Gender Diversity Improves Performance but Reinforces Gendered Roles

Xiaoyue Shan*

Abstract

Does group gender diversity benefit individuals? I examine this question with a field experiment randomizing 3,060 students to small study groups at university entry. Assignment to mixed-gender rather than single-gender groups improves performance and well-being for both men and women: first-year grades increase by about 0.10 SD, well-being by 0.15 SD, and program dropout falls by 6 pp (24%). However, mixed-gender groups also induce more traditional attitudes toward family gender roles. Mechanism analyses suggest that gender diversity fosters collaboration and shifts gender attitudes by reinforcing gendered roles in social interaction: while women coordinate and ask questions, men compete and explain.

Keywords: gender diversity, performance, well-being, gender roles
JEL Codes: C93, D91, I21, I31, J16

*I thank Judd Kessler, Corinne Low, Michel Maréchal, Ulf Zölitz, Anne Brenoe, Pietro Biroli, Jan Feld, Raquel Fernández, Uri Gneezy, Lorenz Goette, Johannes Haushofer, Jessica Pan, Olga Stoddard, Colin Sullivan, Melanie Wasserman, Basit Zafar, and attendees at SITE Experimental Economics, the Workplace of the Future Conference, the BFI Women in Empirical Microeconomics Conference, SOLE, CESifo Area Conference on Behavioral Economics, NYUAD iSEE Workshop, URPP Conference on Boys and Men, BEAT Conference, UCSD, UCLA, UniMelb, UNSW, MonashU, PKU, HKU, FudanU, WZB, NUS, UZH, and UPenn for useful comments. A previously circulated working paper used a small subset of the data analyzed here to study a related question focused on women's minority status; this paper uses a much larger sample and supersedes the earlier version. The experiment conducted for this paper received IRB approval from the Human Subjects Committee of the Faculty of Economics, Business Administration, and Information Technology (OEC IRB) at the University of Zurich. This study is registered in the AEA RCT registry under ID AEARCTR-0004392; see Section 2 for full details about the preregistration. Shan: Department of Economics, National University of Singapore. Email: x.shan@nus.edu.sg.

1 Introduction

Group work is prevalent in educational and labor market settings (Springer et al. 1999; Kozlowski and Ilgen 2006; Wuchty et al. 2007). As women's representation has risen in higher education and workplaces (Goldin 2014; Olivetti and Petrongolo 2016; NCES 2020), mixed-gender groups are now commonplace. These groups allow members to exchange ideas and divide tasks, while also bringing men and women into close contact. Through these interactions, individuals express themselves, build relationships, and form beliefs about how men and women should behave (West and Zimmerman 1987; Butler 1988). Accordingly, mixed-gender groups may affect not only learning and production outcomes, but also the social construction of gender roles. Despite growing descriptive evidence linking gender diversity to team output and firm revenues (Herring 2009; Adams and Ferreira 2009; Hunt et al. 2023), causal evidence on how it affects individual performance and gender role attitudes remains thin—largely because real-world groups rarely provide exogenous variation in gender composition at scale.

Conceptually, group gender diversity may improve individual performance by combining a broader range of skills, which could boost creativity and facilitate knowledge sharing within groups (Hamilton et al. 2003; Hoogendoorn et al. 2013; Nielsen et al. 2017). However, cross-gender interaction may also be hindered by homophily and communication frictions (Cullen and Perez-Truglia 2023; Lazear 1999; Hjort 2014). In addition, the social roles that members assume can affect patterns of collaboration within groups. A large literature documents gender differences in social behaviors, such as agentic versus communal traits, competition, leadership, and self-promotion (Eagly 2013; Niederle and Vesterlund 2007; Eckel et al. 2021; Exley and Kessler 2022). If men and women gravitate toward gender-congruent roles—potentially due to image concerns or marriage market incentives (Bursztyn et al. 2017)—this role sorting may enhance collaboration in mixed-gender groups, while also reinforcing traditional gender norms.

This paper uses a large field experiment to examine the causal effects of group gender diversity on individual performance, well-being, and gender role attitudes. The experiment involves 3,060 university students across seven cohorts (from 2018/19 to 2024/25) in a European business school. At the start of the first semester, these students were randomly assigned to 765 four-person study groups in an introductory economics course—the business school's largest compulsory course.[1] The study groups were created to support learning and adjustment to university life. To strengthen peer-to-peer connections and facilitate collaboration, a few design features were introduced: (1) the group size was chosen to be small to mitigate feelings of anonymity or "social loafing," which can reduce group cohesion (Latané et al. 1979; Liden et al. 2004; Wheelan 2009); (2) groups were provided with free vouchers for an icebreaker activity; and (3) one student from each group was randomly selected as the coordinator to help organize activities on a voluntary basis.

Given the random assignment, study groups vary exogenously in their level of gender diversity. This paper focuses on the effects of mixed-gender versus single-gender groups on student outcomes,

[1] The student population is male-dominated overall, with men making up about 60% of students.

conditioning on students' own gender. The variation in gender composition allows me to further break down the effects of mixed-gender groups into three scenarios: (1) being the majority gender, referring to women in female-majority groups and men in male-majority groups; (2) being the minority gender, referring to women in male-majority groups and men in female-majority groups; and (3) being in a gender-balanced group.[2]

The results show that students assigned to mixed-gender groups achieve 0.14 SD higher grades in the economics course than students assigned to homogeneous groups ($p = 0.003$). Breaking down this overall impact, I find that being in the majority, in the minority, and in a balanced group all directionally improve performance; their effects are not significantly different from one another. Splitting the sample by gender, I find that both men and women perform significantly better when assigned to a mixed-gender rather than homogeneous group. Moreover, the estimated effect of mixed-gender groups does not significantly vary across a wide range of student attributes, such as high school grades, personality traits, baseline gender attitudes, and parents' socioeconomic background. The effect also appears consistent across the seven experimental cohorts.

Greater gender diversity within groups also raises performance in other first-semester courses. Students assigned to mixed-gender groups obtain 0.10 SD higher grades than those assigned to homogeneous groups ($p < 0.001$). These gains are driven by courses commonly taken by all group members, suggesting that students in mixed-gender groups exchange information or collaborate on learning across multiple courses. Furthermore, the performance effects of gender diversity persist in the longer run. Initial assignment to mixed-gender groups increases course grades in the second semester by approximately 0.06 SD. By the start of the second academic year, students assigned to mixed-gender groups are 6 percentage points (or 24%) more likely to remain in their original study program ($p = 0.022$). Taken together, these findings show that gender-diverse peer groups formed at university entry lead to lasting improvements in academic performance and higher program retention.[3]

Do the benefits of gender diversity extend beyond educational attainment? To assess broader welfare impacts, I measure different aspects of subjective well-being. At the end of the first semester, students assigned to mixed-gender groups report significantly higher satisfaction with their group, more optimistic expectations for future academic achievement, and better mental health. The average effect of mixed-gender groups across well-being outcomes is equivalent to 0.15 SD ($p < 0.001$). These improvements also translate into better mental health in the longer run. Students initially assigned to mixed-gender groups are 7.9 percentage points less likely to report moderate or

[2]For completeness, I also use alternative specifications to estimate the effects of a continuous increase in gender diversity: from single-gender groups to gender-imbalanced groups (scenarios 1 and 2) and to gender-balanced groups (scenario 3). To measure the three-level variation in diversity, I calculate the within-group standard deviation of gender. Across outcomes in the paper, the estimated effects of the binary mixed-gender indicator and the continuous gender diversity measure are closely aligned. The results are also very similar when using the inverse Herfindahl–Hirschman Index (HHI) to measure continuous variation in gender diversity.

[3]The improved performance in second-semester courses is unlikely to be driven by selective course-taking, as the effect is concentrated in the compulsory courses. The effects of gender diversity on other course grades and program dropout also apply similarly to both men and women.

severe psychological distress at the end of the second semester ($p = 0.019$).

Turning to gender attitudes, intergroup contact theory predicts that interaction with opposite-gender peers can diminish stereotypes and prejudice (Allport 1954; Pettigrew and Tropp 2006). However, I find that assignment to mixed-gender groups leads students to adopt more traditional attitudes toward family gender roles. At the end of the first semester, students in mixed-gender groups are more likely to agree that it is problematic for a woman to earn more than her partner and that mothers' full-time employment may harm children—controlling for students' views at baseline. These changes in attitudes appear for both men and women in the short run and persist for men about eight months later.

Next, I investigate why group gender diversity has mixed effects on performance, well-being, and gender attitudes. I focus on gender-typed social roles as a crucial mechanism through which gender diversity may enhance within-group collaboration while also inducing traditional gender attitudes. I first show that men and women assume different roles in group interaction, and these differences are more pronounced in mixed-gender groups. I then present suggestive evidence that women are more effective than men in the group coordinator role. I further show that, consistent with gendered role sorting, mixed-gender groups exhibit more active and inclusive social dynamics.

In social role theory, agentic traits—emphasizing assertiveness, independence, and the pursuit of status—are more associated with men, whereas communal traits—emphasizing cooperation, connection, and concern for others—are more associated with women (Bakan 1966; Abele and Wojciszke 2007; Eagly 2013). In line with this framework, I find that women are more likely to report asking questions as their primary role in group interaction, while men are more likely to report explaining course material to others. Women also report a lower tendency to compete than men, as widely documented (Gneezy et al. 2003; Niederle and Vesterlund 2007). Importantly, these gender differences are not visible in homogeneous groups but are particularly salient in mixed-gender groups. This suggests that gender-diverse environments induce men and women to sort into gender-congruent roles, which may reduce frictions and foster collaboration within groups.

I also employ the group coordinator assignment to compare the effects of male versus female coordinators—a role closely aligned with communal traits. While mixed-gender groups with male coordinators perform better than homogeneous groups, the benefit of mixed-gender groups with female coordinators is much larger. Overall, female-coordinated mixed groups perform about 0.11 SD higher than male-coordinated mixed groups ($p = 0.014$). This suggests that women are more effective in the coordinator role, contributing substantially to the performance gains of mixed-gender groups. The finding echoes recent evidence that women are more likely to take on non-promotable tasks or service duties in the workplace (Babcock et al. 2017; Guarino and Borden 2017).

I further examine how gender diversity shapes interaction patterns within study groups. Using rich data on social interaction, I find that students in mixed-gender groups engage in significantly more frequent group-level and pair-level activities than students in homogeneous groups. Moreover, the social dynamics in mixed groups are more inclusive: the frequency of interaction is similar

for both same-gender and opposite-gender pairs, and lower-performing members are less socially isolated than in homogeneous groups.[4] These inclusive dynamics may facilitate knowledge sharing and peer mentoring, which especially benefit academically struggling members. In addition, when asked to report their closest study partners across the whole class, students in mixed-gender groups are much more likely to name assigned group members and opposite-gender peers, underscoring the importance of these groups in shaping broader social networks and weakening gender homophily.

Finally, I show that the effects of gender diversity are not driven by cognitive or noncognitive skill diversity within groups. Instead, gender diversity has distinct effects on performance: it leaves top-performing students in a group unaffected but substantially raises the performance of those at the bottom, thereby narrowing within-group performance gaps. This pattern is consistent with the more inclusive dynamics in mixed-gender groups and also appears in small-scale replications of the experiment in female-dominated environments.[5]

This paper highlights the powerful yet mixed effects of group gender diversity. The findings point to a tension in how educators and organizations design peer groups and structure onboarding. Forming gender-diverse groups can improve social dynamics and yield gains in performance and well-being, but the reinforced traditional attitudes may spill over into gender gaps in future career choices and family decisions. One practical implication is not to avoid gender-diverse groups, but to reduce the extent to which effective collaboration hinges on gender-divided roles. For example, institutions can assign and rotate group responsibilities, explicitly recognize and reward service-oriented contributions, or provide training to strengthen teamwork skills.

This paper contributes to three strands of literature. The first is a growing body of work experimentally analyzing how the gender composition of small groups shapes leadership behavior, voice and influence, credit allocation, group dynamics, and output (Amini et al. 2017; Isaksson 2018; Chen and Houser 2019; Born et al. 2020; Sarsons et al. 2021; Coffman et al. 2021; Fenoll and Zaccagni 2022; Coveney et al. 2025). Most studies in this literature examine short-lived teams formed for laboratory tasks or one-off assignments, which are less suited to investigating real-life outcomes, especially over the longer run. A closely related study by Karpowitz et al. (2024) also assigns university students to study groups, but focuses on how female-majority and male-majority groups differ in individual influence during team-building tasks. My study extends the literature by showing that greater gender diversity in real-world groups can have persistent effects on performance, well-being, and gender attitudes.

This paper also builds on a large literature examining how the gender composition of classrooms affects academic performance. In addition to earlier studies using natural variation to estimate

[4]Students report the frequency of learning or extracurricular activities they had with each assigned peer. This allows me to capture detailed interactions between each pair of students within the group.

[5]I replicated the experiment in two mathematics courses primarily for biomedicine, biochemistry, or biology students at the same university. These courses are female-dominated overall, with about 65% women. A total of 324 students were assigned to 81 groups during the fall of 2024/25. I find similar effects of gender diversity: the final grade directionally increases, the lowest grade obtained by group members rises substantially, and within-group performance gaps narrow noticeably; the latter two effects are marginally significant.

gender peer effects (e.g., Hoxby 2000; Lavy and Schlosser 2011; Hill 2017), recent work also employs random assignment to address the reflection problem in identification (Oosterbeek and Van Ewijk 2014; Booth et al. 2018; Eisenkopf et al. 2015; Gong et al. 2019). My study complements prior work by providing new evidence that gender diversity in small peer groups improves academic performance and reduces program dropout. I also present suggestive evidence on gendered social roles as a novel mechanism through which gender diversity fosters collaborative learning and raises performance.[6]

Lastly, this paper contributes to an expanding literature examining how intergroup contact affects attitudes, trust, perspective taking, and cooperation (Boisjoly et al. 2006; Rao 2019; Finseraas et al. 2019; Mousa 2020; Lowe 2021; Alan et al. 2021; Anderberg et al. 2024). Two recent studies focus on cross-gender contact and are especially related to this paper. Dahl et al. (2021) show that exposure to female peers in boot camps leads men to adopt less traditional gender attitudes in the short run, although the effect fades after about six months. Greenberg et al. (2024) find that integrating women into all-male combat units does not reduce men's performance or behaviors but negatively affects their perceived workplace quality. Compared to these studies based on highly gender-skewed workplaces, this paper shows that in a relatively gender-balanced environment, collaborating with women can shift men toward more traditional gender role attitudes that persist in the longer run.

2 The Experiment

2.1 Institutional background

The setting for the experiment is the business school of a large European university. Each entering cohort at the business school has more than 1,000 students majoring or minoring in economics, business, finance, or information technology. These students are required to complete a series of courses in the first academic year to remain in their study program. Among these courses, introductory economics is the largest course that is compulsory for nearly all first-semester students.

The student body in the economics course is male-dominated, with approximately 60% men.[7] The course has two lectures per week, complemented by optional tutorial sessions. Several problem sets are assigned throughout the semester, but they are not graded. The final grade for the course depends solely on students' performance in a high-stakes exam at the end of the semester. The

[6]This paper differs from prior studies using random assignment in two key aspects—as summarized in Appendix Table A.1. First, compared to earlier work examining larger teaching groups with a median size of around 18 students, this paper analyzes small groups. Second, prior work mainly employs variations within mixed-gender groups to estimate linear effects of increasing the male or female share, while this paper focuses on comparing mixed-gender groups to homogeneous ones. More broadly, this paper relates to studies examining the effects of other peer characteristics such as ability, race, personality, and socioeconomic status on educational or labor market outcomes (e.g., Booij et al. 2017; Bursztyn et al. 2019; Brenøe and Zölitz 2020; Linos et al. 2024; Carrell et al. 2018; Golsteyn et al. 2021; Cools et al. 2022; Zölitz and Feld 2021; Hampole et al. 2021). My results show that the effects of group gender diversity appear independent of and distinct from group diversity in other dimensions.

[7]The underrepresentation of women in the economics student body is widely observed across undergraduate economics programs in the U.S. and Europe (Lundberg 2018; Auriol et al. 2022).

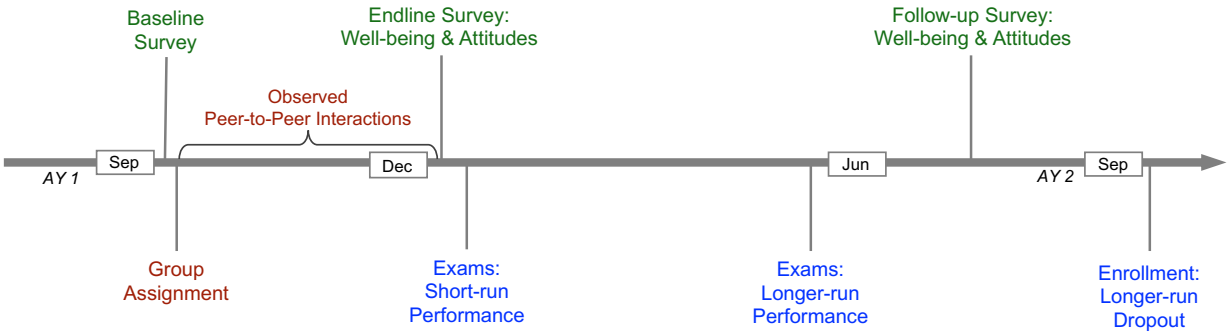
exam comprises mostly multiple-choice questions, making the performance measure fairly objective.

During the first semester, most students take five additional courses at the business school. These courses cover subjects such as business administration, mathematics, finance, and accounting. During the second semester, students also take about five compulsory courses on similar subjects. The grading scale for most courses in the university ranges from 1 to 6, with 1 being the lowest, 6 the highest, and increments of 0.25. The final grades appear on students' transcripts and directly affect their Grade Point Average (GPA). Students need to obtain a grade of at least 4 to pass a course. Those who fail a course may retake it in the following year, but they are allowed only one retake attempt. Given the high stakes of these outcomes, I use course grades and course passing as two measures of performance.

2.2 Experimental design

In the fall semester of 2018/19, study groups were first introduced in the economics course to help new students form academic or social ties, exchange information, and collaborate on learning. Given positive feedback from students, the study groups were also offered in subsequent years. Capitalizing on this opportunity, I conducted the experiment annually from 2018/19 to 2024/25, following roughly the same design.[8] Across the seven cohorts, the economics course was taught by the same professor.

Figure 1. A Timeline of the Experiment



Notes: The figure shows the timeline for each experimental cohort from 2018/19 to 2024/25. The first academic year ("AY 1") consists of two semesters, fall and spring. The lecture period starts in mid-September and ends before Christmas for the fall semester, and spans from mid-February to late May for the spring semester. The figure illustrates the timing of three surveys, the group assignment, measured group interactions, short- and longer-run academic performance, and program dropout by the start of the second academic year ("AY 2").

Baseline survey: Figure 1 shows a timeline of the experiment for each cohort. Before the first lecture of the semester, all students enrolled in the economics course were invited to fill out a "baseline survey." This survey asked questions related to demographics, high school grades, family background, personality traits, and gender attitudes. Table 1 shows summary statistics of baseline

[8]The experiment received IRB approval from the Human Subjects Committee of the Faculty of Economics, Business Administration, and Information Technology (OEC IRB) at the University of Zurich.

characteristics. The overall response rate for the survey was about 82%. I observe student gender (either male or female) from course registration lists, so the gender information is available for all students.

Experiment participation: Students were invited to participate in the experiment during the first lecture of the economics course. They were explicitly informed about the following features of the experiment: (1) participation is completely voluntary and does not influence course grades, and (2) groups are randomly assigned, taking into account students' study programs. A total of 3,060 students from seven cohorts opted to participate and were randomly assigned to 765 four-person groups. The participation rate is approximately 30% to 45%, depending on the cohort. Figure A.1 shows that men and students repeating the course are less likely to participate. Participation is higher among students with greater openness, extraversion, and agreeableness, as well as those with more highly educated parents, but is lower among students with higher conscientiousness and higher parental income. By contrast, the participation decision does not significantly differ by age, high school math or language grades, competitiveness, neuroticism, or speaking German as the first language.[9]

As in many experiments, voluntary participation is necessary due to ethical and administrative constraints. It may also be more appropriate for the research design in this setting: students not interested in interacting with new peers are unlikely to respond to the gender composition of their peers; forcing these students into a group could undermine social cohesion and harm collaboration. However, endogenous participation may limit the external validity of the findings (List 2020). To address this concern, I first examine whether the effects of gender diversity vary across a large set of student characteristics and find little evidence of heterogeneity (see Section 4.1). Next, I apply the method of Andrews and Oster (2019) to assess the external validity of the estimated effects for the broader student population. The results suggest minimal selection bias and tight bounds for adjusted effects (see Section 6.2.1). Additionally, I replicate the experiment in female-dominated environments and provide suggestive evidence that the findings are generalizable beyond the male-dominated setting (see Section 6.2.2).

Group randomization: The group assignment was stratified by three broad categories of study programs: (1) business, economics, and finance majors; (2) informatics majors; and (3) other majors. Because the curricula are similar within each broad category of programs, the stratified randomization facilitates the coordination of group activities and encourages students to study together across multiple shared courses.[10] Starting with the second cohort (2019/20), one student from each group was randomly selected as the group coordinator to help organize activities on a

[9]The official language of instruction for the course is German. Consistent with this, speaking German as the first language is associated with significantly higher performance in the course.

[10]Due to uneven numbers of participating students across categories of study programs, about 20 out of the 765 study groups were composed of students from multiple categories. All results remain essentially the same (and even slightly stronger) when excluding these groups. There are also a few groups composed of three or fewer students—these groups are not included in the analysis.

voluntary basis. This further creates variation in the coordinator gender for mixed-gender groups.

After group assignment, students received emails providing information about their group, suggestions for group activities, and reminders about working together for problem sets. Figure A.2 shows a sample email, which announces the assigned group members and coordinator and highlights two basic principles of the study groups: voluntary participation and respectful behavior. Students were also encouraged to create a WhatsApp group for easier communication and coordination.[11] To foster peer-to-peer interaction, each group was provided with vouchers worth about $20 for four drinks at a university bar—this serves as an icebreaker event for group members to get to know each other. Overall, 57% of groups collected the vouchers.

Endline & follow-up surveys: At the end of the first semester, before the final exams, all students in the course were invited to complete an "endline survey." The aim of this survey was to measure non-academic outcomes and to explore potential mechanisms underlying the effects of group gender diversity. Specifically, the survey asked students to report their interactions with group members, expectations about future achievement, general competitiveness, gender attitudes, and mental health—see Section 5 for details on the survey instruments. The overall response rate was about 60%. Importantly, the group gender composition does not affect the response rate among men, women, or the full sample (see Table A.2, Panel A).

I also conducted a brief "follow-up survey" at the end of the first academic year, during the summer break. This survey was mainly used to examine longer-run effects on gender attitudes and mental health. Likely due to the suboptimal timing of the survey, the overall response rate was only around 35%. Nevertheless, I find no evidence that group gender composition affects the likelihood of men or women responding to the survey (see Table A.2, Panel B).

To account for the relatively high attrition, I show that the estimated effects on non-academic outcomes in the short and longer run are robust to inverse probability weighting (IPW) and to sensitivity tests under different assumptions about missing data (Kling and Liebman 2004). The Lee (2009) bounds on the estimated effects are also relatively tight. Notably, all academic outcomes are measured using administrative data and are observed for all students.

Preregistration: This study was registered on the AEA RCT Registry under ID AEARCTR-0004392.[12] This research paper largely follows the preregistration regarding the experimental design, the treatment, and outcomes. As registered, the experiment randomly assigns students to groups of four, the key treatment is the variation in group gender composition, and this paper analyzes all registered primary and secondary outcomes—except when data limitations prevent analysis.[13]

[11] For the first experimental cohort, I also created a WhatsApp group for each study group. Figure A.3 shows screenshots of conversations that students had in these groups, providing evidence that some groups indeed met for social events or discussed course material together.

[12] The registration was modified once in 2022 to make the preanalysis plan publicly available.

[13] This paper does not analyze major choice or degree completion due to data limitations. These outcomes are not observed for students enrolled in a major program outside of the business school. Observing these outcomes also takes years, especially for the recent experimental cohorts. I thus leave this to future work.

To be fully transparent, this paper deviates from the preregistration in three respects. First, the preregistration and preanalysis plan followed the literature and hypothesized a linear relationship between student performance and the share of women (or men). However, the data reveal a clear nonlinear relationship (see Section 4.1). To accommodate this departure, I follow Banerjee et al. (2020), who emphasize that the final research article should be "written and judged as a distinct object from the results of the preanalysis plan" to allow for knowledge creation. Thus, instead of adhering to the original hypothesis, this paper employs more appropriate models to analyze the effects of gender composition. Given that multiple models are tested, I present the effects of both the female share and the mixed-gender indicator on primary academic outcomes and apply multiple-testing corrections to all estimates. Appendix Table A.3 shows that the effects of mixed-gender groups are robust to corrections that control the family-wise error rate (FWER) or the false discovery rate (FDR).

Second, the data used in this paper cover a much longer period than originally preregistered. Rather than only using data through 2019/20, this paper includes data from five additional cohorts (2020/21 to 2024/25). I collected these additional data to increase the precision of estimates, especially given disruptions caused by the COVID-19 pandemic. Appendix Figure A.4 shows that the main finding of this paper holds even when restricting the analysis to the initial cohorts, and excluding any single cohort from the sample does not change the results. Furthermore, this paper includes data from replications of the experiment in different settings to assess the generalizability of the findings (see Section 6.2.2).

Third, this paper analyzes additional outcomes such as mental health and gender role attitudes, which were added after the preregistration to examine broader impacts of peer groups. To mitigate concerns about multiple hypothesis testing across outcomes, I show that the estimated academic and non-academic effects of mixed-gender groups are also robust to multiple-testing corrections (see Appendix Table A.4). The idea of randomly selecting a group coordinator also emerged after the preregistration as a way to better facilitate group activities. I use this additional randomization to understand why mixed-gender groups improve student learning.

3 Descriptive Statistics and Empirical Strategy

3.1 Summary statistics

Panel A of Table 1 summarizes the baseline characteristics of students in the sample. While gender and course-repeating status are obtained from administrative data and are observable for all students, the other characteristics are measured in the baseline survey and are only available for students who responded to the relevant questions.

Among all students in the sample, 43% are female, and 11.5% are retaking the course. For students who answered the survey questions, 87% report German as their first language, the median age is 20, and the average final math and language grades in high school are 4.6 and 4.7, respectively.

The survey also elicits the Big Five personality traits, using the 15-item version from the German Socio-Economic Panel (SOEP; see John et al. 1991; Gerlitz and Schupp 2005). The personality traits of students in the sample are comparable to those of the German population of similar age.[14]

Table 1. Descriptive Statistics of Baseline Characteristics and Academic Outcomes

	(1) N	(2) Mean	(3) SD	(4) Minimum	(5) Median	(6) Maximum
Panel A: Baseline Characteristics						
Female	3,060	0.428	0.495	0	0	1
Repeating course	3,060	0.115	0.319	0	0	1
Age	2,569	20.24	1.317	18(−)	20	22(+)
High school math grade	2,569	4.552	0.800	1	4.5	6
High school language grade	2,569	4.749	0.600	1	4.5	6
First language is German	2,569	0.873	0.333	0	1	1
Openness	2,542	4.877	1.082	1	5	7
Conscientiousness	2,542	4.785	0.960	1	4.7	7
Extraversion	2,542	4.752	1.248	1	4.7	7
Agreeableness	2,542	5.499	0.891	2	5.7	7
Neuroticism	2,542	4.064	1.310	1	4	7
Competitiveness	2,542	6.876	2.003	1	7	10
Traditional gender attitudes	2,535	2.231	0.834	1	2.25	5
Parents' income level	2,520	3.335	0.922	1 (lowest)	3	5 (highest)
Parents with college degree	2,520	0.871	0.850	0 (neither)	1	2 (both)
Panel B: Academic Outcomes						
Final grade in the economics course	3,060	3.979	1.231	1 (lowest)	4.25	6 (highest)
Passing the economics course	3,060	0.608	0.488	0	1	1
Final grade in other first-semester courses	12,322	4.260	1.139	1 (lowest)	4.5	6 (highest)
Final grade in second-semester courses	9,365	4.220	1.134	1 (lowest)	4.5	6 (highest)
Program dropout by second year	2,580	0.253	0.435	0	0	1

Notes: Panel A shows summary statistics for students' baseline characteristics. The variables *Female* and *Repeating Course* are obtained from administrative data. The other characteristics are measured by different questions in the baseline survey. Since the questions are optional, the number of observations varies, depending on the response rate. Panel B shows summary statistics for students' academic outcomes. Due to data restrictions, the second-semester grades and program dropout are not available for the 2024/25 cohort, so the sample size is smaller for these outcomes.

I also measure general competitiveness based on a survey question developed and validated by Buser et al. (2024). The average competitiveness score is 6.88 on a scale of 1 to 10, which is similar to that in Buser et al. (2024). In addition, students report their agreement with four statements (adapted from the World Values Survey) related to traditional gender roles in the family and society, on a scale of 1 (strongly disagree) to 5 (strongly agree). I use the average of these agreement scores as an overall measure of traditional gender attitudes—see Section 4.4.2 for details. As for family background, the average parental income level is 3.3 on a scale of 1 (lowest) to 5 (highest), and the

[14] According to the 2019 SOEP data, the average scores of the five personality traits for respondents aged 18 to 22 are: openness 4.71 (versus 4.88 in my sample), conscientiousness 5.19 (versus 4.79), extraversion 4.87 (versus 4.75), agreeableness 5.4 (versus 5.5), and neuroticism 3.77 (versus 4.06). Each trait is measured with three items on a Likert scale of 1 to 7.

average student has 0.87 parents with a college degree.[15]

Panel B of Table 1 shows summary statistics for students' academic outcomes: the final grade in the economics course, an indicator for successfully passing the course (i.e., obtaining a grade of 4 or above); final grades for other courses taken during the first or second semester; and whether students drop out of their initial study program by the start of the second academic year. The average economics grade in the sample is 3.98, with a standard deviation of 1.23, and the overall likelihood of passing the course is 61%. Students obtain higher grades on average in other courses in the first and second semesters, with a mean of around 4.5. By the start of the second academic year, 25% of students have dropped out of their original study program at the business school.

