Effects of team diversity on individual performance and voice: A field experiment of group composition by gender and language*

Valentina Contreras[a], Chiara Orsini[b,a,d], Berkay Özcan[ac], and Johann Koehler[a]
[a]The London School of Economics and Political Science
[b]The University of Sheffield [c]New York University Abu Dhabi [d]IZA

Abstract

We present results from a field experiment that tests the effects of varying gender and linguistic group composition on performance and on group-members' perception that their voice is heard when completing complex collaborative work within a low scrutiny environment. We randomize individuals enrolled in a postgraduate course populated by mostly women and non-native English speakers into small teams within larger, exogenously assigned seminar groups. Groups are tasked with complex and deliberative research assignments over three months. Using administrative and survey data, we find that a higher share of women in seminar groups significantly benefits the academic performance of group members—an effect driven by a positive effect on female native English speakers — while a greater proportion of women in small teams improves non-native language speakers' perception of being heard.

Keywords: Team Dynamics, Gender, Linguistic diversity, Peer effects, Higher Education, Field experiment

* We are grateful for helpful comments and suggestions from Stephen Jenkins, Sarah Brown, Almudena Sevilla, Lindsey Macmillan, Tania Burchardt, Thomas Biegert, and the participants of the LSE's Social Policy quantitative reading group, the 1st Diversity and Human Capital Workshop of EDGE Network at the University of Exeter, the second NXNW workshop, and EALE-Bergen. All errors are our own.

1. Introduction

As women and migrants increasingly enter higher education and labor markets, the nature and profile of professional work correspondingly changes. Those changes will be particularly felt in group work, where newer entrants must collaborate with the old guard. But as the make-up of collaborative workgroups changes, particularly in their gender- and linguistic-diversity, it becomes increasingly important to understand how the members of those groups fare.[1] Yet, intuitions about the effects of workplace diversity on individual outcomes point in mixed directions. For firms, increased diversity might bring the benefits of a broader range of views and skills, or it might introduce new communication difficulties and friction (Lazear 1999). Likewise, in universities, students may gain from exposure to a diverse student body, or diversity might disturb instruction's pace and flow (Diette & Uwaifo Oyelere 2012). However, testing the causal effect of a group's linguistic and gender diversity is challenging because people often sort themselves into homophilous groups; that self-selection, in turn, threatens valid estimation of group composition's effects. Consequently, without causal identification, it remains unclear what to expect from workplace diversity or how to manage it effectively.

We therefore conducted the first *field experiment* to test how gender and linguistic diversity in group composition affects individual performance and perceptions of voice in a dual-layered group setting.[2] The study unfolded in two consecutive iterations of a compulsory one-term postgraduate research methods course with cohorts of ~180 Master's students, 80% of whom were women and 51% non-native English speakers. Students were mature learners with an average age of 24.6, many with prior work experience. In this course, university administrators assigned students to "seminar" groups of approximately 16 people each. Following this exogenous assignment, we then randomly allocated individuals into four smaller "teams" of about four students each within their seminars.[3]

We estimate the causal effect of gender and linguistic diversity on two main outcomes: individual performance and perceptions of voice. For individual performance, we examine how group composition

[1] Consider three shifts in particular: first, women's participation in tertiary education nearly doubled from 2000 to 2014 (UNESCO 2022), and female-to-male workforce participation in OECD countries rose from 70% in 2000 to 78% in 2024 (World Bank 2025). Second, OECD countries received an average of eight migrants per thousand inhabitants, and international students made up ~13% of Master's and ~22% of PhD enrolments (OECD 2020). Third, jobs requiring high social interaction grew nearly 12 percentage points as a share of all jobs in the USA (Weidman and Deming 2021). The modern workforce must therefore adapt to the growing need to collaborate across gender and language differences.

[2] Ethical approval for this study was obtained by the LSE and the University of Sheffield.

