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 ‡

June 25, 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. 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, PhD Student at KU Leuven & Pontificia Universidad Javeriana. (email: jaime.polanco@javeriana.edu.co)
†KU Leuven & UNU-Merit, Maastricht University. (email: kristof.dewitte@kuleuven.be).
‡Pontificia Universidad Javeriana. (email: gbernal@javeriana.edu.co).

1 Introduction

The decision of pursuing post-secondary studies carries significant implications, shaping not only future career paths but also long-term earning potential. National Center for Science and Engineering Statistics (2023) highlights the substantial income advantage (35%) associated with careers in science, technology, engineering, and mathematics (STEM) compared to fields like humanities, social sciences, and education. While financial incentives are crucial, other factors such as the development of critical thinking, problem-solving, and communication skills significantly influence educational choices, particularly in humanities (Alesina, Giuliano, & Nunn, 2013; Clifford D. Evans, 2006; Evans & Diekman, 2009; Lent, Brown, & Hackett, 1994).

In Colombia, a developing country in South America, these implications become especially prominent given the significant number of students (approximately half a million) completing secondary education annually. Despite a higher rate of women completing university degrees, there is a noticeable gender gap in the selection of academic disciplines. Specifically, in the 7.5 years following secondary education, an average of 33% of women and 29% of men pursue and complete a college academic program. However, only 16% of female university graduates choose careers in mathematics, engineering, science, architecture, construction, and related fields, compared to 24% of male graduates (ICFES, 2019).

Among the factors influencing these decisions, research consistently shows that men are more likely to enter and perform well in competitive environments compared to women (Gneezy, 2003; Vesterlund, 2011). This disparity has been attributed to factors such as overconfidence, attitudes towards competition, and gender-task stereotypes (Riener, 2010). Importantly, these differences are not inherent and can be influenced by factors such as early childhood education and societal expectations (Vesterlund, 2011). Moreover, despite demonstrating the gap between women and men in competitiveness, it is shown that in mixed environments, the competitiveness gap between men and women becomes greater (Gneezy, 2003). This implies that while women compete in different topics and maintain the gap, the presence of men in the competition exacerbates this gap.

Given this context, the learning environment plays a crucial role in shaping students' career choices, especially in STEM fields (Pregaldini, Backes-Gellner, & Eisenkopf, 2020). The impact of classroom demography on girls' STEM performance and persistence highlights the importance of considering the social context in fostering interest and success in STEM subjects.

Further, self-efficacy and academic performance or early STEM experiences are key factors influencing STEM degree choices (Bottia, Stearns, Mickelson, Moller, & Valentino, 2015; Pregaldini et al., 2020). Psychological, social, and cultural mechanisms also play a significant role in driving gender differences in STEM choices (Degol, 2020; Tyler-Wood, Johnson, & Cockerham, 2018).

Our study aims to explore the impact of gender composition in classrooms on post-secondary study decisions, focusing on the influence male students may exert on female students' academic aspirations through social interactions. Drawing on previous research highlighting male students' tendency towards higher levels of competitiveness(Buser, Niederle, & Oosterbeek, 2014; Kohen & Nitzan, 2022), we hypothesize that a greater proportion of male students in a classroom may negatively affect female students' propensity to pursue university majors in mathematics, engineering, and sci-

ence. This study addresses an important gap in the literature and contributes to our understanding of the social dynamics shaping educational outcomes. Furthermore, our research is the first to establish a connection between gender school composition and post-secondary study preferences in a developing country.

The literature suggests that differences in competitiveness between female and male students emerge as early as the first 4 to 5 years of age (Sutter & Glätzle-Rützler, 2010). Male students exhibit a greater perception of motor and spatial competitiveness, while female students tend to lean towards a greater perception of verbal competitiveness (Shahar Gindi & Pilpel, 2019). These characteristics are significant because male students are more inclined towards careers in STEM, while female students tend to gravitate towards social and human sciences.

To empirically investigate this phenomenon, we utilize datasets from the Integrated School Enrollment System in Colombia (SIMAT) and the National Higher Education Information System (SNIES). These datasets allow us to track individuals over time and explore the majors they 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 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.

We employ a rigorous 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 includes 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 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 three primary sources: the Integrated School Enrollment System in Colombia (SIMAT), the National Higher Education Information System (SNIES), and the Formal Education Survey (EDUC).

The Integrated School Enrollment System in Colombia (SIMAT) serves as a fundamental resource in our research, providing detailed records of student enrollment and academic progress throughout their educational journey. This system offers a longitudinal perspective, enabling us to track individuals over time and analyze trends in educational outcomes.

Similarly, the National Higher Education Information System (SNIES) offers invalu-

able insights into post-secondary education in Colombia. With its extensive database of higher education institutions, programs, and student enrollment data, SNIES allows us to explore the majors students choose after completing secondary school. By examining enrollment patterns and graduation rates, we can better understand the dynamics of post-secondary education and its implications for future career paths.

To complement our analysis of individual-level data, we turn to the Formal Education Survey (EDUC), which provides detailed information about school attributes and educational environments. By examining factors such as school size, resources, and student demographics, we gain a more holistic understanding of the contextual nuances shaping the relationship between gender, school composition, and post-secondary study decisions.

In table 1, we present the distribution of students across different gender compositions in secondary schools. This table illustrates the proportion of male students in each school, ranging from 0 (representing single-gender female schools) to 1 (representing single-gender male schools). It is important to note that the values at the extremes (0 and 1) denote single-gender schools, which are excluded from the descriptive analysis. These single-gender schools lack sample variation within the school, thereby limiting their utility when calculating the relationship between studying a major and gender as a conditioning factor.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.377	0.462	0.457	0.546	1.000

Table 1: Proportion of Male Students in the Last Year of Secondary Schools

From Figure 1, we observe a typical distribution representing the gender composition within classrooms, revealing an approximate mean with 46% women. In it, we identify two main challenges. Firstly, single-gender schools transitioning to mixed schools encounter what we term 'corner decisions.' Secondly, students attending mixed schools, where classrooms have varying gender compositions, alter their post-secondary career choices due to gender reinforcement dynamics within the classroom.

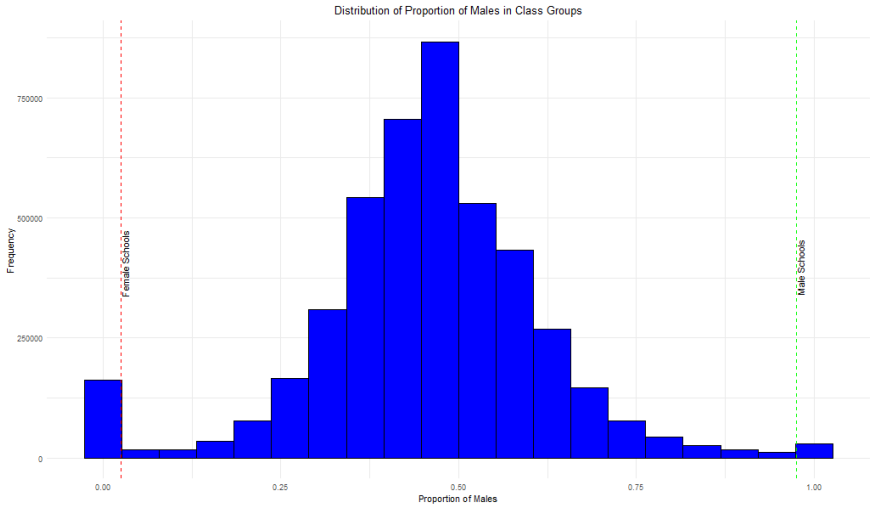


Figure 1: Distribution of Proportion of Males in Class Groups

3 Empirical Strategy

This study delves into the influence exerted by the gender composition within secondary schools on students' decisions regarding their choice of university majors. To achieve this, we undertake an examination of participation rates across ten distinct fields of study , each representing potential post-secondary career paths. Our investigation employs two distinct methodologies to ensure comprehensive analysis, while also exploring the underlying mechanisms driving these relationships.

3.1 Mechanism

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 [1].

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.

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)).

The investigation conducted by Sutter and Glätzle-Rützler (2010) revealed a notable gender-based divergence in competitiveness, wherein males exhibited a heightened

[1] 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.

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 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.

$$\text{Proportion of male students in a classroom} \Rightarrow \text{Female competitiveness} \Rightarrow \text{University major choice}$$

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.

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.

3.2 Methodology 1: Event-Study Regressions

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.

$$\hat{\tau} = \bar{Participation}^{P}_{\text{after transition}} - \bar{Participation}^{P}_{\text{before transition}} \tag{1}$$

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

Given your focus on the participation of students in major P before and after a transition, the adapted formula might look something like this (based on Sun and Abraham (2021)):

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

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 provide 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.

To address potential biases intrinsic to fixed effect estimation, our approach hinges on leveraging the transition from single-sex schools to coeducational settings as a natural intervention influencing student behavior. This unanticipated shift provides a unique opportunity for a robust examination of its effects.

Identification Strategy

Our study aims to ascertain the causal effect of Colombian schools transitioning from single-sex to coeducational settings on students' choice of university majors post-graduation. This is accomplished through the utilization of a staggered difference-in-differences (S-DiD) design, which involves a comparison between schools that have already undergone the transition and those that have not yet done so as of 2020.

The effectiveness of the S-DiD design relies on several critical assumptions. Firstly, we assume parallel trends in the absence of the transition, implying that the trends in students' choice of major P would have evolved similarly between schools that transitioned and those that have not yet transitioned, under the influence of common unobserved factors affecting major choice across schools.

Additionally, we assume the absence of concurrent shocks differentially affecting treated and untreated schools over the study period, apart from the transition itself. The inclusion of time fixed effects helps in adjusting for broader secular trends. Furthermore, we assume no anticipation of the transition's impact on student behavior until its actual occurrence, and test for any anticipatory effects by examining leads of the treatment indicator.

Moreover, the irreversibility assumption posits that once a school transitions, it remains coeducational throughout the study period, with no schools reverting to single-sex education within the sample. We also rely on the presence of overlapping cohorts within each school at any given time, enabling the observation of treated and untreated cohorts concurrently to identify the effect.

Finally, we assume the stability of student and school compositions over time. School fixed effects are incorporated to adjust for time-invariant compositional differences across schools.

The staggered timing of schools' transitions provides variation in treatment status over time. By comparing outcomes between treated and untreated schools before and after the transitions, and conditioning on school and time fixed effects, we can effectively isolate the causal effect of the transition, contingent upon the validity of the aforementioned assumptions.

3.3 Methodology 2: Probability Estimation based on Gender Composition

In this secondary analysis, we compute the probability that a secondary school student, denoted as i, selects a particular university major c based on the gender composition within their classroom. For instance, a gender composition of 0.2 signifies that a secondary female student, denoted as i, studied in a classroom where 20% of the students were male, while the remaining 80% were female.

To gauge the likelihood of a secondary school student opting for a specific university major contingent upon the classroom gender composition, we segment the groups based on specific proportions of male students within each classroom.

Moreover, as an integral part of our analytical framework, we have adapted the methodology proposed by Imbens and Kalyanaraman (2012). This methodology, focusing on minimizing the mean squared error to identify similar groups based on their outcome variable, has been tailored to segment a range of fractions of male students in a classroom who make similar study decisions. This adaptation is founded on the Binary Cross Entropy loss function (Mao, Mohri, and Zhong (2023)), with a detailed explanation provided in Appendix A.3.