3.2 Empirical strategy

To estimate the impact of group gender diversity on academic performance and other student outcomes, I use the following model:

$$Y_{ig} = \beta \cdot \textit{Mixed-Gender}_g + \gamma_1 \cdot \textit{Female}_{ig} + \gamma_2 R_{ig} + \delta X_{ig} + \varepsilon_{ig}, \tag{1}$$

where Y_{ig} is the outcome of student i in group g. The key variable of interest, $\textit{Mixed-Gender}_g$, indicates whether group g is mixed-gender, with homogeneous groups as the omitted category. I also estimate the effects of three mixed-gender scenarios by replacing the overall mixed-gender indicator with three indicators: *Majority Gender in Group* represents the scenario of being in the majority (i.e., women in female-majority groups and men in male-majority groups); *Minority Gender in Group* represents being in the minority (i.e., women in male-majority groups and men in female-majority groups); and *Gender-Balanced Group* indicates being in groups with two men and two women. For completeness, I also use an alternative specification replacing the mixed-gender indicator with a continuous diversity measure, *Group Gender Diversity*, which takes three values: 0 for homogeneous groups; 0.86 for gender-imbalanced groups; and 1 for gender-balanced groups.[16]

The specification always controls for students' own gender, captured by $\textit{Female}_{ig}$, which is a dummy variable indicating whether student i is female. Following standard practice (Glennerster and Takavarasha 2013), the baseline specification additionally controls for the strata fixed effects, captured by the vector R_{ig}: a series of dummies indicating each cohort-by-program stratum. For robustness, I also show results from specifications that control for student characteristics observed at

[15] Panel A of Table A.5 summarizes the baseline characteristics of men and women separately. Although I find gender differences in characteristics such as language grades, competitiveness, neuroticism, and gender attitudes, men and women do not significantly differ in math grades or speaking German as first language, which are strong predictors of performance. Section 5.3 shows that the effects of gender diversity are not driven by group composition in other observable student attributes.

[16] The diversity measure is equal to the group standard deviation of a binary variable indicating female (or male) students. For ease of interpretation, I scale this measure so that 0 refers to the least diverse groups, and 1 refers to the most diverse groups. Alternatively, I construct a Herfindahl-based measure of gender diversity, which is equal to one minus the Herfindahl–Hirschman index: $1 - s_m^2 - s_w^2$, where s_m is the share of men in a group, and s_w is the share of women. After rescaling, this measure has three values: 0, 0.75, and 1. These two measures of diversity are highly correlated, with a correlation coefficient of 0.986, and produce essentially the same results.

baseline, captured by the vector X_{ig}, which consists of an indicator for German as the first language, a course-repeating indicator, high school math and language grades, parental income level, number of college-educated parents, the Big Five personality traits, competitiveness, and baseline gender attitudes. ε_{ig} is the error term; I use robust standard errors clustered at the study group level.

The key identification assumption for the estimation is that group gender composition is orthogonal to student characteristics that may affect performance or other outcomes. Given that students are randomly assigned to groups, this assumption should hold. To check whether the randomization was successful, I conduct a series of balance tests. First, I test whether each student characteristic observed at baseline is jointly predicted by group gender composition, captured by four indicators for 1–4 women in a group (with 0 as the omitted category). Panel A of Table 2 shows the F-statistics and p-values derived from the joint tests of these gender composition dummies. Among the 14 characteristics, only the German language indicator is predicted by gender composition—an effect marginally significant. Together, these results highlight that student characteristics are overall balanced across groups with different gender compositions.

Panel B of Table 2 tests the balance of student characteristics between homogeneous groups and mixed-gender groups—when combining all mixed groups (Column 2) and when splitting them into three scenarios (Columns 3–5). The results show that students in homogeneous groups report higher openness and have higher parental income than students in mixed-gender groups. However, there are no statistically significant differences along the other 12 dimensions. Overall, the baseline characteristics do not jointly predict the likelihood of being assigned to a homogeneous group versus any type of mixed-gender group. Panel C further shows that student characteristics are mostly similar across the three scenarios of mixed-gender groups. Again, the baseline characteristics do not jointly predict the likelihood of being assigned to one mixed-gender scenario rather than another.

Appendix Table A.6 shows the results of these balance tests separately for men and women. Although a few baseline characteristics appear different across gender compositions, the vast majority of these differences do not reach statistical significance at the 5% level. When tested jointly, the baseline characteristics do not significantly predict the likelihood of men or women being assigned to any type of gender composition versus another. Appendix Table A.7 further shows results from the Leuven tests of randomization. Among the 15 baseline characteristics (including gender), only course-repeating status significantly varies across study groups. This may be due to the nature of the binary variable, which takes the value 0 for about 90% of the sample. Moreover, the results of this paper are robust when controlling for all observable student characteristics and when excluding the repeating students (see Appendix Figure A.5).

Table 2. Balance Tests of Baseline Characteristics

	(1) A.	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		B. Effects of Homogeneous Groups vs.				C. Between Mixed Groups		
	Effects of #Women Dummies	All Mixed Groups	Minority Gender	Gender-Balanced	Majority Gender	Balanced vs. Minority	Balanced vs. Majority	Minority vs. Majority
Repeating course	0.550	0.003	0.029	-0.004	-0.001	0.038**	0.009	-0.003
	[0.699]	(0.017)	(0.023)	(0.020)	(0.019)	(0.020)	(0.017)	(0.024)
Age	0.212	-0.041	-0.024	-0.019	-0.050	-0.024	-0.022	0.040
	[0.932]	(0.073)	(0.108)	(0.084)	(0.080)	(0.088)	(0.061)	(0.088)
High school math grade	0.656	0.004	-0.004	0.023	-0.004	0.039	0.010	0.005
	[0.623]	(0.049)	(0.068)	(0.055)	(0.053)	(0.052)	(0.039)	(0.052)
High school language grade	0.947	0.001	-0.032	-0.014	0.028	0.022	0.062**	0.006
	[0.436]	(0.031)	(0.047)	(0.035)	(0.035)	(0.042)	(0.029)	(0.044)
First language is German	2.103*	0.016	0.018	0.004	0.009	0.022	-0.007	-0.034
	[0.079]	(0.018)	(0.026)	(0.021)	(0.019)	(0.023)	(0.015)	(0.024)
Openness	1.325	0.104*	0.023	0.173***	0.078	-0.054	-0.069	-0.057
	[0.259]	(0.057)	(0.085)	(0.065)	(0.060)	(0.071)	(0.049)	(0.082)
Conscientiousness	0.453	0.011	-0.002	0.017	0.025	0.014	0.005	-0.085
	[0.770]	(0.052)	(0.078)	(0.057)	(0.057)	(0.058)	(0.044)	(0.064)
Extraversion	1.220	0.045	0.167	0.055	0.006	0.104	-0.020	-0.113
	[0.301]	(0.072)	(0.103)	(0.080)	(0.076)	(0.083)	(0.057)	(0.084)
Agreeableness	0.581	0.015	-0.056	0.006	0.048	-0.076	0.005	-0.008
	[0.677]	(0.048)	(0.066)	(0.053)	(0.054)	(0.054)	(0.042)	(0.061)
Neuroticism	0.440	0.014	0.053	0.028	-0.024	0.051	-0.052	-0.149*
	[0.779]	(0.072)	(0.097)	(0.080)	(0.077)	(0.082)	(0.060)	(0.082)
Competitiveness	0.297	0.087	0.169	0.070	0.122	0.128	0.013	-0.089
	[0.880]	(0.112)	(0.165)	(0.125)	(0.120)	(0.135)	(0.095)	(0.133)
Traditional gender attitudes	0.311	0.033	0.079	0.027	0.032	0.056	-0.004	0.022
	[0.871]	(0.045)	(0.064)	(0.051)	(0.048)	(0.052)	(0.040)	(0.057)
Parents with college degree	0.104	-0.004	-0.082	0.030	0.017	-0.061	-0.004	-0.053
	[0.981]	(0.048)	(0.074)	(0.055)	(0.052)	(0.057)	(0.040)	(0.062)
Parents' income level	1.291	0.111**	0.033	0.121**	0.117**	-0.030	0.015	0.014
	[0.272]	(0.050)	(0.069)	(0.057)	(0.056)	(0.060)	(0.046)	(0.067)
Observations	3,060	3,060	889	1,504	1,667	1,393	2,171	1,556
Joint significance (p-value)	–	0.589	0.462	0.336	0.770	0.677	0.858	0.156

Notes: Panel A tests the joint effect of four group gender composition dummies (indicating groups with 1, 2, 3, or 4 women) on each baseline characteristic, with the F-statistics and p-values (in brackets) reported. Panel B tests differences in baseline characteristics between students in homogeneous groups and students in different types of mixed-gender groups—the table reports the estimated coefficient and standard error of the "Homogeneous Group" variable. Panel C tests differences between each pair of mixed-gender scenarios. Each estimate in Columns 2–8 is derived from one ordinary least squares (OLS) regression or linear probability model (LPM), in which the dependent variable is a baseline characteristic and the independent variable is an indicator for a specific type of gender composition. All regressions control for students' own gender and randomization strata (cohort-by-program) fixed effects. Standard errors in parentheses are clustered at the group level. For the comparisons in each column, I also conduct joint tests of whether all baseline characteristics predict the treatment assignment; the p-value in the bottom row indicates the joint significance of all baseline characteristics.

4 Results

4.1 Effects on economics performance

I first examine how group gender diversity affects students' performance in the economics course—the most immediate outcome that study groups are likely to influence. Column 1 of Table 3 shows

that being assigned to a mixed-gender rather than homogeneous group improves the final grade in economics by 0.14 SD. This effect is equivalent to 0.17 grade points on a scale of 1 to 6 and is statistically significant at the 1% level ($p = 0.003$). Column 2 breaks down this overall impact and estimates the effects of being in the majority, in a balanced group, and in the minority. All three mixed-gender scenarios directionally improve course performance. Specifically, being the majority gender in a group raises final grades by 0.15 SD ($p = 0.004$), being in a balanced group raises grades by 0.14 SD ($p = 0.006$), and being the minority gender raises grades by 0.09 SD ($p = 0.200$). Although the effect of being in the minority is not statistically significant, it does not significantly differ from the effects of the other two scenarios either. These estimates highlight a non-monotonic effect of the female (or male) share on course performance: the grade increases when moving from being the minority gender to being the majority gender, but it decreases substantially when moving from majority-gender groups to single-gender groups.

Column 3 of Table 3 estimates a linear effect of increasing the level of gender diversity using a continuous diversity measure ranging from 0 (homogeneous groups) to 1 (gender-balanced groups). The results show that moving from the least-diverse to the most-diverse group raises course grades by 0.15 SD. This estimate is comparable to the overall effect of mixed-gender groups and is statistically significant at the 1% level ($p = 0.003$).[17] Columns 4–5 show the effect of mixed-gender groups separately for women and men. Assignment to a mixed-gender group raises women's final grades by 0.18 SD ($p = 0.032$) and men's final grades by 0.15 SD ($p = 0.009$).[18]

Because the economics course is compulsory for students' study programs, successfully passing the course is a high-stakes outcome. Column 6 of Table 3 shows that students in mixed-gender groups are overall 6.5 percentage points (or 11%) more likely to pass the course than those in homogeneous groups ($p = 0.005$). Column 7 shows that this effect is also of similar magnitude across the three mixed-gender scenarios. Appendix Figure A.7 further plots the effect of mixed-gender groups across the grade distribution: assignment to a mixed-gender group increases the likelihood of achieving grades above different thresholds across the distribution, with particularly strong effects around the passing threshold.

How large is the effect of group gender diversity? An increase in course grades of approximately 0.14 SD is sizable compared to the predictive effects of baseline characteristics. For example, it is equivalent to the effect of a 0.5 SD increase in high school math grades or a 1.7 SD increase in conscientiousness (see Appendix Table A.9). The effect is also comparable to that of other interventions and policies aimed at improving academic performance, such as financial aid, incentive schemes, and sleep interventions (e.g., Goldrick-Rab et al. 2016; Angrist et al. 2009; Barrow et al. 2014; Giuntella et al. 2024).

[17] The similarity between the estimated effects of the binary mixed-gender indicator and the continuous gender diversity measure is perhaps not surprising, since the two diversity measures are highly correlated ($\rho > 0.95$). Appendix Figure A.6 shows the estimated effects of the continuous gender diversity measure on main outcomes examined in this paper.

[18] Appendix Table A.8 further shows that for women, the effect of mixed-gender groups is similar across female-majority, balanced, and female-minority groups. For men, the effects of male-majority and balanced groups are of comparable magnitude, while the effect of male-minority groups is not significantly different from zero.

Table 3. Effects of Group Gender Diversity on Economics Performance

Dependent Variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	A. Standardized Final Economics Grade					B. Course Passing	
Analysis Sample:	All	All	All	Women	Men	All	All
Mixed-Gender Group	0.139***			0.183**	0.152***	0.065***	
	(0.046)			(0.085)	(0.058)	(0.023)	
	[0.003]			[0.032]	[0.009]	[0.005]	
Majority Gender in Group		0.150***					0.071***
		(0.051)					(0.025)
		[0.004]					[0.005]
Gender-Balanced Group		0.144***					0.064**
		(0.052)					(0.026)
		[0.006]					[0.014]
Minority Gender in Group		0.086					0.043
		(0.067)					(0.033)
		[0.200]					[0.193]
Group Gender Diversity			0.149***				
			(0.050)				
			[0.003]				
Female	-0.226***	-0.218***	-0.227***			-0.125***	-0.122***
	(0.036)	(0.038)	(0.036)			(0.018)	(0.019)
Observations	3,060	3,060	3,060	1,311	1,749	3,060	3,060
R-squared	0.033	0.033	0.033	0.021	0.023	0.028	0.029
Cohort-by-Program FEs	✓	✓	✓	✓	✓	✓	✓
Raw Dep. Var. Mean	3.979	3.979	3.979	3.806	4.109	0.608	0.608
Minority = Balanced (p-value)		0.326					0.476
Minority = Majority (p-value)		0.292					0.357
Balanced = Majority (p-value)		0.894					0.751

Notes: Each column represents one OLS regression or LPM. The dependent variable in Panel A is the final economics grade standardized across the whole sample, with mean 0 and SD 1. The dependent variable in Panel B is a binary indicator for successfully passing the course (i.e., obtaining a grade of 4 or above). The row *Raw Dep. Var. Mean* reports the mean of the raw dependent variable. *Group Gender Diversity* is a continuous measure taking three values: 0 (homogeneous groups), 0.86 (imbalanced groups), and 1 (balanced groups)—see footnote 16 for details. All other independent variables are binary indicators: *Mixed-Gender Group* indicates any groups with a mixture of men and women; *Majority Gender in Group* indicates men in male-majority groups and women in female-majority groups; similarly, *Minority Gender in Group* indicates women in male-majority groups and men in female-majority groups. I also test the differences between each pair of these three variables and report the p-values in the bottom rows. All regressions control for cohort-by-program fixed effects. Standard errors in parentheses are clustered at the group level; p-values are in brackets. $^{*}p<.1$, $^{**}p<.05$, $^{***}p<.01$.

Robustness: Appendix Table A.9 shows that the effect of mixed-gender groups remains highly robust in both magnitude and precision when controlling for a rich set of student attributes, including high school grades, German as first language, age, personality traits, gender attitudes, and parental background. The effect is even slightly larger when controlling for the group gender ratio. The effect of mixed-gender groups also appears consistent across the seven cohorts in the sample. Appendix Figure A.4 shows that the effect estimated separately for each cohort is relatively noisy but consistently positive in sign. Excluding any single cohort from the sample does not meaningfully change the overall effect. By contrast, including data from additional cohorts substantially increases the precision of the estimate and leads it to converge toward a stable magnitude. This pattern of convergence reflects the value of replications in enhancing the validity and reliability of research

findings (Camerer et al. 2016; Gertler et al. 2018).

Heterogeneity: The results above show that group gender diversity benefits both men and women. To test for heterogeneity more systematically, I split the sample into two subsamples by each observed baseline characteristic and separately estimate the effect of mixed-gender groups on the economics grade. Appendix Figure A.8 plots the estimated effects for all subsamples, as well as the p-values from tests of differences between all pairs of estimates. To account for multiple hypothesis testing, I further correct the p-values using the step-down approach of Romano and Wolf (2005*a*,*b*, 2016). The results show that mixed-gender groups directionally improve course grades across almost all 30 subsamples, and none of the differences reach statistical significance at the 5% level.[19]

These results alleviate concerns about students' self-selection into the experiment. Although the participation decision is predicted by various baseline characteristics, the estimated effect of group gender diversity does not significantly vary by these characteristics. The lack of heterogeneity provides suggestive evidence that voluntary participation in the experiment is unlikely to systematically amplify or dampen the effect of gender diversity. Section 6.2 discusses the external validity of the findings in greater detail.

4.2 Effects on performance in other first-semester courses

Next, I examine whether study groups also influence performance in other first-semester courses. Conceptually, spillovers on other courses are ambiguous. On the one hand, improved learning habits or knowledge in economics may benefit performance in other courses. On the other hand, study groups formed for economics could crowd out effort or attention devoted to other subjects. To assess potential spillovers, I use administrative data on grades for other courses taken during the first semester at the business school.

Column 1 of Table 4 shows that assignment to a mixed-gender rather than homogeneous group in the economics course significantly improves grades in other first-semester courses by 0.098 SD ($p < 0.001$). Decomposing this effect, Column 2 shows that being the majority gender, the minority gender, or in a balanced group all lead to positive and statistically significant improvements in other course grades, with magnitudes ranging from 0.08 to 0.11 SD. Columns 3–4 split the sample by gender and show that the spillover effect is stronger and more precisely estimated for men, while for women it remains positive and marginally significant.

The positive spillovers may arise through two channels: (1) collaborative learning or information exchange among group members for other shared courses; and (2) direct knowledge spillovers across courses—for example, improved understanding of economics helps students learn related subjects. Column 5 of Table 4 shows that the spillovers are concentrated in the common courses shared by all group members. This is consistent with the first channel: group members study together for multiple courses, and as a result, the benefits of mixed-gender groups emerge simultaneously in

[19] Appendix Figure A.9 shows similar patterns when using course passing as the dependent variable.

economics and other common courses. As for the second channel, Appendix Table A.10 shows that the spillover effects are strongest in banking and finance but not present in business administration. Mixed-gender groups also improve performance in mathematics, accounting, and informatics. These patterns do not seem consistent with the idea that spillovers are stronger in subjects more closely related to economics, suggesting that the second channel is less plausible.

Columns 6 and 7 of Table 4 present similar effects of gender diversity on course passing. Overall, assignment to mixed-gender groups in the economics course raises the likelihood of passing other courses by 2.9 percentage points ($p = 0.007$). For courses commonly taken by group members, this effect amounts to about 5 percentage points. Appendix Table A.11 further shows that the spillover effects are robust to controlling for observable student characteristics at baseline and are more pronounced for compulsory courses that students must take during the first semester.

Table 4. Effects of Mixed-Gender Groups on Performance in Other First-Semester Courses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable:	A. Standardized Grades for Other Courses					B. Course Passing	
Analysis Sample:	All	All	Women	Men	All	All	All
Mixed-Gender Group	0.098***		0.077*	0.137***	0.005	0.029***	-0.009
	(0.024)		(0.044)	(0.029)	(0.046)	(0.011)	(0.020)
	[<0.001]		[0.079]	[<0.001]	[0.914]	[0.007]	[0.641]
Mixed-Gender Group × Common Course					0.133**		0.054**
					(0.054)		(0.023)
					[0.014]		[0.020]
Majority Gender in Group		0.111***					
		(0.026)					
		[<0.001]					
Gender-Balanced Group		0.076***					
		(0.027)					
		[0.005]					
Minority Gender in Group		0.109***					
		(0.034)					
		[0.001]					
Female	-0.217***	-0.213***			-0.219***	-0.086***	-0.087***
	(0.018)	(0.019)			(0.018)	(0.008)	(0.008)
Observations	12,225	12,225	4,794	7,384	12,225	12,225	12,225
R-squared	0.109	0.109	0.147	0.094	0.110	0.076	0.077
Cohort–Program–Course FEs	✓	✓	✓	✓	✓	✓	✓
Raw Dep. Var. Mean	4.260	4.260	4.126	4.346	4.260	0.739	0.739

Notes: Each column represents one OLS regression or LPM. The observations are other first-semester courses taken by students in the sample. The dependent variable is either the final course grade standardized across the full sample (Panel A) or an indicator for passing the course (Panel B). All models control for cohort-by-program-by-course fixed effects and cluster standard errors at the group-by-course level. *Common Course* is an indicator for courses commonly taken by all group members; 68% of the observations correspond to common courses. The models in Columns 5 and 7 additionally control for the common course indicator. The estimated effects of three mixed-gender scenarios in Column 2 are not significantly different from one another. Standard errors are in parentheses; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

4.3 Effects on longer-run performance and program dropout

The results above provide robust evidence that group gender diversity improves student performance during the first semester at university. These effects may carry over into the longer run, given that learning is cumulative and that early academic success may build confidence and stimulate subsequent investment. I examine two sets of longer-run outcomes: (1) performance in courses taken during the second semester; and (2) remaining enrolled at the start of the second academic year, indicating whether a student has dropped out of their original major or minor study program at the business school.

Panel A of Table 5 reports the estimated effects of mixed-gender groups on second-semester course performance. Column 1 includes the full set of courses taken by students in the sample. Initial assignment to mixed-gender groups raises final grades for these courses by 0.06 SD, an effect that is marginally significant ($p = 0.052$). After excluding a small set of observations corresponding to elective courses with fewer than 5 students enrolled, Column 2 shows that the estimate increases slightly and becomes statistically significant at the 5% level. The effect further increases when restricting the sample to compulsory courses. Column 3 shows that mixed-gender groups improve final grades for these courses by 0.09 SD ($p = 0.022$).

The split-gender analyses in Columns 4–5 show that the longer-run effect applies to both genders: the point estimate is larger for women but is more precise for men. Column 6 further shows that all three mixed-gender scenarios lead to directionally higher grades relative to homogeneous groups, and the differences across the three scenarios are not statistically significant. These performance gains are unlikely to be driven by endogenous course selection. Appendix Table A.12 shows that students' course enrollment decisions and the likelihood of group members enrolling in common courses do not differ significantly between mixed-gender and homogeneous groups.

Turning to program dropout, Column 7 of Table 5 shows that students assigned to mixed-gender rather than homogeneous groups at university entry are 6 percentage points (or 24%) less likely to drop out of their original study program by the start of the second academic year ($p = 0.022$). Columns 8–11 restrict the sample to non-repeating students, for whom the dropout measure is even cleaner: they must remain enrolled in the second year unless they have dropped out of their study program. For these students, initial assignment to mixed-gender groups lowers the dropout rate by 7.7 percentage points (see Column 8). This effect is both economically and statistically significant: it corresponds to 32% of the average dropout rate and is statistically significant at the 1% level. Splitting the sample by gender, Columns 9–10 show that the effect applies to both genders and is particularly strong for women: the dropout rate decreases by 12.4 percentage points (or 42%) for women and 6.1 percentage points (or 30%) for men.[20]

Finally, Column 11 of Table 5 reports the effects of three mixed-gender scenarios on dropout.

[20] When including all students regardless of the repeating status, the effect of mixed-gender groups is equivalent to 9.7 and 5.3 percentage points for women and men, respectively; both estimates are statistically significant at the 10% level.

Table 5. Effects of Mixed-Gender Groups on Longer-Run Course Performance and Program Dropout

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Dependent Variable:	A. Standardized Grade for Courses in Second-Semester						B. Program Dropout by Second Year				
Analysis Sample:	All	Main	Compulsory Courses				All	Non-Repeating Students			
	Courses	Courses	All	Women	Men	All		All	Women	Men	All
Mixed-Gender Group	0.059*	0.060**	0.086**	0.156*	0.088**		-0.060**	-0.077***	-0.124**	-0.061**	
	(0.030)	(0.030)	(0.038)	(0.081)	(0.042)		(0.026)	(0.026)	(0.056)	(0.030)	
	[0.052]	[0.047]	[0.022]	[0.055]	[0.037]		[0.022]	[0.003]	[0.027]	[0.041]	
Majority Gender in Group						0.104**					-0.098***
						(0.040)					(0.028)
						[0.010]					[<0.001]
Gender-Balanced Group						0.066					-0.047
						(0.042)					(0.029)
						[0.117]					[0.106]
Minority Gender in Group						0.068					-0.083**
						(0.052)					(0.035)
						[0.191]					[0.018]
Female	-0.166***	-0.169***	-0.161***			-0.152***	0.084***	0.083***			0.078***
	(0.022)	(0.022)	(0.027)			(0.029)	(0.019)	(0.019)			(0.020)
Observations	9,215	9,087	6,115	2,246	3,843	6,115	2,580	2,292	967	1,325	2,292
R-squared	0.098	0.093	0.094	0.097	0.097	0.094	0.026	0.034	0.032	0.031	0.037
Cohort–Program–Course FEs	✓	✓	✓	✓	✓	✓	–	–	–	–	–
Cohort–Program FEs	–	–	–	–	–	–	✓	✓	✓	✓	✓
Raw Dep. Var. Mean	4.220	4.209	4.114	4.006	4.178	4.178	0.253	0.239	0.292	0.200	0.200
Minority = Balanced (p-value)						0.972					0.214
Minority = Majority (p-value)						0.405					0.610
Balanced = Majority (p-value)						0.233					0.016

Notes: Each column represents one OLS regression or LPM. The dependent variable in Panel A is the final grade for courses taken during the second semester at university, standardized across all student-by-course observations. The dependent variable in Panel B is a binary indicator for dropping out of the original study program by the start of the second academic year (one year after group assignment). Due to data restrictions, the analysis sample in this table does not include the 2024/25 cohort. In Columns 8–11, the sample excludes students who were retaking the economics course. The models in Panel A control for cohort-by-program-by-course fixed effects and cluster standard errors at the group-by-course level. The models in Panel B control for cohort-by-program fixed effects and cluster standard errors at the group level. Standard errors are in parentheses; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Being the majority gender in a group has the strongest impact, corresponding to a 9.8 percentage point decrease in dropout ($p < 0.001$). This effect is not significantly different from the effect of being the minority gender (8.3 percentage points) but is significantly larger than the effect of gender-balanced groups (4.7 percentage points). Again, the effects of being in the minority and being in a balanced group are not significantly different from each other. Despite variation in the magnitude, these estimates suggest that the effects of mixed-gender groups are largely consistent across the three scenarios.

Taken together, the results in this section highlight that gender-diverse peer groups formed early in university have positive effects on students' course performance throughout the first year. These improvements in academic achievement also translate into a lower likelihood of dropping out of the original study program, potentially facilitating a more successful educational trajectory. These conclusions remain unchanged when estimating the effects of the continuous gender diversity measure (see Appendix Figure A.6).

4.4 Effects on subjective well-being and gender role attitudes

While group gender diversity benefits students' educational attainment, it remains unclear whether these benefits are confined to the academic domain or also extend to non-academic outcomes. To assess broader welfare effects, I examine how gender diversity influences students' subjective well-being in the short and longer run. I further analyze how cross-gender interaction in study groups shapes gender role attitudes during a formative period at university.

4.4.1 Subjective well-being in the short run

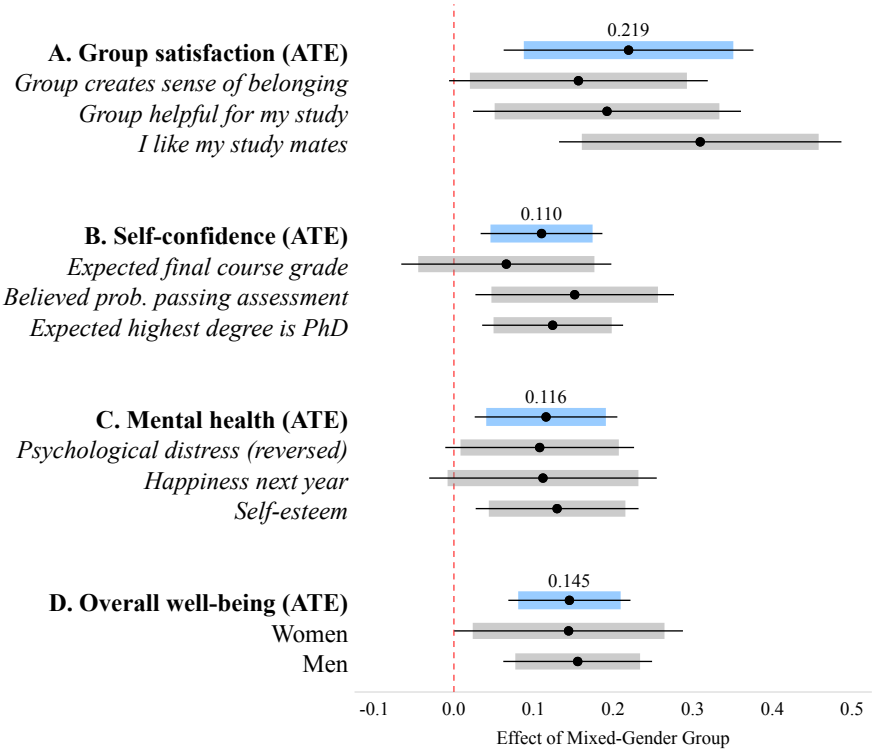
I measure three aspects of subjective well-being at the end of the first semester, before the final exams: (1) satisfaction with the study group, (2) academic self-confidence, and (3) mental health. Each aspect of well-being is measured with multiple survey questions (details below). To mitigate concerns about multiple hypothesis testing, I construct an index for each dimension. Specifically, each item within a dimension is first standardized to have mean zero and standard deviation one; I then take the average of these standardized items to obtain the index. I further construct an overall well-being index by calculating the average of all standardized items across three dimensions. The average standardized effects based on these indices are very similar to those estimated using the seemingly unrelated regression framework, which accounts for covariance across estimates (Clingingsmith et al. 2009; Kling et al. 2004)—see Appendix Table A.13.

