., Mathematics, Physics, Chemistry, Biology, Environmental Science, Geology, Astronomy, Statistics)
- 5. Social Sciences and Humanities (e.g., Sociology, Anthropology, History, Political Science, Geography, Literature, Philosophy, Religious Studies, Linguistics, Communication Studies)
- 6. Agronomy, Veterinary and related Careers (e.g., Agronomy, Animal Science, Veterinary Medicine, Zoology, Horticulture, Fisheries and Aquaculture)
- 7. Education Sciences (e.g., Early Childhood Education, Special Education, Educational Psychology, Education in Mathematics, Education in Sciences)
- 8. Health Sciences (e.g., Nursing, Dentistry, Pharmacy, Physical Therapy, Occupational Therapy, Public Health, Nutrition, Biomedical Sciences, Health Administration,)
- 9. No Studies (The student does not continue with professional studies)

Table A.1: Correlation between the gender composition in a class and the likelihood of a student choosing a career

	Dependent variable:								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Proportion of males within a class group	-0.931^{***}	0.982^{***}	-0.200^{***}	-0.436^{***}	-0.996^{***}	-0.011	-0.707^{***}	-1.119^{***}	0.445^{***}
	(0.010)	(0.009)	(0.026)	(0.026)	(0.012)	(0.026)	(0.021)	(0.015)	(0.006)
Constant	-1.828^{***}	-2.535^{***}	-4.285^{***}	-4.176^{***}	-2.407^{***}	-4.335^{***}	-3.658^{***}	-2.732^{***}	0.385^{***}
	(0.005)	(0.005)	(0.013)	(0.012)	(0.006)	(0.012)	(0.010)	(0.007)	(0.003)
Observations	4,486,601	4,486,601	4,486,601	4,486,601	4,486,601	4,486,601	4,486,601	4,486,601	4,486,601
Log Likelihood	-1,412,835.000	-1,562,897.000	-299,879.800	-300,436.700	-947,917.100	-308,616.400	-411,800.100	-723,508.200	-2,921,328.000
Akaike Inf. Crit.	2,825,674.000	3,125,798.000	599,763.700	600,877.400	1,895,838.000	617,236.900	823,604.200	1,447,020.000	5,842,660.000
Note:									*p<0.1; **p<0.05; ***p<0.01

A.2 Optimal bandwidth estimation based on Binary cross-entropy

In order to analyze the probability that a secondary school student chooses an area of knowledge P to pursue post-secondary studies. We analyze the probability according to different gender compositions in the classrooms. Therefore we assume that exists a fixed value that allow to subset by $\exists X_{\text{optimal}}$ The formal representation using mathematical notation for the partitioning of the range into fixed intervals: Let X_i be a subset that belongs to the gender composition with values between [0,1], we can say that, X_1 is a subset that goes from $min(X)$ to $min(X) + X_i$, consequently X_2 is a subset which goes from X_1 to $X_1 + X_i$, and sequentially until X_n goes from X_{n-1} to $max(X)$.

Let X_{optimal} be a fixed interval representing the space between each subset.

The subsets X_i can be defined as:

$$\begin{aligned}
X_1 &= [0, X_{\text{optimal}}) \\
X_2 &= [X_{\text{optimal}}, 2X_{\text{optimal}}) \\
X_3 &= [2X_{\text{optimal}}, 3X_{\text{optimal}}) \\
&\dots \\
X_n &= [(n-1)X_{\text{optimal}}, nX_{\text{optimal}})
\end{aligned}$$

These representations X_i cover the entire range in fixed intervals of X_{optimal} and define distinct subsets, each representing an interval of size X_{optimal} within the overall range.

To estimate X_{optimal} we modify the methodology proposed in Imbens and Kalyanaraman (2012), In it, the key step is to replace the mean squared error (MSE) criterion with a BCE-based criterion.

The key outcome we are trying to predict is a binary variable indicating whether a student chooses a particular area of study (e.g. science, humanities etc) or not. Let's call this $Y_i \in 0, 1$.

$Y_i = 1$ means student i chose that area of study $Y_i = 0$ means they did not choose that area. Our regression discontinuity model is estimating the probability $p_i = P(Y_i = 1|X_i)$ that the student chooses that area, conditioned on the gender composition in classrooms X_i.

Let's call this estimated probability $m(X_i)$, which depends on the bandwidth h.

The BCE loss for a single data point measures how well our model is estimating this probability. It is:

$$\text{BCE}_i = \begin{cases} -\log(m(X_i)), & \text{if } Y_i = 1 \;\; -\log(1 - m(X_i)), \\ \text{if } Y_i = 0 \end{cases}$$

Penalizes underestimating probability if actual outcome is 1 Penalizes overestimating probability if actual outcome is 0 We then define the overall expected BCE loss over the distribution of (X_i, Y_i) as:

$$\text{BCE}(h) = E[-Y_i \log(m(X_i)) - (1 - Y_i) \log(1 - m(X_i))]$$

Minimizing this BCE(h) gives the optimal bandwidth for our RD model.

A.3 more...