[3] From this point on, when referring to our empirical setting, we use *team* to denote the smaller group of four students and *seminar group* for the larger group of sixteen students.

shapes each member's achievement on complex, research-oriented tasks. Second, we focus on individual perceptions of how much a group member's voice was heard during group deliberations, as the supposed benefits of diversity depend on whether their contributions are meaningfully considered.[4]

Throughout the term, students completed two weekly research tasks that applied lecture concepts to specific policy problems: The first task required members of a four-person team to confer privately and agree upon answers to interpretively complex policy research questions, which they then submitted as a joint team response.[5] Teams then convened publicly with one to three other teams during a weekly in-person seminar meeting (comprising roughly sixteen students) to complete a second activity. Like the earlier task, the second activity also required deliberation over complex policy research questions; however, unlike the earlier task, seminar activities required each team to present and defend its answers against those submitted by other teams.[6]

Although neither the team discussions nor seminar participation were directly graded, students faced a strong incentive to engage: the final exam assessed the same skills but without the benefit of group interaction. Therefore, complex collaboration and deliberation remained the key pedagogical focus of both the small team component comprising four students and the larger seminars comprising multiple teams. However, the two course components involved different group dynamics. In small teams, private completion of the weekly tasks required navigating gender and linguistic differences to sustain an intensive, ten-week collaboration within a fixed group. Frictions arising within one week had to be resolved by the next; friendships and allegiances had to be brokered; and hierarchies navigated. These backstage social dynamics stood in contrast to, and interacted with, the frontstage dynamics that played out once teams subsequently assembled in the seminar classroom. In the seminar, students publicly navigated group dynamics that were both *more* and *less* exposed than in team settings: seminars were *more* exposed because their contributions unfolded in front of the instructor and other classmates; they were also *less* exposed because social retreat within a larger crowd could go unnoticed. These contrasting dynamics were therefore likely to produce different effects: we expect public seminars to promote

[4] Prior research shows that feeling heard enhances individuals' sense of control (Folger, 1977) and increases satisfaction and motivation at work (Greenberger & Strasser, 1986). In contrast, feeling ignored or dismissed (Folger, 1977; Pinder & Harlos, 2001) can frustrate employees and undermine the benefits of collaboration in diverse groups.

[5] As an example, teams assessed which analysis of a contentious social policy issue (*e.g.*, racial discrimination in police violence) was more persuasive, and they then defended their decision in a jointly authored statement visible to classmates.

[6] As an example, in one week, teams classified 15 studies using the five-level Cochrane Hierarchy of Evidence (from correlational study to randomized trial). Disagreements were recorded for seminar discussion (see appendix A.1).

individual performance as larger groups broadened the base for feedback and refinement, whereas private team settings were more likely to foster a sense of voice and inclusion.

However, variation in group composition complicates these expected effects. Homophilous groups may facilitate smoother discussion, sharpen deliberative skills, and foster networks of like-minded peers; or, conversely, they may intensify exclusionary dynamics and stifle discussion. Gender and linguistic diversity bring these mixed expectations into sharper focus. On one hand, diverse work-groups may benefit from the inclusion of more women, who are often associated with higher supportive and collaborative skills (Jemieson et al. 1995; and Manne, 2017). On the other hand, tasks requiring high language proficiency may impose communication barriers that non-native speakers struggle to overcome (Rodriguez and Cruz 2009; Stebleton et al. 2010).

The set up with two levels of group interactions—gender and linguistic diversity on one hand, public seminars and private team deliberation on the other—promises substantive and professional contributions to knowledge. Substantively, we can test main and interaction effects of different forms of disadvantage, which may operate in distinct ways. Setting those differences alongside one another could sharpen disadvantage, or they may blunt it. For instance, if addressing gender disadvantage yields positive effects but addressing linguistic disadvantage does not—or *vice versa*—then efforts to form groups with a 'critical mass' of disadvantaged workers may need rethinking. Indeed, our findings affirm this challenge: a higher share of women in seminars increases the likelihood of both women and native speakers earning a distinction grade in the program overall and in the dissertation, this effect is driven by female native English speakers. In contrast, in smaller team settings, a higher share of women increases the likelihood that non-native speakers feel their voice was heard during group discussion. This latter finding suggests that it is in smaller group dynamics where more women can foster inclusivity for those with potential language barriers.