Group satisfaction: Students report their agreement with three statements regarding their study group on a 7-point Likert scale: (1) "In general, the group is helpful for my studies"; (2) "The group gives me a sense of belonging at [university name]"; and (3) "On the whole, I like my study mates in the group." Panel A of Figure 2 shows that students in mixed-gender groups are systematically more likely to agree with these statements, indicating higher satisfaction with their group. The

estimated average treatment effect (ATE) of mixed-gender groups on the three group satisfaction measures is equivalent to 0.22 SD ($p = 0.006$).

Self-confidence: I measure students' expectations regarding three future academic outcomes: (1) the final grade they expect to obtain in the economics course; (2) the perceived likelihood of successfully completing all compulsory courses in their study program (i.e., passing the assessment); and (3) the highest educational degree they expect to attain, including a doctoral degree, master's degree, or lower. For the third outcome, I focus on a binary indicator for expecting a doctoral degree. Panel B of Figure 2 shows that students in mixed-gender groups report significantly more optimistic expectations about future achievement. On average, assignment to mixed-gender groups raises academic self-confidence by 0.11 SD ($p = 0.005$).

Figure 2. Short-Run Effects on Subjective Well-Being



Notes: Panel A plots the estimated effects of mixed-gender groups on overall group satisfaction and its specific components: "In general, the group is helpful for my studies"; "The group gives me a sense of belonging at [university name]"; and "On the whole, I like my study mates in the group." Panel B shows the effects on academic self-confidence, an index combining three aspects: the expected final grade for the economics course, the perceived likelihood of passing the assessment for the study program, and whether they expect a doctoral degree as the highest educational degree. Panel C shows the effects on overall mental health, combining three outcomes: a reversed measure of psychological distress (anxiety and depression), the expected happiness during the next academic year, and self-esteem. Panel D plots the effects on overall well-being, combining all outcomes across the three dimensions. All items within each dimension are standardized across the whole sample to have mean 0 and SD 1. The overall estimates in each panel reflect the average standardized treatment effects (ATE). All regressions control for cohort-by-program fixed effects, gender, and observable baseline characteristics. Standard errors are clustered at the group level. Error bars indicate 95% and 90% confidence intervals.

Mental health: I elicit three aspects of mental health: (1) psychological distress, measured with the four-item Patient Health Questionnaire on anxiety and depression (Kroenke et al. 2009); (2) the expected happiness in the next academic year, reported on a 10-point Likert scale where 10 means "very happy"; and (3) self-esteem, which is measured using items from the Rosenberg (1965) scale or the one-item scale from Robins et al. (2001), depending on the cohort.[21] Panel C of Figure 2 shows that assignment to mixed-gender groups leads to directional improvements in all three components of mental health and has an average effect of 0.12 SD ($p = 0.011$).

Lastly, Panel D of Figure 2 shows the average effect of mixed-gender groups on well-being, combining all measures of group satisfaction, self-confidence, and mental health. Overall, assignment to a mixed-gender rather than homogeneous group improves well-being by 0.15 SD ($p < 0.001$). This effect is almost identical in magnitude for men and women.[22] These findings highlight that gender diversity not only enhances academic performance but also improves subjective well-being in the short run. These two sets of effects are potentially intertwined and may reinforce each other over time: prior work emphasizes that positive emotions and mental health can elevate productivity and performance (Compte and Postlewaite 2004; Oswald et al. 2015); conversely, improvements in learning outcomes could also boost self-confidence and mental health.

4.4.2 Gender role attitudes in the short run

Next, I analyze how group gender diversity affects attitudes toward gender roles. I elicit students' agreement with the following statements at both baseline and endline: (1) "If a woman earns more money than her partner, this might cause problems"; (2) "When a mother is working full time, the child develops less well"; (3) "A university degree is more important for a man than for a woman"; and (4) "On the whole, men make better business executives than women do."[23] Students answer these questions on a Likert scale of 1 (strongly disagree) to 5 (strongly agree), with higher values indicating more traditional attitudes. The first two statements concern gender roles in the family, while the last two concern broader roles in society, related to educational attainment and business leadership. Overall, students hold more traditional attitudes toward family gender roles and more progressive attitudes toward gender roles in society: about 50% of students either agree with or express neutral views on statements 1–2, whereas only about 30% do so for statements 3–4 (see

[21] For the 2019/20 and 2020/21 cohorts, I use four items from the Rosenberg (1965) scale: "On the whole, I'm satisfied with myself"; "I certainly feel useless at times" [reversed]; "I take a positive attitude toward myself"; and "All in all, I am inclined to feel that I am a failure" [reversed]. Students report their agreement with these statements on a 5-point Likert scale. For later cohorts, I instead use a single-item self-esteem scale that has been tested and validated in various contexts (Brailovskaia and Margraf 2020): "Is the following statement true of you? *I have high self-esteem.*" The answer scale ranges from 1 (not very true of me) to 10 (very true of me).

[22] Appendix Table A.14 reports detailed estimates of how mixed-gender groups affect overall well-being. The results show that the effects for men, women, and the combined sample are robust to the inclusion or exclusion of baseline controls. Assignment to any type of mixed-gender group has a positive effect on well-being, although the benefit of being in the minority is slightly smaller in magnitude.

[23] The statements on gender roles are drawn from the World Values Survey. Minor changes in the wording were made to better accommodate the student population and ensure comparability between the English and German versions of the survey. These questions are available starting with the 2019/20 cohort.

Appendix Figure A.10 for the full distribution of responses).

Table 6 reports how mixed-gender groups affect students' agreement with different gender role statements at endline, controlling for their baseline agreement. Panel A focuses on traditional gender roles in the family and shows that students' overall agreement with these statements increases by 0.16 SD when assigned to mixed-gender rather than homogeneous groups ($p = 0.002$). This effect appears for both statements concerning women's relative earnings and mothers' full-time employment, and applies to both men and women. By contrast, Panel B shows that students' views about gender roles in educational attainment and business leadership are not significantly affected: the estimated effects of mixed-gender groups are directionally negative but close to zero.[24]

These results highlight that close contact between men and women in peer groups can reinforce traditional views about family gender roles. This impact does not extend to gender roles in educational attainment or business leadership, likely because most students already hold progressive views in these domains—representing relatively entrenched social norms.

Table 6. Short-Run Effects on Gender Role Attitudes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Agreement with ...	A. Traditional Gender Roles in Family				B. Traditional Gender Roles in Society			
	Woman not earns more	Mother not works full time	Overall Agreement		College degree for men	Business executive for men	Overall Agreement	
Mixed-Gender Group	0.139**	0.114**	0.155***		-0.054	-0.030	-0.050	
	(0.059)	(0.051)	(0.049)		(0.054)	(0.050)	(0.049)	
	[0.019]	[0.026]	[0.002]		[0.319]	[0.555]	[0.302]	
Mixed-Gender Group × Female				0.179***				-0.081
				(0.069)				(0.064)
				[0.009]				[0.203]
Mixed-Gender Group × Male				0.144**				-0.036
				(0.065)				(0.065)
				[0.028]				[0.584]
Female	0.060	-0.088**	-0.040	-0.071	-0.209***	-0.234***	-0.247***	-0.207**
	(0.048)	(0.042)	(0.042)	(0.091)	(0.043)	(0.039)	(0.039)	(0.089)
Observations	1,571	1,571	1,571	1,571	1,571	1,571	1,571	1,571
R-squared	0.367	0.511	0.498	0.498	0.401	0.519	0.547	0.547
Baseline Attitudes	✓	✓	✓	✓	✓	✓	✓	✓
Raw Dep. Var. Mean	2.443	2.562	2.503	2.503	1.953	2.026	1.990	1.990

Notes: Each column represents one OLS regression. The dependent variables are standardized measures of agreement with four statements: "If a woman earns more money than her partner, this might cause problems"; "When a mother is working full-time, the child develops less well"; "A university degree is more important for a man than for a woman"; and "On the whole, men make better business executives than women do." The raw agreement score ranges from 1 (strongly disagree) to 5 (strongly agree). In Columns 3–4 and 7–8, I use the average agreement score for two statements in each panel (also standardized). All regressions control for cohort-by-program fixed effects, the baseline level of the dependent variable, and other observable baseline characteristics. Standard errors in parentheses are clustered at the group level; p-values are in brackets. In Columns 4 and 8, the differences between the two gender-specific estimates are not statistically significant. *p <.1, **p <.05, ***p <.01.

[24] Appendix Table A.15 presents consistent results when using binary measures of traditional attitudes, i.e., agreeing or disagreeing with different statements.

4.4.3 Mental health and gender role attitudes in the longer run

I further examine whether the effects on well-being and gender attitudes persist over time, using a brief follow-up survey conducted during the summer break after the first academic year. This survey elicits students' psychological distress and gender attitudes using the same instruments as in the endline survey. The analysis sample for the longer-run effects is relatively small because the follow-up survey was administered only to the 2021/22–2024/25 cohorts and the summer response rate was relatively low.

Panel A of Table 7 presents the estimated effects of mixed-gender groups on two measures of mental health: the total score for anxiety and depression (ranging from 0 to 12); and an indicator for moderate or severe psychological distress (i.e., having a score above 6). Column 1 shows that assignment to mixed-gender groups does not affect the continuous distress score. However, Column 2 shows that it significantly reduces the likelihood of students reporting moderate or severe distress by 7.9 percentage points ($p = 0.019$). Column 3 further shows that this effect does not significantly differ by gender, although it is larger in magnitude for women.

Panel B of Table 7 estimates the effects on gender role attitudes. Column 4 shows that, as in the short run, students assigned to mixed-gender groups hold significantly more traditional views about family gender roles than those assigned to homogeneous groups (0.17 SD; $p = 0.033$). However, unlike in the short run, the longer-run effect differs significantly by gender. Column 5 shows that assignment to mixed-gender groups increases men's agreement with traditional family gender roles by about 0.30 SD, but no longer significantly affects women's agreement. Thus, the effect persists—and even intensifies—for men, but fades for women. These gender-differentiated dynamics are not driven by changes in the survey sample: Appendix Figure A.11 shows similar dynamic patterns when using a common sample of students who responded to both the endline and follow-up surveys.

As for broader gender roles in society, Column 6 shows that the overall effect of mixed-gender groups is still close to zero. Column 7 shows that the gender-specific effects mirror the patterns for family gender roles: men's attitudes become more traditional, whereas women's become more progressive. However, the gender gap in the effect is not statistically significant ($p = 0.109$).

Collectively, the results in this section show that the effects of mixed-gender groups extend beyond academic outcomes. Because these groups are formed during a critical onboarding period and bring men and women into close contact, they also generate persistent effects on students' well-being and gender role attitudes. These findings are consistent with prior work documenting strong and lasting impacts of peer groups formed in schools and universities (Chetty et al. 2011; Duflo et al. 2011; Carrell et al. 2018; Thiemann 2022). For robustness, Appendix Table A.4 reports the multiple-testing corrected p-values for the estimated effects of mixed-gender groups on all academic and non-academic outcomes in the short and longer run. Appendix Figure A.6 shows similar findings when using the continuous measure of group gender diversity.

Table 7. Longer-Run Effects on Mental Health and Gender Role Attitudes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	A. Mental Health			B. Gender Role Attitudes			
	Distress Score	Moderate/Severe Distress		Traditional Gender Roles in Family		Traditional Gender Roles in Society	
Mixed-Gender Group	-0.057	-0.079**		0.171**		0.054	
	(0.272)	(0.033)		(0.080)		(0.074)	
	[0.833]	[0.019]		[0.033]		[0.470]	
Mixed-Gender Group			-0.125*		-0.048		-0.095
× Female			(0.072)		(0.121)		(0.100)
			[0.084]		[0.691]		[0.343]
Mixed-Gender Group			-0.053		0.295***		0.138
× Male			(0.036)		(0.105)		(0.102)
			[0.142]		[0.005]		[0.176]
Female	0.370*	0.022	0.087	-0.068	0.241	-0.373***	-0.162
	(0.209)	(0.025)	(0.083)	(0.065)	(0.158)	(0.066)	(0.143)
Observations	754	754	754	754	754	754	754
R-squared	0.170	0.094	0.096	0.439	0.442	0.479	0.480
Baseline Attitudes	–	–	–	✓	✓	✓	✓
Raw Dep. Var. Mean	2.914	0.100	0.100	2.445	2.445	1.803	1.803
Female = Male (p-value)			0.384		0.037		0.109

Notes: Each column represents one LPM or OLS regression. The observations include students in the 2021/22–2024/25 cohorts who responded to a follow-up survey conducted at the end of the first academic year. Panel A analyzes effects on psychological distress measured with the 4-item PHQ scale for anxiety and depression. Column 1 examines the continuous distress score (on a scale of 0 to 12), and Columns 2–3 examine the likelihood of having moderate or severe distress (i.e., the score being above 6). Panel B analyzes gender role attitudes measured with the same four questions as in Table 6. Columns 4–5 examine the average agreement with traditional family gender roles, while Columns 6–7 examine agreement with gender roles for educational attainment and business leadership—the agreement score is standardized to have mean 0 and SD 1. All regressions control for cohort-by-program fixed effects, baseline characteristics, and in Panel B, the baseline level of the dependent variable. Standard errors in parentheses are clustered at the group level; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Attrition in survey responses: Given that the non-academic outcomes are observable only for students who responded to the endline or follow-up survey, there may be bias due to endogenous survey response. To mitigate this concern, I conduct four sets of tests. First, I show that the gender composition of the assigned group does not significantly affect the likelihood of men or women responding to the endline or follow-up survey (see Appendix Table A.2). Second, I apply inverse probability weighting (IPW) when estimating the effects on survey-measured outcomes. Appendix Figure A.12 shows that the estimated effects on all well-being and attitude outcomes remain largely unchanged. Third, I conduct sensitivity tests under different assumptions about the missing values of students in homogeneous (control) and mixed-gender (treatment) groups. Following prior work (Kling and Liebman 2004; Blattman et al. 2020), I estimate the effects when assuming that the missing values take the mean ±0.05 SD or ±0.1 SD of the outcome. Appendix Table A.16 shows that assuming a 0.1 SD gap between control and treatment students with missing values does not change the conclusions. When imposing a more extreme assumption that non-responding students in the control versus treatment groups have a 0.2 SD gap in the outcome, I find that the attenuated

effects retain the same sign and mostly remain significant at the 5% level. Finally, I estimate the Lee (2009) bounds for the main effects on well-being and gender role attitudes. Appendix Table A.17 shows that the derived upper and lower bounds are relatively narrow and consistently carry the same sign as the original estimates.

5 Mechanisms

The results so far provide robust and consistent evidence that increasing gender diversity within peer groups significantly improves student performance and well-being, while also inducing more traditional attitudes toward family gender roles. To shed light on these strong yet mixed effects of gender diversity, I use rich survey data and randomization designs to explore the underlying mechanisms. In Section 5.1, I examine the roles that men and women assume in group interaction, and whether mixed-gender groups induce more gender-divided roles. This role sorting may strengthen traditional gender norms but also facilitate collaboration among group members. Section 5.2 thus tests how gender diversity affects patterns of interaction within the group and broader social networks. Lastly, Section 5.3 tests whether gender diversity improves performance through skill diversity or other observed dimensions of diversity.

5.1 Gendered social roles

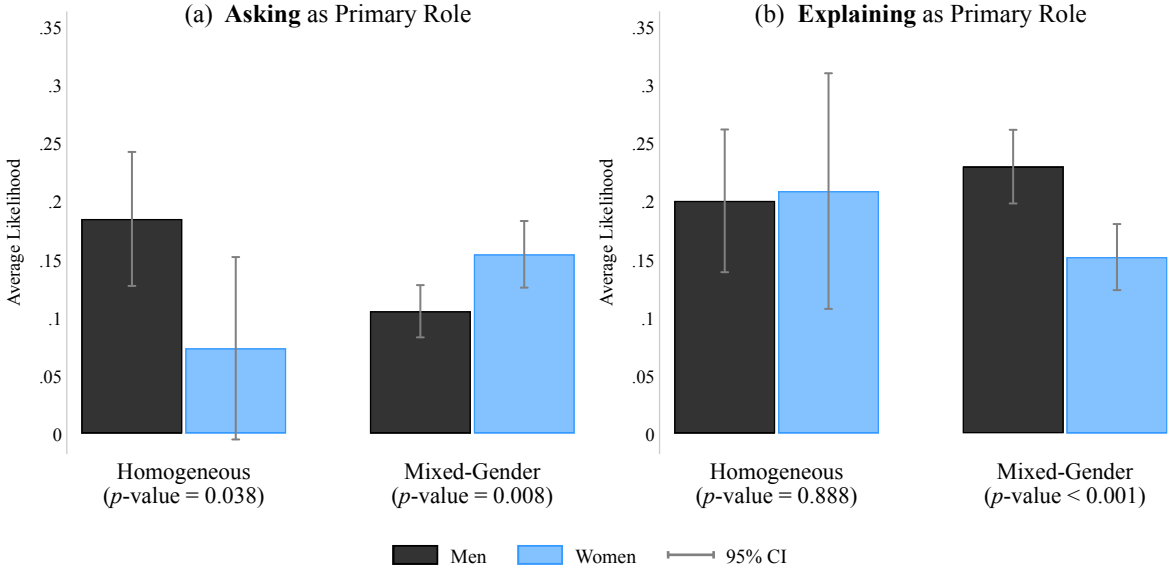
A large literature in social psychology distinguishes between agentic and communal traits as two core dimensions of social behavior (Bakan 1966; Abele and Wojciszke 2007). Agentic behavior emphasizes assertiveness, independence, and the pursuit of achievement and status—commonly expressed through actions such as taking charge, competing for influence, and self-promotion. Communal behavior, by contrast, emphasizes cooperation, connection, and concern for others—expressed through warmth, relationship maintenance, and providing support. Social role theory further highlights systematic gender differences in these roles: men are more strongly associated with agentic traits, whereas women are more strongly associated with communal traits (Eagly et al. 2000; Eagly and Wood 2012). Consistent with these insights, economics research documents robust evidence of gender differences in competition, leadership behavior, and self-promotion (Gneezy et al. 2003; Niederle and Vesterlund 2007; Alan et al. 2020; Eckel et al. 2021; Exley and Kessler 2022). The coexistence of agentic and communal roles in social contexts may generate interpersonal complementarity, which facilitates coordination and collaboration (Helgeson 1994).

Asking versus explaining: I first investigate what roles students primarily assume in group interaction, using a simple question in the endline survey: "When studying with others, what was your primary role in the group?" Students can choose from several options, including explaining something to others, listening to others, asking questions, discussing something with others, and

ensuring a good atmosphere.[25] Among these roles, clear gender differences emerge in the asking versus explaining roles: while women more often report asking questions, men more often report explaining course material to others (see Appendix Figure A.17). In a group setting, these two roles are associated differently with agentic and communal traits. Members who predominantly explain may demonstrate competence and expertise, potentially raising their status relative to others. In contrast, members who predominantly ask questions may acknowledge their own vulnerability and facilitate others' participation. Thus, the observed gender differences in asking and explaining appear broadly consistent with social role theory.

Importantly, these gender differences are not visible in homogeneous groups but are particularly salient in mixed-gender groups. Figure 3 shows that in homogeneous groups, women are actually less likely than men to report asking questions as their primary role, and there is no significant gender gap in the explaining role. In mixed-gender groups, however, women are 4.9 percentage points (or 47%) more likely than men to primarily ask questions, and men are 7.8 percentage points (or 51%) more likely to primarily explain course material to others. These gender gaps in asking and explaining are precisely estimated ($p = 0.008$ and $p < 0.001$).

Figure 3. Gender Gaps in Asking versus Explaining in Homogeneous and Mixed-Gender Groups



Notes: The figure plots the likelihood that a student reports "asking questions" or "explaining something to others" as their primary role in group interaction. The dependent variables on the y-axis are residualized by controlling for cohort-by-program fixed effects and all student characteristics observed at baseline. The residuals are then shifted by adding back the sample mean of the raw variable. The p-values indicate the significance of gender differences.

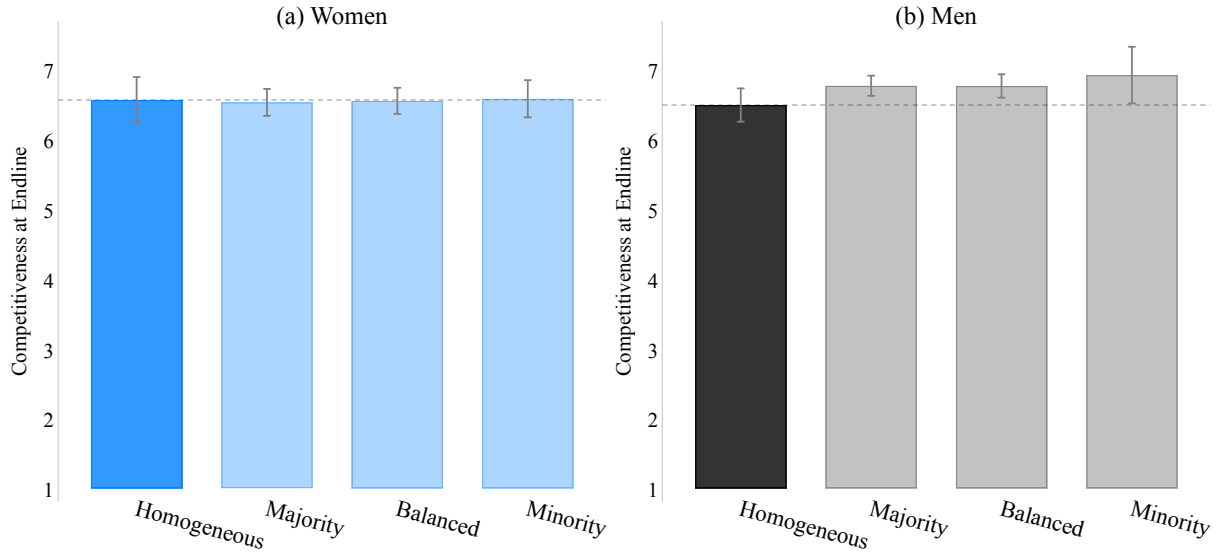
General competitiveness: Prior work shows that gender differences in competitiveness are more pronounced in mixed-gender rather than single-gender tournaments (Gneezy et al. 2003; Booth and

[25]The question also includes an open-ended text entry option. About 10% of students choose this option and provide responses such as "I have multiple roles" and "I do not have a typical role." The analysis focuses on students who select one of the five provided options.

Nolen 2012). Nevertheless, it remains unclear whether the gender composition of collaborative peer groups affects students' general preference for competition. To test this, I measure baseline and endline competitiveness using a survey question adapted from Buser et al. (2024): "In general, how competitive do you consider yourself?" Students respond on a scale ranging from 1 (not competitive at all) to 10 (very competitive).

Figure 4 plots competitiveness measured at the end of the first semester, separately by student gender and group gender composition. Panel (a) shows that women's endline competitiveness is very similar across different gender compositions. By contrast, Panel (b) shows that men assigned to mixed-gender groups report significantly higher competitiveness than men in homogeneous groups. Overall, assignment to mixed-gender groups raises men's competitiveness by about 0.3 points on a 1–10 scale or 0.16 SD (see Appendix Table A.18).

Figure 4. Endline Competitiveness by Group Gender Composition



Notes: The figure plots the residualized endline competitiveness—after controlling for baseline competitiveness and cohort-by-program fixed effects—by student gender and group gender composition. The residuals are further shifted by adding a constant to have the same mean as the raw measure. Error bars indicate 95% confidence intervals.

Coordinator role: After study groups were assigned, one student in each group was randomly selected to serve as the group coordinator. The coordinators were asked to voluntarily help create a chat group and coordinate group activities. As these tasks align with communal traits, women may take on the coordinator role more naturally and smoothly. I employ the coordinator assignment to test whether mixed-gender groups with a female coordinator perform better than mixed-gender groups with a male coordinator.

Column 1 of Table 8 compares mixed-gender groups with female or male coordinators against homogeneous groups. The results show that mixed-gender groups with a male coordinator improve economics grades by 0.11 SD, while mixed-gender groups with a female coordinator improve grades

by 0.17 SD. The difference between these effects is marginally significant ($p = 0.097$). Columns 2–3 show that this pattern applies similarly to both women and men. Column 4 shows that when excluding group coordinators themselves, the effect of coordinator gender is larger and more precisely estimated ($p = 0.045$). This means that female coordinators have stronger positive externalities on other group members' performance than male coordinators. The specification in Column 5 further restricts the sample to mixed-gender groups and includes fixed effects for group gender composition. The results show that assignment to a female rather than male coordinator raises grades by 0.11 SD ($p = 0.014$) when comparing mixed-gender groups with the same composition. These results provide robust evidence that the performance gains of mixed-gender groups are more concentrated in those with female coordinators, suggesting that women are on average more effective than men in the coordinator role.

Table 8. Effects of Mixed-Gender Groups on Economics Performance by Coordinator Gender

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Standardized Grade in Economics				
	All	Women	Men	Non-Coordinator	Mixed Groups
Mixed-Gender Group	0.174***	0.227**	0.186***	0.185***	0.111**
× Female Coordinator	(0.053)	(0.091)	(0.069)	(0.064)	(0.045)
	[0.001]	[0.013]	[0.008]	[0.004]	[0.014]
Mixed-Gender Group	0.106**	0.139	0.114*	0.089	
× Male Coordinator	(0.052)	(0.105)	(0.065)	(0.064)	
	[0.041]	[0.185]	[0.079]	[0.163]	
Coordinator	-0.007	-0.032	0.011		-0.003
	(0.043)	(0.079)	(0.064)		(0.045)
	[0.865]	[0.682]	[0.864]		[0.951]
Observations	2,812	1,206	1,606	2,109	2,328
R-squared	0.036	0.024	0.021	0.038	0.036
Female = Male (p-value)	0.097	0.234	0.247	0.045	–

Notes: Each column represents one OLS regression. The sample includes students in the 2019/20–2024/25 cohorts, since the first cohort did not have a coordinator assignment. *Female/Male Coordinator* indicates whether a mixed-gender group is coordinated by a woman/man. *Coordinator* is a binary indicator equal to one if a student is the assigned coordinator of their group. Column 4 restricts the sample to students who are not coordinators; Column 5 restricts the sample to students in mixed-gender groups. All regressions control for cohort-by-program fixed effects, and if applicable, student gender. Column 5 additionally controls for gender composition fixed effects (i.e., dummies for the number of women in a group). The bottom row shows the p-values indicating the significance of differences between female and male coordinators. Standard errors in parentheses are clustered at the group level; p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Taken together, the results in this section document consistent patterns of gender differences in social roles. While women tend to assume communal roles, such as asking questions and coordinating group activities, men tend to assume agentic roles, such as explaining and competing. These differences are especially pronounced in mixed-gender groups, implying that students sort into gender-congruent roles when engaging in cross-gender interactions. This role sorting may strengthen traditional norms about what men and women "should" do—as documented in Section 4.4.2. At the same time, the complementarity of social roles between men and women may reduce frictions and

enhance collaboration within groups, which could further improve learning outcomes, strengthen students' sense of belonging, and improve overall well-being.

5.2 Social interaction and group dynamics

Prior studies suggest that cross-gender interaction can be hampered by homophily (McPherson et al. 2001) or greater communication costs (Lazear 1999; Lyons 2017). On the other hand, mixed-gender groups may collaborate effectively if men and women take on complementary roles. Given these opposing channels, I use rich survey data to test how gender diversity affects students' interaction with group members as well as networking behavior beyond the group.

5.2.1 Interactions within the group

Interaction frequency: I first examine how group gender diversity affects the frequency of group-level and pair-level interactions among group members. Figure 5(a) shows how group-level activities vary with gender composition, using administrative records of voucher collection and survey measures of joint activities. The figure reveals a consistent pattern: mixed-gender groups—especially balanced groups—are more likely to collect vouchers for the icebreaker activity, have a WhatsApp group, have any joint meetings, and meet together at least three times.[26]

Besides joint group activities, students may also interact one-on-one to varying degrees. I elicit the frequency of activities that students had with each assigned peer on a 6-point Likert scale: never, less than once per month, once per month, two to three times per month, once a week, and multiple times per week. Students report the frequency separately for learning and extracurricular activities. For simplicity, I treat the reported frequency as a continuous variable and calculate the average frequency of academic and extracurricular activities as an overall measure of pairwise interactions.[27]

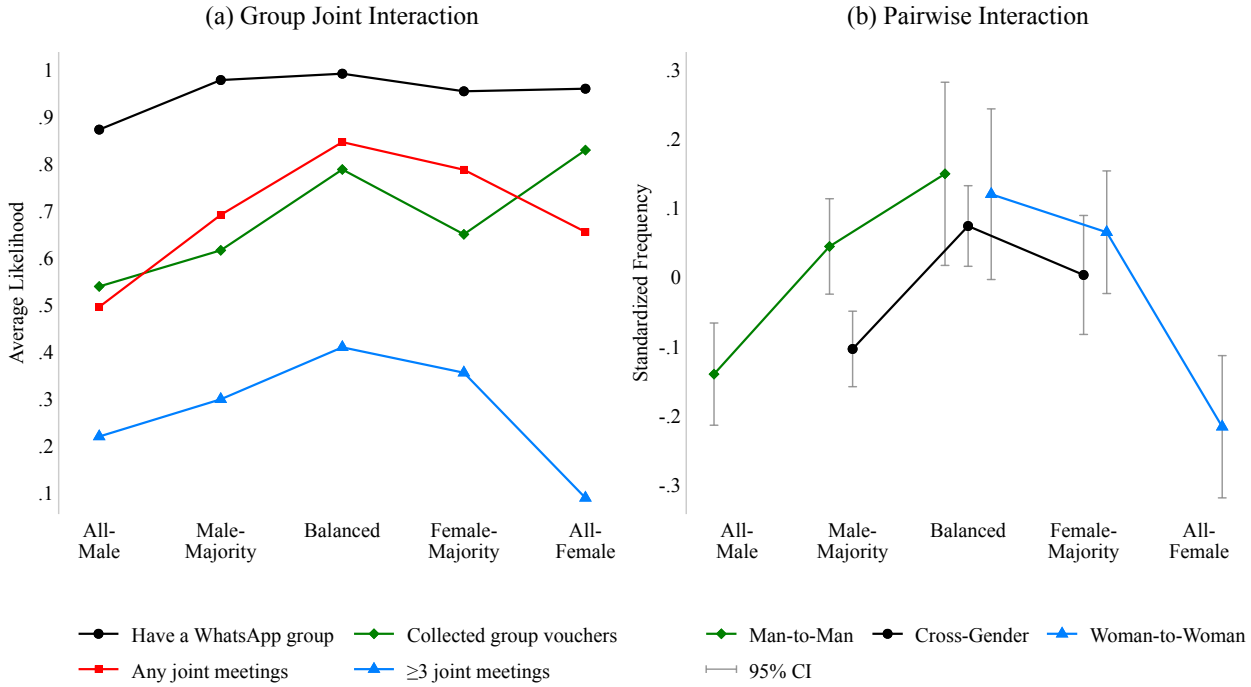
Figure 5(b) shows that the overall intensity of pairwise activities also increases with gender diversity: interactions are least frequent in homogeneous groups, more frequent in imbalanced groups, and most frequent in balanced groups. This pattern applies uniformly to male–male, female–female, and male–female pairs. Within mixed-gender groups, cross-gender interactions are generally less active than same-gender interactions, reflecting gender homophily in social networks.