1. Define the BCE Loss Function:

The key outcome we are trying to predict is a binary variable indicating whether a student chooses a particular area of study (e.g. science, humanities etc) or not. Let's call this $Y_i \in 0, 1$.

$Y_i = 1$ means student i chose that area of study $Y_i = 0$ means they did not choose that area. Our regression discontinuity model is estimating the probability $p_i = P(Y_i = 1|X_i)$ that the student chooses that area, conditioned on the gender composition in classrooms X_i.

Let's call this estimated probability $m(X_i)$, which depends on the bandwidth h.

The BCE loss for a single data point measures how well our model is estimating this probability. It is:

$$\text{BCE}_i = \begin{cases} -\log(m(X_i)), & \text{if } Y_i = 1 \ -\log(1 - m(X_i)), \\ \text{if } Y_i = 0 \end{cases}$$

Penalizes underestimating probability if actual outcome is 1 Penalizes overestimating probability if actual outcome is 0 We then define the overall expected BCE loss over the distribution of (X_i, Y_i) as:

$$\text{BCE}(h) = E[-Y_i \log(m(X_i)) - (1 - Y_i) \log(1 - m(X_i))]$$

Minimizing this BCE(h) gives the optimal bandwidth for our RD model.

2. Approximate BCE: We have defined the BCE loss as:

$$\text{BCE}(h) = E[-Y_i \log(m(X_i)) - (1 - Y_i) \log(1 - m(X_i))]$$

However, we cannot directly optimize this BCE(h) to find the best bandwidth h. The expectation over (Xi, Yi) pairs and dependence on the regression function m(Xi) is too complicated.

So we take a Taylor expansion of BCE(h) around the point h=0. This allows us to approximate BCE(h) for small values of h (which is the relevant range for bandwidth selection).

Specifically:

We assume higher order terms are negligible. After substituting the derivatives, this second order approximation takes the form:

$$\text{AMSEBCE}(h) = C_1 h^4 (m''(c) - m'' - (c))^2 + \frac{C_2}{Nh}$$

Where:

C_1, C_2 depend on moments of Y distribution and kernel m'' and m''_- are second derivatives of the regression function This $\text{AMSE}_{\text{BCE}}(h)$ can now be optimized tractably to find the best bandwidth h. It maintains the key structure and tradeoff between variance and bias squared terms.

Minimize Approximate BCE: In the previous step, we derived the approximate BCE loss function:

$$\text{AMSEBCE}(h) = C_1 h^4 (m''(c) - m'' - (c))^2 + \frac{C_2}{Nh}$$

This approximate loss maintains the core structure from the MSE case - having a bias squared term that increases with h and a variance term that decreases with h.

Our goal now is to find the value of h that minimizes this loss, balancing the bias-variance tradeoff. We can find this by taking the derivative with respect to h and setting it equal to zero:

$$\frac{d}{dh}\text{AMSEBCE}(h) = 4C_1 h^3 (m''(c) - m'' - (c))^2 - \frac{C_2}{Nh^2}$$

Setting this equal to zero gives us the optimal bandwidth that minimizes the approximate BCE:

$$h_{\text{opt, BCE}} = \left(\frac{C_2}{4C_1}\right)^{\frac{1}{5}} N^{-\frac{1}{5}}$$

We get a very similar expression as in the MSE case, with the leading constant now depending on the BCE-based constants C_1 and C_2.

This $h_{\text{opt, BCE}}$ minimizes the approximate expected BCE loss over the distribution of data. Using this bandwidth will give us the regression discontinuity model that best trades off bias vs variance in terms of BCE.

Estimate the Bandwidth

We derived the formula for the optimal bandwidth that minimizes the approximate BCE criterion:

$$h_{\text{opt, BCE}} = \left(\frac{C_2}{4C_1}\right)^{\frac{1}{5}} N^{-\frac{1}{5}}$$

The issue is this still relies on unknown population quantities - namely the constants C_1, C_2 and the second derivatives of the regression function $m''(c)$ and $m''_-(c)$.

So the final step is to estimate these unknowns from the data, in order to obtain a data-driven bandwidth estimate. There are a few options for doing this estimation:

Use pilot estimates: Obtain initial/crude estimates of C_1, C_2, m'', m''_- using some pilot bandwidth h_{pilot}. These don't need to be very precise. Moment approximations: Approximate moments of Y distribution and kernel to get estimates of C_1, C_2 without directly estimating them. Iterative/cross-validation: Obtain estimates of the derivatives m'', m''_- using some initial h. Then solve for $\hat{\mathrm{h}}$ opt. Iterate with updated derivative estimates. Either way, once we plug in these estimates, we get a feasible bandwidth formula:

$$\hat{h}_{\text{opt, BCE}} = \left(\frac{\hat{C}_2}{4\hat{C}_1}\right)^{\frac{1}{5}} N^{-\frac{1}{5}}$$

This estimated $\hat{h}_{\text{opt, BCE}}$ consistently estimates the optimal bandwidth and maintains the same asymptotic properties as if the true unknowns were used.