Our analysis also offers professional insights. Similar structures exist in research and development departments in private and public organizations, where small teams first brainstorm before presenting new ideas to a larger group. Although such interactions matter, it is not clear *ex-ante* whether homophilous teams enhance individuals' sense of being heard or promote excellence in individual performance. Given that many of our participants were mature students with prior professional experience, our results are especially relevant wherever workplace or classroom diversification intersects with tasks involving collaborative, high-level problem-solving and innovation—especially when diversification cuts across gender or linguistic lines. Examples include course design in educational

settings and team formation in industry and the public sector. In such contexts, our findings suggest assigning non-native language speakers to teams with a critical mass of women may foster greater inclusion. More broadly, when excellence in innovation is the goal, our results suggest that groups with more women tend to yield stronger individual performance outcomes.

2. Contributions to the literature

Our study contributes to the literature on diversity in team settings by drawing on a uniquely rich field experimental design. We are, to our knowledge, the first to examine the effects of gender and linguistic diversity in a female-dominated, international setting with mature participants. These individuals engage in research-oriented tasks over several weeks, with repeated interactions in small private teams and larger public-facing groups. The low-scrutiny, real-world environment allows us to observe not only performance but also how participants perceive their ability to contribute and be heard.

These features collectively enable us to advance the literature in *three* broad directions: First, we disaggregate the effects of diversity across two dimensions of group composition—gender and language—and examine a broader set of outcomes beyond productivity, including voice being heard. Second, our field-based design enhances external validity by embedding the experiment in a realistic setting of policy-oriented research collaboration. Third, we distinguish between "frontstage" and "backstage" peer effects in small teams and in larger groups, offering a more fine-grained picture of how group dynamics shape inclusion and influence.

Our first contribution is to clarify how diversity affects team collaboration by separating the effects of gender and language composition, and by looking beyond performance outcomes to consider whether individuals feel heard in group work. Existing research in professional and educational settings tends to focus on a single aspect of group composition – typically gender – and on narrow outcome measures such as productivity or test scores. In workplace settings, where field experiments remain rare, studies often examine gender composition effects on performance[7] but overlook linguistic diversity and perceptions of voice or inclusion (see the review by Knyazeva, Knyazeva and Naveen, 2021). Our study

[7] An important exception is the paper by Dahl et al. (2020), which, in addition to performance and satisfaction with the service, focuses on outcomes such as field of study, occupation, workplace gender composition, and short and longer-term gender attitudes.

brings these dimensions together in a single field experiment and provides new insight into how participants experience collaborative group work.[8]

Educational settings have offered somewhat richer evidence on peer diversity but the findings are mixed and largely based on school-age children in primary and secondary schools, raising questions about their relevance for adult environments. Research on gender composition, for instance, shows that a higher share of girls in school classrooms results in improved test performance (Hoxby 2000; Lavy & Schlosser 2011)[9], though results on longer-term outcomes and aspirations are more equivocal (*e.g.*, Anelli & Peri 2019; Black et al. 2013; Schneeweis & Zweimüller 2012). Studies on linguistic diversity show similarly varied effects: Diette & Uwaifo Oyelere (2012) report heterogeneous effects in the USA, with small positive effects for low and mid-achieving native students' scores, but small negative effects for those at the top; in contrast, Geay et al. (2013) find no negative effect from the presence of non-native English speakers in UK primary school performance. Research on the effects of immigrant student shares on academic outcomes also show mixed results: Gould et al. (2009) and Jensen & Rasmussen (2011) find negative effects in Israel and Denmark; Schneeweis (2015) observes negative effects only for migrant students themselves in Austria.[10]

Taken together, prior research in professional and educational settings generates no clear consensus about linguistic diversity's effects in group settings, but evidence suggests that being in the minority is bad for women. We know little about how gender and linguistic composition determine group members' sense of inclusion and engagement. Also, prior work has concentrated on diversity's effects in narrowly defined tasks with easily measured outcomes in specific jobs (Owan, 2014). We know far less about how teams fare, beyond productivity, when they undertake complex problem-solving tasks (Azmat, 2019). Our study addresses these gaps by examining the effects of groups working on policy-oriented research tasks in an international course-setting, and by focusing on individuals' sense of being heard alongside performance.