[26] Appendix Table A.19 reports regression results corresponding to the figure. The estimated effects of the binary mixed-gender indicator and the continuous gender diversity measure are uniformly positive in direction and mostly reach statistical significance at the 10% or 5% level.

[27] Specifically, each student answers two sets of questions in the endline survey: (1) "How often have you studied for [course name] together with each of your study mates? For example, you can attend tutorials together, work on problem sets and old exams, go through lecture slides together, etc."; and (2) "How often have you joined social events together with each of your study mates? For example, you can do sports, watch movies, have drinks and dinner together, etc." The survey questions focus on meaningful meetings but exclude online or low-intensity interactions, such as communications through chat groups, social media, emails, or brief encounters on campus. Therefore, the measured intensity of pairwise interactions should represent a lower bound. Comparing the responses between pairs of group members, I find that over 90% of pairs report frequencies that are the same or just one level apart (for both types of activities). This suggests that students are generally truthful when reporting their interaction with peers.

Nevertheless, cross-gender interactions become more frequent when moving from imbalanced to balanced groups. Overall, these results highlight that increasing gender diversity strengthens peer-to-peer connections within the group.

Figure 5. Frequency of Group-Level and Pair-Level Interactions by Gender Composition



Notes: Panel (a) plots the likelihood of study groups engaging in different group activities by gender composition: having a WhatsApp group; collecting the group activity vouchers; meeting jointly at least once; and meeting jointly at least three times. Voucher collection was recorded by the research team and therefore observed for all groups, and other activities are reported by students in the 2023/24–2024/25 cohorts. Panel (b) shows the standardized frequency of meaningful pairwise interactions within the group, using pairwise observations ($N = 5,504$), separately for same-gender and cross-gender pairs (see footnote 27 for details). Error bars indicate 95% confidence intervals.

Regression results in Appendix Table A.20 show that the overall frequency of pairwise interaction is 0.24 SD higher in mixed-gender groups than in homogeneous groups ($p < 0.001$). This effect is mainly driven by learning activities but is also visible for extracurricular activities. Increasing gender diversity also significantly raises the frequency of a student's interaction with the closest peer in the group (i.e., the maximum frequency across peers). Furthermore, Appendix Figure A.14 reports largely consistent results when estimating the effects on the likelihood of having different frequency levels of academic or extracurricular activities with group members.

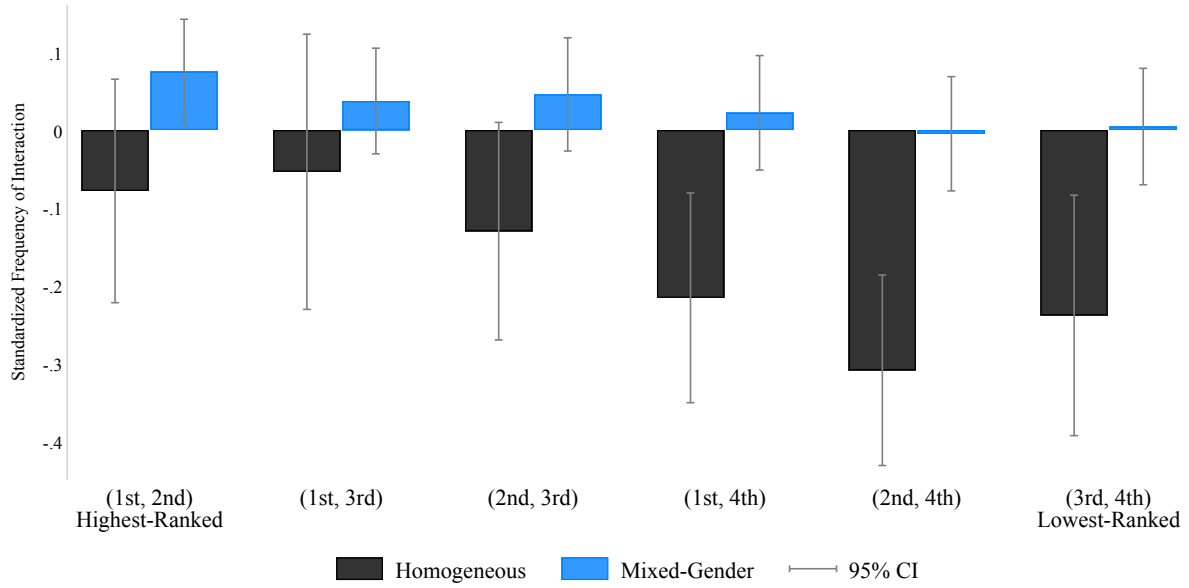
Interaction by performance rank: To capture potential social learning or knowledge spillover within groups, I decompose pairwise interactions by students' performance rank, with the top performer in a group ranked first. Figure 6 shows that in homogeneous groups, the interaction intensity is substantially lower for pairs including the fourth-ranked member—i.e., pairs labeled as (1st, 4th), (2nd, 4th), and (3rd, 4th). This means that the lowest-performing members are more isolated than higher-performing members in homogeneous groups. By contrast, in mixed-gender

groups, the lowest-performing members are similarly engaged in social interaction as other group members. These patterns remain largely consistent when using the rank of *predicted* rather than actual performance (see Appendix Figure A.15).

Regression results in Appendix Table A.20 show that the positive effect of mixed-gender groups on pairwise interaction is approximately 0.13 SD larger for pairs involving the lowest-performing members. When further splitting these pairs by the gender composition of two students in a pair, Appendix Figure A.16 shows that mixed-gender groups increase the social engagement of both male and female lowest-performing members, and their interactions with higher-ranked men and women are similarly active.

These results provide robust evidence that gender-diverse groups collaborate more closely and exhibit more inclusive social dynamics. In particular, the academically struggling members in these groups are more socially engaged, which can facilitate knowledge sharing, peer mentoring, or role model effects. Consistent with these patterns, Section 5.3 below shows that the performance effects of gender diversity are especially strong for low-performing group members, and the within-group performance inequality is substantially narrower in mixed-gender groups.

Figure 6. Frequency of Pairwise Interaction by Performance Rank within a Group



Notes: The figure plots the standardized frequency of pairwise interactions by the within-group rank of final grade in economics. For example, (1st, 2nd) denotes pairs of students ranked first and second within a group, while (3rd, 4th) denotes pairs ranked third and fourth.

5.2.2 Broader social network

After showing that gender diversity fosters more active and inclusive group dynamics, I further investigate whether these effects translate into changes in the broader social network. To this end, I ask students to report the study partners with whom they interacted most frequently during the

semester. They can voluntarily list up to five students from the entire class, covering nearly the full cohort. This question was administered to the 2021/22–2024/25 cohorts, and 761 students (40%) answered the question and listed at least one study partner.[28]

Panel A of Table 9 shows that the total number of listed study partners does not differ significantly between students in mixed-gender groups and those in homogeneous groups. Panel B examines whether the listed partners are from the assigned group ("in-group") or outside the group ("out-group"). Overall, 36% of students report at least one assigned peer as their closest study partner. This means that although most students eventually form broader connections, the assigned groups help some students form close study partnerships. This is especially true for students assigned to mixed-gender groups. Columns 2–4 show that assignment to mixed-gender groups significantly raises the proportion of in-group partners by 9.4 percentage points (or 47%) and the likelihood of listing at least one in-group partner by 12.7 percentage points (or 35%).

Panel C of Table 9 examines the gender composition of listed study partners. Overall, only 29% of the reported partners are of the opposite gender, consistent with strong gender homophily documented among college students in the U.S. (Jackson et al. 2022). Nevertheless, Columns 5–6 show that assignment to mixed-gender groups significantly reduces homophily: the share of opposite-gender partners increases by 12.5 percentage points, and the likelihood of having any opposite-gender partner increases by 24.7 percentage points ($p < 0.001$). Importantly, this does not reduce the chance of having any same-gender partners (see Column 7). Moreover, the reported opposite-gender study partners are not confined to the assigned group: the likelihood of reporting at least one opposite-gender partner outside the group also directionally increases (see Column 8).

Taken together, the results in this section highlight the influence that study groups can have on freshmen's social networks. Because students typically enter university without established peer connections, the assigned group provides a starting point for finding study partners and expanding social ties. The influence is particularly strong for mixed-gender groups, where students interact frequently and inclusively with each other. This helps explain why these groups can generate persistent effects on students' well-being, attitudes, and learning outcomes. Prior work shows that romantic relationships or marriage incentives can mediate peer effects and shape behavior in social environments (Hill 2015; Bursztyn et al. 2017). These forces could also contribute to improved social dynamics in mixed-gender groups, but I do not have enough data to empirically test this.[29]

[28] On average, students listed 3.14 study partners, and 26% listed five partners. Reassuringly, I find no evidence that group gender composition affects the likelihood of men or women answering the question. To reduce measurement error, the survey question was coded to prompt full student names after the entry of a few characters.

[29] For the first cohort in 2018/19, I asked students in the endline survey whether they were in a romantic relationship, and, if so, whether their partner was from the assigned group. Out of the 155 students who answered this question, 45% reported being in a romantic relationship, but none reported that their partner was from the assigned group. Appendix Table A.21 shows that the effects of group gender diversity do not significantly differ by relationship status. Given the small sample size, these results are only suggestive.

Table 9. Effects of Mixed-Gender Groups on Self-Reported Study Partners

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A.	B. In-Group vs. Out-Group			C. Same-Gender vs. Opposite-Gender			
	Number of study partners	Share of in-group partners	Any in-group partners	Any out-group partners	Share of opposite-gender partners	Any opposite-gender partners	Any same-gender partners	Any out-group opp-gender partners
Mixed-Gender Group	0.145	0.094***	0.127**	-0.044***	0.125***	0.247***	-0.006	0.071
	(0.139)	(0.033)	(0.061)	(0.016)	(0.026)	(0.049)	(0.023)	(0.049)
	[0.297]	[0.004]	[0.038]	[0.006]	[<0.001]	[<0.001]	[0.790]	[0.145]
Female	-0.286***	0.023	0.052	-0.000	0.092***	0.112***	-0.040**	0.114***
	(0.106)	(0.023)	(0.036)	(0.017)	(0.023)	(0.037)	(0.019)	(0.037)
Observations	761	761	761	761	761	761	761	761
R-squared	0.022	0.063	0.051	0.028	0.078	0.071	0.035	0.035
Dep. Var. Mean	3.142	0.202	0.360	0.941	0.287	0.566	0.933	0.419

Notes: The table examines the number and composition of close study partners reported by students in cohorts of 2021/22–2024/25. In the endline survey, students voluntarily list up to five classmates with whom they interacted most frequently during the semester. Each column represents one OLS regression or LPM. Panel A examines the effects of mixed-gender groups on the total number of study partners listed. Panel B examines whether listed partners are from the assigned study group ("in-group") or from outside the group ("out-group"). The dependent variables include the share of listed partners being from the group, and indicators for having any in-group or any out-group partner. Panel C examines the gender composition of reported study partners. The dependent variables include the share of opposite-gender partners and indicators for having any opposite-gender partner, any same-gender partner, and any opposite-gender partner from outside the assigned group. All regressions control for cohort-by-program fixed effects. Standard errors are in parentheses and clustered at the group level; p-values are in brackets. $^{*}p<.1$, $^{**}p<.05$, $^{***}p<.01$.

5.3 Different diversity dimensions

The diversity literature highlights a key mechanism through which gender diversity can enhance creativity and productivity: gender-diverse groups bring together a broader range of skills and knowledge. To test this mechanism, I employ the experimental design that also exogenously varied the skill composition of study groups. This allows me to directly measure group skill diversity and test whether it contributes to the estimated gender-diversity effects.

I first present more detailed results on how gender diversity affects group members' performance. In Appendix Table A.22, I use 765 group-level observations to estimate the effects of gender diversity on four aggregate performance measures: group average grade, minimum grade, maximum grade, and within-group grade inequality. Consistent with the student-level analysis, I find that gender diversity significantly raises the average grade obtained by group members. Interestingly, this average effect is primarily driven by increased performance at the bottom rather than at the top: while leaving the group maximum grade unaffected, gender diversity raises the group minimum grade by about 0.39 points on a 1–6 scale, or 0.33 SD ($p = 0.001$). This *lifted floor*, combined with a *stable*

ceiling, in turn leads to a 0.30 SD reduction in performance inequality within groups ($p = 0.002$).[30] Table A.22 further shows that these effects of gender diversity are robust to controlling for measures of diversity in cognitive skill (e.g., high school math and language grades) and noncognitive skill (e.g., personality traits and competitiveness), as well as diversity in other observables (e.g., age, parental background, and gender attitudes).[31] These results suggest that the performance effects of gender diversity are unlikely to be driven by skill diversity or other observable dimensions of diversity.

Next, I examine whether skill diversity affects performance in similar ways as gender diversity. I construct aggregate indicators of cognitive and noncognitive skills and then compute their within-group standard deviations as diversity measures.[32] For comparison, I also estimate the effects of nationality diversity, which is commonly examined in the literature. Since nationality is a categorical variable, I use one minus the Herfindahl concentration index to measure nationality diversity.[33] Figure 7 plots the estimated effects of gender, skill, and nationality diversity—all rescaled to range from 0 to 1—on different performance outcomes at the group level. The first row reproduces the effects of gender diversity reported in Table A.22.

The second row of Figure 7 shows that diversity in cognitive skill does not affect average performance but has opposite effects on top and bottom performers within groups: while the group minimum grade significantly decreases, the maximum grade increases. Accordingly, the performance gaps within groups significantly increase with cognitive skill diversity. The third and fourth rows of Figure 7 show that noncognitive skill diversity and nationality diversity have similar effects on performance. They both decrease the group minimum grade and enlarge within-group performance gaps. In addition, nationality diversity significantly reduces group members' average performance, consistent with prior work documenting negative effects of ethnic or nationality diversity on workplace output (Hjort 2014; Lyons 2017; Marx et al. 2021).[34]

[30] Appendix Figure A.13 shows the effect of mixed-gender groups separately for grades ranked first (highest), second, third, and fourth (lowest) within groups. The results show that the effect monotonically increases in magnitude as the within-group rank drops. This suggests that the relatively lower-performing group members benefit more from gender diversity.

[31] I calculate the within-group standard deviation of each characteristic as the diversity measure. In addition to the diversity measures, the analyses also control for the average value of observable characteristics—to capture both diversity effects and level effects. For binary attributes such as course-retaking status, I include only the diversity measure, because the standard deviation and the average are mechanically correlated.

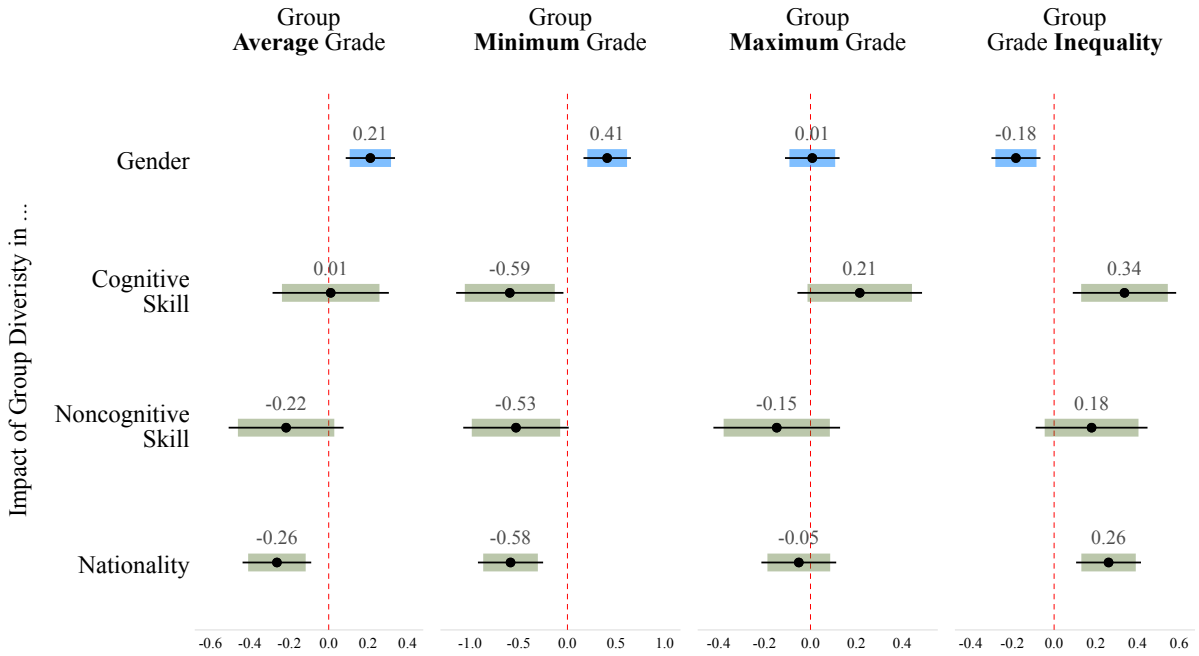
[32] Cognitive skill is measured by the predicted final grade based on high school math and language grades, speaking German as first language, and course-repeating status. Noncognitive skill is measured by the predicted grade based on competitiveness and the Big Five personality traits. I estimate the prediction model using the non-experimental sample of students without study groups and apply it to students in the experimental sample; results are very similar when I instead fit the model using the experimental sample.

[33] Specifically, nationality diversity is calculated as $1-\sum_k s_k^2$, where s_k denotes the share of students with nationality k in a group. Nationality information is obtained from administrative records and is available for 99% of students in the sample. Overall, 22% of students are international, representing a wide range of countries. For simplicity, nationalities with less than 5 observations are combined into a single category—the results are quite robust to this threshold.

[34] The effects of skill diversity also echo Carrell et al. (2013), who show that ability-diverse groups do not necessarily engage in beneficial knowledge sharing; instead, high-ability and low-ability students separate and form homogeneous subgroups.

Collectively, these results highlight that gender diversity has distinctive effects on group members' performance compared to skill or nationality diversity. While gender diversity significantly benefits the lowest-performing group members, other forms of diversity tend to harm them. Gender diversity also narrows performance inequality within groups, while other forms of diversity tend to widen it. Understanding why different dimensions of diversity have such divergent effects provides an important avenue for future research.

Figure 7. Effects of Gender, Skill, and Nationality Diversity on Group-Level Performance Measures



Notes: The figure shows how different dimensions of diversity affect aggregate measures of performance in the economics course, using group-level observations. The dependent variables use raw measures of grades on a scale of 1 to 6. See footnotes 32 and 33 for details on how skill diversity and nationality diversity are measured. All diversity measures are rescaled to range from 0 to 1. Each estimate is derived from one OLS regression, which controls for cohort-by-program fixed effects and uses robust standard errors. When estimating the effects of gender diversity or skill diversity, the regressions additionally control for the proportion of women and group average skill, respectively. Error bars indicate 95% and 90% confidence intervals.

6 Discussion

6.1 Are study groups beneficial?

The analyses in this paper focus on students who voluntarily participate in the experiment and are randomly assigned to a study group. Because participation is self-selected, I cannot causally identify the effect of having a study group per se on academic performance. That said, I present *descriptive* evidence on how study groups correlate with performance.

Appendix Figure A.18 shows that for both men and women, participating in the experiment and having a study group correlates with higher grades in economics. Overall, having a study

group is associated with a 0.15 SD increase in the final grade ($p < 0.001$). When controlling for all student characteristics observed at baseline, the estimated benefit remains sizable and statistically significant, corresponding to a 0.07 SD increase ($p = 0.004$). Although these estimates may still reflect selection on unobservables, they provide suggestive evidence that having a study group can be beneficial for student performance.

6.2 External validity

This paper uses a large field experiment to estimate the causal impact of gender diversity on student outcomes. While the findings are internally valid for the experimental sample, the extent to which they apply to the larger student population in the setting and generalize to other settings remains unclear. I assess the external validity of the findings along these two dimensions.

6.2.1 Generalizability beyond the experimental sample

To assess whether the experimental estimates generalize beyond the students who opted into the experiment, I implement the sensitivity tests in Andrews and Oster (2019). The idea is that participants may differ from the broader student population in ways that also matter for the treatment effect of group gender diversity. This approach estimates heterogeneity of the average treatment effect (ATE) within the experimental sample and combines this with differences in baseline characteristics between participants and the broader target sample to form an "observable selection bias" correction. This correction is then scaled by a sensitivity parameter Ψ to allow for selection on unobservables. I compute the implied ATE for the target sample at $\Psi = 1$ (when selection on unobservables is as strong as selection on observables), bounds for $\Psi \in [1, 2]$, as well as the value of $\Psi(0)$, which would make the target ATE equal to zero. These analyses are conducted among all students—with or without a random study group—who completed the baseline survey and therefore have a large set of observables.

The results in Appendix Table A.23 show that across all primary outcomes analyzed in the paper, the estimated external-validity adjustments are modest and do not change the qualitative conclusions. For economics grades, the baseline ATE among participants—who have completed the baseline survey—is 0.115 SD, and the observable-adjusted estimate is even slightly larger: 0.129 SD, with bounds [0.129, 0.143] corresponding to $\Psi \in [1, 2]$. The value of $\Psi(0)$ is -8.3, implying that selection on unobservables would have to operate in the opposite direction and be substantially more important than selection on observables to deliver an ATE of zero in the population. For program dropout, the baseline effect (6.9 percentage points) is also quite similar to the observable-adjusted ATE (6.0 percentage points), with similarly tight bounds. The same pattern applies to survey-measured well-being and family gender role attitudes in the short and longer run. The adjusted ATEs for these outcomes are closely aligned with the baseline effects, the $\Psi \in [1, 2]$ bounds are quite tight, and the absolute values of $\Psi(0)$ are sizable (ranging from 20 to 653)—meaning that the ATE is very unlikely to be zero for the broader student population.

6.2.2 Generalizability beyond the male-dominated setting

I further examine whether the results of this paper are confined to the baseline male-dominated setting or possibly generalize to other environments. I conducted small-scale replications of the experiment in female-dominated settings during fall 2024. In total, 324 students were randomly assigned to 81 study groups across two mathematics courses at the same university. These courses primarily serve students majoring in biomedicine, biology, and biochemistry. As these fields are female-dominated, women make up approximately 65% of enrolled students. The gender gap in performance is also reversed in these courses: women outperform men by about 0.15 SD.

Despite these contextual differences, the performance effects of gender diversity are broadly consistent with those observed in the baseline experiment. Appendix Table A.24 shows that assignment to mixed-gender groups directionally increases the group average grade. This effect is driven by a substantial increase in the group minimum grade, equivalent to 0.75 points on a 1–6 scale, or roughly 0.50 SD ($p = 0.060$). Performance inequality within mixed-gender groups also declines by about 0.50 SD, an effect marginally significant ($p = 0.055$). These results are qualitatively and quantitatively similar when using the continuous measure of group gender diversity. Given the limited sample size, these estimates should be interpreted with caution. Nevertheless, they provide suggestive evidence that the benefits of gender-diverse groups extend beyond male-dominated or male-advantaged environments.

7 Conclusion

Modern organizations and firms increasingly rely on teams and groups to organize learning, research, and production. In educational settings, teachers often assign students to project groups, classmates naturally form study groups to work on problem sets and prepare for exams, and many institutions create orientation groups to help students adapt to new academic and social environments. In workplaces, employees are routinely assigned to functional teams, task forces, committees, or self-managing teams to coordinate tasks and make joint decisions. Social groups based on shared hobbies or interests are also common in both contexts. A natural question for educators, employers, and policymakers is whether increasing gender diversity within these groups benefits individual performance and well-being.

I investigate this question using a large-scale field experiment that randomly assigns 3,060 students to homogeneous or mixed-gender study groups at university entry. I find robust evidence that greater gender diversity within these groups significantly raises course performance throughout the first year and lowers the program dropout rate by the second year. Beyond academic benefits, assignment to mixed-gender groups also improves students' subjective well-being in the short and longer run. However, cross-gender contact in these groups leads students to adopt more traditional attitudes toward family gender roles. This attitudinal change is observed for both men and women in the short run and persists longer for men.

Evidence from rich survey data points to gendered social roles as a key mechanism underlying these strong yet mixed effects of gender diversity. In social interaction, men tend to assume agentic roles, such as explaining and competing, whereas women tend to assume communal roles, such as coordinating group activities and asking questions. These gender differences are especially pronounced in mixed-gender groups, suggesting that students sort into gender-stereotypical roles when interacting with the opposite gender. This role sorting may foster more active and inclusive group dynamics, while also reinforcing traditional gender norms.

As one of the first studies providing causal evidence that gender diversity raises performance and well-being, an important consideration is the generalizability of the results. I show that the effects of gender diversity apply to students with different baseline characteristics, remain robust to including a wide range of group-composition controls, and appear consistent across seven cohorts. Moreover, similar patterns emerge in both male-dominated and female-dominated environments. Although this paper focuses on higher education, the underlying mechanisms may also apply to collaborative groups in workplace contexts, providing a promising avenue for future research.

The findings offer practical guidance for educators and organizations. When forming peer groups, using gender-diverse rather than homogeneous compositions can improve social dynamics, productivity, and subjective well-being. Nevertheless, these benefits may come with unintended changes in gender-typed behaviors and gender attitudes. Understanding the welfare implications of these changes requires further work on how gender diversity affects long-term outcomes, such as labor market trajectories, family decisions, and life satisfaction.

References

Abele, Andrea E, and Bogdan Wojciszke. 2007. "Agency and communion from the perspective of self versus others." *Journal of Personality and Social Psychology*, 93(5): 751.

Adams, Renée B, and Daniel Ferreira. 2009. "Women in the boardroom and their impact on governance and performance." *Journal of Financial Economics*, 94(2): 291–309.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay. 2021. "Building social cohesion in ethnically mixed schools: An intervention on perspective taking." *The Quarterly Journal of Economics*, 136(4): 2147–2194.

Alan, Sule, Seda Ertac, Elif Kubilay, and Gyongyi Loranth. 2020. "Understanding gender differences in leadership." *The Economic Journal*, 130(626): 263–289.

Allport, Gordon W. 1954. "The nature of prejudice."

Amini, Makan, Mathias Ekström, Tore Ellingsen, Magnus Johannesson, and Fredrik Strömsten. 2017. "Does gender diversity promote nonconformity?" *Management Science*, 63(4): 1085–1096.

Anderberg, Dan, Gordon B Dahl, Christina Felfe, Helmut Rainer, and Thomas Siedler. 2024. "Diversity and Cooperation." *Working Paper*.

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American statistical Association*, 103(484): 1481–1495.

Andrews, Isaiah, and Emily Oster. 2019. "A simple approximation for evaluating external validity bias." *Economics Letters*, 178: 58–62.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and services for college achievement: Evidence from a randomized trial." *American Economic Journal: Applied Economics*, 1(1): 136–163.

Auriol, Emmanuelle, Guido Friebel, Alisa Weinberger, and Sascha Wilhelm. 2022. "Underrepresentation of women in the economics profession more pronounced in the United States compared to heterogeneous Europe." *Proceedings of the National Academy of Sciences*, 119(16): e2118853119.

Babcock, Linda, Maria P Recalde, Lise Vesterlund, and Laurie Weingart. 2017. "Gender differences in accepting and receiving requests for tasks with low promotability." *American Economic Review*, 107(3): 714–747.

Bakan, David. 1966. "The duality of human existence: An essay on psychology and religion."

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann. 2020. "In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics." *Working Paper*.

Barrow, Lisa, Lashawn Richburg-Hayes, Cecilia Elena Rouse, and Thomas Brock. 2014. "Paying for performance: The education impacts of a community college scholarship program for low-income adults." *Journal of Labor Economics*, 32(3): 563–599.

Blattman, Christopher, Nathan Fiala, and Sebastian Martinez. 2020. "The long-term impacts of grants on poverty: Nine-year evidence from Uganda's youth opportunities program." *American Economic Review: Insights*, 2(3): 287–304.

Boisjoly, Johanne, Greg J Duncan, Michael Kremer, Dan M Levy, and Jacque Eccles. 2006. "Empathy or antipathy? The impact of diversity." *American Economic Review*, 96(5): 1890–1905.

Booij, Adam S, Edwin Leuven, and Hessel Oosterbeek. 2017. "Ability peer effects in university: Evidence from a randomized experiment." *The Review of Economic Studies*, 84(2): 547–578.

Booth, Alison, and Patrick Nolen. 2012. "Choosing to compete: How different are girls and boys?" *Journal of Economic Behavior & Organization*, 81(2): 542–555.

Booth, Alison, Lina Cardona-Sosa, and Patrick Nolen. 2018. "Do single-sex classes affect academic achievement? An experiment in a coeducational university." *Journal of Public Economics*, 168: 109–126.