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

$$\log\left(\frac{P(Y^c_{i,s,t}=1)}{1-P(Y^c_{i,s,t}=1)}\right) = \beta_1 \times Gender_{i,s,t} + \beta_2 \times X^c_{i,s,t} + \gamma_t + \gamma_s + \varepsilon_{i,s,t} \qquad (3)$$

Where $Y^c_{i,s,t}$ 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.

The table 2 provides a comprehensive summary of key statistics derived from BCE bootstraping estimation, facilitating inference on the optimal distance concerning the selection of university majors based on gender composition within secondary schools. Each row of the table corresponds to a distinct outcome or university major, while the columns offer insights into the mean Binary Cross Entropy (BCE) score, standard deviation, and optimal distance. This optimal distance serves as a critical parameter in the estimation process across various subsets of gender composition settings.

The mean BCE score represents the average predictive accuracy of the model in estimating the probability of students choosing a specific major, with lower values indi-

Table 2: Summary of bootstraped BCE estimation

Outcome	Mean BCE	Standard Deviation	Optimal Distance
Law	0.1189	0.0117	0.0998
Medicine	0.1	0.0108	0.1109
No Studies	0.7555	0.0062	0.0832
Health Sciences	0.1585	0.0309	0.1109
Education Sciences	0.1313	0.0184	0.1109
Agronomy, Veterinary & Related	0.1202	0.0117	0.1109
Social Sciences & Humanities	0.2107	0.0262	0.1109
Mathematics & Natural Sciences	0.0988	0.0036	0.1109
Fine Arts	0.1041	0.0107	0.1109
Eng., Arch & Related	0.4616	0.0884	0.1109
Economics & Business Related	0.389	0.0538	0.0832

cating better predictive performance. The standard deviation reflects the variability in BCE scores across observations for each major, providing insights into the consistency of model predictions.

Additionally, the optimal distance represents the threshold of gender composition within classrooms that minimizes entropy in schooling decisions for each major. This distance signifies the proportion of male students within a classroom that exerts the most significant influence on students' choices of university majors. For instance, a higher optimal distance suggests that gender composition plays a more substantial role in determining students' decisions regarding that major.

These statistics serve as valuable indicators of the relationship between gender composition and university major selection, shedding light on the nuanced dynamics influencing students' educational trajectories. They provide essential context for understanding the impact of classroom demographics on academic choices and inform policymakers and educators seeking to promote gender diversity and equity within educational settings.

4 Results

The likelihood of female students selecting specific university majors varies significantly based on the gender composition within their classrooms, as evidenced by the series of figures presented.

Female Students Choosing not to Pursue Further Studies

Furthermore, the odds ratio for female students choosing not to pursue further studies also varies with changes in the male composition fraction in the classroom. The overall sample odd ratio for female students not pursuing any further studies is 0.101. However, when the proportion of male students is less than 14%, the odds ratio decreases to around -0.2. This indicates that the likelihood of female students not continuing their studies increases as the proportion of male students in the classroom

increases. This relationship is visually depicted in Figure 2, illustrating how the likelihood of a female student choosing not to pursue further studies shifts in response to variations in the gender composition of the classroom.

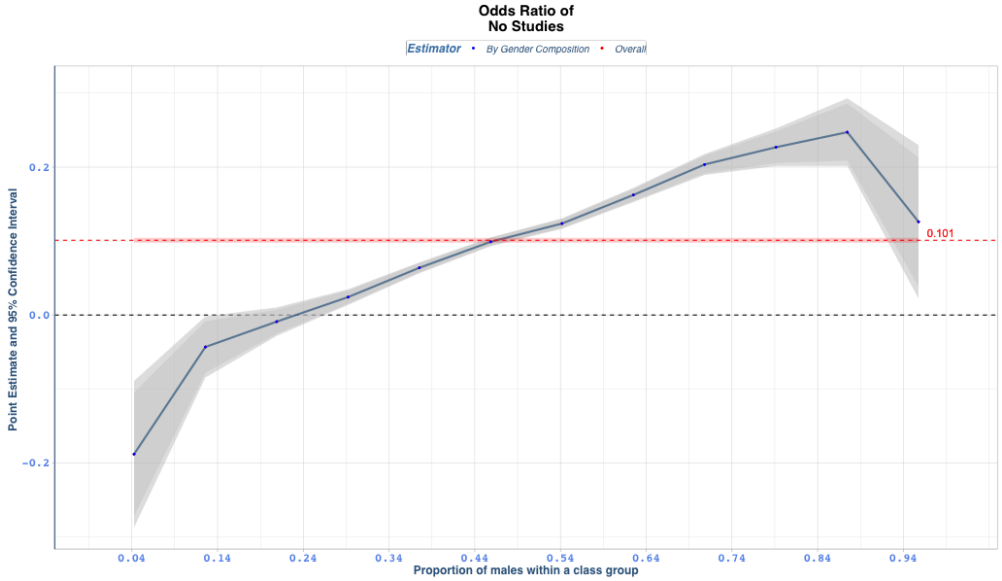


Figure 2: Likelihood of a female student choosing not to pursue further studies

Additionally, when a female school transitions to a coeducational school, the proportion of women who cease studying in the two years following the change tends to increase. This trend is demonstrated in Figure 3, highlighting the impact of school gender composition changes on female students' study decisions."

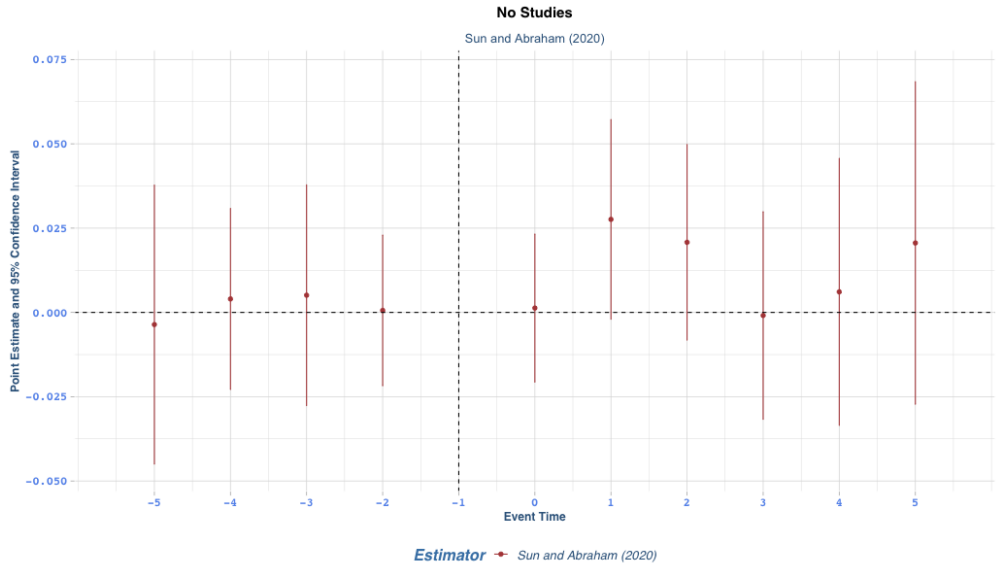


Figure 3: Proportion of Female Students Not Pursuing Further Studies in Schools Transitioning from Female to Coeducational

Female Students Choosing STEM Studies

We found that the odds ratio of female students choosing a STEM career reaches 0.92, indicating a significant difference compared to male students in similar classroom compositions. The relationship between classroom gender composition and female students' propensity to pursue STEM studies is depicted in Figure 4, which illustrates how the likelihood of a female student choosing a STEM-related career varies across different proportions of male students in the classroom.

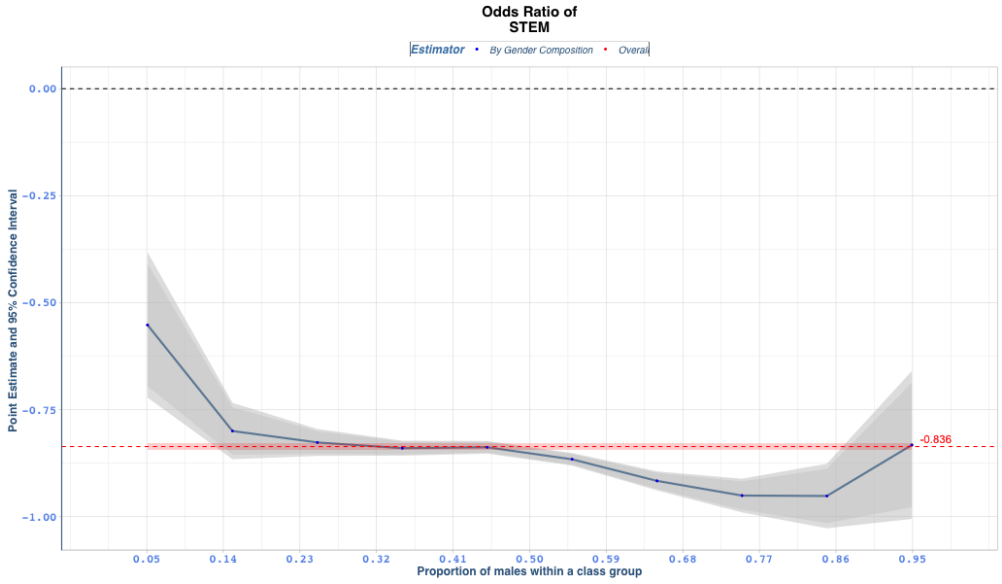


Figure 4: Likelihood of a female student choosing a STEM-related career

When the proportion of male students in the classroom increases, the likelihood of female students opting for STEM-related careers decreases. This finding aligns with existing theories that suggest a correlation between classroom composition, gender dynamics, and career aspirations (Buser et al., 2014; Gneezy, 2003). Specifically, in mixed-gender environments, the disparity in competitiveness between men and women tends to widen, with higher levels of competitiveness often associated with a greater inclination towards STEM careers (Buser et al., 2014).