[8] Studies in this area have examined how leaders' gender shape workplaces. For instance, Alan et al. (forthcoming) using data on white-collar professionals in Turkiye, find that female leaders foster less segregation and lower quit rates. Also, Born et al. (2022) study "leadership" randomizing participants into sex-varied teams solving hypothetical survival scenarios. They find men are more likely to assume leadership, regardless of team composition.

[9] Fennoll & Zaccagni (2022) study high schoolers in northern Italy solving math problems in random vs. self-formed teams. Female-dominated teams underperform when randomized, but the gap vanishes when teams self-select.

[10] Using data from Turkiye, Alan et al. (2023) show that teacher prejudice increases peer violence toward refugee children and weakens inter-ethnic ties.

Second, we contribute to the limited body of *field experiments* on how team communication unfolds in diverse teams. Much of the existing literature relies on short laboratory-based experiments with tasks involving strangers, where collaboration is limited to brief, one-off artificial engagement. These studies often find that gender norms shape participation and influence (*e.g.*, Coffman et al. 2021; Hardt, Mayer and Rincke 2023). Field experiments in more natural settings clarify whether those findings prove durable: For instance, Karpowitz et al. (2023) show that women in the numerical minority are disproportionally less likely to be rated as influential in team deliberations or to be chosen as a spokesperson for their team relative to women in female-majority teams.[11]

Our study extends this work by observing how diverse groups function over time among mature, international participants—imminent entrants to the workforce—engaged in a research-oriented task that allowed repeated, organic interactions. In particular, our analysis helps make sense not only of how group dynamics determine individual performance *and* team members' perception of inclusion in collaborative work in a female-dominated setting; it also disentangles how those effects cut differently for women and for international students who may face steeper language or cultural barriers.

Third, we contribute to the literature on *peer group effects* by leveraging a dual-layered setting. This dual-layered setting—consisting of the private, backstage completion of collaborative tasks in small teams (4–5 members) and then the public, frontstage completion of further tasks in mid-sized groups (8–16 members)—yields a more differentiated picture of group effects than prior research.

Indeed, prior research typically defines peer exposure at the level of entire university cohorts or classrooms and reports mixed findings. For instance, Braakmann & McDonald (2018) find heterogeneous effects of *cohort-level diversity* on academic outcomes among students exposed to different gender, ethnicity, and socioeconomic backgrounds. Chevalier et al. (2020) show that linguistic diversity *within classrooms* benefits non-native English speakers in UK universities. Oosterbeek & Van Ewijk (2014) find no significant effect of peer gender composition at the classroom level. Feld and Zölitz (2021) show that gender composition in large business school sections influences major choices and labour market outcomes. These divergent findings suggest that context may matter not just in degree (i.e. small teams, classrooms or entire cohorts), but in kind —depending on the scale. If context matters, and especially if it structures diversity's effects differently in different settings, then we need to know how.

[11] Karpowitz et al. (2023) present results from two studies: one from a program where women are the minority, and another from a program where women are the majority.

Our study directly addresses this by testing the effects of gender and linguistic diversity on two separate sets of group dynamic—one among small teams, and then another among mid-sized seminars.

Together, these contributions offer new insight into how gender and linguistic diversity shape individual outcomes in collaborative research settings—across both private and public interaction—under conditions that resemble real-world collaborative work more closely than existing designs.

3. Data

We draw on administrative and survey data to examine the relationship between student characteristics, academic performance, and perception of being heard in a graduate-level course. The administrative data comprises records from the university's Registrar that include final course grade and dissertation grades, seminar teachers' and academic advisors' characteristics, as well as self-reported information on demographic characteristics, admissions information and previous academic background.

We also collected survey data in two stages for each cohort. First, students completed a baseline questionnaire in the first week of the term. Second, students completed an endline questionnaire in the final week of term, several weeks before the final exam, ensuring that their perceptions of team dynamics and their own contributions were not influenced by exam performance. The endline survey repeated some items from the baseline survey, and also contained items that captured respondents' reflections about the group-work to which they had contributed, including self-assessments of their contribution to the collaborative group-work, as well as their perception that their voice was heard during group discussions.