Born, Andreas, Eva Ranehill, and Anna Sandberg. 2020. "Gender and willingness to lead: Does the gender composition of teams matter?" *The Review of Economics and Statistics*, 1–46.

Brailovskaia, Julia, and Jürgen Margraf. 2020. "How to measure self-esteem with one item? Validation of the German single-item self-esteem scale (G-SISE)." *Current Psychology*, 39(6): 2192–2202.

Brenøe, Anne Ardila, and Ulf Zölitz. 2020. "Exposure to more female peers widens the gender gap in STEM participation." *Journal of Labor Economics*, 38(4): 1009–1054.

Bursztyn, Leonardo, Georgy Egorov, and Robert Jensen. 2019. "Cool to be smart or smart to be cool? Understanding peer pressure in education." *The Review of Economic Studies*, 86(4): 1487–1526.

Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. 2017. "'Acting wife': Marriage market incentives and labor market investments." *American Economic Review*, 107(11): 3288–3319.

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek. 2024. "Can competitiveness predict education and labor market outcomes? Evidence from incentivized choice and survey measures." *Review of Economics and Statistics*, 1–45.

Butler, Judith. 1988. "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory." *Theatre Journal*, 40(4): 519–531.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, et al. 2016. "Evaluating replicability of laboratory experiments in economics." *Science*, 351(6280): 1433–1436.

Carrell, Scott E, Bruce I Sacerdote, and James E West. 2013. "From natural variation to optimal policy? The importance of endogenous peer group formation." *Econometrica*, 81(3): 855–882.

Carrell, Scott E, Mark Hoekstra, and Elira Kuka. 2018. "The long-run effects of disruptive peers." *American Economic Review*, 108(11): 3377–3415.

Chen, Jennjou, and Tsui-Fang Lin. 2023. "Gender Composition and Team Performance: Evidence from a Randomized Experiment." SSRN Working Paper 4462453. Posted May 29, 2023.

Chen, Jingnan, and Daniel Houser. 2019. "When are women willing to lead? The effect of team gender composition and gendered tasks." *The Leadership Quarterly*, 30(6): 101340.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *The Quarterly Journal of Economics*, 126(4): 1593–1660.

Clingingsmith, David, Asim Ijaz Khwaja, and Michael Kremer. 2009. "Estimating the impact of the Hajj: religion and tolerance in Islam's global gathering." *The Quarterly Journal of Economics*, 124(3): 1133–1170.

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov. 2021. "Gender stereotypes in deliberation and team decisions." *Games and Economic Behavior*, 129: 329–349.

Compte, Olivier, and Andrew Postlewaite. 2004. "Confidence-enhanced performance." *American Economic Review*, 94(5): 1536–1557.

Contreras, Valentina, Chiara Orsini, Berkay Özcan, and Johann Koehler. 2025. "Effects of team diversity on individual performance and voice: A field experiment of group composition by gender and language." *Labour Economics*, 102763.

Cools, Angela, Raquel Fernández, and Eleonora Patacchini. 2022. "The asymmetric gender effects of high flyers." *Labour Economics*, 79: 102287.

Coveney, Max, Pilar García Gómez, and Teresa Bago d'Uva. 2025. "Gender and performance in collaboration: Evidence from random student teams." Tinbergen Institute Discussion Paper.

Cullen, Zoe, and Ricardo Perez-Truglia. 2023. "The old boys' club: Schmoozing and the gender gap." *American Economic Review*, 113(7): 1703–1740.

Dahl, Gordon B, Andreas Kotsadam, and Dan-Olof Rooth. 2021. "Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams." *The Quarterly Journal of Economics*, 136(2): 987–1030.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya." *American Economic Review*, 101(5): 1739–1774.

Eagly, Alice H. 2013. *Sex differences in social behavior: A social-role interpretation.* Psychology Press.

Eagly, Alice H, and Wendy Wood. 2012. "Social role theory." *Handbook of Theories of Social Psychology*, 2(9): 458–476.

Eagly, Alice H, Wendy Wood, and Amanda B Diekman. 2000. "Social Role Theory of Sex Differences and Similarities: A Current Appraisal." *The Developmental Social Psychology of Gender*, 123.

Eckel, Catherine, Lata Gangadharan, Philip J Grossman, and Nina Xue. 2021. "The gender leadership gap: Insights from experiments." In *A research agenda for experimental economics*. 137–162. Edward Elgar Publishing.

Eisenkopf, Gerald, Zohal Hessami, Urs Fischbacher, and Heinrich W Ursprung. 2015. "Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland." *Journal of Economic Behavior & Organization*, 115: 123–143.

Exley, Christine L, and Judd B Kessler. 2022. "The gender gap in self-promotion." *The Quarterly Journal of Economics*, 137(3): 1345–1381.

Fenoll, Ainoa Aparicio, and Sarah Zaccagni. 2022. "Gender mix and team performance: Differences between exogenously and endogenously formed teams." *Labour Economics*, 79: 102269.

Finseraas, Henning, Torbjørn Hanson, Åshild A Johnsen, Andreas Kotsadam, and Gaute Torsvik. 2019. "Trust, ethnic diversity, and personal contact: A field experiment." *Journal of Public Economics*, 173: 72–84.

Gerlitz, Jean-Yves, and Jürgen Schupp. 2005. "Zur Erhebung der Big-Five-basierten persoenlichkeitsmerkmale im SOEP." *DIW Research Notes*, 4: 2005.

Gertler, Paul, Sebastian Galiani, and Mauricio Romero. 2018. "How to make replication the norm." *Nature*, 554(7693): 417–419.

Giuntella, Osea, Silvia Saccardo, and Sally Sadoff. 2024. "Sleep: Educational Impact and Habit Formation." *Working Paper*.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide.* Princeton, NJ:Princeton University Press.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. "Performance in competitive environments: Gender differences." *The Quarterly Journal of Economics*, 118(3): 1049–1074.

Goldin, Claudia. 2014. "A grand gender convergence: Its last chapter." *American Economic Review*, 104(4): 1091–1119.

Goldrick-Rab, Sara, Robert Kelchen, Douglas N Harris, and James Benson. 2016. "Reducing income inequality in educational attainment: Experimental evidence on the impact of financial aid on college completion." *American Journal of Sociology*, 121(6): 1762–1817.

Golsteyn, Bart HH, Arjan Non, and Ulf Zölitz. 2021. "The impact of peer personality on academic achievement." *Journal of Political Economy*, 129(4): 1052–1099.

Gong, Jie, Yi Lu, and Hong Song. 2019. "Gender peer effects on students' academic and noncognitive outcomes: Evidence and mechanisms." *Journal of Human Resources*, 0918–9736R2.

Greenberg, Kyle, Melanie Wasserman, and Anna Weber. 2024. "The Effects of Gender Integration on Men: Evidence from the U.S. Military." *Working Paper.*

Guarino, Cassandra M, and Victor MH Borden. 2017. "Faculty service loads and gender: Are women taking care of the academic family?" *Research in higher education*, 58(6): 672–694.

Hamilton, Barton H, Jack A Nickerson, and Hideo Owan. 2003. "Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation." *Journal of Political Economy*, 111(3): 465–497.

Hampole, Menaka, Francesca Truffa, and Ashley Wong. 2021. "Peer effects and the gender gap in corporate leadership: Evidence from MBA students." *Working Paper.*

Helgeson, Vicki S. 1994. "Relation of agency and communion to well-being: Evidence and potential explanations." *Psychological Bulletin*, 116(3): 412.

Herring, Cedric. 2009. "Does diversity pay?: Race, gender, and the business case for diversity." *American Sociological Review*, 74(2): 208–224.

Hill, Andrew J. 2015. "The girl next door: The effect of opposite gender friends on high school achievement." *American Economic Journal: Applied Economics*, 7(3): 147–177.

Hill, Andrew J. 2017. "The positive influence of female college students on their male peers." *Labour Economics*, 44: 151–160.

Hjort, Jonas. 2014. "Ethnic divisions and production in firms." *The Quarterly Journal of Economics*, 129(4): 1899–1946.

Hoogendoorn, Sander, Hessel Oosterbeek, and Mirjam Van Praag. 2013. "The impact of gender diversity on the performance of business teams: Evidence from a field experiment." *Management Science*, 59(7): 1514–1528.

Hoxby, Caroline M. 2000. "Peer effects in the classroom: Learning from gender and race variation." *NBER Working Paper*.

Hunt, Vivian, Sundiatu Dixon-Fyle, Celia Huber, María del Mar Martínez Márquez, Sara Prince, and Ashley Thomas. 2023. "Diversity Matters Even More: The Case for Holistic Impact." McKinsey & Company. Report; published December 5, 2023; accessed 2026-02-13.

Isaksson, Siri. 2018. "It takes two: Gender differences in group work." *Working Paper*.

Jackson, Matthew O., Stephen M. Nei, Erik Snowberg, and Leeat Yariv. 2022. "The Dynamics of Networks and Homophily." National Bureau of Economic Research NBER Working Paper 30815.

John, Oliver P, Eileen M Donahue, and Robert L Kentle. 1991. "Big five inventory." *Journal of Personality and Social Psychology*.

Karpowitz, Christopher F, Stephen D O'Connell, Jessica Preece, and Olga Stoddard. 2024. "Strength in numbers? Gender composition, leadership, and women's influence in teams." *Journal of Political Economy*, 132(9): 3077–3114.

Kling, Jeffrey R, and Jeffrey B Liebman. 2004. "Experimental analysis of neighborhood effects on youth." *Available at SSRN 600596*.

Kling, Jeffrey R, Jeffrey B Liebman, Lawrence F Katz, and Lisa Sanbonmatsu. 2004. "Moving to opportunity and tranquility: Neighborhood effects on adult economic self-sufficiency and health from a randomized housing voucher experiment." *Available at SSRN 588942*.

Kozlowski, Steve WJ, and Daniel R Ilgen. 2006. "Enhancing the effectiveness of work groups and teams." *Psychological Science in the Public Interest*, 7(3): 77–124.

Kroenke, Kurt, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2009. "An ultra-brief screening scale for anxiety and depression: the PHQ–4." *Psychosomatics*, 50(6): 613–621.

Latané, Bibb, Kipling Williams, and Stephen Harkins. 1979. "Many hands make light the work: The causes and consequences of social loafing." *Journal of Personality and Social Psychology*, 37(6): 822.

Lavy, Victor, and Analia Schlosser. 2011. "Mechanisms and impacts of gender peer effects at school." *American Economic Journal: Applied Economics*, 3(2): 1–33.

Lazear, Edward P. 1999. "Culture and language." *Journal of Political Economy*, 107(S6): S95–S126.

Lee, David S. 2009. "Training, wages, and sample selection: Estimating sharp bounds on treatment effects." *The Review of Economic Studies*, 76(3): 1071–1102.

Liden, Robert C, Sandy J Wayne, Renata A Jaworski, and Nathan Bennett. 2004. "Social loafing: A field investigation." *Journal of Management*, 30(2): 285–304.

Linos, Elizabeth, Sanaz Mobasseri, and Nina Roussille. 2024. "Intersectional peer effects at work: The effect of white coworkers on black women's careers." *Management Science*.

List, John A. 2020. "Non est disputandum de generalizability? A glimpse into the external validity trial." National Bureau of Economic Research.

Lowe, Matt. 2021. "Types of contact: A field experiment on collaborative and adversarial caste integration." *American Economic Review*, 111(6): 1807–1844.

Lundberg, Shelly. 2018. "The 2018 report of the committee on the status of women in the economics profession." *American Economic Association*.

Lyons, Elizabeth. 2017. "Team production in international labor markets: Experimental evidence from the field." *American Economic Journal: Applied Economics*, 9(3): 70–104.

Marx, Benjamin, Vincent Pons, and Tavneet Suri. 2021. "Diversity and team performance in a Kenyan organization." *Journal of Public Economics*, 197: 104332.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual Review of Sociology*, 27(1): 415–444.

Mousa, Salma. 2020. "Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq." *Science*, 369(6505): 866–870.

NCES. 2020. "Digest of Education Statistics, 2020." *https://nces.ed.gov/programs/digest/2020menu_tables.asp*, Online; accessed 2024-09-11.

Niederle, Muriel, and Lise Vesterlund. 2007. "Do women shy away from competition? Do men compete too much?" *The Quarterly Journal of Economics*, 122(3): 1067–1101.

Nielsen, Mathias Wullum, Sharla Alegria, Love Börjeson, Henry Etzkowitz, Holly J Falk-Krzesinski, Aparna Joshi, Erin Leahey, Laurel Smith-Doerr, Anita Williams Woolley, and Londa Schiebinger. 2017. "Gender diversity leads to better science." *Proceedings of the National Academy of Sciences*, 114(8): 1740–1742.

Olivetti, Claudia, and Barbara Petrongolo. 2016. "The evolution of gender gaps in industrialized countries." *Annual Review of Economics*, 8(1): 405–434.

Oosterbeek, Hessel, and Reyn Van Ewijk. 2014. "Gender peer effects in university: Evidence from a randomized experiment." *Economics of Education Review*, 38: 51–63.

Oswald, Andrew J, Eugenio Proto, and Daniel Sgroi. 2015. "Happiness and productivity." *Journal of Labor Economics*, 33(4): 789–822.

Pettigrew, Thomas F, and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology*, 90(5): 751.

Rao, Gautam. 2019. "Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools." *American Economic Review*, 109(3): 774–809.

Robins, Richard W, Holly M Hendin, and Kali H Trzesniewski. 2001. "Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale." *Personality and Social Psychology Bulletin*, 27(2): 151–161.

Romano, Joseph P, and Michael Wolf. 2005*a*. "Exact and approximate stepdown methods for multiple hypothesis testing." *Journal of the American Statistical Association*, 100(469): 94–108.

Romano, Joseph P, and Michael Wolf. 2005*b*. "Stepwise multiple testing as formalized data snooping." *Econometrica*, 73(4): 1237–1282.

Romano, Joseph P, and Michael Wolf. 2016. "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing." *Statistics & Probability Letters*, 113: 38–40.

Rosenberg, Morris. 1965. "Rosenberg self-esteem scale." *Journal of Religion and Health*.

Sarsons, Heather, Klarita Gërxhani, Ernesto Reuben, and Arthur Schram. 2021. "Gender differences in recognition for group work." *Journal of Political Economy*, 129(1): 101–147.

Springer, Leonard, Mary Elizabeth Stanne, and Samuel S Donovan. 1999. "Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis." *Review of educational research*, 69(1): 21–51.

Thiemann, Petra. 2022. "The persistent effects of short-term peer groups on performance: Evidence from a natural experiment in higher education." *Management Science*, 68(2): 1131–1148.

West, Candace, and Don H Zimmerman. 1987. "Doing gender." *Gender & Society*, 1(2): 125–151.

Wheelan, Susan A. 2009. "Group size, group development, and group productivity." *Small Group Research*, 40(2): 247–262.

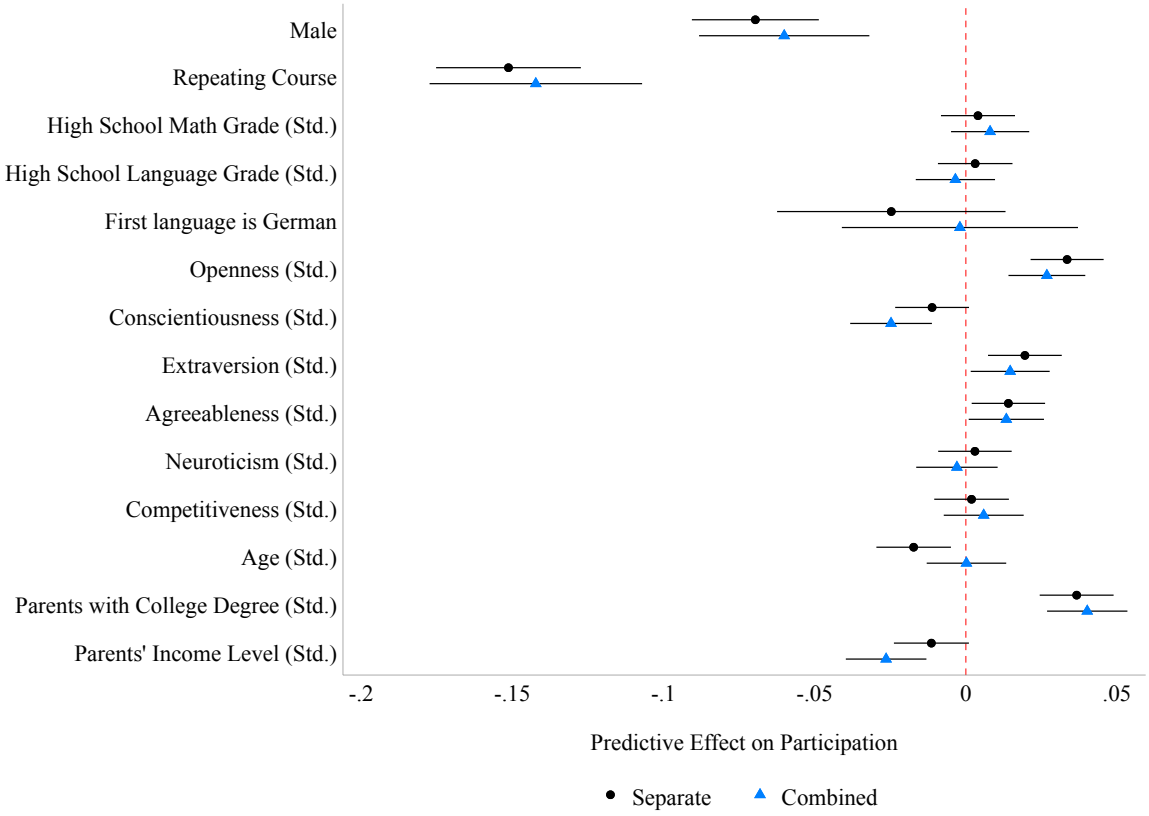
Wuchty, Stefan, Benjamin F Jones, and Brian Uzzi. 2007. "The increasing dominance of teams in production of knowledge." *Science*, 316(5827): 1036–1039.

Zölitz, Ulf, and Jan Feld. 2021. "The effect of peer gender on major choice in business school." *Management Science*, 67(11): 6963–6979.

A Additional Figures and Tables

A.1 Additional Figures

Figure A.1. How Baseline Characteristics Predict Participation in the Experiment



Notes: The figure shows the predictive effects of various baseline characteristics on the likelihood of participating in the experiment and having a random study group. The estimates in black circles are derived from 14 separate linear probability models (LPM), and the estimates in blue triangles are derived from one LPM that includes all baseline characteristics as independent variables. The independent variable is a binary indicator for participating in the experiment (mean = 0.36). The models control for program-by-year fixed effects and use robust standard errors. Error bars indicate 95% confidence intervals.

Figure A.2. Template of the Email Announcing Group Assignment

Hallo Xiaoyue,

Herzlich willkommen zu deiner Studiengruppe! Welcome to your study group!

Deine Gruppennummer ist **${GroupCode}**. Your group number is ${GroupCode}.
Hier sind die Namen, Handynummern und Emailadressen von der anderen Mitglieder deiner Lerngruppe:
Here are the names, cell phone numbers and email addresses of the other members of your study group:
${PeerName1}: ${PeerPhone1}; ${PeerEmail1}; ${PeerEmail_private1}
${PeerName2}: ${PeerPhone2}; ${PeerEmail2}; ${PeerEmail_private2}
${PeerName3}: ${PeerPhone3}; ${PeerEmail3}; ${PeerEmail_private3}

Gruppenkoordinator Group coordinator
Von allen Gruppenmitgliedern haben wir zufällig **${Coordinator}** als den vorübergehenden Gruppenkoordinator ausgewählt. Ihr könnt später auch einen anderen Gruppenkoordinator bestimmen.
Of all the group members, we have randomly selected ${Coordinator} as the temporary group coordinator. You can also choose a different group coordinator later.
Wenn du Koordinator bist, dann bitten wir dich jetzt eine **WhatsApp-Gruppe** zu erstellen, damit ihr euer erstes Gruppentreffen einfach organisieren könnt. Ihr könnt euer erstes Treffen natürlich auch per Email vereinbaren.
If you are the coordinator, we ask you to create a WhatsApp group now so that you can easily organize your first group meeting. You can of course also arrange your first meeting by email.

Lerngruppentreffen Study group meetings
Wir empfehlen euch, dass ihr euch mindestens einmal in der Woche **persönlich oder online** trefft. Ihr könnt gemeinsam Vorlesungsfolien durchgehen und Übungsaufgaben lösen. Wenn ihr euch persönlich treffen möchtet, könnte der Lichthof im Hauptgebäude oder die VWL Bibliothek ein guter Ort sein.
We recommend that you meet at least once a week in person or online. You can go through lecture slides and solve exercises together. If you would like to meet in person, the atrium in the main building or the economics library could be a good place.

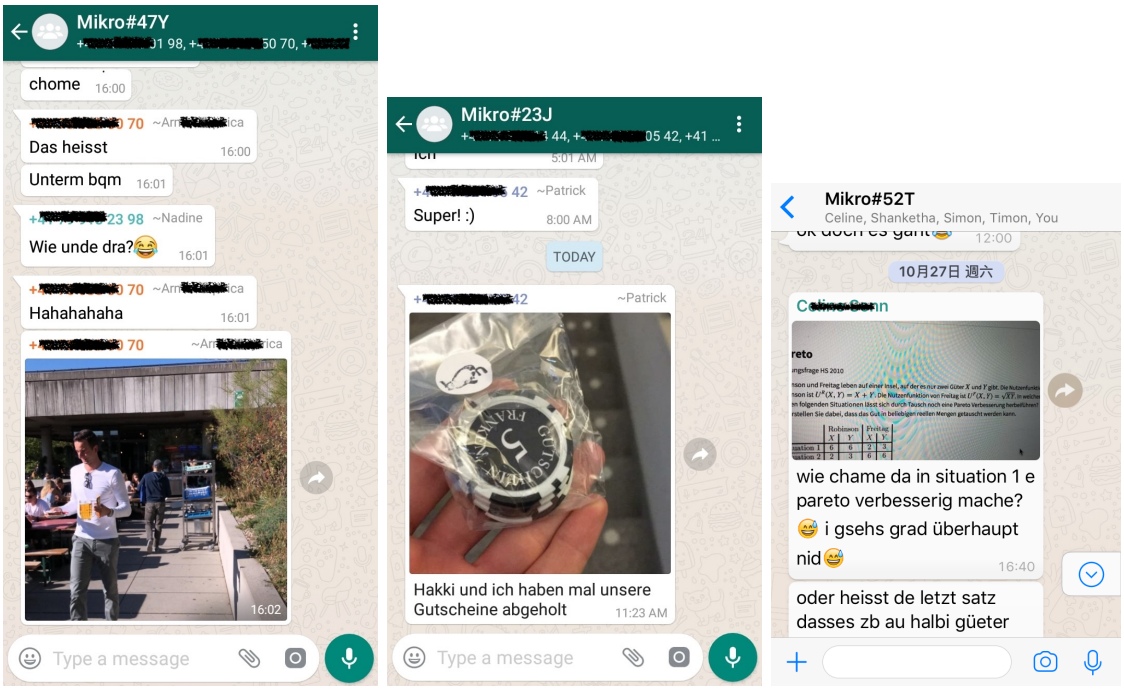
Verhaltensregeln Rules of conduct
(1) Uns ist es wichtig, dass ihr respektvoll miteinander umgeht und versucht euch gegenseitig zu helfen.
It is important to us that you treat each other with respect and try to help each other.
(2) Eure Teilnahme an Lerngruppen ist vollkommen freiwillig. Es kann also vorkommen, dass ein Gruppenmitglied an einem bestimmtem Treffen nicht teilnehmen kann. In diesem Fall ermutigen wir euch die übrigen Mitglieder zu treffen.
Your participation in study groups is completely voluntary. It may therefore happen that a group member is unable to attend a particular meeting. In this case, we encourage you to meet the other members.

Wir wünschen euch eine erfolgreiche Zusammenarbeit! We wish you a successful collaboration!

Liebe Grüsse,
Das XXX Support Team

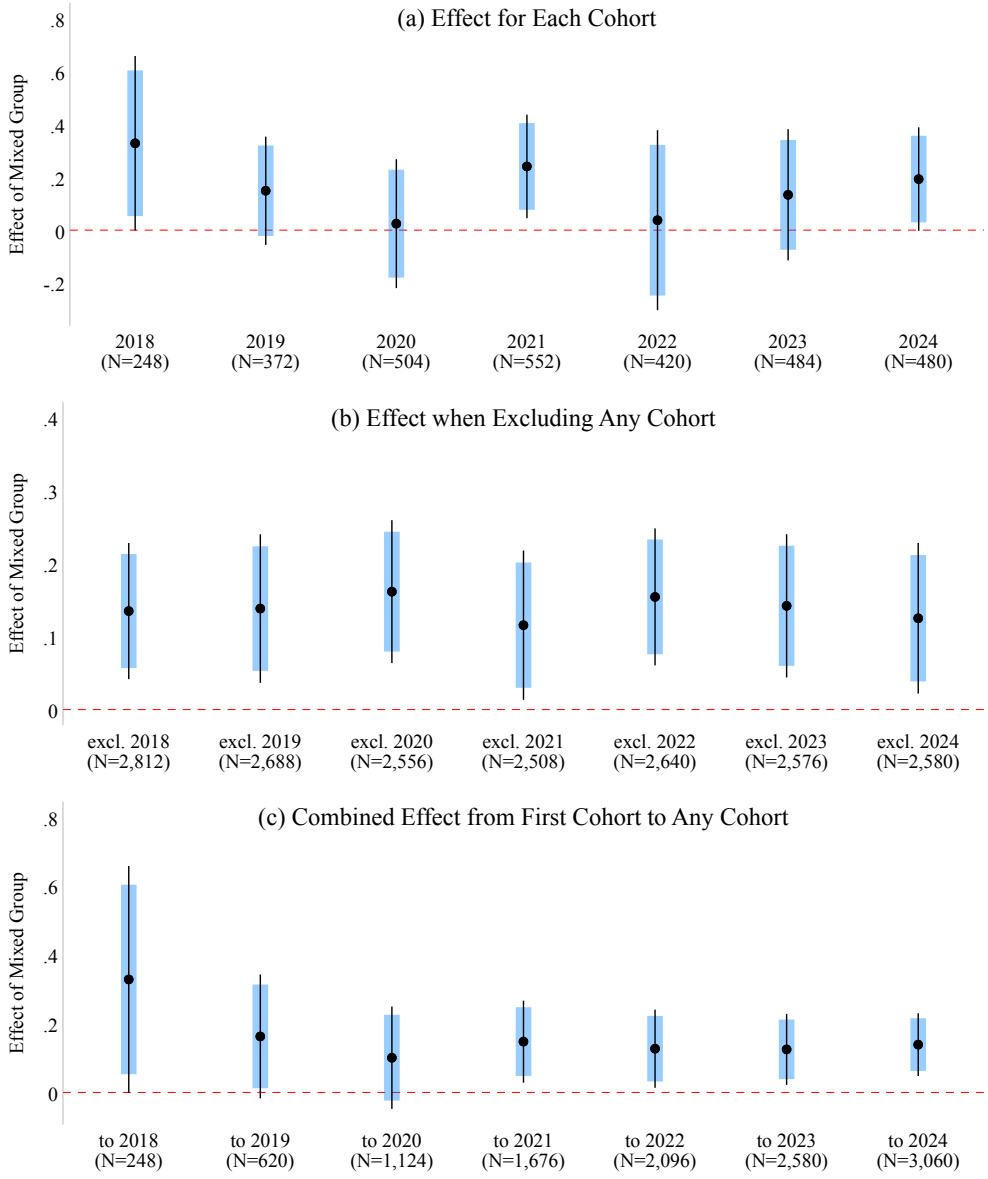
Notes: The German text is the original version, while the English text in gray is the translated version. This email template was used after the onset of the COVID-19 pandemic, highlighting the option of online meetings. During the pandemic, the email also contained information regarding social distancing regulations. Otherwise, the content of the email remained the same over time.

Figure A.3. Screenshots of Communications in WhatsApp Groups (in German)



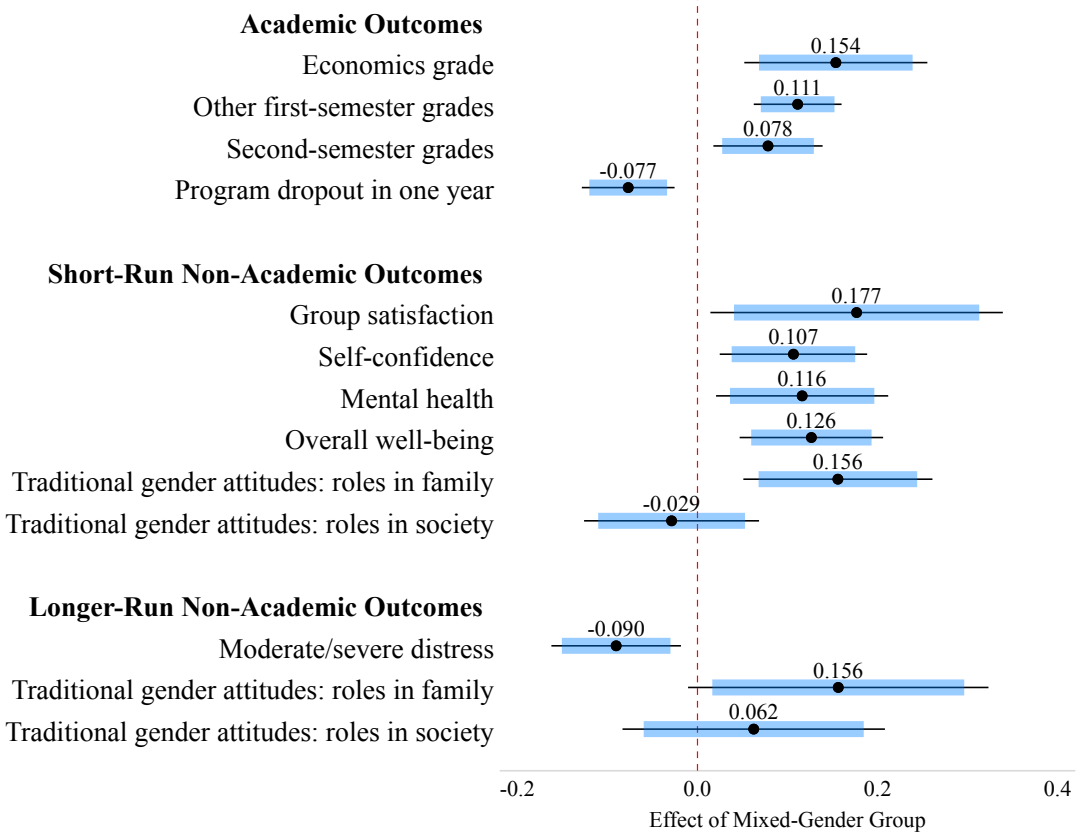
Notes: The left figure shows that group members are having joint social events; the middle figure shows that the group has collected the free vouchers; the right figure shows that the group is discussing a problem set.