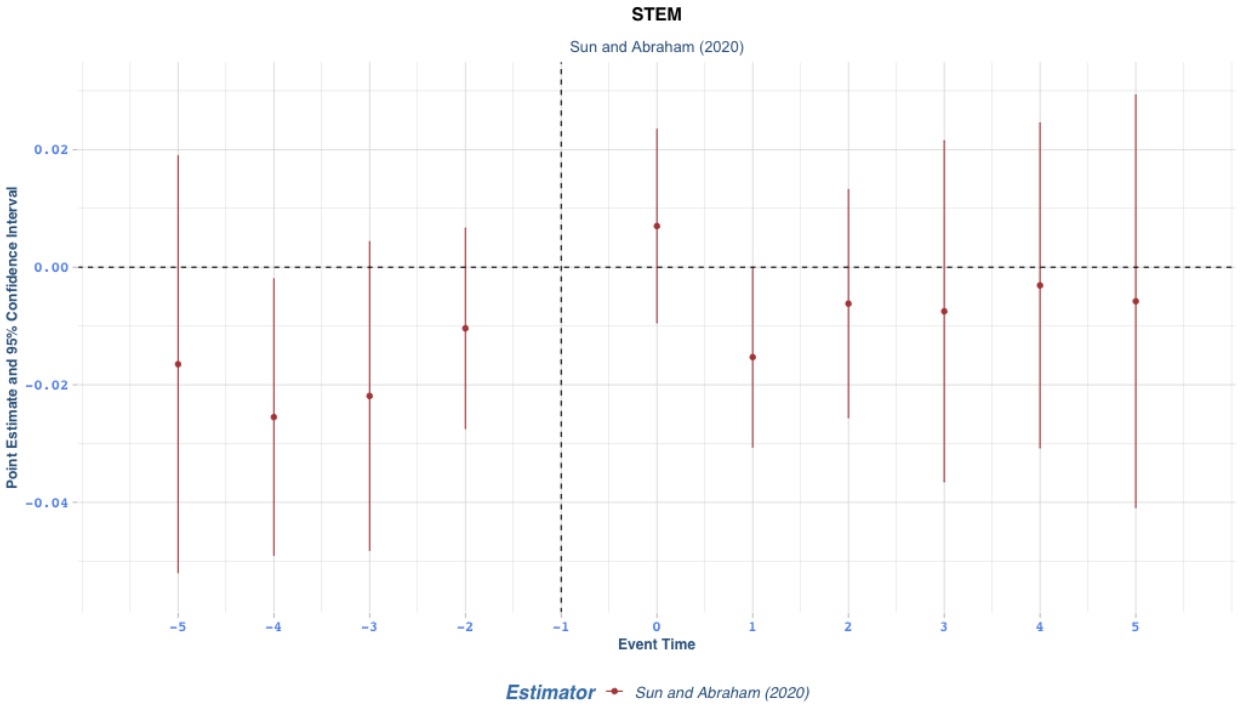


Figure 5: Proportion of Female Students Choosing STEM Majors in Schools Transitioning from Female to Coeducational

5 Conclusion

This study provides valuable insights into the impact of classroom gender composition on students' university major decisions. Our analysis reveals significant variations in female students' choices across academic disciplines, influenced by the presence of male peers in secondary school classrooms.

These findings underscore the complex interplay of gender dynamics in educational environments. Specifically, a higher proportion of male students correlates with increased interest among female students in traditionally male-dominated fields like engineering, architecture, and mathematics. Conversely, greater gender diversity in classrooms leads female students to exhibit a stronger preference for humanities and social sciences.

These results suggest that exposure to male peers may foster competitiveness and aspirations among female students, potentially narrowing gender gaps in STEM fields. However, they also highlight enduring gender-stereotypical preferences, with female students leaning towards communication-oriented disciplines in classrooms with more male representation.

Interestingly, minimal gender differences are observed in fields such as agriculture, veterinary sciences, and medicine, indicating limited influence of classroom gender compositions on major selection within these domains.

The findings underscore the importance of early educational experiences in shaping career trajectories and addressing gender disparities. By cultivating inclusive and gender-balanced learning environments, educational institutions can mitigate stereo-

typical pressures and promote equitable distribution of interests among students.

Moving forward, further research is warranted to elucidate the multifaceted mechanisms underlying these patterns, including peer influences, role modeling, and sociocultural norms. Longitudinal studies tracking students' evolving aspirations could provide deeper insights into the dynamics of gender composition and its long-term effects.

In conclusion, our analysis highlights the nuanced relationship between classroom gender composition and students' post-secondary schooling decisions. While specific majors exhibit varied responses to gender dynamics, overall, our findings suggest that gender-diverse environments may encourage female students to pursue higher education opportunities beyond secondary school.

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

Bottia, M. C., Stearns, E., Mickelson, R. A., Moller, S., & Valentino, L. (2015). Growing the roots of stem majors: Female math and science high school faculty and the participation of students in stem. *Economics of Education Review*, *45*, 14-27. Retrieved from https://www.sciencedirect.com/science/article/pii/S0272775715000035 doi: https://doi.org/10.1016/j.econedurev.2015.01.002

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

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

Degol, M.-T. W. J. (2020). Motivational pathways to stem career choices: Using expectancy–value perspective to understand individual and gender differences in stem fields. *Developmental Review*.

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.

Gneezy, M. . R. A., Uri ; Niederle. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*.

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

ICFES, I. (2019, April). *Saber 11° 2019-2.* Retrieved from https://www.datos.gov.co/Educaci-n/Saber-11-2019-2/rnvb-vnyh/about_data (Datos anonimizados de las pruebas de Saber 11 de Calendario A del año 2019)

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

Mao, A., Mohri, M., & Zhong, Y. (2023, 23–29 Jul). Cross-entropy loss functions: Theoretical analysis and applications. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 23803–23828). PMLR. Retrieved from https://proceedings.mlr.press/v202/mao23b.html

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

Pregaldini, D., Backes-Gellner, U., & Eisenkopf, G. (2020). Girls' preferences for stem and the effects of classroom gender composition: New evidence from a natural experiment. *Journal of Economic Behavior and Organization*.

Riener, N. D. G. . G. (2010). Explaining gender differences in competitiveness: Gender-task stereotypes. *Jena Economic Research Papers*.

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

Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, *225*(2), 175-199. Retrieved from https://www.sciencedirect.com/science/article/pii/S030440762030378X (Themed Issue: Treatment Effect 1) doi: https://doi.org/10.1016/j.jeconom.2020.09.006

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`

Tyler-Wood, T., Johnson, K., & Cockerham, D. (2018). Factors influencing student stem career choices: Gender differences. *Journal of Research in STEM Education*.

Vesterlund, M. N. . L. (2011). Gender and competition. *Annual Review of Economics*.

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 Results by Fields of Study

The varying preferences of female students across different fields of study in post-secondary schooling decisions can be illuminated by three key factors: social dynamics, role models and mentors, and perceived stereotypes and bias.

In areas such as Economics, Business, Social Sciences/Humanities, Education Sciences, Health Sciences (except Medicine), Medicine, and Law, female students exhibit a higher preference (See more details on preferred majors for female students in Subsection A.1.1). This preference may be influenced by the presence of supportive social dynamics within these fields, where female students feel a sense of belonging and encouragement. Additionally, the availability of visible role models and mentors in these disciplines could inspire female students to pursue further studies.

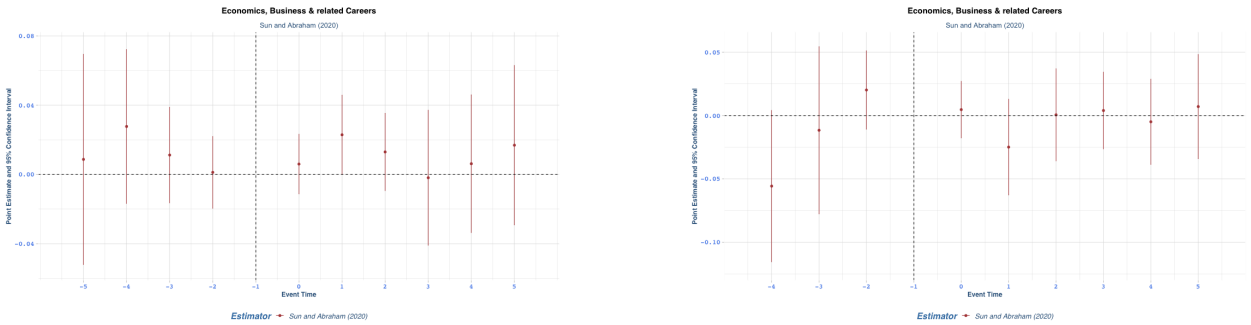
Conversely, in fields such as Engineering, Architecture, Mathematics and Natural Sciences, female students exhibit a lower preference, potentially influenced by entrenched stereotypes suggesting these disciplines are more suited for males(For further details, refer to Subsection A.1.2).. This bias, compounded by the competitiveness often observed in male students, tends to favor the choice of STEM-related careers, further dissuading female participation.

Interestingly, some areas such as Fine Arts and Agronomy/Veterinary-related fields demonstrate a similar level of preference between male and female students(For more insights, refer to Subsection A.1.3). This parity suggests that in disciplines where social dynamics are more balanced, and stereotypes and biases are less pronounced, both genders may feel equally encouraged to pursue further studies.

A.1.1 Fields of Study Preferred by Female Students

In this section, we analyze the fields of study where female students demonstrate a higher preference compared to male students.

1. **Economics, Business and Related Majors**: The Sun and Abraham (2021) estimation, depicted in the figure A.1 illustrates that the change in gender at school apparently does not significantly affect the decision to participate in careers related to administration and economics. However, it is crucial to note, as mentioned previously, that this observation only reflects the minimum diversity compositions of the new gender in the former single-gender schools. Therefore, the study depicted in Figure A.2 becomes essential, as it considers several gender compositions and their relationship on career choices.



((a)) Ex female schools ((b)) Ex male schools

Figure A.1: Changes in the Proportion of Students Choosing Economics and Business Related Majors in Schools Transitioning from Single-Sex to Coeducational

The odds ratio for female students selecting majors in economics and business-related fields exhibits variation in response to changes in the male composition fraction within the classroom (See Figure A.2). As the proportion of male students increases, the odds ratio demonstrates fluctuation, reflecting the nuanced influence on female students' decisions regarding these majors. Higher proportions of male students in the classroom correlate with elevated odds ratios, suggesting a potential inclination among female students toward these majors in more gender-diverse environments. Notably, the analysis reveals that the odds ratio for the entire sample (12,120 students, depicted by the red line) peaks when the composition of male students in a classroom reaches approximately 29.22%, with a substantial difference of 25 percentage points (0.64 - 0.39). This finding underscores the significant impact of classroom demographics on female students' preferences and underscores the importance of fostering gender diversity within educational settings.

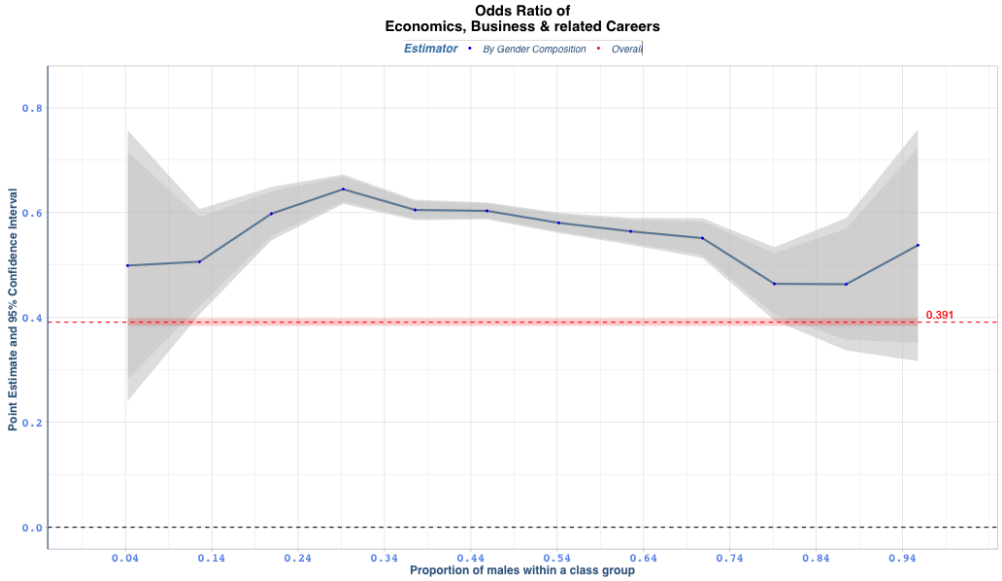
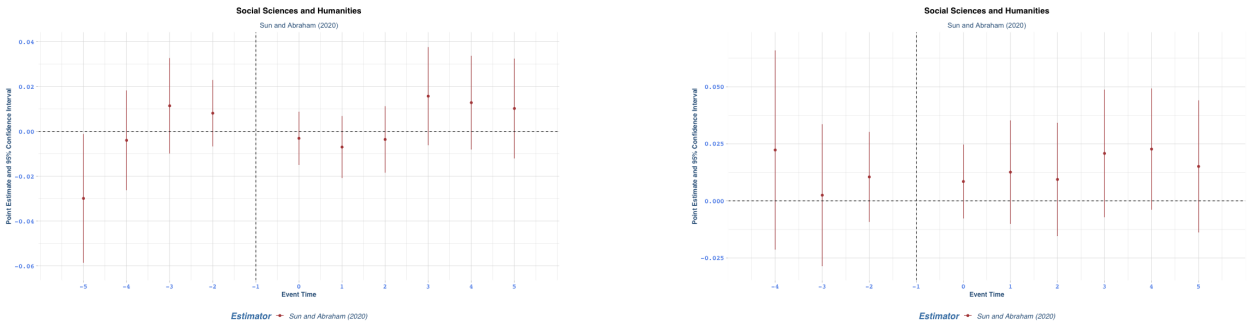