3.1 Administrative data

Table 1 presents summary statistics for the administrative data for both cohorts of students. Panel (a) of Table 1 shows that the students' average age is 24.6, 79.5% of students are women, and 49% of the students are native English speakers. Women account for 84.2 % of the non-native speakers, and 75.6 % of the native speakers, while native speakers make up 60.0 % of the male subsample versus 46.6 % of the female subsample. Panel (b) presents statistics about the students' previous academic backgrounds. Most students held a Bachelor's degree as their highest qualification (81.9%), and 18% held a prior Master's degree before starting the MSc. Panel (b) also shows that 33% of students have either a completed or pending qualification from a university in the United Kingdom. Panel (c) of Table 1 presents three sets of academic outcomes: Exam Mark corresponds to average performance in SP401's course-specific exam, which accounted for the totality of the course grade; Dissertation Distinction

corresponds to the proportion of students who earned the highest grade classification in their year-long capstone thesis; and Programme Distinction corresponds to the proportion of students who earned the highest grade classification across all the coursework in the degree. The average exam mark for women, men, native English speakers, and non-native English speakers. Men and women performed similarly (earning, on average, a mark equal to 68/100); however, there were differences by native language. Native English speakers earned an average mark of 72.5 while non-native English speakers earned an average of 66.7. Panel (c) also shows that only 26% of non-native English speakers earned distinction in their dissertation, whereas 38.95% of native English speakers did so. Finally, only 20.9% of non-native English speakers gained distinction in the degree, whereas 42% of native English speakers did so.

3.2 Survey data

Table 2 presents summary statistics of the data we collected through the two surveys.

3.2.1 Baseline

The baseline survey collected self-reported data on students' native language, their usual role in group work, and their familiarity with relevant subjects. We then used this data to derive key variables for our analyses. For instance, for non-native English speakers, we derived a measure of the linguistic "distance" between the student's native language (as reported in the baseline survey) and the English language. The variable allows us to capture heterogeneity between non-native English speakers. We use Chiswick & Miller (2005)'s linguistic distance scale, which ranges from 1 to 3, in 0.25 increments, with three being the most similar to English. In our sample, students' distance scores covered the full range from 1 to 3, with a mean of 1.9, reflecting substantial linguistic diversity. Appendix A.2 outlines details on students' country of origin (Table A.2.1), native language, and their correspondent measure of linguistic distance to English (Table A.2.2). Among non-native English speakers with available distance scores, Chinese is the most frequent native language (27% of students, mean distance = 1.5), followed by Spanish (14%, distance = 2.25), French and Italian (8%, distance = 2.5), and Hindi (8 %, distance = 1.75).

Furthermore, we collected information on students' familiarity with relevant subjects and their expected final mark in the course. Most students reported some experience with research methods, with similar average familiarity with qualitative and quantitative methods. Interestingly, there is homogeneity in the average mark members of all subgroups expected to earn in the exam.

3.2.2 Endline survey

The endline survey asked students about the team dynamics that they perceived over the preceding three months, as well as to predict their future contribution to team-based work. The bottom panel of Table 2 presents the mean and standard deviation values by group for the relevant endline survey data.

The survey included three questions related to the students' perception of their "voice" in team interactions. Prior literature on group dynamics defines voice as "as informal and discretionary communication by an employee of ideas, suggestions, concerns, information about problems, or opinions about work-related issues to persons who might be able to take appropriate action, with the intent to bring about improvement or change" (Morrison, 2014). Students rated their agreement (on a scale from 0 to 10, with 5 measuring "neither agree or disagree") with the statement *"My voice was heard during group discussions",* and with two follow up items: *"Working in teams for SP401 made me more confident than before in voicing my view in future interactions"*, and *"Working in teams for SP401 made me more confident than before that my view will be heard in future interactions"*. Table 2 shows summary statistics for these outcomes. Most students agreed that their voice was heard during discussions, with women and native English speakers particularly more likely to agree. On the two follow-up items, respondents, on average neither agreed nor disagreed; however, women and non-native English speakers again tended to agree more than their counterparts.