Figure A.4. Effects of Mixed-Gender Groups on Economics Grades across Cohorts



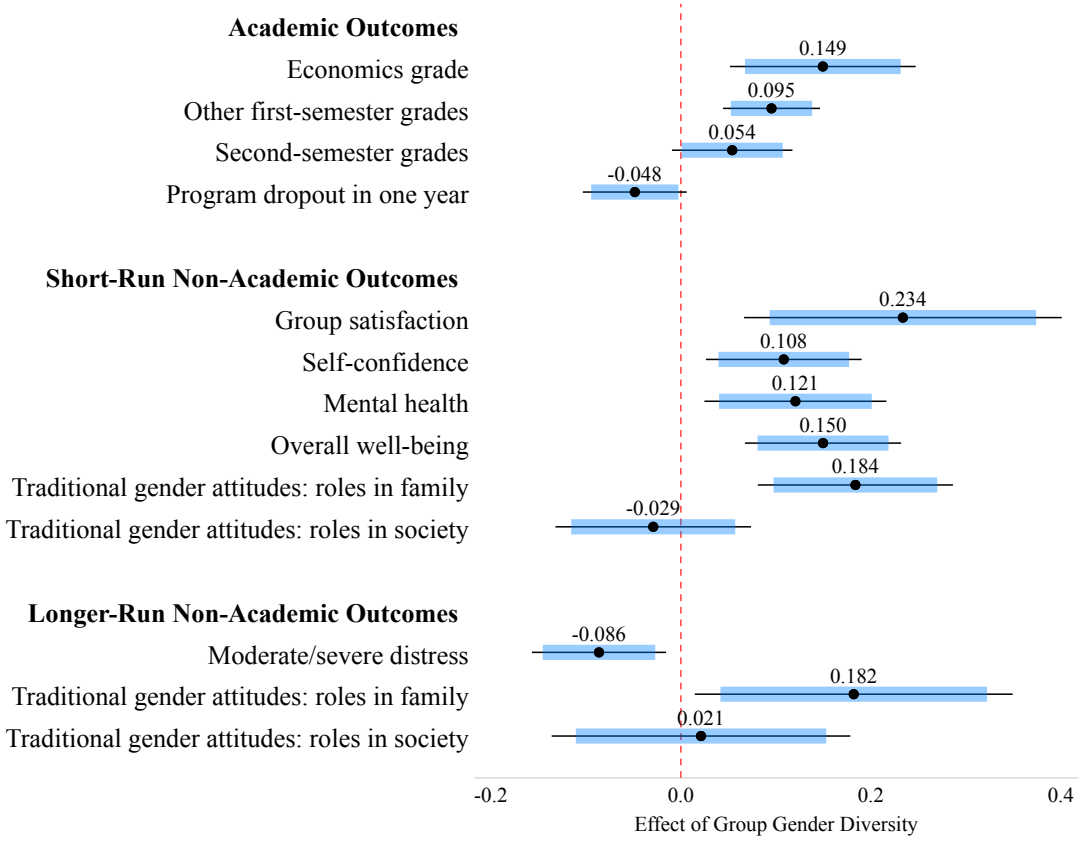
Notes: Panel (a) shows the estimated effects of being assigned to a mixed-gender group on standardized economics grade *separately* for each experimental cohort; Panel (b) shows the effects when *excluding* each of the seven cohorts from the sample; Panel (c) shows the effects when *including* the sample of students from the first cohort to each of the later cohorts. Each point estimate is derived from one OLS regression. All regressions control for gender and cohort-by-program fixed effects. Standard errors are clustered at the group level. Error bars indicate 95% and 90% confidence intervals of the estimates.

Figure A.5. Effects of Mixed-Gender Groups Removing Repeating Students



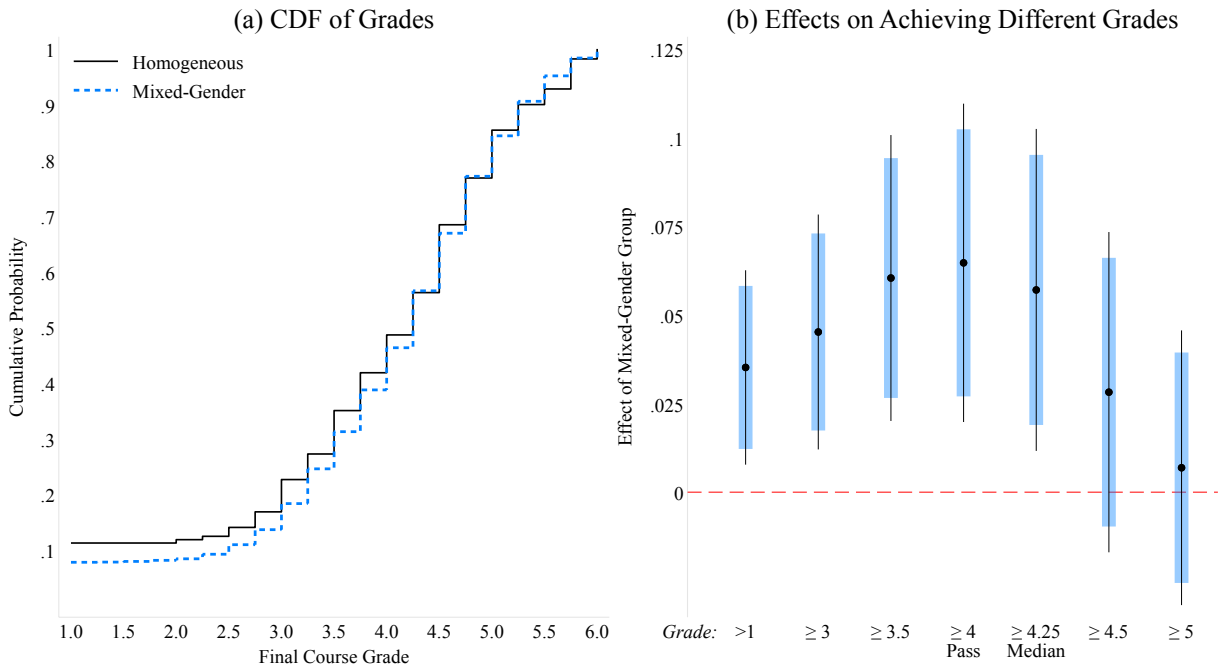
Notes: The figure plots the estimated effects of assignment to mixed-gender rather than homogeneous groups on the primary academic and non-academic outcomes analyzed in this paper, when removing the small set of students repeating the economics course during the experiment. Each estimate is derived from one OLS regression or LPM. The specifications follow analyses in Table 3, Table 4, Figure 2, Table 6, and Table 7. All non-binary dependent variables are standardized across the whole sample (including repeaters and non-repeaters—for ease of comparison to results in the paper). Error bars indicate 90% and 95% confidence intervals.

Figure A.6. Effects of Group Gender Diversity on Academic and Non-Academic Outcomes



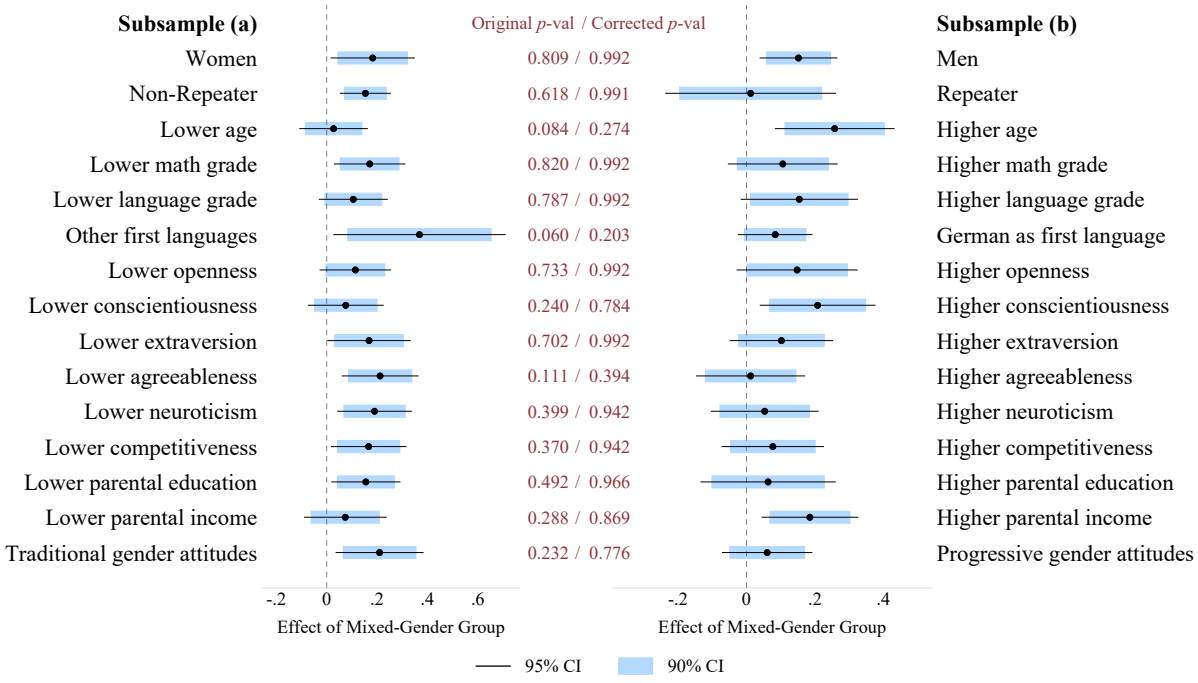
Notes: The figure plots the estimated effects of the continuous group gender diversity measure taking three values (0, 0.86, and 1) on the primary academic and non-academic outcomes analyzed in this paper. Each estimate is derived from one OLS regression or LPM. The specifications follow analyses in Table 3, Table 4, Figure 2, Table 6, and Table 7. All non-binary dependent variables are standardized across the whole sample. Error bars indicate 90% and 95% confidence intervals.

Figure A.7. Distribution of Grades and Effects of Mixed-Gender Groups across the Grade Distribution



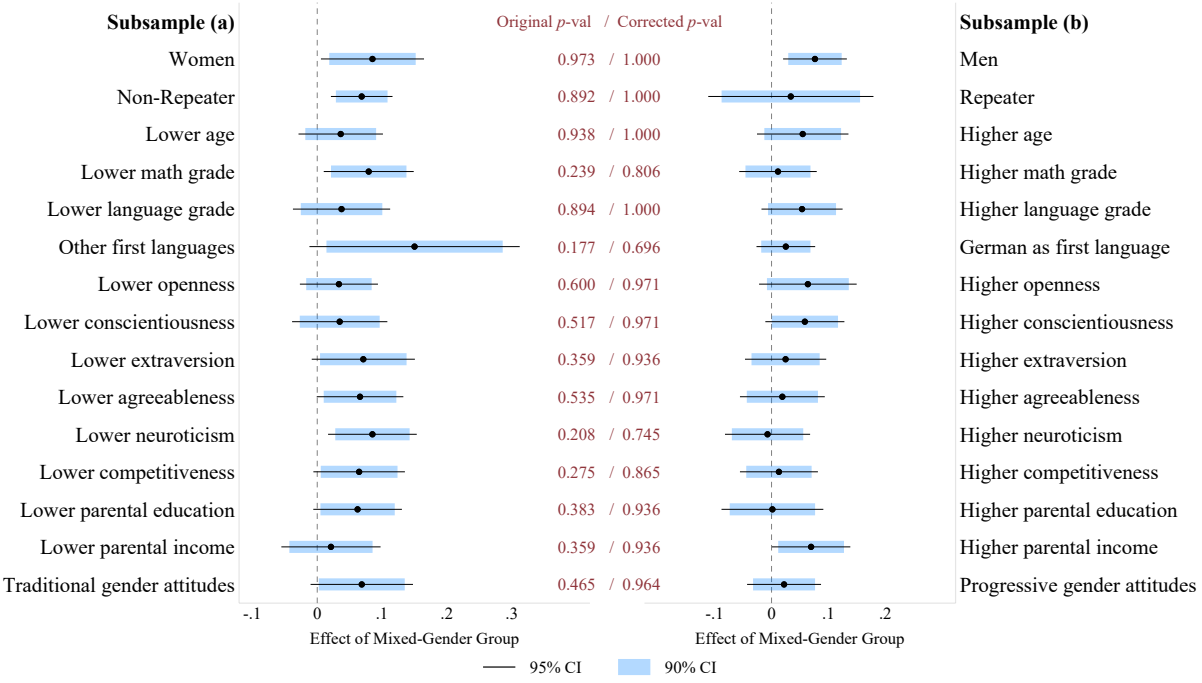
Notes: Panel (a) shows the CDF of the final economics grade, separately for homogeneous groups and mixed-gender groups. Panel (b) shows the effects of mixed-gender groups on the likelihood of achieving different thresholds of grades for the economics course. Each estimate in Panel (b) is derived from one linear probability model that controls for student gender and cohort-by-program fixed effects. Standard errors are clustered at the group level. Error bars indicate 95% and 90% confidence intervals.

Figure A.8. Impact of Mixed-Gender Groups on Economics Grades by Student Characteristics



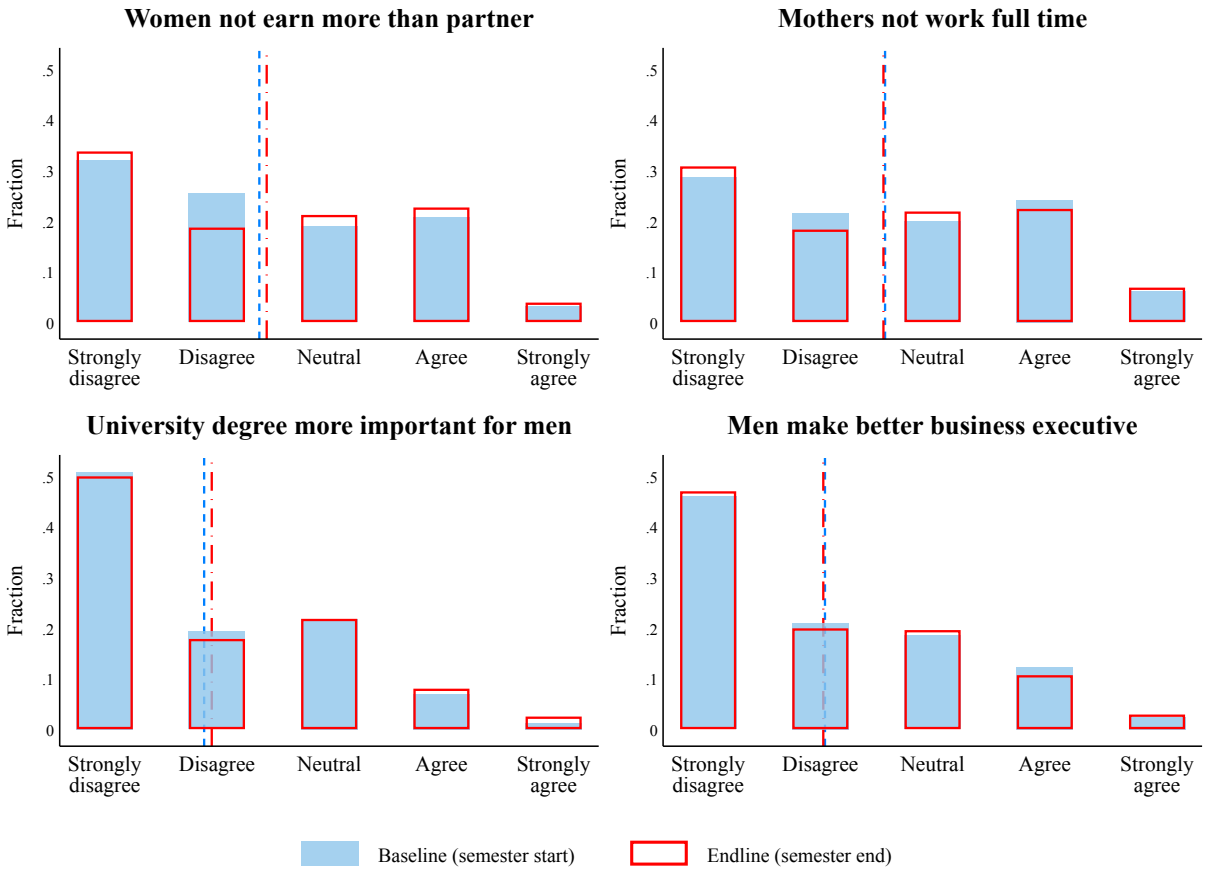
Notes: I split students into two subsamples based on each baseline characteristic and separately estimate the effect of mixed-gender groups on standardized economics grades using the specification in Column 1 of Table 3. For baseline characteristics measured on a continuous scale, such as math grades and personality traits, Panel (a) reports the estimated effects for students with below-median ("Lower") values, while Panel (b) reports the effects for students with above-median ("Higher") values. For each characteristic, I additionally estimate an interacted model using the full sample, including the interaction term "Mixed-Gender Group × Subsample A." The p-value associated with the interaction term is displayed in the middle column of the figure (labeled "original p-val"). The "corrected p-values" are obtained by applying multiple-testing corrections simultaneously to the 30 estimates, including both the mixed-gender group indicator and the interaction term from each regression. Specifically, I use the step-down procedure of Romano and Wolf (2005*a*,*b*, 2016), which controls the family-wise error rate by resampling group clusters within randomization strata 1,000 times.

Figure A.9. Impact of Mixed-Gender Groups on Course Passing by Student Characteristics



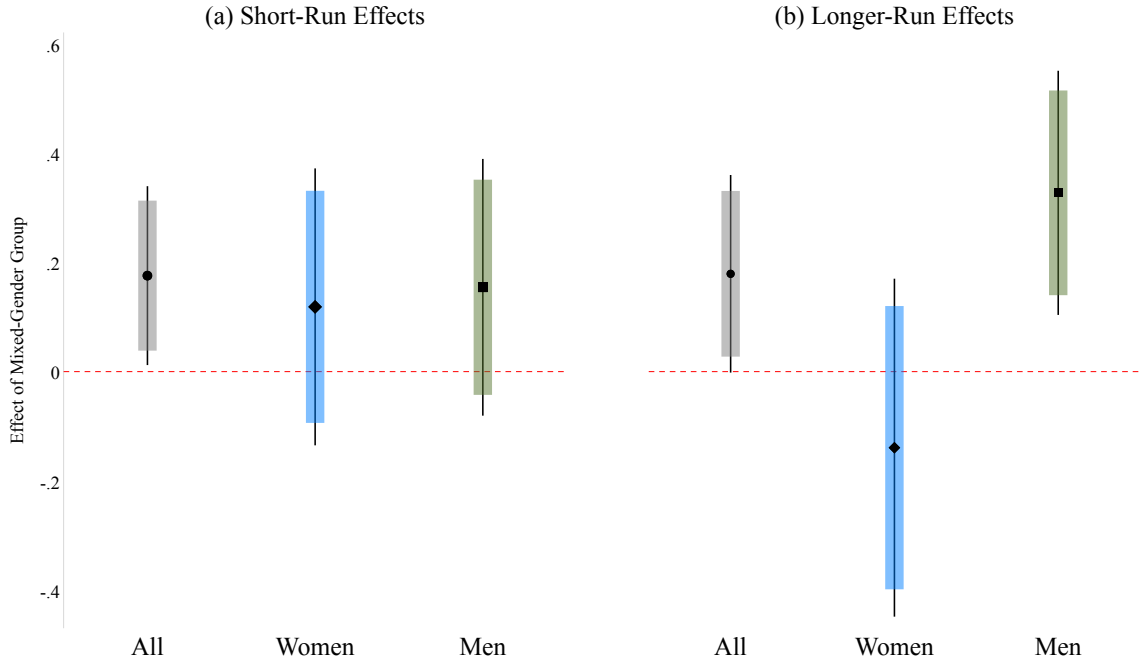
Notes: The figure replicates the results in Appendix Figure A.8 when using the course passing as the outcome. Each point estimate is derived from one LPM for one subsample. The p-values in the middle panel are derived from tests of differences between the two estimates in a row. I also correct the p-values using the Romano-Wolf approach. See the notes of Figure A.8 for details.

Figure A.10. Distribution of Gender Role Attitudes



Notes: The figure shows the distribution of responses to different statements about gender roles: (1) "If a woman earns more money than her partner, this might cause problems"; (2) "When a mother is working full time, the child develops less well"; (3) "A university degree is more important for a man than for a woman"; and (4) "On the whole, men make better business executives than women do." Students report their agreement with each statement on a scale of 1 (strongly disagree) to 5 (strongly agree). For each statement, I show the distribution of responses at baseline and at endline; with the average agreement score plotted as a vertical dashed line.

Figure A.11. Short and Longer-Run Effects on Attitudes toward Family Gender Roles



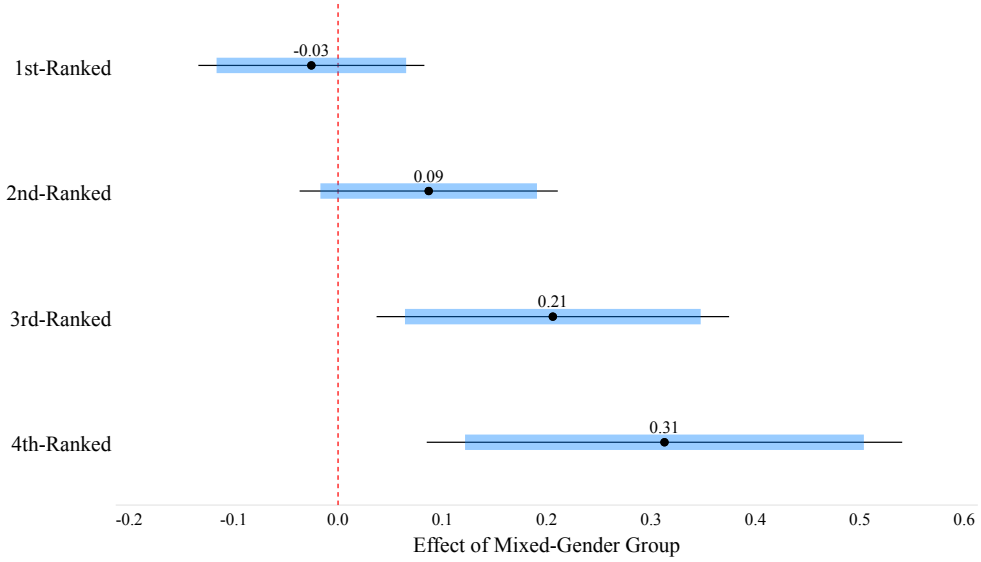
Notes: The figure shows the effects of mixed-gender groups on traditional attitudes toward family gender roles using a common sample of students who reported attitudes in both surveys ($N = 674$). The outcome variable is standardized among this sample to have a mean of 0 and SD of 1. Panels (a) and (b) show the short-run and longer-run effects, respectively. Each estimate is derived from one OLS regression, which controls all observed baseline characteristics, baseline agreement with traditional gender roles, student gender (for the combined sample), and cohort-by-program fixed effects. Standard errors are clustered at the group level. Error bars indicate 95% and 90% confidence intervals.

Figure A.12. Effects on Non-Academic Outcomes with Inverse Probability Weighting for Survey Response



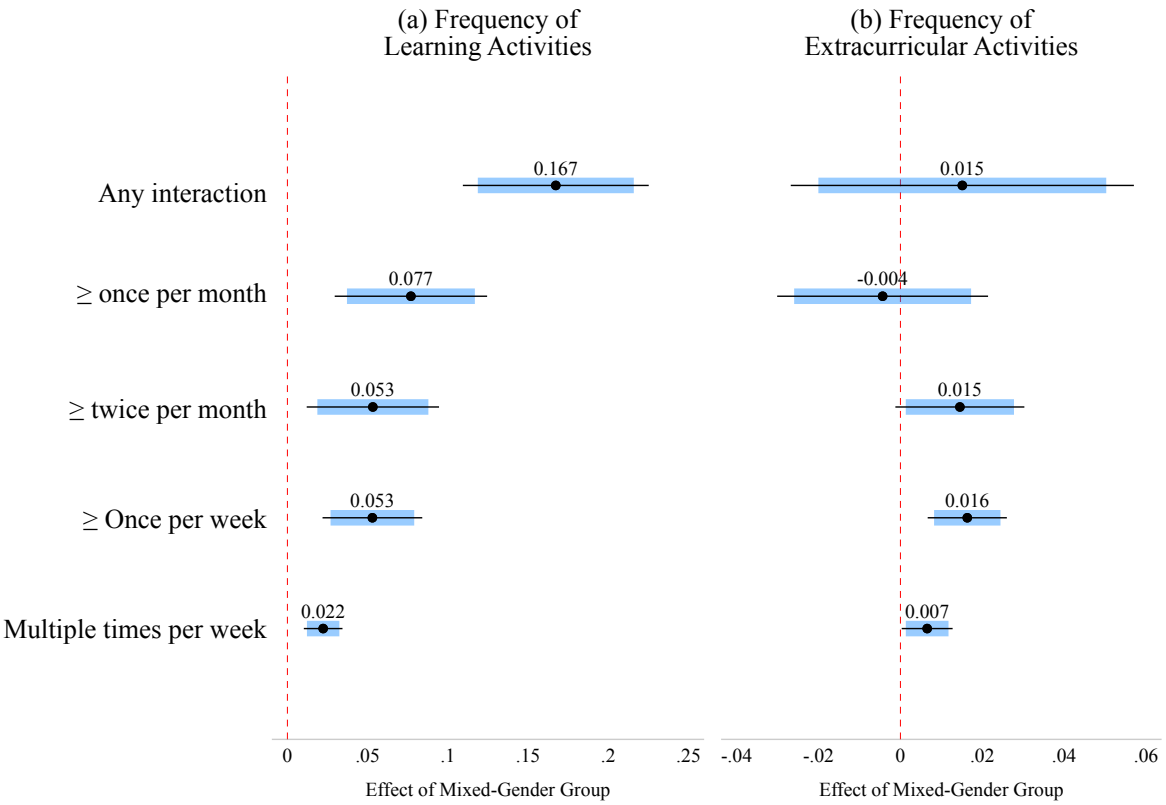
Notes: The figure shows the effects of mixed-gender groups on non-academic outcomes measured in the endline and follow-up surveys. Compared to the original estimates presented in Figure 2, Table 6 and Table 7, the estimates in this figure are derived from regressions that include the inverse probability weighting. The probability is the predicted likelihood of answering the endline or follow-up survey based on all baseline characteristics and cohort-by-program fixed effects. Refer to the original figure or table for details about the regression specification. All outcomes are standardized to have a mean of 0 and SD of 1. Error bars indicate 95% and 90% confidence intervals.

Figure A.13. Effects of Mixed-Gender Groups by Performance Rank within Groups



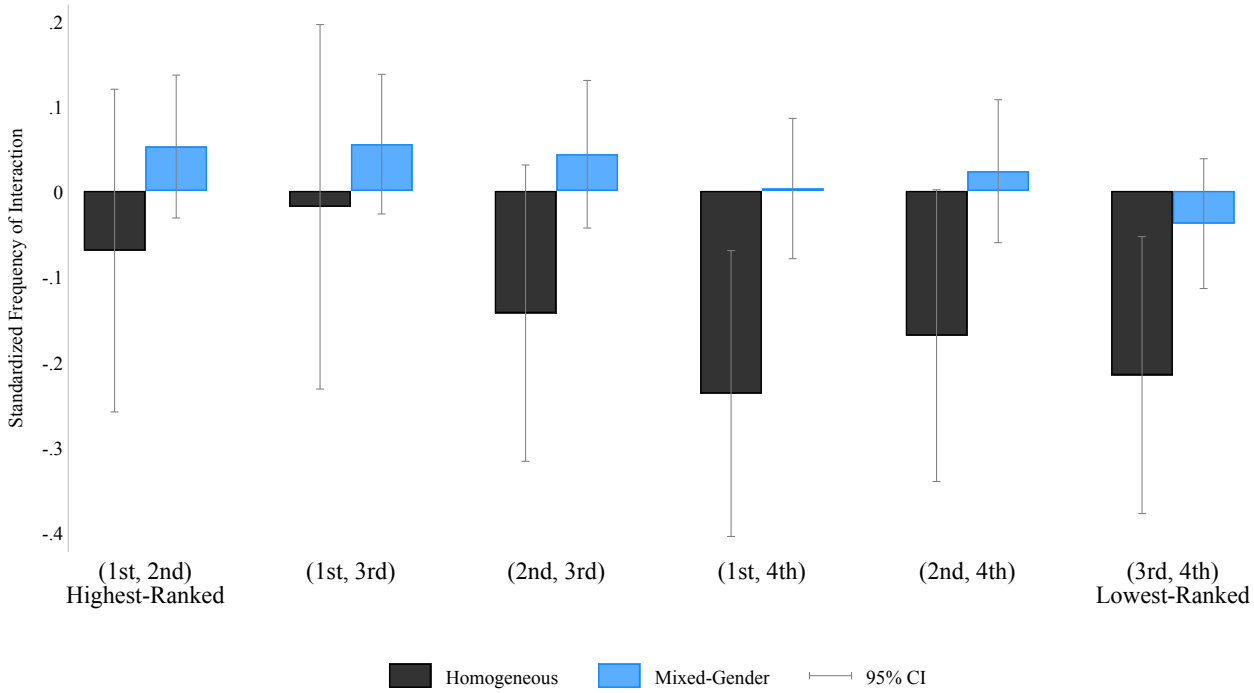
Notes: The figure shows the estimated effects of mixed-gender groups on the raw final grade for economics, separately for group members whose final grades rank first (highest), second, third, and fourth (lowest) in a group. Each estimate is derived from one OLS regression which controls for program-by-cohort fixed effects and student gender and uses robust standard errors. Error bars indicate 90% and 95% confidence intervals. The pattern is similar when splitting the sample by the rank of predicted grades.