Figure A.2: Likelihood of a female student choosing economics/business-related majors

2. **Social Sciences/Humanities Majors**: The odds ratio for female students opt-

ing for majors in social sciences and humanities also varies with the male composition fraction in the classroom. Interestingly, unlike some other fields, higher proportions of male students are associated with higher odds ratios, indicating a potential preference among female students for these majors in more gender-diverse environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.3: Changes in the Proportion of Students Choosing Social Sciences in Schools Transitioning from Single-Sex to Coeducational

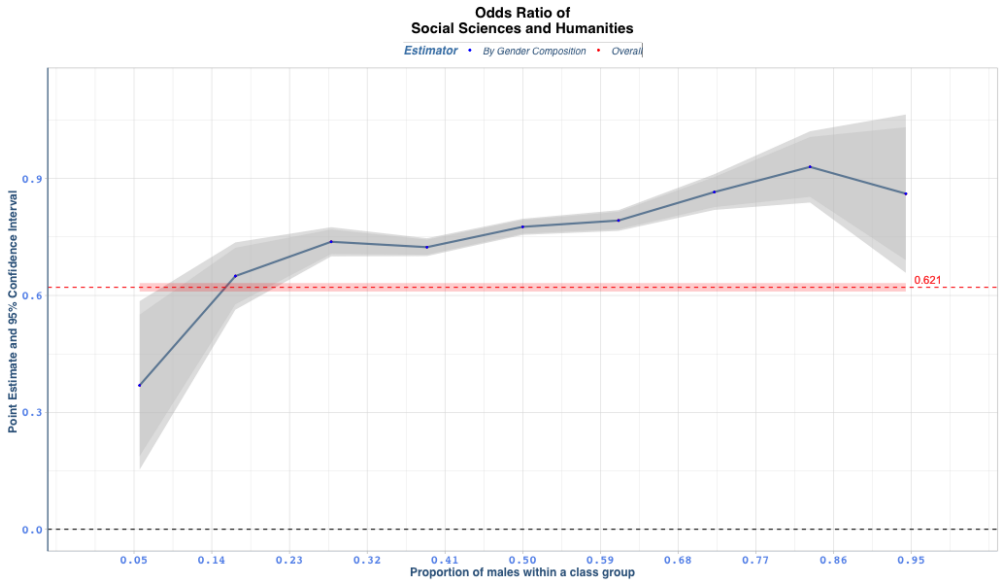
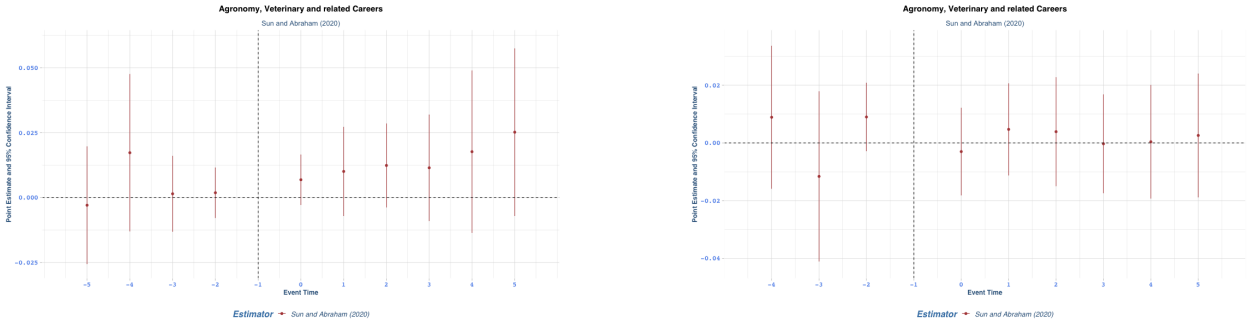


Figure A.4: Likelihood of a female student choosing social sciences/humanities majors

3. **Education Sciences Majors**: The odds ratio for female students choosing majors in education sciences demonstrates variability based on the male composition fraction in the classroom. Interestingly, higher proportions of male students are associated with higher odds ratios, indicating a potential preference among female students for these majors in more gender-diverse environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.5: Changes in the Proportion of Students Choosing Education Sciences Majors in Schools Transitioning from Single-Sex to Coeducational

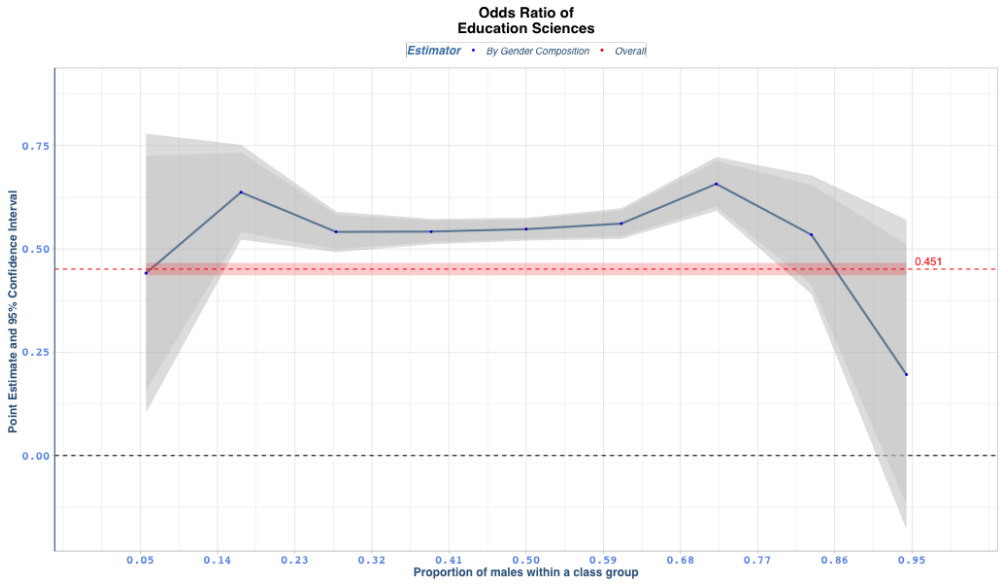
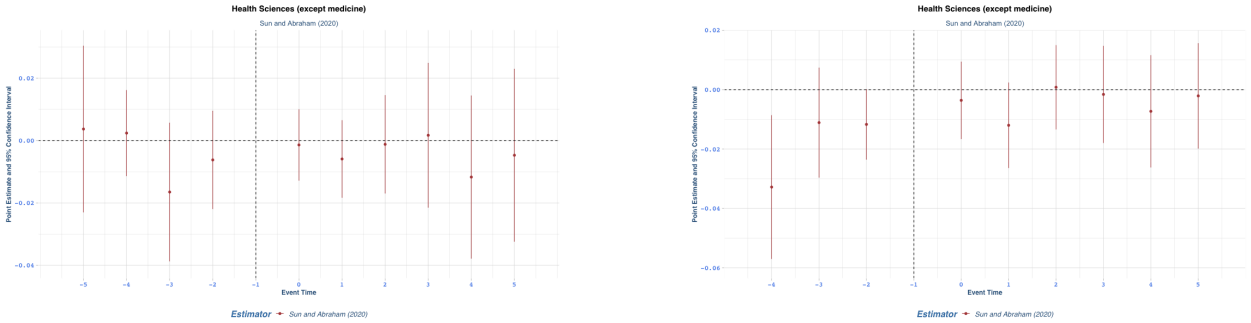


Figure A.6: Likelihood of a female student choosing education sciences majors

4. **Health Sciences Majors (Except Medicine)**: The odds ratio for female students opting for majors in health sciences exhibits variability with changes in the male composition fraction in the classroom. Similar to some other fields, higher proportions of male students are associated with higher odds ratios, suggesting a potential preference among female students for these majors in more gender-diverse environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.7: Changes in the Proportion of Students Choosing Health Science Majors (Except Medicine) in Schools Transitioning from Single-Sex to Coeducational

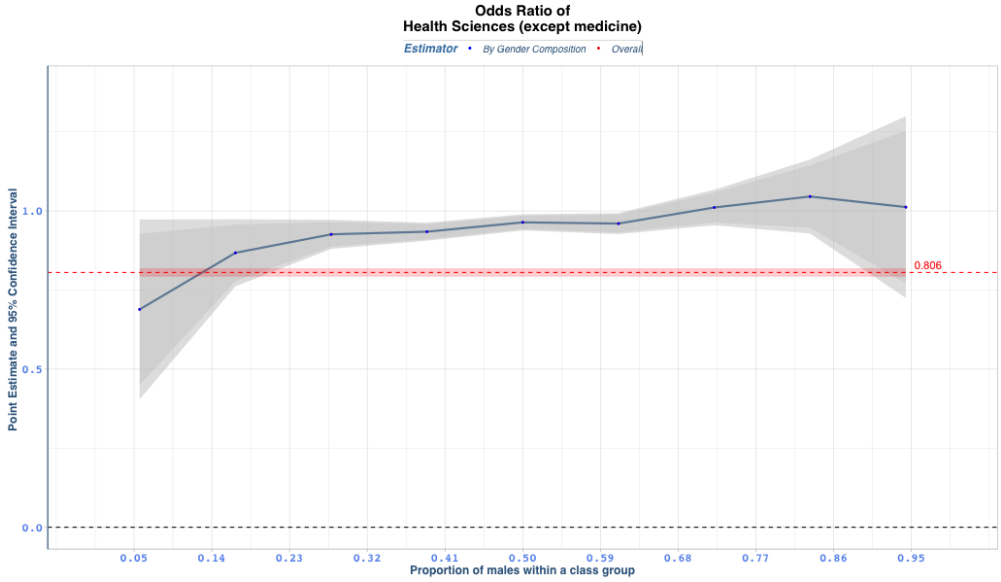
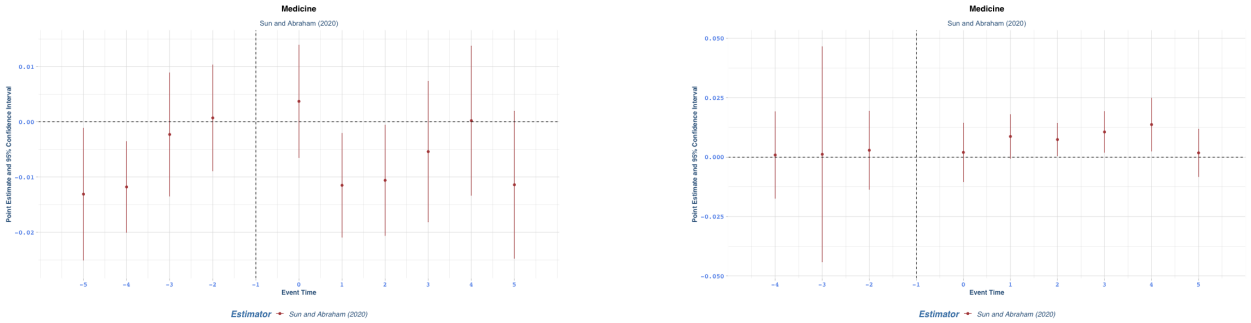