4 Empirical strategy

4.1 Random Assignment

Our field experiment mitigates selection problems because group formation is exogenous to students' characteristics and outcomes of interest at every level: to the course, the seminar, and the team. First, there is no self-selection at the course level: SP401 was compulsory for all students in the programme. Second, students were allocated to seminars by the course administrator without regard to gender, language background, or other observable traits. Students could request changes in their allocated seminar group only in cases of a timetable clash. In practice, this applied to a very small number of students in each cohort. Because SP401 was the largest compulsory course in the programme, its seminar timetable was prioritized in scheduling. Administrators also aimed to place SP401 seminars at the end of the week when most other course lectures had already taken place, reducing the likelihood of conflict. When clashes did occur, affected students were reallocated to one of the remaining available seminars, without reference to their preferences or characteristics. To maintain balance across seminars,

administrators sometimes randomly reassigned a few additional students after the re-allocation. Taken together, these practices mean that any departures from pure random allocation at the seminar level were rare and not systematically related to the outcomes we study.

Third, and most crucially, to identify causal effects of the team composition, we further randomized students within each seminar into teams of approximately four students each. As seminar size varied between cohorts, there are either two or four teams per seminar, but the total number of teams remained constant across cohorts. This randomization created teams of varying gender composition and share of native English speakers. Table A2.3 shows a breakdown of teams and individuals by gender and language composition.

To corroborate that assignment to seminars was as good as random, we use a regression-based test. The test examines the within-group correlation between each individual's characteristics of interest (native language and gender) and the average characteristic of their peers within the reference group (seminar).[12] Table 3 reports the test statistics and two-sided *p*-values for assignment to seminars, as well for our randomization to smaller teams (within seminars). *P*-values are large for all tests, indicating insufficient evidence to reject the null hypothesis of random assignment to seminars and teams. These results provide evidence that the allocation to seminars and teams was exogenous to the student's gender and native English speaker status.

Additionally, we test whether the proportion of women and native English speakers in seminar groups is systematically correlated with individual characteristics. Table 4 presents a series of balancing tests assessing whether the variations in the share of females and native English speakers are associated with the individual's gender, age, previous UK studies experience, highest level of education, and familiarity with quantitative and qualitative methods, expected mark, and an indicator for usual role in the team (leader). Across the 16 tests performed, no correlation is significant at the 5% level and only one appears to be significantly different from zero at the 10% level. These results suggests that the allocation of

[12] Specifically, we use the test proposed by Jochmans (2023). The regression-based test described in Jochmans (2023) is a test for the (conditional) random assignment of individuals in urns to peer groups. In this test, the dependent variable is a characteristic of the individual, and the independent variable is the average characteristic of the individual's peers. The test controls for fixed effects at the urn level and additional covariates. The idea is that if conditional random assignment cannot be rejected, then the coefficient on the peer-group average should not be statistically different from zero. Although the idea of a regression-based test of random assignment traces back to Sacerdote (2001), Jochmans (2023) provides a corrected *t*-statistic that is robust to varying urn and peer group sizes, accommodates designs where peer groups are not mutually exclusive, and is robust to arbitrary forms of heteroskedasticity.

students to seminar groups was unlikely to be systematically influenced by these individual characteristics.

4.2 Estimation

We use a linear model to estimate the causal impact of the proportion of women and the proportion of native speakers in a seminar group and in a team on all the outcomes of interest. Manski (1993) introduced the original model to estimate peer effects, which attributes outcomes to individual characteristics and the characteristics of a group to which an individual belongs. We extend that model to capture the effects on each of the outcomes of interest of the variation in gender composition and the share of native English speakers as follows:

$$Y_{igs} = \alpha + \beta_1 NS_{-ig} + \beta_2 NS_{-gs} + \beta_3 W_{-ig} + \beta_4 W_{-gs} + \gamma X_{is} + C_i + \epsilon_{igs} \qquad (1)$$