Figure A.14. Effects on Likelihood of Pairwise Interaction Reaching Different Frequency Levels



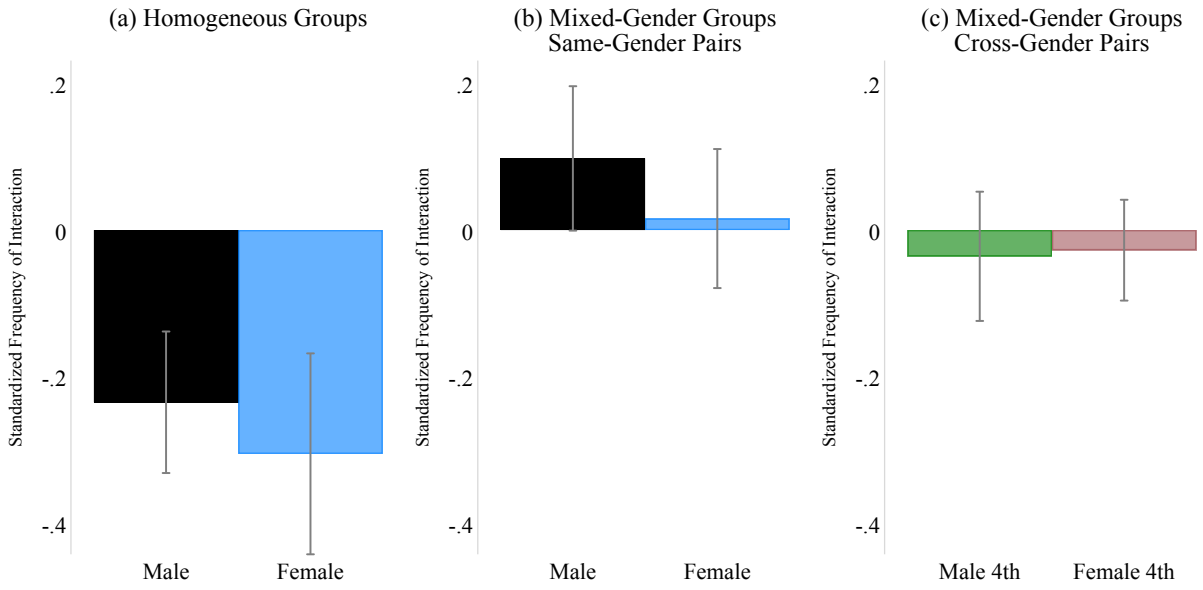
Notes: The figure shows the effects of mixed-gender groups on the likelihood of having different frequencies of pairwise interaction. See footnote 27 for detail about the survey questions on pairwise activities. Each estimate is derived from one linear probability model that uses 5,504 student-peer-level observations. The models control for cohort-by-program fixed effects, student gender, peer gender, and an indicator for whether the peer is of the same gender. Standard errors are clustered at the student level. Error bars indicate 95% and 90% confidence intervals.

Figure A.15. Frequency of Pairwise Interaction by Rank of Predicted Performance



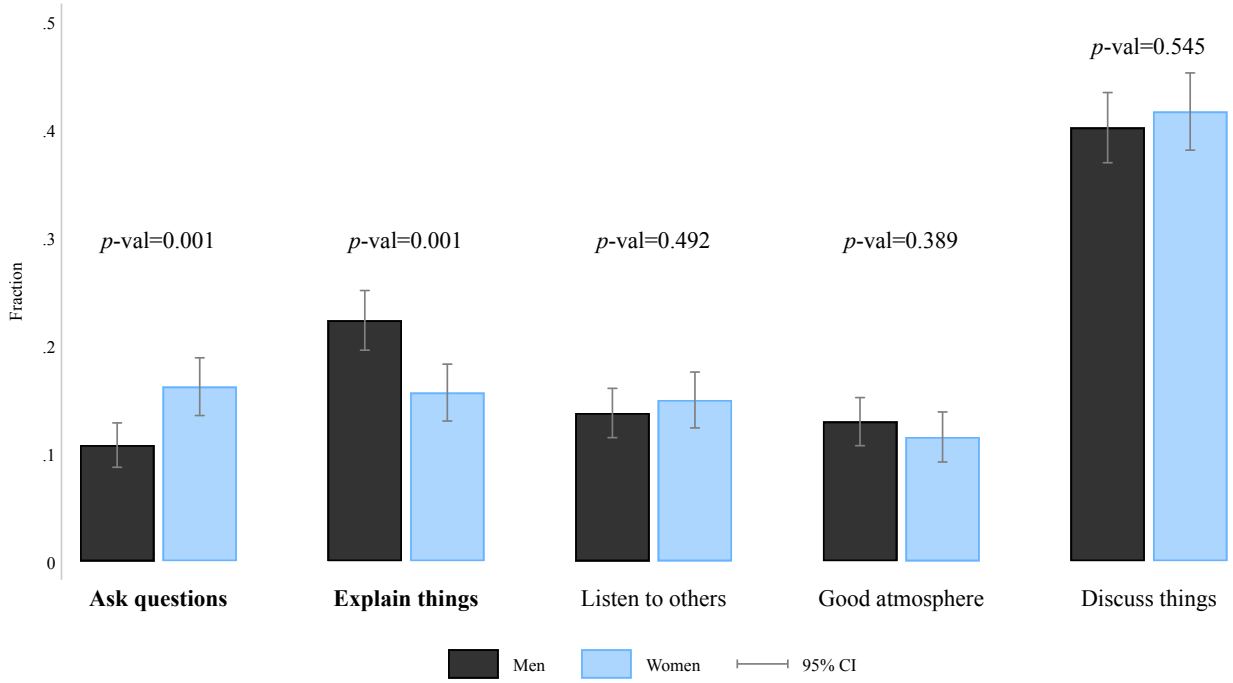
Notes: The figure shows the standardized frequency of pairwise interactions by the pair's ranking of predicted economics grade within a group. For example, (1st, 2nd) represents pairs of students whose predicted final grades are ranked the first and second in a group, and (3rd, 4th) represents pairs ranked the third and fourth in a group. The predicted grade is derived using all observable student characteristics at baseline.

Figure A.16. Interaction with 4th-Ranked Member by Group and Pair Gender Composition



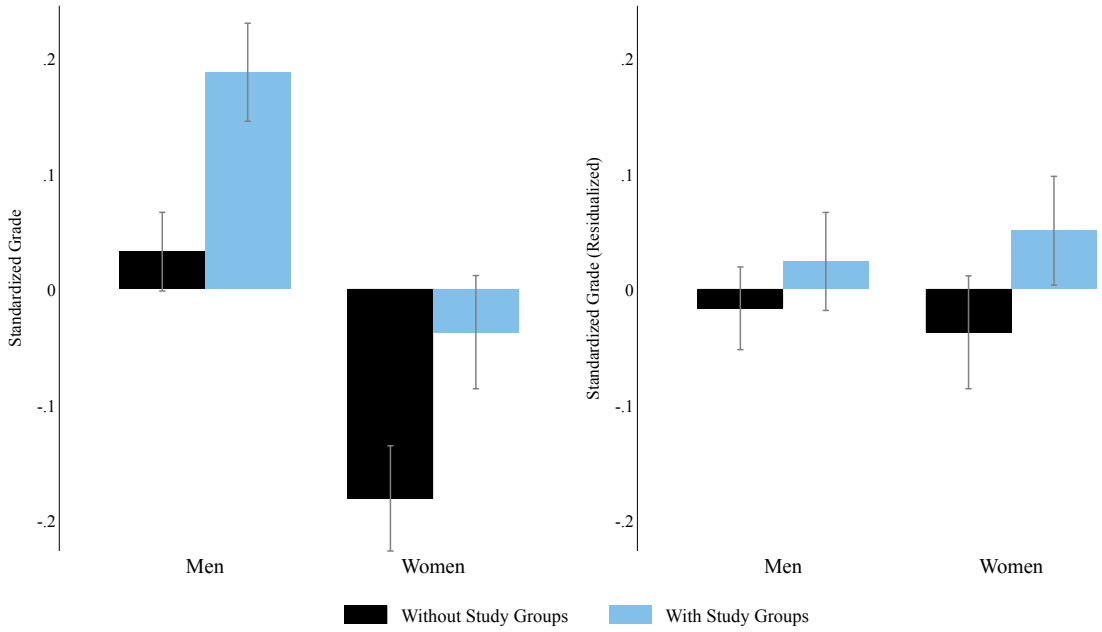
Notes: The figure shows the standardized frequency of pairwise interaction between fourth-ranked members and other members, separately by the group and pair gender composition. In homogeneous groups, the pairs are male–male or female–female. In mixed-gender groups, the pairs can be same-gender or opposite gender, with the fourth-ranked being male or female. Error bars indicate 95% confidence intervals.

Figure A.17. The Primary Role in Group Work by Gender



Notes: In the endline survey for the 2019/20–2024/25 cohorts, students report what their primary role is when studying with others. The figure shows the proportion of men and women who choose the following roles: explaining things to others; listening to others; asking questions; discussing things with others; and ensuring a good atmosphere. The p-values are derived from tests of gender differences in reporting each primary role.

Figure A.18. Economics Performance by Experiment Participation and Gender



Notes: The figure plots the standardized final grade in economics by student gender, separately for students not participating in the experiment (without study groups) and students assigned to groups in the experiment. The right panel shows the residualized grades after controlling for program-by-cohort fixed effects and observable student characteristics at baseline. Error bars indicate 95% confidence intervals.

A.2 Additional Tables

Table A.1. Summary of Studies using Randomized Groups to Examine Effects of Gender Composition

Authors (year) and Journal	Sample and Setting	Group Size	# of Groups	Composition (% Female)	Primary Outcomes
Panel A: Large-Sized, Mixed-Gender Groups					
Hampole et al. (2021) *Working Paper*	MBA sections U.S.	≈ 60	148	Mean=0.34 SD=0.04	Career outcomes
Gong et al. (2019) *Journal of Human Resources*	Middle school classes, China	≈ 43	208	Mean=0.34 SD=0.08	Course performance
Oosterbeek and Van Ewijk (2014) *Economics of Education Review*	Univ. teaching sessions, NL	≈ 39	22	Mean=0.3 ∈[0.14, 0.51]	Dropout & performance
Zölitz and Feld (2021) *Journal of Labor Economics*	Univ. teaching sessions, NL	≈ 14	2,559	Mean=0.38 SD=0.14	Major choice
Chen and Lin (2023) *Working Paper*	Univ. course teams, Taiwan	≈ 12	170	Mean=0.58 SD=0.14	Team performance
Hoogendoorn et al. (2013) *Management Science*	Univ. start-up teams, NL	≈ 12	45	Mean=0.44 ≈[0.2, 0.6]	Team sales & profits
Panel B: Large-Sized, Single-/Mixed-Gender Groups					
Eisenkopf et al. (2015) *Journal of Economic Behavior & Organization*	High school classes, CH	≈ 20	39	All-female female-majority	Course performance
Booth et al. (2018) *Journal of Public Economics*	Univ. classes in U.K.	≈ 18	37	All-male all-female male-majority	Course performance
Contreras et al. (2025) *Labour Economics*	Univ. teaching groups, U.K.	≈ 16	≈ 24	All-female female-majority	Student performance
Panel C: Small-Sized Groups					
Karpowitz et al. (2024) *Journal of Political Economy*	Univ. study groups, U.S.	5 or 6	220	Male-majority female-majority	Individual influence
Fenoll and Zaccagni (2022) *Labour Economics*	School math camp teams, IT	6	81	Male-majority female-majority balanced	Team/student performance
Coveney et al. (2025) *Working Paper*	Univ. course teams, NL	2	3,200* repeated	Male-male female-female male-female	Pair task performance
This Paper	Study groups in university, Europe	4	765	All-male All-female male-majority female-majority balanced	Performance, dropout & well-being in short & longer run

Notes: *Coveney et al. (2025) assign students repeatedly into different pairs across three teaching sessions.

Table A.2. Effects of Group Gender Composition on Survey Response

	(1)	(2)	(3)	(4)	(5)	(6)
Analysis sample:		All		Women	Men	All
Panel A: Endline Survey Response						
Joint Effect of #Women Dummies: F-statistic	0.939	0.671				
Joint Effect of #Women Dummies: p-value	0.441	0.612				
Mixed-Gender Group			0.031	0.076	0.011	
			(0.026)	(0.051)	(0.030)	
			[0.232]	[0.134]	[0.720]	
Majority Gender in Group						0.041
						(0.027)
						[0.130]
Gender-Balanced Group						0.028
						(0.028)
						[0.324]
Minority Gender in Group						0.003
						(0.035)
						[0.940]
Average Response Rate	0.575	0.623	0.559	0.606	0.596	0.644
Panel B: Follow-Up Survey Response						
Joint Effect of #Women Dummies: F-statistic	1.183	0.968				
Joint Effect of #Women Dummies: p-value	0.317	0.424				
Mixed-Gender Group			0.006	0.001	0.022	
			(0.029)	(0.053)	(0.037)	
			[0.829]	[0.987]	[0.549]	
Majority Gender in Group						0.026
						(0.032)
						[0.421]
Gender-Balanced Group						-0.021
						(0.031)
						[0.500]
Minority Gender in Group						0.013
						(0.041)
						[0.757]
Average Response Rate	0.346	0.382	0.326	0.363	0.372	0.407
Baseline Characteristics	-	✓	✓	✓	✓	✓

Notes: The table tests whether group gender composition affects the endline survey response of all students, men, and women. Each column in each panel is estimated with a linear probability model in which the dependent variable is a binary indicator for completing the endline or follow-up survey. Columns 1–2 test the joint effect of gender composition (dummies for 0–3 women) and report the F-statistics and p-values. Other columns estimate the effects of mixed-gender groups compared to homogeneous groups. All specifications also control for student own gender (for the combined sample) and cohort-by-program fixed effects. The specifications in Columns 2–6 further control for all student characteristics observed at baseline. Standard errors in parentheses are clustered at the group level; p-values are in brackets.

Table A.3. Effects of Female Proportion vs. Mixed-Gender Indicator on Primary Academic Outcomes with Multiple-Testing Corrections

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Final Grade in Economics		Other First-Semester Course Grades		Second-Semester Course Grades		Program Dropout In One Year	
Panel A: Effects of Female Proportion								
Proportion of Women	-0.064	-0.061	-0.031	-0.035	-0.117	-0.100	0.065	0.055
	(0.084)	(0.085)	(0.041)	(0.042)	(0.052)	(0.054)	(0.044)	(0.046)
Original p-value	0.451	0.478	0.454	0.409	0.024**	0.062*	0.138	0.238
FWER p-value	0.590	0.590	0.590	0.589	0.013**	0.062*	0.177	0.374
FDR q-value	0.219	0.219	0.219	0.219	0.033**	0.053*	0.097*	0.136
Observations	3,060	2,520	12,225	10,299	9,215	8,005	2,580	2,193
R-squared	0.030	0.185	0.108	0.244	0.098	0.198	0.024	0.091
Panel B: Effects of Mixed-Gender Groups								
Mixed-Gender Group	0.139	0.134	0.099	0.095	0.061	0.104	-0.060	-0.071
	(0.046)	(0.051)	(0.024)	(0.026)	(0.030)	(0.032)	(0.026)	(0.028)
Original p-value	0.003***	0.008***	<0.001***	<0.001***	0.046**	0.001***	0.022**	0.012**
FWER p-value	0.001***	0.003***	0.001***	0.001***	0.044**	0.001***	0.011**	0.006***
FDR q-value	0.010**	0.021**	0.001***	0.003***	0.051*	0.006***	0.033**	0.024**
Observations	3,060	2,520	12,225	10,299	9,215	8,005	2,580	2,193
R-squared	0.033	0.187	0.109	0.245	0.098	0.198	0.026	0.094
Baseline controls	-	✓	-	✓	-	✓	-	✓

Notes: The table shows the estimated effects of group gender ratio (Panel A) and group gender diversity (Panel B) on the main academic outcomes. Each column in each panel represents one OLS regression or LPM. The models in Columns 1–2 and 5–6 use student-level observations, control for gender and cohort-by-program fixed effects, and cluster standard errors at the group level. The models in Columns 3–4 use student-by-course observations and follow the specification used in Table 4, with standard errors clustered at the group-by-course level. The middle rows show the original p-values of the estimates, the corrected p-values controlling the family-wise error rate (FWER), and the sharpened q-values controlling the false discovery rate (FDR). The multiple testing corrections are applied *simultaneously* to all 12 estimates in the table. The FWER corrected p-values are derived using the approach of Romano and Wolf (2005*a*,*b*, 2016) through a resampling procedure. For student-level estimations, I use a bootstrapping procedure that draws 1,000 resamples of group clusters within the program-cohort strata. For student-course-level estimations in Columns 3–6, the bootstrapping procedure draws 1,000 resampling of group clusters within program-cohort-course strata—392 (or 1.6% of) observations are dropped because they are singletons within the stratum. The FDR corrected q-values are derived using the approach of Anderson (2008). Clustered standard errors for the original estimates are in parentheses. $^{*}p<.1$, $^{**}p<.05$, $^{***}p<.01$.

Table A.4. Academic and Non-Academic Effects of Mixed-Gender Groups with Multiple Testing Corrections

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A. Short-Run and Longer-Run Academic Outcomes							
	Final Grade in Economics		Other First-Semester Course Grades		Second-Semester Course Grades		Program Dropout In One Year	
Original estimate	0.139	0.134	0.099	0.095	0.061	0.104	-0.060	-0.071
	(0.046)	(0.051)	(0.024)	(0.026)	(0.030)	(0.032)	(0.026)	(0.028)
Original p-value	0.003***	0.008***	<0.001***	<0.001***	0.046**	0.001***	0.022**	0.012**
FWER p-value	0.006***	0.010**	0.001***	0.003***	0.032**	0.003***	0.021**	0.011**
FDR q-value	0.007***	0.012**	0.001***	0.002***	0.022**	0.005***	0.017**	0.013**
Baseline characteristics	-	✓	-	✓	-	✓	-	✓

	B. Short-Run Non-Academic Outcomes					
	Group Satisfaction	Self-Confidence	Mental Health	Overall Well-Being	Gender Attitudes	
					Family Roles	Society Roles
Original estimate	0.219	0.110	0.116	0.145	0.163	-0.026
	(0.080)	(0.060)	(0.039)	(0.046)	(0.039)	(0.050)
Original p-value	0.006***	0.005***	0.011**	<0.001***	0.002***	0.596
FWER p-value	0.008***	0.008***	0.011**	0.001***	0.003***	0.717
FDR q-value	0.010**	0.009***	0.013**	0.001***	0.005***	0.118
Baseline characteristics	✓	✓	✓	✓	✓	✓

	C. Longer-Run Non-Academic Outcomes			
	Distress Score	Moderate/ Severe Distress	Gender Attitudes	
			Family Roles	Society Roles
Original estimate	-0.057	-0.079	0.168	0.050
	(0.272)	(0.033)	(0.078)	(0.069)
Original p-value	0.833	0.019**	0.033**	0.470
FWER p-value	0.787	0.021**	0.026**	0.704
FDR q-value	0.162	0.017**	0.020**	0.097*
Baseline characteristics	✓	✓	✓	✓

Notes: The table shows multiple-testing corrected p-values of the main estimates of the paper using two correction methods: (1) the approach of Romano and Wolf (2005*a*,*b*, 2016), which controls the family-wise error rate (FWER) using a resampling procedure; and (2) the approach of Anderson (2008), which controls the false discovery rate (FDR) and reports sharpened q-values. I apply multiple testing corrections *simultaneously* to all 18 estimates. To derive the Romano-Wolf corrected p-values, I use a bootstrapping procedure that draws 1,000 resampling of group clusters within the program-cohort strata (for student-level outcomes) and program-cohort-course strata (for student-course-level outcomes). The student-course-level estimations in Panel A, Columns 3–6, drop 392 or 1.6% observations because they are singletons within program-cohort-course strata. Clustered standard errors for the original estimates are in parentheses. $^{*}p$ <.1, $^{**}p$ <.05, $^{***}p$ <.01.

Table A.5. Summary Statistics of Baseline Characteristics by Gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Men			Women			Difference
	N	Mean	SD	N	Mean	SD	p-value
Repeating course	1,749	0.10	0.30	1,311	0.13	0.34	0.008
Age	1,457	20.35	1.29	1,112	20.11	1.34	<0.001
High school math grade	1,457	4.56	0.80	1,112	4.54	0.80	0.517
High school language grade	1,457	4.67	0.60	1,112	4.85	0.58	<0.001
First language is German	1,457	0.88	0.32	1,112	0.86	0.34	0.193
Openness	1,441	4.84	1.07	1,101	4.93	1.10	0.036
Conscientiousness	1,441	4.68	0.95	1,101	4.93	0.96	<0.001
Extraversion	1,441	4.71	1.23	1,101	4.81	1.27	0.055
Agreeableness	1,441	5.42	0.91	1,101	5.61	0.86	<0.001
Neuroticism	1,441	3.68	1.22	1,101	4.56	1.26	<0.001
Competitiveness	1,441	7.13	1.96	1,101	6.55	2.01	<0.001
Traditional gender attitudes	1,436	2.40	0.88	1,099	2.01	0.71	<0.001
Parents' income level	1,428	3.37	0.94	1,092	3.29	0.90	0.040
Parents with college degree	1,428	0.89	0.85	1,092	0.85	0.85	0.234

Notes: The table shows summary statistics of each background characteristic separately for the sample of men and women and the p-value for the test of gender difference.

Table A.6. Balance Tests by Gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	A. Men							B. Women						
	Homogeneous Groups vs.				Between Mixed Groups			Homogeneous Groups vs.				Between Mixed Groups		
	All	Minor.	Gender	Major.	Balanced vs.		Min. vs.	All	Minor.	Gender	Major.	Balanced vs.		Min. vs.
	Mixed	Gender	Balance	Gender	Minor.	Major.	Major.	Mixed	Gender	Balance	Gender	Minor.	Major.	Major.
Repeating course	0.002	0.032	-0.028	0.005	0.064**	0.031	0.052**	0.007	0.014	0.016	0.005	0.018	-0.020	0.047
	(0.021)	(0.028)	(0.025)	(0.022)	(0.025)	(0.020)	(0.025)	(0.033)	(0.048)	(0.038)	(0.036)	(0.026)	(0.026)	(0.032)
Age	-0.053	0.082	-0.089	-0.063	0.119	0.040	0.036	0.039	-0.004	0.146	-0.019	-0.089	-0.109	0.014
	(0.088)	(0.188)	(0.107)	(0.092)	(0.140)	(0.079)	(0.139)	(0.142)	(0.190)	(0.163)	(0.149)	(0.115)	(0.101)	(0.124)
High school math grade	-0.038	-0.094	0.050	-0.078	-0.076	-0.066	-0.048	0.055	0.001	-0.075	0.121	0.107	0.133	-0.050
	(0.061)	(0.123)	(0.076)	(0.065)	(0.090)	(0.053)	(0.087)	(0.091)	(0.126)	(0.100)	(0.094)	(0.066)	(0.058)	(0.070)
High school language grade	0.016	-0.049	0.010	0.029	-0.037	0.050	-0.080	0.004	-0.031	0.005	0.023	0.062	0.072	-0.079
	(0.044)	(0.100)	(0.054)	(0.048)	(0.068)	(0.038)	(0.066)	(0.044)	(0.079)	(0.048)	(0.049)	(0.055)	(0.044)	(0.057)
First language is German	0.048**	0.035	0.034	0.044*	0.017	0.003	0.007	-0.045	-0.011	-0.057*	-0.052*	0.025	-0.032	0.071**
	(0.023)	(0.047)	(0.029)	(0.025)	(0.036)	(0.021)	(0.035)	(0.029)	(0.042)	(0.032)	(0.032)	(0.031)	(0.024)	(0.033)
Openness	0.134*	0.015	0.215**	0.105	-0.212*	-0.107	-0.081	0.038	0.034	0.041	0.021	0.054	0.007	0.070
	(0.074)	(0.157)	(0.091)	(0.076)	(0.120)	(0.069)	(0.117)	(0.093)	(0.141)	(0.109)	(0.108)	(0.092)	(0.083)	(0.105)
Conscientiousness	0.024	-0.004	0.031	0.029	-0.037	0.031	-0.112	-0.004	0.026	-0.006	0.016	0.035	0.003	0.048
	(0.064)	(0.135)	(0.077)	(0.070)	(0.100)	(0.060)	(0.099)	(0.096)	(0.146)	(0.109)	(0.109)	(0.075)	(0.075)	(0.086)
Extraversion	-0.021	-0.012	-0.072	-0.001	0.074	0.054	-0.021	0.123	0.294*	0.299**	0.010	0.145	-0.109	0.199*
	(0.090)	(0.167)	(0.104)	(0.098)	(0.128)	(0.074)	(0.125)	(0.117)	(0.167)	(0.127)	(0.127)	(0.113)	(0.092)	(0.115)
Agreeableness	-0.009	-0.124	-0.038	0.003	-0.165*	-0.013	-0.151*	0.063	0.110	0.050	0.099	-0.040	0.052	-0.069
	(0.068)	(0.131)	(0.082)	(0.073)	(0.090)	(0.057)	(0.089)	(0.063)	(0.097)	(0.084)	(0.070)	(0.071)	(0.069)	(0.079)
Neuroticism	-0.008	0.005	0.079	-0.067	-0.014	-0.020	-0.063	0.047	0.111	-0.045	0.008	0.097	-0.072	0.183
	(0.084)	(0.161)	(0.103)	(0.091)	(0.127)	(0.081)	(0.125)	(0.137)	(0.145)	(0.137)	(0.149)	(0.108)	(0.093)	(0.115)
Competitiveness	0.127	0.107	0.115	0.141	0.010	0.033	-0.005	0.045	-0.010	-0.025	0.044	0.177	-0.005	0.162
	(0.136)	(0.244)	(0.162)	(0.145)	(0.211)	(0.125)	(0.201)	(0.202)	(0.253)	(0.225)	(0.221)	(0.175)	(0.155)	(0.178)
Traditional gender attitudes	0.049	0.083	0.035	0.061	0.008	-0.005	0.078	-0.022	-0.014	-0.071	0.013	0.042	0.027	0.009
	(0.058)	(0.113)	(0.068)	(0.064)	(0.096)	(0.056)	(0.095)	(0.079)	(0.122)	(0.099)	(0.075)	(0.061)	(0.053)	(0.067)
#Parents wt. college degree	0.001	-0.110	0.067	0.006	-0.180**	-0.045	-0.194**	0.023	-0.024	-0.010	0.044	0.031	0.055	-0.046
	(0.056)	(0.116)	(0.067)	(0.060)	(0.092)	(0.053)	(0.095)	(0.095)	(0.145)	(0.114)	(0.100)	(0.074)	(0.065)	(0.080)
Parents' income level	0.131**	0.112	0.153*	0.136*	-0.108	-0.018	-0.056	0.063	-0.037	0.010	0.063	0.038	0.054	-0.064
	(0.065)	(0.132)	(0.079)	(0.071)	(0.105)	(0.062)	(0.100)	(0.092)	(0.133)	(0.102)	(0.099)	(0.077)	(0.073)	(0.086)
Observations	1,428	375	671	904	523	1,052	757	1,092	335	522	460	632	757	569
Joint significance (p-value)	0.452	0.955	0.245	0.517	0.320	0.596	0.710	0.898	0.853	0.277	0.912	0.817	0.409	0.213

Notes: The table tests balance in baseline characteristics across different gender compositions separately for the sample of men (Panel A) and women (Panel B). Each point estimate is derived from one OLS regression or LPM, in which the dependent variable is a baseline characteristic, and the independent variable is an indicator for a certain gender composition. Columns 1–4 and 8–11 show the estimated effects of being assigned to a homogeneous group, compared to all mixed-gender groups or specific gender-mixture scenarios. The remaining columns show estimated effects of one type of gender mixture compared to another. All regressions control for cohort-by-program fixed effects. Robust standard errors in parentheses are clustered at the group level. For each column, I also conduct a joint test of whether baseline characteristics predict the gender composition assignment; the p-value in the bottom row indicates the joint significance of all baseline characteristics.

Table A.7. Leuven Test of Randomization

	(1) Joint Effects of Group Dummies	(2)
	F-statistic	p-value
Female	0.918	0.924
Repeating course	1.449	<0.001
Age	0.894	0.961
High school math grade	0.984	0.601
High school language grade	1.042	0.254
First language is German	0.936	0.852
Openness	0.900	0.949
Conscientiousness	0.879	0.977
Extraversion	0.964	0.715
Agreeableness	0.912	0.924
Neuroticism	1.056	0.189
Competitiveness	1.012	0.424
Traditional gender attitudes	1.023	0.357
Parents with college degree	1.017	0.392
Parents' income level	1.014	0.407

Notes: The table tests whether student characteristics are randomly assigned across study groups. For each baseline characteristic, I first regress it on cohort-by-program dummies that capture the randomization strata fixed effects. I then derive the residuals and regress them on a series of study group dummies. Finally, I test the joint significance of these group dummies. The table reports the F-statistics and p-values derived from these joint tests.