Figure A.8: Likelihood of a female student choosing health sciences majors

5. **Medicine Major**: The odds ratio for female students selecting majors in medicine fluctuates with changes in the male composition fraction in the classroom. Higher proportions of male students are associated with lower odds ratios, suggesting potential barriers or deterrents for female students in pursuing medicine majors in more male-dominated environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.9: Changes in the Proportion of Students Choosing A Major in Medicine from Schools Transitioning from Single-Sex to Coeducational

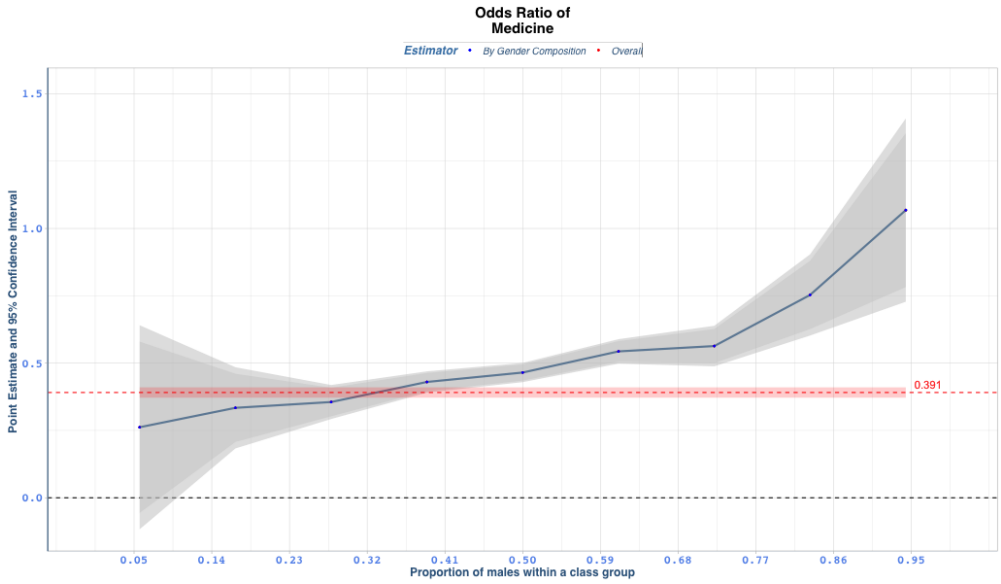
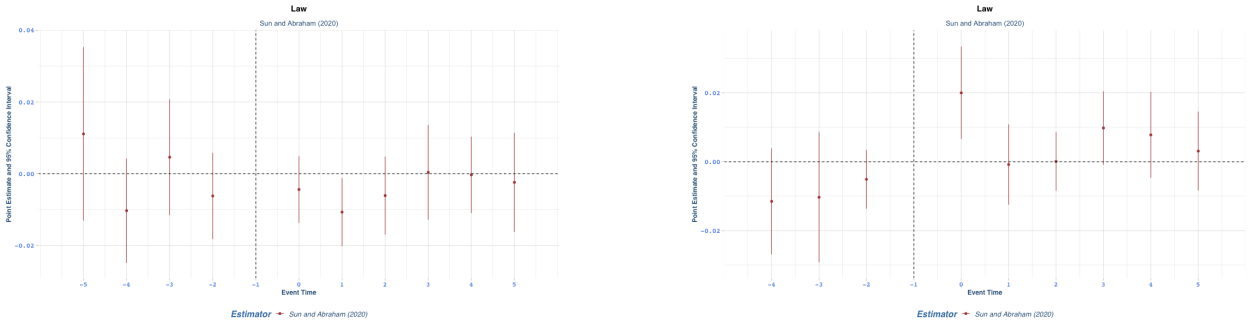


Figure A.10: Likelihood of a female student choosing medicine majors

6. **Law Majors**: The odds ratio for female students opting for majors in law demonstrates variability with changes in the male composition fraction in the classroom. Higher proportions of male students are associated with higher odds ratios, suggesting a potential preference among female students for these majors in more gender-diverse environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.11: Changes in the Proportion of Students Choosing a Major in LAW in Schools Transitioning from Single-Sex to Coeducational

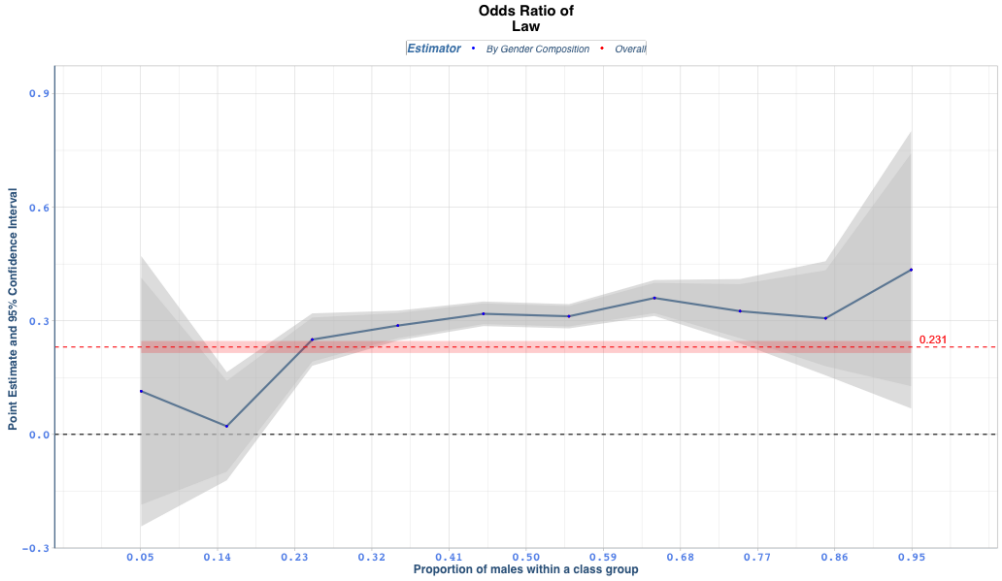
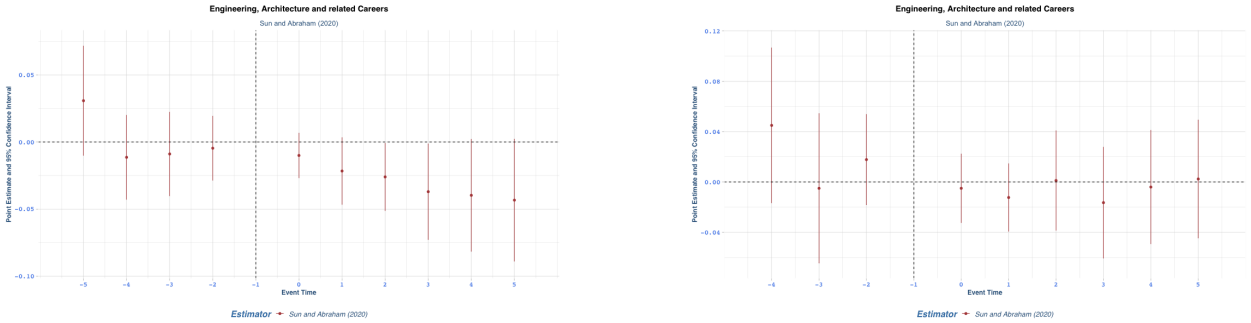


Figure A.12: Likelihood of a female student choosing law majors

A.1.2 Fields of Study Less Preferred by Female Students

This section examines the fields of study where female students display a lower preference compared to male students.

1. **Engineering/Architecture Related Majors**: The odds ratio for female students opting for majors in engineering and architecture also exhibits variability based on the male composition fraction in the classroom. Interestingly, as the proportion of male students increases, the odds ratio tends to decrease, suggesting a potential deterrent effect on female students' interest in these fields. This trend highlights the importance of considering classroom demographics in understanding gender disparities in engineering and architecture-related majors.



((a)) Ex female schools ((b)) Ex male schools

Figure A.13: Changes in the Proportion of Students Choosing Engineering, Architecture and Related Majors in Schools Transitioning from Single-Sex to Coeducational

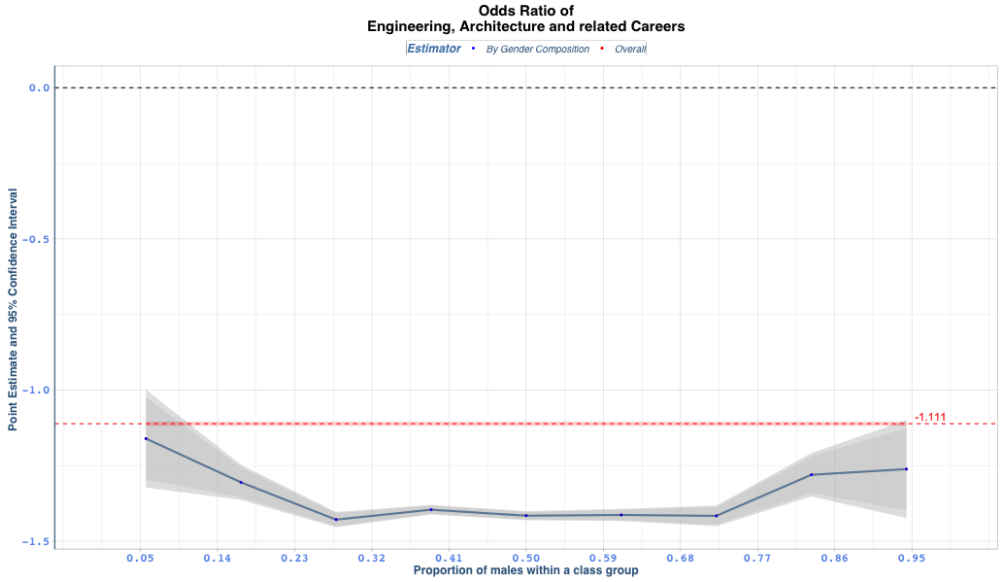
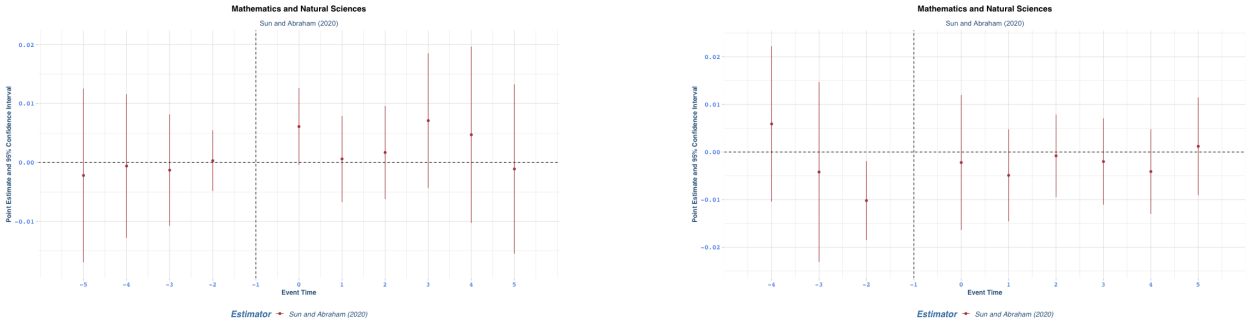


Figure A.14: Likelihood of a female student choosing engineering/architecture-related majors

2. **Mathematics and Natural Sciences Majors**: Similar to other fields, the odds ratio for female students selecting majors in mathematics and natural sciences fluctuates with changes in the male composition fraction. Higher proportions of male students in the classroom are associated with lower odds ratios, suggesting potential barriers or deterrents for female students in pursuing these majors in more male-dominated environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.15: Changes in the Proportion of Students Choosing Mathematics, Natural Sciences and Related Majors in Schools Transitioning from Single-Sex to Coeducational

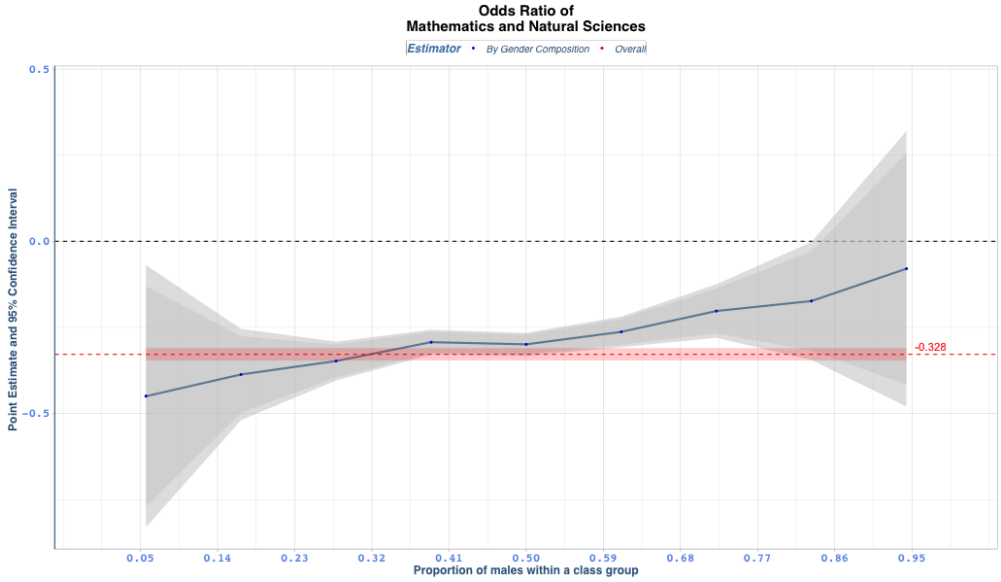
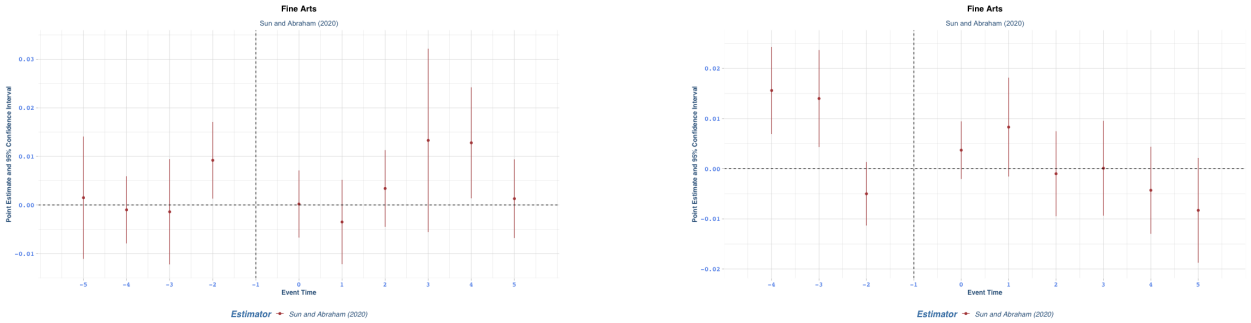


Figure A.16: Likelihood of a female student choosing mathematics/natural sciences majors

A.1.3 Fields of Study with Similar Preferences Between Female and Male Students

Here, we explore the fields of study where both female and male students exhibit similar preferences.

1. **Fine Arts Majors**: The odds ratio for female students choosing majors in fine arts demonstrates a less consistent pattern compared to other fields. While there is some fluctuation in the odds ratio with changes in the male composition fraction, the overall trend is less pronounced. Female students seem to exhibit varying preferences for fine arts majors regardless of the gender composition in the classroom.



((a)) Ex female schools ((b)) Ex male schools

Figure A.17: Changes in the Proportion of Students Choosing Fine Arts and Related Majors in Schools Transitioning from Single-Sex to Coeducational

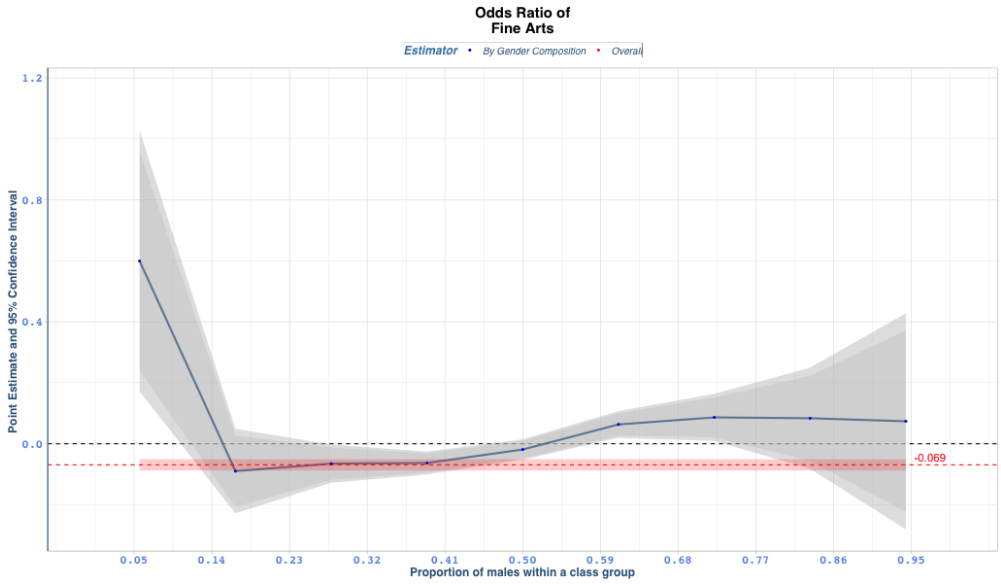
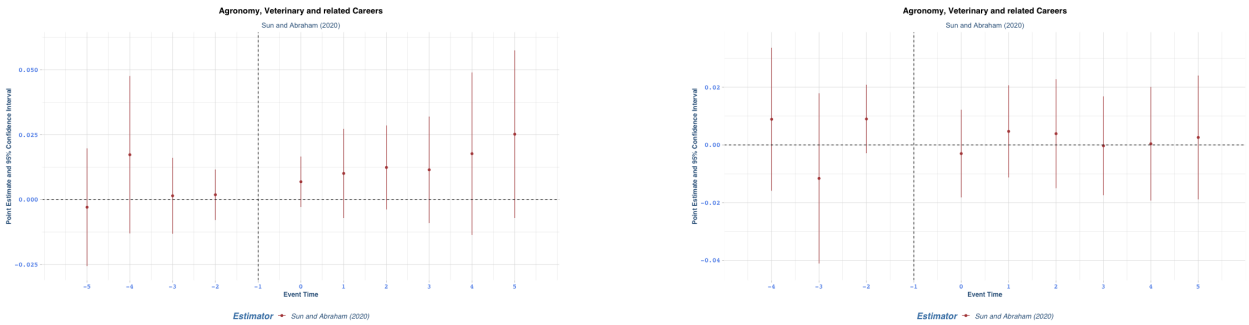


Figure A.18: Likelihood of a female student choosing fine arts majors

2. **Agronomy/Veterinary Related Majors**: The odds ratio for female students selecting majors in agronomy and veterinary-related fields shows fluctuations with changes in the male composition fraction in the classroom. Higher proportions of male students are associated with lower odds ratios, suggesting potential barriers or deterrents for female students in pursuing these majors in more male-dominated environments.



((a)) Ex female schools ((b)) Ex male schools

Figure A.19: Changes in the Proportion of Students Choosing Agronomy, Veterinary, and Related Majors in Schools Transitioning from Single-Sex to Coeducational

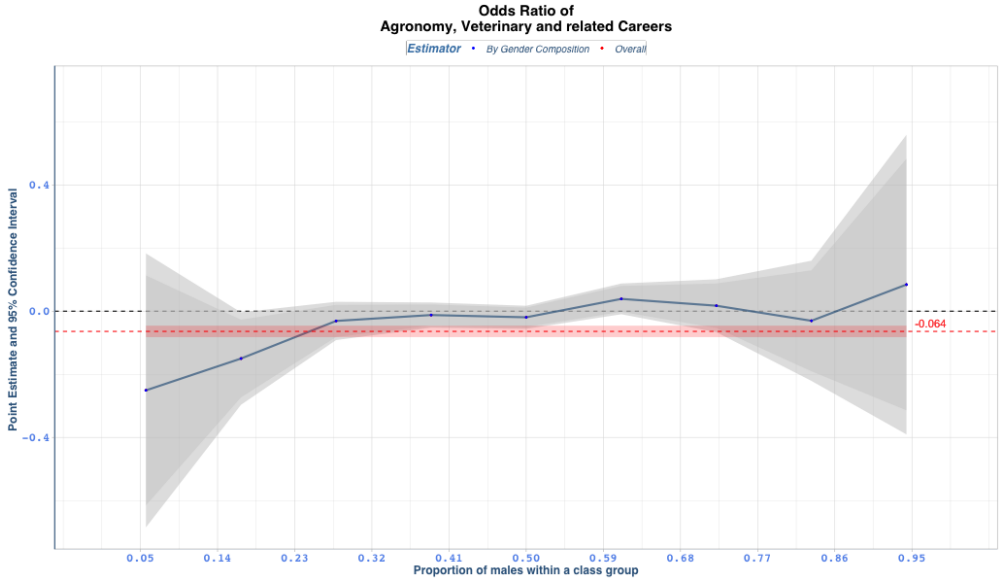


Figure A.20: Likelihood of a female student choosing agronomy/veterinary-related majors

A.2 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.3 Optimal bandwidth estimation based on Binary cross-entropy

In order to analyze the probability that a secondary school student chooses an field of study 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.

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 $\hat{\text{h}}$ 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.

A.4 Optimal Distance for Different Major Categories

In this section, we present the graphs illustrating the optimal distance for different major categories based on the Bayesian Cross Entropy (BCE) metric. Each graph corresponds to a specific academic major category.

Here are the texts for each graph, explaining the optimal distance and its significance:

1. Agronomy and Veterinary Related Majors: As depicted in Figure A.21, the optimal distance that minimizes the entropy in schooling decisions regarding the selection of Agronomy and Veterinary Related Majors is 0.1109.