Table A.8. Impact of Gender Diversity on Economics Performance by Gender

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Analysis Sample:	A. Women				B: Men			
Dependent Variable:	Grade	Passing	Grade	Passing	Grade	Passing	Grade	Passing
Mixed-Gender Group	0.225**	0.085**			0.187***	0.076***		
	(0.105)	(0.040)			(0.071)	(0.029)		
	[0.032]	[0.036]			[0.009]	[0.008]		
Majority Gender in Group			0.211*	0.084*			0.212***	0.081***
			(0.116)	(0.045)			(0.077)	(0.030)
			[0.070]	[0.061]			[0.006]	[0.008]
Gender-Balanced Group			0.229**	0.084*			0.195**	0.080**
			(0.112)	(0.044)			(0.082)	(0.033)
			[0.041]	[0.055]			[0.017]	[0.016]
Minority Gender in Group			0.260**	0.091*			-0.058	0.022
			(0.126)	(0.051)			(0.135)	(0.051)
			[0.040]	[0.077]			[0.669]	[0.659]
Observations	1,311	1,311	1,311	1,311	1,749	1,749	1,749	1,749
R-squared	0.021	0.014	0.021	0.014	0.023	0.024	0.026	0.025
Cohort-by-Program FEs	✓	✓	✓	✓	✓	✓	✓	✓
Dep. Var. Mean	3.806	0.535	3.806	0.535	4.109	0.662	4.109	0.662
Dep. Var. SD	1.222	–	1.222	–	1.221	–	1.221	–
Minority = Balanced (p-value)			0.727	0.870			0.047	0.224
Minority = Majority (p-value)			0.629	0.885			0.032	0.211
Balanced = Majority (p-value)			0.837	0.999			0.807	0.979

Notes: Each column represents one OLS regression or LPM. While Panel A uses the sample of women, Panel B focuses on the sample of men. The dependent variable is either the *raw* final economics grade—for ease of comparison—or an indicator for passing the course. All models control for cohort-by-program fixed effects. Standard errors in parentheses are clustered at the group level; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.9. Effects on Economics Performance with Different Sets of Controls

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Standardized Final Economics Grade				
Mixed-Gender Group	0.139***	0.134***	0.125**	0.134***	0.147***
	(0.046)	(0.050)	(0.050)	(0.051)	(0.051)
	[0.003]	[0.007]	[0.013]	[0.008]	[0.004]
Proportion of Women in Group					-0.109
					(0.085)
					[0.198]
Female	-0.226***	-0.235***	-0.252***	-0.294***	-0.271***
	(0.036)	(0.036)	(0.039)	(0.038)	(0.042)
Std. High School Math Grade		0.306***	0.277***	0.269***	0.269***
		(0.020)	(0.020)	(0.019)	(0.019)
First Language is German		0.225***	0.205***	0.176***	0.175***
		(0.053)	(0.053)	(0.053)	(0.053)
Std. Openness			-0.123***	-0.118***	-0.118***
			(0.019)	(0.019)	(0.019)
Std. Conscientiousness			0.072***	0.081***	0.081***
			(0.021)	(0.021)	(0.021)
Std. Traditional Gender Attitudes				-0.045**	-0.045**
				(0.018)	(0.018)
Parents' Income Level				0.078***	0.078***
				(0.022)	(0.022)
Observations	3,060	2,569	2,542	2,520	2,520
R-squared	0.033	0.142	0.162	0.187	0.188
Cohort-by-Program FEs	✓	✓	✓	✓	✓
Educational Backgrounds		✓	✓	✓	✓
Personality Traits			✓	✓	✓
Parental Backgrounds & Age				✓	✓

Notes: The table shows the estimated effects of mixed-gender groups on final economics grade with different sets of baseline controls. Column 1 only includes student gender and cohort-by-program fixed effects. Column 2 additionally controls for high school math and language grades, German as first language, and course-repeating status. Column 3 further controls for the Big Five personality traits and general competitiveness at baseline. Column 4 also controls for age, parents' income and education levels, and traditional gender role attitudes. Finally, Column 5 also controls for the proportion of women within a group. All columns use OLS regressions and use the standardized final grade as the dependent variable. Standard errors in parentheses are clustered at the study group level; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.10. Spillover Effects on Other First-Semester Courses by Subject

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Final Course Grade				
	Banking & Finance	Mathematics	Accounting	Informatics	Business Admin.
Mixed-Gender Group	0.217***	0.144**	0.146**	0.118*	0.017
	(0.071)	(0.066)	(0.062)	(0.069)	(0.051)
	[0.002]	[0.031]	[0.019]	[0.089]	[0.740]
Female	-0.254***	-0.327***	-0.251***	-0.297***	-0.126***
	(0.046)	(0.048)	(0.044)	(0.055)	(0.041)
Observations	1,855	2,427	1,995	3,202	2,599
R-squared	0.026	0.035	0.046	0.123	0.034
Dep. Var. Mean	4.354	3.852	4.565	4.407	4.143
Dep. Var. SD	1.006	1.152	1.034	1.265	1.003

Notes: Each column represents one OLS regression, in which the dependent variable is the raw final grade obtained for first-semester courses related to different subjects: finance, mathematics, accounting, informatics, and business. Each regression controls for the course-by-program-by-cohort fixed effects. Because some subjects have only one course, the standard errors in parentheses are clustered at the group level. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.11. Effects of Mixed-Gender Groups on Performance in Other First-Semester Courses

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Course Observations:	All Courses	Compulsory Courses					
Dependent Variable:	Standardized Final Grade				Course Passing		
Mixed-Gender Group	0.094***	0.128***	0.117***		0.041***	0.028**	
	(0.026)	(0.026)	(0.027)		(0.011)	(0.012)	
	[<0.001]	[<0.001]	[<0.001]		[<0.001]	[0.018]	
Mixed-Gender Group × Female				0.143***			0.078***
				(0.050)			(0.029)
				[0.004]			[0.007]
Mixed-Gender Group × Male				0.125***			0.033***
				(0.029)			(0.012)
				[<0.001]			[0.007]
Female	-0.300***	-0.234***	-0.315***	-0.251***	-0.092***	-0.116***	-0.134***
	(0.020)	(0.019)	(0.020)	(0.054)	(0.009)	(0.010)	(0.030)
	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Observations	10,299	11,232	9,502	11,232	11,232	9,502	11,232
R-squared	0.245	0.104	0.245	0.104	0.072	0.154	0.072
Cohort–Program–Course FEs	✓	✓	✓	✓	✓	✓	✓
Baseline Characteristics	✓	–	✓	–	–	✓	–
Female = Male (p-value)				0.752			0.148

Notes: Each column represents one OLS regression or LPM. The observations in Column 1 include all courses taken by all students during the first semester, and the models in other columns focus on compulsory courses taken during the first semester. The dependent variable is either the final course grade standardized across the whole sample (Columns 1–4) or an indicator for passing a course (Columns 5–7). All models control for cohort-by-program-by-course fixed effects and cluster standard errors at the group-by-course level. Even-numbered columns additionally control for all observable baseline characteristics summarized in Table 1. Standard errors are in parentheses; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.12. Effects of Mixed-Gender Groups on Course Enrollment in the Second Semester

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Student-Level Enrollment		Group-Level: Enrolled in Any Course Common to...					
	Any Course	# Courses	≥ 2 Members		≥ 3 Members		≥ 4 Members	
Mixed-Gender Group	0.000	-0.000	0.047	0.010	-0.036	-0.034	0.004	0.029
	(0.019)	(0.095)	(0.050)	(0.052)	(0.051)	(0.051)	(0.049)	(0.049)
	[0.990]	[0.999]	[0.352]	[0.853]	[0.477]	[0.508]	[0.930]	[0.559]
Proportion of Women				0.250***		-0.016		-0.167**
				(0.083)		(0.081)		(0.076)
				[0.003]		[0.846]		[0.029]
Female	-0.040***	-0.294***						
	(0.015)	(0.075)						
Observations	2,580	2,580	645	645	645	645	645	645
R-squared	0.051	0.244	0.123	0.136	0.131	0.131	0.149	0.155
Dep. Var. Mean	0.843	3.648	0.377	0.377	0.529	0.529	0.384	0.384

Notes: The table tests whether being assigned to a mixed-gender group influences students' course enrollment decisions in the second semester (in spring). Due to data restrictions, the analysis includes 2,580 students in 645 groups—excluding the last cohort in 2024/25. Columns 1–2 examine the effects on a student being enrolled in any course and the number of courses enrolled. Columns 3–8 examine the effects on a group enrolling in any course that is common to 2, 3, or 4 group members, with or without controlling for the proportion of women in a group. All models control for cohort-by-program fixed effects. The models in Columns 1–2 cluster standard errors at the group level; Columns 3–8 use robust standard errors. Standard errors are in parentheses; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.13. Average Standardized Effects of Mixed-Gender Groups on Subjective Well-being

	(1) Group Satisfaction	(2) Self-confidence	(3) Mental Health	(4) Overall Well-being
Mixed-Gender Group	0.218***	0.114***	0.118***	0.150***
	(0.079)	(0.037)	(0.044)	(0.037)
	[0.006]	[0.002]	[0.007]	[0.000]
Baseline Characteristics	✓	✓	✓	✓

Notes: The table uses the seemingly unrelated regression framework to calculate the average standardized treatment effects on group satisfaction, self-confidence, mental health, and overall well-being (Kling et al. 2004; Clingingsmith et al. 2009). Section 4.4.1 provides details of the specific survey instruments used to measure each aspect of well-being. All regressions control for cohort-by-program fixed effects, student gender, and observed baseline characteristics. Standard errors in parentheses are clustered at the group level. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.14. Effects of Mixed-Gender Groups on Overall Well-being in the Short Run

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Standardized Index for Overall Well-being							
	All	All	Women	Women	Men	Men	All	All
Mixed-Gender Group	0.121***	0.145***	0.112*	0.144**	0.153***	0.155***		
	(0.036)	(0.039)	(0.060)	(0.073)	(0.045)	(0.048)		
	[0.001]	[0.000]	[0.062]	[0.049]	[0.001]	[0.001]		
Majority Gender in Group							0.140***	0.161***
							(0.038)	(0.041)
							[0.000]	[0.000]
Gender-Balanced Group							0.112***	0.136***
							(0.040)	(0.043)
							[0.005]	[0.002]
Minority Gender in Group							0.072	0.108**
							(0.049)	(0.051)
							[0.140]	[0.036]
Female	-0.152***	-0.117***					-0.143***	-0.110***
	(0.024)	(0.027)					(0.024)	(0.027)
Observations	1,759	1,569	781	703	978	866	1,759	1,569
R-squared	0.062	0.176	0.043	0.139	0.060	0.210	0.064	0.177
Baseline Characteristics	–	✓	–	✓	–	✓	–	✓
Minority = Balanced (p-value)							0.340	0.507
Minority = Majority (p-value)							0.070	0.152
Balanced = Majority (p-value)							0.367	0.407

Notes: Each column represents one OLS regression. The dependent variable is a normalized index measuring well-being at the end of the first semester—see Section 4.4.1. All regressions control for cohort-by-program fixed effects. Standard errors in parentheses are clustered at the group level. p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.15. Effects on Binary Measures of Traditional Gender Role Attitudes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Women not earns more than partner		Mother not works full time		College degree more for men		Business executives more for men	
	≠Disagree	Agree	≠Disagree	Agree	≠Disagree	Agree	≠Disagree	Agree
Mixed-Gender Group	0.063**	0.072***	0.060**	0.051*	-0.029	-0.014	-0.019	0.010
	(0.031)	(0.026)	(0.026)	(0.027)	(0.027)	(0.022)	(0.029)	(0.022)
	[0.041]	[0.006]	[0.023]	[0.058]	[0.288]	[0.527]	[0.498]	[0.661]
Female	0.030	0.047**	-0.074***	0.009	-0.109***	-0.015	-0.111***	-0.020
	(0.026)	(0.023)	(0.024)	(0.021)	(0.022)	(0.015)	(0.021)	(0.015)
Observations	1,571	1,571	1,571	1,571	1,571	1,571	1,571	1,571
R-squared	0.272	0.275	0.404	0.365	0.306	0.201	0.394	0.309
Cohort-by-Program FEs	✓	✓	✓	✓	✓	✓	✓	✓
Baseline Characteristics	–	✓	✓	✓	✓	✓	✓	✓
Baseline Attitudes	–	✓	✓	✓	✓	✓	✓	✓
Dep. Var. Mean	0.476	0.265	0.510	0.292	0.323	0.105	0.331	0.136

Notes: The table shows the effects of mixed-gender groups on the likelihood of not disagreeing with (in odd columns) or agreeing with (even-numbered columns) four statements about gender roles: "If a woman earns more money than her partner, this might cause problems"; "When a mother is working full-time, the child develops less well"; "A university degree is more important for a man than for a woman"; and "On the whole, men make better business executives than women do." All regressions control for the baseline level of the outcome variable, cohort-by-program fixed effects, and all observable characteristics at baseline. Standard errors in parentheses are clustered at the group level. p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.16. Effects on Survey Outcomes under Different Assumptions about Missing Data

	(1)	(2)	(3)	(4)	(5)
	Attenuated Effect		Original Effect	Strengthened Effect	
Outcome Variable	±0.1 SD	±0.05 SD		±0.05 SD	±0.1 SD
Overall well-being (short run)	0.094***	0.112***	0.145***	0.147***	0.165***
	(0.024)	(0.024)	(0.039)	(0.024)	(0.024)
	[<0.001]	[<0.001]	[<0.001]	[<0.001]	[<0.001]
Family gender role attitudes (short run)	0.081**	0.117***	0.155***	0.189***	0.225***
	(0.037)	(0.037)	(0.049)	(0.037)	(0.037)
	[0.032]	[0.002]	[0.002]	[<0.001]	[<0.001]
Moderate/severe distress (longer run)	-0.022**	-0.045***	-0.079**	-0.091***	-0.113***
	(0.010)	(0.010)	(0.033)	(0.010)	(0.010)
	[0.024]	[<0.001]	[0.019]	[<0.001]	[<0.001]
Family gender role attitudes (longer run)	0.011	0.082***	0.171**	0.225***	0.296***
	(0.027)	(0.027)	(0.080)	(0.027)	(0.027)
	[0.686]	[0.002]	[0.033]	[<0.001]	[<0.001]

Notes: The table shows the effects of the mixed-gender groups on primary survey outcomes, under different assumptions about the outcomes of participants not responding to the endline or follow-up survey. Column 3 shows the original estimates, using the same sample and specification as in Figure 2, Table 6, and Table 7. Columns 1–2 and 4–5 show the estimates when assuming non-responding students' outcomes to be equal to the mean plus or minus 0.05 or 0.1 SD of the outcome. For well-being and attitudes outcomes, Columns 1–2 assume the missing values to take values above the mean for students in homogeneous groups (control) and below the mean for students in mixed-gender groups (treatment), so that the estimated treatment effects are attenuated. By contrast, Columns 4–5 assume the opposite—the control group below the mean, and the treatment group above the mean—so that the estimated effects are stronger than the original estimates. For the distress outcome, the assumption is reversed for missing values of students in treatment and control, so that Columns 1–2 still represent attenuated effects, and Columns 4–5 represent strengthened effects. Standard errors in parentheses are clustered at the study group level. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.17. Lee Bounds on Effects of Mixed-Gender Groups on Survey Outcomes

(1)	(2)	(3)	(4)
	Original	Lee Bounds	
Outcome Variable	Estimate	Lower Bound	Upper Bound
	(Clustered SE)	(Bootstrap SE)	(Bootstrap SE)
Overall well-being	0.122***	0.111***	0.135***
(short run)	(0.036)	(0.042)	(0.044)
	[0.001]	[0.008]	[0.002]
Family gender role attitudes	0.197***	0.166*	0.219**
(short run)	(0.072)	(0.090)	(0.089)
	[0.007]	[0.066]	[0.014]
Moderate/severe distress	-0.068**	-0.140***	-0.054
(longer run)	(0.034)	(0.051)	(0.041)
	[0.050]	[0.007]	[0.188]
Family gender role attitudes	0.247**	0.080	0.413***
(longer run)	(0.107)	(0.158)	(0.155)
	[0.021]	[0.612]	[0.008]

Notes: This table presents the estimated effects of assignment to mixed-gender groups on various outcomes, along with the Lee (2009) bounds for these effects. Column 2 shows the original estimates of the mixed-gender treatment effects on different outcomes for students who responded to the endline or follow-up survey. Columns 3 and 4 report the lower and upper Lee bounds, along with clustered bootstrap standard errors, estimated using the full baseline sample with and without completing the endline/follow-up survey. The original estimates are slightly different from those in the paper because the Lee bounds calculation does not include any control variables. To increase comparability with the specification used in the paper, I first regress each outcome on student gender and randomization strata dummies, then derive and standardize the residuals across the entire baseline sample. The bootstrap standard errors are derived by resampling group clusters within randomization strata and estimating the bounds 1,000 times. The results are quantitatively and qualitatively similar when using raw outcomes instead of residualized outcomes. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.18. Effects of Gender Diversity on General Competitiveness

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: Standardized Competitiveness at Endline					
	All	Women	Men	All	Women	Men
Mixed-Gender Group	0.099*	-0.023	0.156**			
	(0.050)	(0.094)	(0.065)			
	[0.051]	[0.804]	[0.017]			
Gender Diversity				0.101*	-0.029	0.160**
				(0.054)	(0.101)	(0.070)
				[0.059]	[0.770]	[0.023]
Female	-0.122***			-0.122***		
	(0.043)			(0.043)		
Observations	1,738	784	954	1,738	784	954
R-squared	0.433	0.437	0.430	0.433	0.437	0.429
Baseline competitiveness	✓	✓	✓	✓	✓	✓
Other baseline characteristics	✓	✓	✓	✓	✓	✓

Notes: The table estimates the effects of group gender diversity on general competitiveness at endline. The dependent variable is standardized across the whole sample to have a mean of 0 and SD of 1. Each column represents one OLS regression. All regressions control for student characteristics observed at baseline, including baseline competitiveness. Standard errors in parentheses are clustered at the group level. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.19. Effects of Group Gender Diversity on Group-Level Activities

	(1)	(2)	(3)	(4)	(5)
	Voucher Collection		Self-Reported Group Activities		
	All Cohorts	Non-Covid Cohorts	Have a chat group	Met together at least once	Met together ≥ three times
Panel A: Effects of Mixed-Gender Groups					
Mixed-Gender Group	0.058	0.092*	0.076**	0.229***	0.189***
	(0.047)	(0.051)	(0.033)	(0.081)	(0.072)
	[0.215]	[0.072]	[0.024]	[0.005]	[0.009]
Observations	765	639	616	565	565
R-squared	0.097	0.113	0.043	0.067	0.049
Panel B: Effects of Group Gender Diversity					
Gender Diversity	0.084*	0.125**	0.085**	0.270***	0.222***
	(0.050)	(0.055)	(0.035)	(0.086)	(0.078)
	[0.096]	[0.023]	[0.017]	[0.002]	[0.005]
Observations	765	639	616	565	565
R-squared	0.099	0.116	0.044	0.073	0.053

Notes: Columns 1–2 use group-level observations to examine effects of gender-diverse groups on voucher collection. Column 2 excludes the 2020/21 cohort directly affected by the COVID-19 pandemic, which caused disruptions in teaching and imposed restrictions on social events. Columns 3–5 examine self-reported group activities among the 2023/24 and 2024/25 cohorts. The dependent variables include: indicator for having a WhatsApp group, indicator for meeting together at least once during the semester, and indicator for meeting together at least three times. All columns use linear probability models that control for cohort-by-program fixed effects. The models in Columns 1–2 use robust standard errors. The models in Columns 3–5 additionally control for student gender, clustering standard errors at the group level. p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.20. Effects of Group Gender Diversity on Pairwise Interactions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Dependent Variable: Standardized Frequency of Pairwise Interactions						
	All Activities	Learning Activities	Extrac. Activities.	Maximum Interaction	All Pairs	Higher-Ranked	With 4th-Ranked
Panel A: Effects of Mixed-Gender Groups							
Mixed-Gender Group	0.235***	0.271***	0.066	0.174*	0.175***	0.211**	0.346***
	(0.056)	(0.057)	(0.055)	(0.092)	(0.067)	(0.101)	(0.088)
	[<0.001]	[<0.001]	[0.231]	[0.058]	[0.009]	[0.036]	[<0.001]
Mixed-Gender Group					0.129**		
× Pair with 4-th Ranked					(0.065)		
					[0.048]		
Pair with 4-th Ranked					-0.175***		
					(0.058)		
					[0.002]		
Observations	5,504	5,504	5,504	1,835	5,504	2,923	2,581
R-squared	0.030	0.026	0.030	0.036	0.032	0.033	0.033
Panel B: Effects of Group Gender Diversity							
Gender Diversity	0.276***	0.305***	0.102*	0.207**	0.212***	0.241**	0.386***
	(0.060)	(0.061)	(0.061)	(0.099)	(0.072)	(0.107)	(0.095)
	[<0.001]	[<0.001]	[0.095]	[0.037]	[0.003]	[0.025]	[<0.001]
Gender Diversity					0.135*		
× Pair with 4-th Ranked					(0.072)		
					[0.059]		
Pair with 4-th Ranked					-0.171***		
					(0.058)		
					[0.003]		
Observations	5,504	5,504	5,504	1,835	5,504	2,923	2,581
R-squared	0.031	0.027	0.030	0.036	0.032	0.033	0.034
Own & Peer Gender	✓	✓	✓	✓	✓	✓	✓
Same-Gender Pair	✓	✓	✓	✓	✓	✓	✓

Notes: Each column in each panel reports estimates from one OLS regression. The dependent variable is the standardized frequency of pairwise interactions (mean 0, SD 1) reported at endline. The unit of observation is a student–peer pair in Columns 1–3 and 5–7. Column 1 uses all reported pairwise interactions and pools learning and extracurricular activities. Columns 2 and 3 restrict the dependent variable to learning activities and extracurricular activities, respectively. Column 4 uses student-level observations and measures each student's maximum interaction frequency with any assigned peer. Column 5 uses all student–peer pairs. Column 6 restricts to pairs with a higher-ranked peer, and Column 7 restricts to pairs with the 4th-ranked peer. All regressions control for cohort-by-program fixed effects, the student's own gender, the peer's gender, and an indicator for same-gender pairs. Standard errors are clustered at the student level in Columns 1–3 and 5–7, and at the group level in Column 4. Standard errors are in parentheses; p-values are in brackets. *p <.1, **p <.05, ***p <.01.

Table A.21. Effects on Economics Performance by Relationship Status

	(1)	(2)	(3)	(4)
Dependent Variable:	Std. Grade	Course Passing	Std. Grade	Course Passing
Mixed-Gender Group × Single	0.497	0.075		
	(0.387)	(0.099)		
	[0.204]	[0.451]		
Mixed-Gender Group × In a Relationship	-0.202	0.057		
	(0.552)	(0.292)		
	[0.716]	[0.846]		
Gender Diversity × Single			0.581	0.117
			(0.412)	(0.114)
			[0.163]	[0.312]
Gender Diversity × In a Relationship			-0.155	0.094
			(0.562)	(0.293)
			[0.784]	[0.749]
In a Relationship	0.764	0.050	0.732	0.052
	(0.704)	(0.347)	(0.661)	(0.321)
Observations	155	155	155	155
R-squared	0.043	0.059	0.044	0.060

Notes: The table uses the sample of students in the 2018/19 cohort who voluntarily reported their relationship status in the endline survey. "Single" is a dummy variable indicating single status, while "In a Relationship" indicates being in a romantic relationship. The dependent variable is either the standardized final grade in economics or successfully passing the course. All regressions control for program fixed effects and student gender. In each column, the two interaction terms are not significantly different from each other. Standard errors in parentheses are clustered at the group level. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.22. Effects on Group-Level Performance with or without Group Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Group-Level Measures of Raw Economics Grade							
	Average Grade		Minimum Grade		Maximum Grade		Grade Inequality	
Panel A: Effects of Mixed-Gender Groups								
Mixed-Gender Group	0.195***	0.125*	0.388***	0.379**	0.004	-0.019	-0.173***	-0.192**
	(0.059)	(0.072)	(0.115)	(0.156)	(0.056)	(0.072)	(0.056)	(0.078)
	[0.001]	[0.083]	[0.001]	[0.015]	[0.940]	[0.797]	[0.002]	[0.014]
Observations	765	514	765	514	765	514	765	514
R-squared	0.099	0.241	0.082	0.162	0.058	0.231	0.073	0.147
Panel B: Effects of Group Gender Diversity								
Gender Diversity	0.211***	0.132*	0.408***	0.401**	0.008	-0.024	-0.183***	-0.210**
	(0.064)	(0.078)	(0.124)	(0.166)	(0.061)	(0.077)	(0.060)	(0.083)
	[0.001]	[0.091]	[0.001]	[0.016]	[0.897]	[0.760]	[0.002]	[0.012]
Observations	765	514	765	514	765	514	765	514
R-squared	0.099	0.241	0.082	0.162	0.058	0.231	0.072	0.147
Mean of Dep. Var.	3.979	4.049	2.637	2.773	5.067	5.099	1.100	1.055
SD of Dep. Var.	0.631	0.610	1.171	1.126	0.563	0.554	0.528	0.512
Group-Level Controls:								
Diversity in high school grades	–	✓	–	✓	–	✓	–	✓
Diversity in personality traits	–	✓	–	✓	–	✓	–	✓
Diversity in other observables	–	✓	–	✓	–	✓	–	✓
Mean of non-binary traits	–	✓	–	✓	–	✓	–	✓

Notes: The table uses group-level observations to estimate the effects of gender diversity on group minimum, average, and maximum grades, as well as grade inequality, which is measured as the standard deviation of grades among four group members. Panel A estimates the effects of mixed-gender groups versus homogeneous groups; Panel B uses the continuous gender diversity measure ranging from 0 to 1. Even-numbered columns additionally include a rich set of group-level controls constructed from group members' observable baseline characteristics—available for 514 groups with non-missing values for these characteristics. The group-level controls include both the average value and the standard deviation of each baseline characteristic among the four group members, with the standard deviation capturing group diversity along that characteristic and the average value netting out level effects. All regressions control for cohort-by-program fixed effects and the proportion of women, which is included to control for level effects of gender composition (i.e., systematic gender gaps in performance); the results are robust to excluding the female proportion. Robust standard errors are in parentheses; p-values are in brackets. $^{*}p<.1$, $^{**}p<.05$, $^{***}p<.01$.

Table A.23. External Validity of Treatment Effects

(1) Outcome	(2) Baseline effect (95% CI)	(3) Observable-adjusted ATE (95% CI; $\Psi = 1$)	(4) ATE Bounds $\Psi \in [1, 2]$	(5) Ψ (ATE=0)
Final grade in economics	0.115** (0.023, 0.208)	0.129*** (0.034, 0.225)	[0.129, 0.143]	-8.3
Program dropout	-0.069** (-0.125, -0.012)	-0.060** (-0.120, -0.0003)	[-0.060, -0.051]	7.8
Overall well-being (short run)	0.1306*** (0.058, 0.203)	0.1308*** (0.058, 0.203)	[0.1308, 0.1310]	-653.0
Family gender role attitudes (short run)	0.218*** (0.071, 0.366)	0.208*** (0.062, 0.353)	[0.197, 0.208]	20.9
Moderate/severe distress (longer run)	-0.071** (-0.141, -0.001)	-0.069** (-0.138, -0.0003)	[-0.069, -0.067]	34.2
Family gender role attitudes (longer run)	0.259** (0.047, 0.470)	0.258** (0.041, 0.474)	[0.256, 0.258]	215.2

Notes: The table presents estimates from the external validity analyses using the approach of Andrews and Oster (2019). The estimation sample includes students who completed the baseline survey, including both experiment participants and nonparticipants. Column 1 lists the outcomes; all non-binary outcomes are standardized to have a mean of 0 and SD of 1. Column 2 reports the baseline effect of mixed-gender groups estimated among experiment participants, with 95% confidence intervals. Column 3 reports the observable-adjusted average treatment effect (ATE) for the target broader population, corresponding to $\Psi = 1$, with 95% confidence intervals. $\Psi = 1$ assumes selection on unobservables is as strong as selection on observables in explaining the participation decision. Observables include the baseline covariate set (age, Germanas first language, high school math and language grades, personality traits, competitiveness, gender attitudes, repeater status, and parents' education and income). Column 4 reports bounds on the implied target ATE for $\Psi \in [1, 2]$; $\Psi = 2$ assumes that selection on unobservables is twice as strong as selection on observables. Column 5 reports $\Psi(0)$, the value of Ψ that sets the implied target treatment effect to zero. Confidence intervals are normal-based intervals from a cluster bootstrap that resamples groups within program-by-cohort strata for 1,000 replications. $^*p < .1$, $^{**}p < .05$, $^{***}p < .01$.

Table A.24. Effects of Gender Diversity on Group-Level Performance in Female-Dominated Settings

	(1) Group Average Grade	(2) Group Minimum Grade	(3) Group Maximum Grade	(4) Within-Group Grade Inequality
Panel A: Effects of Mixed-Gender Groups				
Gender-Mixed Group	0.297	0.754*	-0.153	-0.385*
	(0.232)	(0.394)	(0.162)	(0.197)
	[0.203]	[0.060]	[0.348]	[0.055]
Group Observations	81	81	81	81
R-squared	0.211	0.115	0.234	0.110
Panel B: Effects of Group Gender Diversity				
Gender Diversity	0.361	0.760*	-0.142	-0.374*
	(0.258)	(0.432)	(0.185)	(0.217)
	[0.165]	[0.083]	[0.447]	[0.088]
Group Observations	81	81	81	81
R-squared	0.213	0.109	0.232	0.101
Dep. Var. Mean	3.048	1.068	4.574	1.618
Dep. Var. SD	0.851	1.369	0.682	0.656

Notes: The table estimates the effects of group gender diversity on different group-level performance measures, using 81 group observations from replications of the experiment in two math courses during fall 2024. The dependent variables are raw measures of group average grade, minimum grade, maximum grade, and grade inequality (within-group standard deviation of grades). The raw final grades for these courses also range from 1 to 6. The study groups are randomly assigned within courses and within two broad categories of majors (majors related bio-/chemistry and bio-/medicine vs. other majors). All regressions control for the proportion of women in a group, the course fixed effects, and major category fixed effects. Robust standard errors are in parentheses. p-values are in brackets. $^{*}p < .1$, $^{**}p < .05$, $^{***}p < .01$.