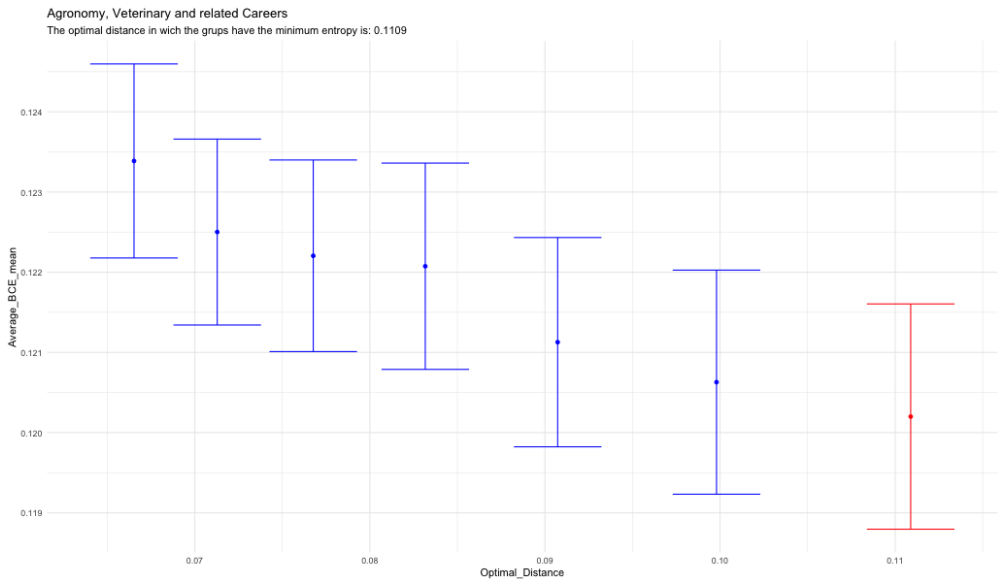


Figure A.21: Optimal Distance for Agronomy and Veterinary Related Majors

2. Economics and Business Related Majors: Figure A.22 illustrates that the optimal distance for minimizing entropy in schooling decisions related to Economics and Business majors is 0.0832.

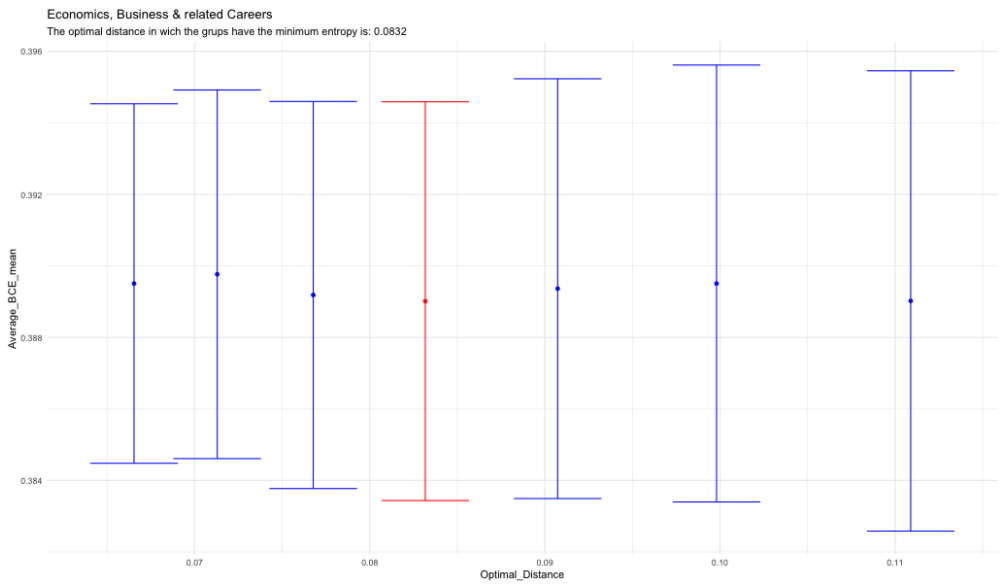


Figure A.22: Optimal Distance for Economics and Business Related Majors

3. Education Sciences Majors: Examining Figure A.23, we find that the optimal distance for Education Sciences majors, which minimizes entropy in schooling decisions, is 0.1109.

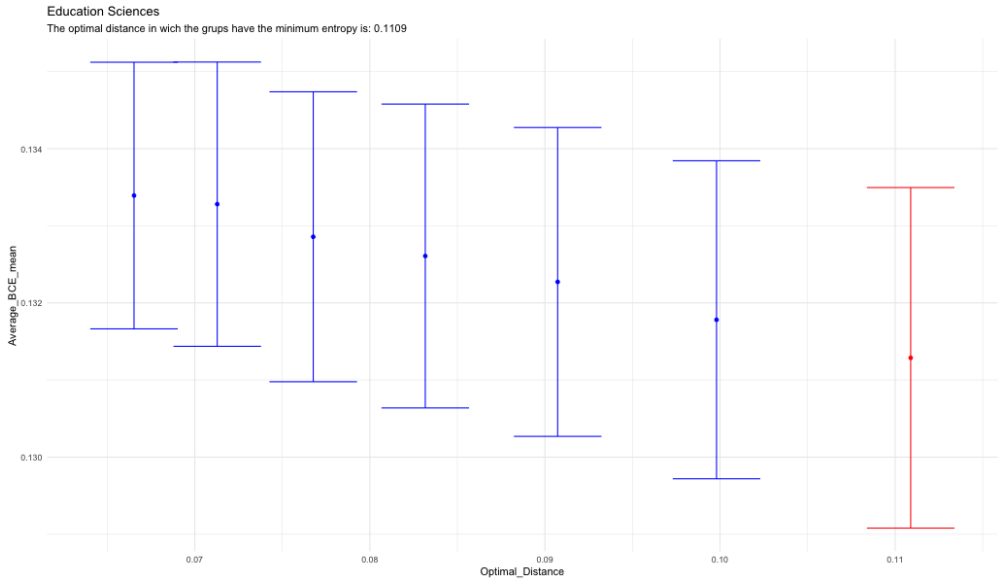


Figure A.23: Optimal Distance for Education Sciences Majors

4. Engineering and Architecture Related Majors: Figure A.24 presents the optimal distance of 0.1109 for minimizing entropy in schooling decisions concerning Engineering and Architecture Related Majors.

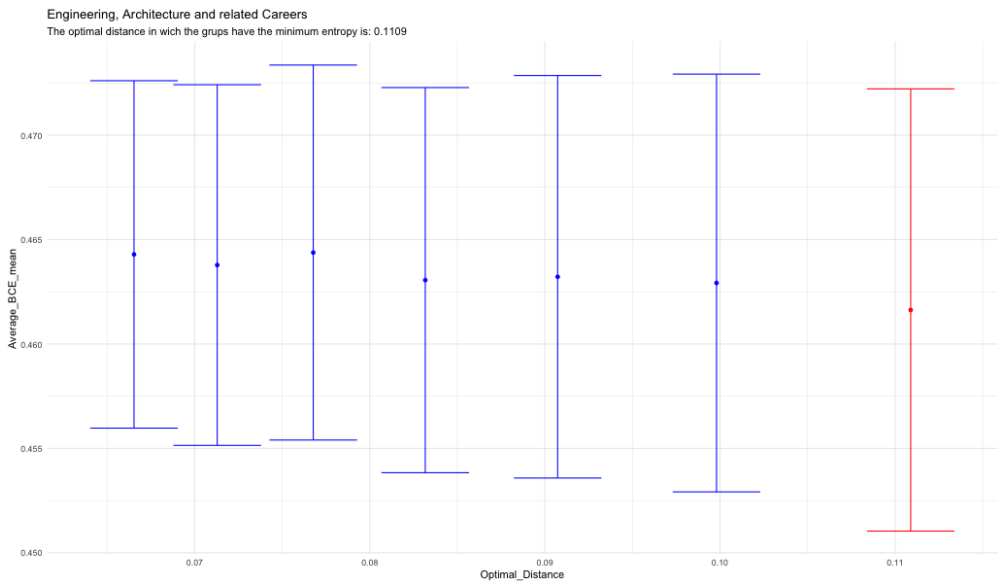


Figure A.24: Optimal Distance for Engineering and Architecture Related Majors

5. Fine Arts Majors: In Figure A.25, we observe that the optimal distance for minimizing entropy in schooling decisions pertaining to Fine Arts majors is 0.1109.

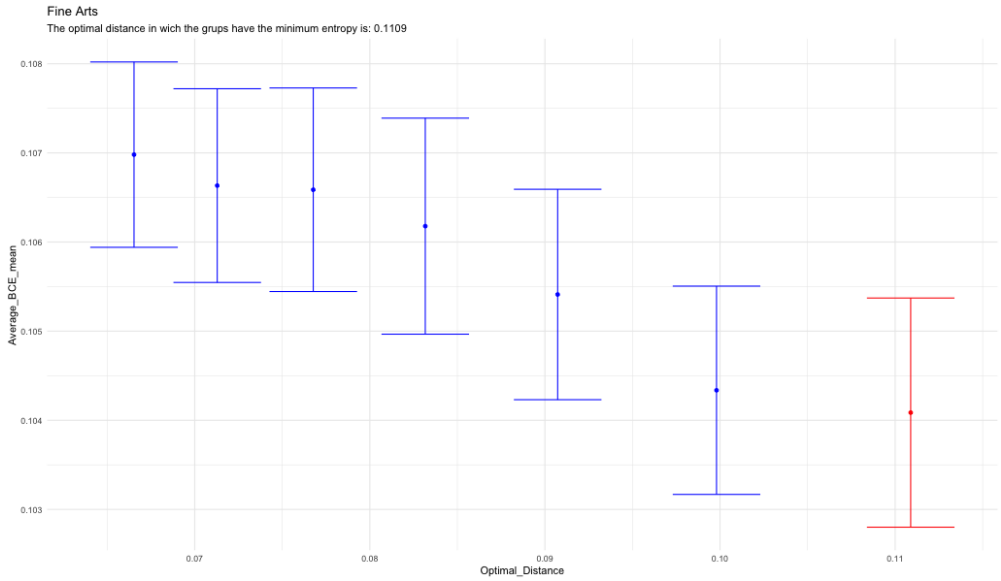


Figure A.25: Optimal Distance for Fine Arts Majors

6. Health Sciences Majors: The optimal distance of 0.0832, as shown in Figure A.26, minimizes entropy in schooling decisions regarding Health Sciences majors.

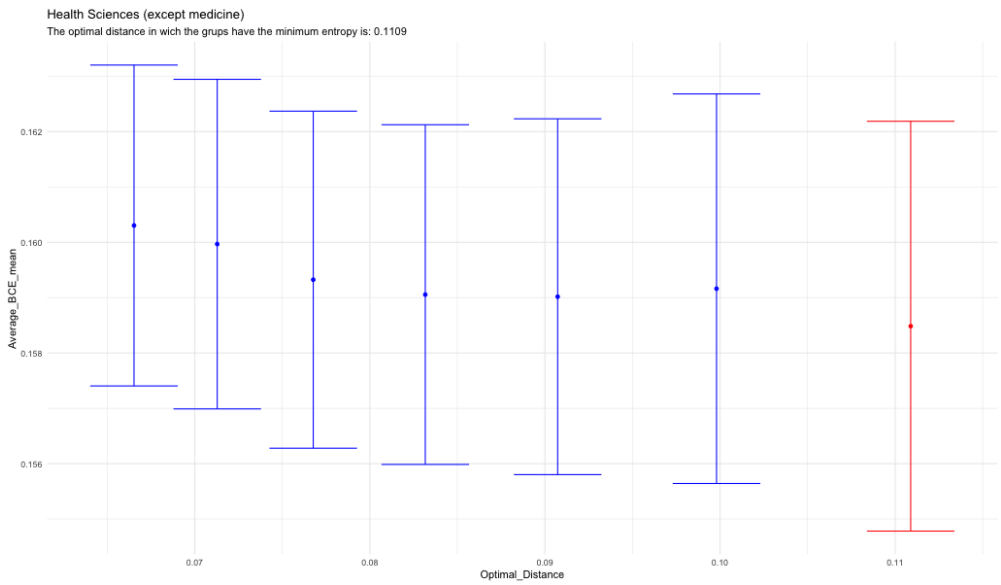


Figure A.26: Optimal Distance for Health Sciences Majors

7. Law Majors: Figure A.27 illustrates that the optimal distance for minimizing entropy in schooling decisions regarding Law majors is 0.0998.

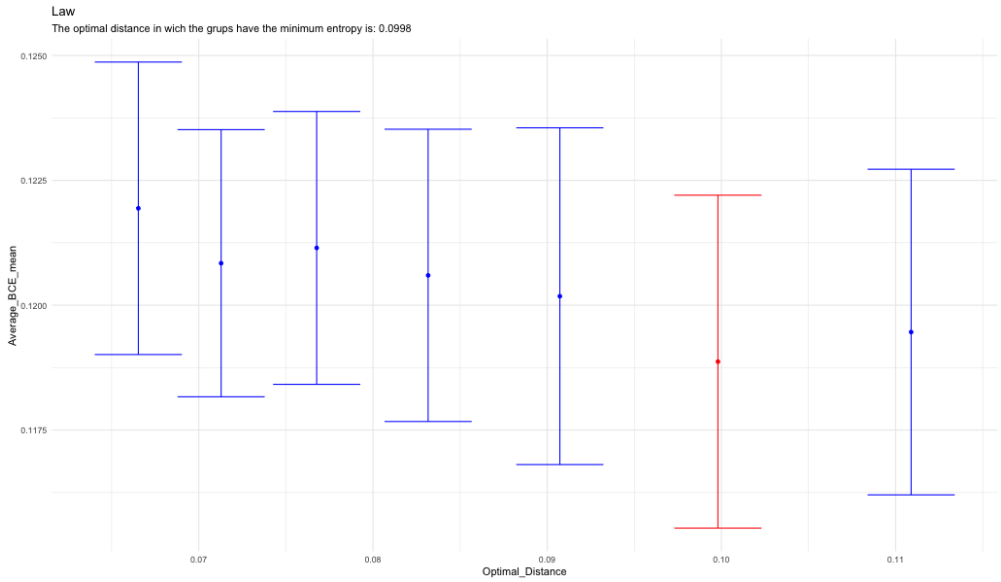


Figure A.27: Optimal Distance for Law Majors

8. Mathematics and Natural Sciences Majors: Examining Figure A.28, we find that the optimal distance for Mathematics and Natural Sciences majors, which minimizes entropy in schooling decisions, is 0.1109.

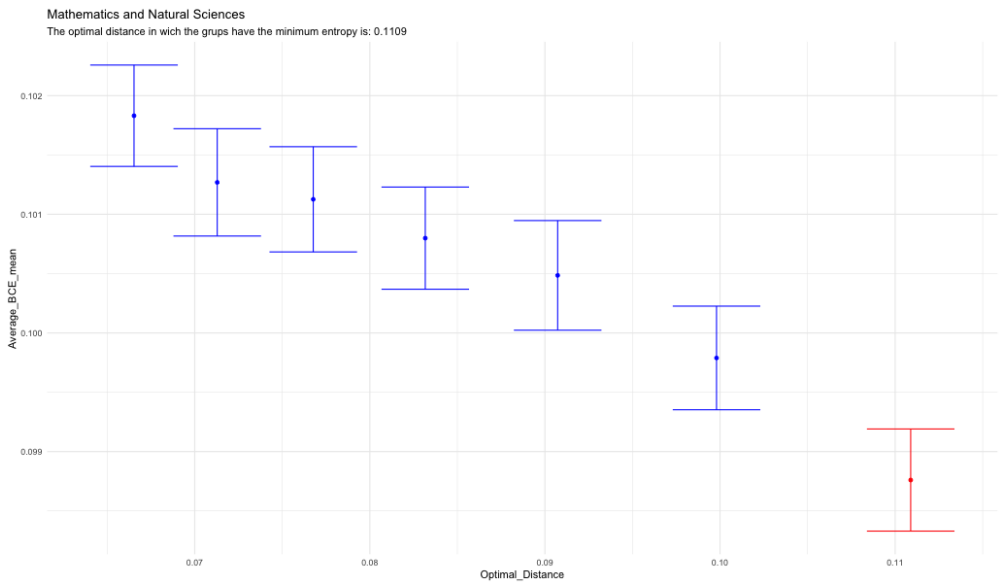


Figure A.28: Optimal Distance for Mathematics and Natural Sciences Majors

9. Medicine Majors: As depicted in Figure A.29, the optimal distance that minimizes the entropy in schooling decisions regarding the selection of Medicine majors is 0.1109.

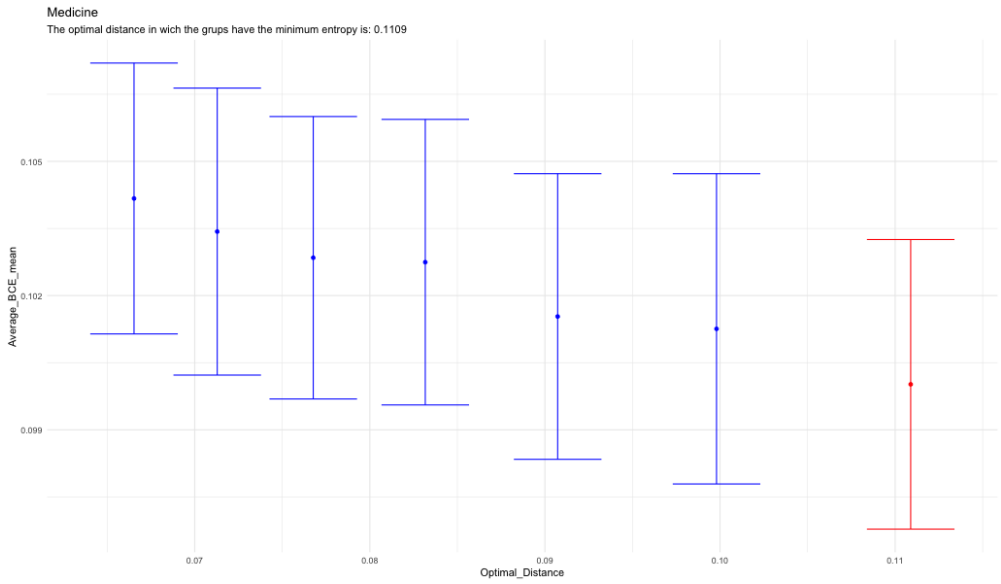


Figure A.29: Optimal Distance for Medicine Majors

10. No Studies: Figure A.30 presents the optimal distance of 0.0832 for minimizing entropy in schooling decisions concerning not pursuing further studies.

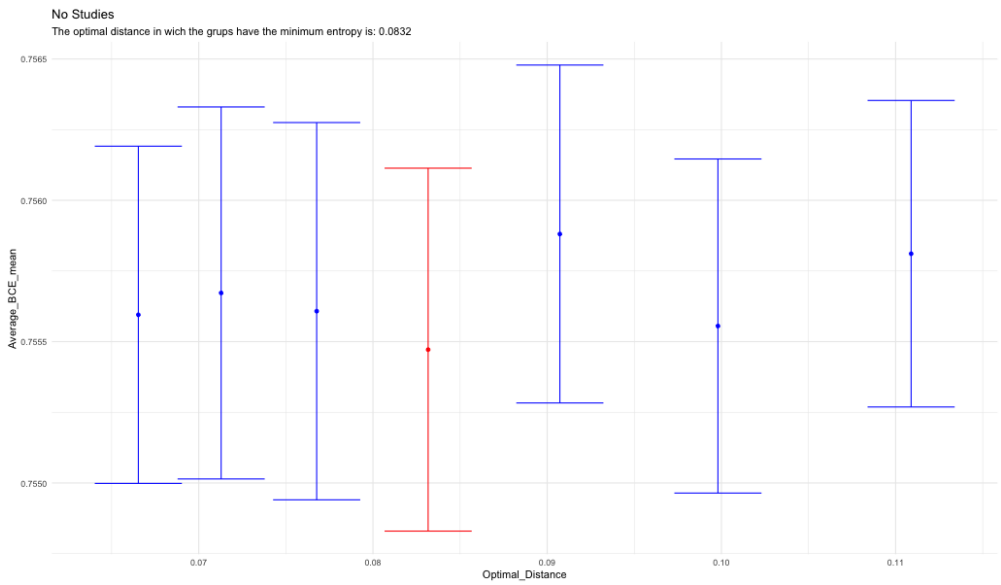


Figure A.30: Optimal Distance for Not Pursuing Further Studies

11. Social Sciences and Humanities Majors: The optimal distance of 0.1109, as shown in Figure A.31, minimizes entropy in schooling decisions regarding Social Sciences and Humanities majors.

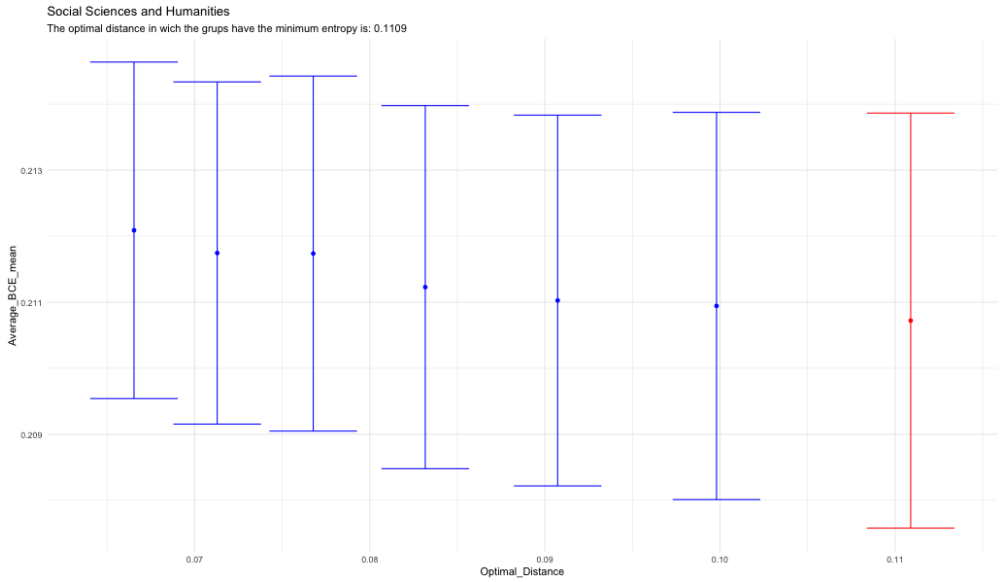


Figure A.31: Optimal Distance for Social Sciences and Humanities Majors