where Y_{igs} is the outcome of interest for student i in team g and seminar s; NS_g is the proportion of native English speakers in team g excluding student i; and W_g is the gender composition of team g excluding student i. NS_{gs} and W_{gs} are the proportions of native English speakers and of women in seminar s, excluding students from student i's own team g. X_{is} is a vector of control variables including age, familiarity with course-relevant subjects, dummies for English as first language and gender, the student's highest level of education, prior UK study experience, and seminar teacher and academic adviser characteristics (gender, and native speaker status). Seminar teachers and academic advisers are also assigned by the program administrators without reference to any student or adviser characteristics. Seminar teachers are allocated to seminar timeslots based on their availability, prior to student enrolment. Academic advisers are entirely randomly assigned by the administrators. Additionally, when estimating equation (1) for the subsample of non-native speakers, we include Chiswick & Miller (2005)'s language distance scores in X_{is}. C_i is cohort fixed effect.

Note that teams are a subunit of the seminar. Therefore, seminar composition varies with the composition of the teams. Thus, excluding students in the same team when calculating seminar-level measures of the proportions of women and native English speakers in the seminar (W_{gs} and NS_{gs}) helps avoid multicollinearity. More importantly, as students cannot self-select into seminar groups or teams, the gender and language compositions at both the team and seminar levels (W_g, W_{gs}, NS_g and NS_{gs}) are exogenous to student outcomes. Therefore, the coefficients of interest β1, β2, β3, and β4, can be interpreted causally as the effects of gender and language diversity on the outcomes of interest.

In addition to the baseline specification, we estimate an interacted model to test whether the effects of both team and seminar composition vary by student gender and native speaker status. Specifically, we interact the composition measures (W_g, W_{gs}, NS_g and NS_{gs}) with indicators for whether the student is female or a native English speaker, as follows:

$$Y_{igs} = \alpha + \beta_1 NS_g + \beta_2 NS_{gs} + \beta_3 W_g + \beta_4 W_{gs} + \delta_1\left(NS_g \times Z_i\right) + \delta_2\left(NS_{gs} \times Z_i\right) + \delta 3\left(W_g \times Z_i\right) + \delta_4\left(W_{gs} \times Z_i\right) + \gamma X_{ig} + C_i + \varepsilon_{igs} \quad (2)$$

where Z_i is a vector that includes indicators for the student's gender and native English speaker status.

Finally, to test for potential non-linearities in the effects of group composition, we estimate a third specification where the proportions of women and native English speakers are replaced with vectors of categorical indicators. This allows us to capture potential threshold effects or diminishing returns to diversity. The model takes the following form:

$$Y_{igs} = \alpha + \theta_1 NS_{-ig}^{cat} + \theta_2 NS_{-gs}^{cat} + \theta_3 W_{-ig}^{cat} + \theta_4 W_{-gs}^{cat} + \gamma X_{is} + C_i + \varepsilon_{igs} \quad (3)$$

Where $NS_{-ig}^{cat}, NS_{-gs}^{cat}, W_{-ig}^{cat}, W_{-gs}^{cat}$ are vectors of categorical indicators denoting whether the share of native English speakers or women in the team, or seminar falls into "low," "medium," or "high" ranges. These shares are calculated excluding the individual student from the team-level measures and excluding the student's own team from the seminar-level measures. For gender composition, we define the categories as low (≤50%), medium (>50% and <100%), and high (=100%). For language composition, we use cutoffs of 35% and 75% to define low, medium, and high categories. This approach allow us to explore whether effects are concentrated in particular ranges of group diversity.

We estimate these equations for two sets of outcomes: individual academic performance and perceptions of voice in team interactions. Individual academic performance outcomes include examination marks, distinction in dissertation, and distinction in the overall programme, for which we estimate equation 1 using ordinary least squares (OLS). Distinction is the highest classification students can achieve in the UK system, signifying a mark equal to or greater than 70%, and is measured as a binary outcome: 1 if the student earns a distinction, 0 otherwise. Standard errors are clustered at the cohort and seminar level. To estimate the effects on perceptions of voice we use endline survey responses to the item "Voice was heard", measured on a scale of 0-10. We apply OLS to estimate equation (1) for all dependent variables. A more detailed description of outcomes of interest is provided in the Online Appendix OA.1

To ensure that the peer groups we study are those influencing students' academic performance, we collected data on peer interactions outside the mandatory course activities. In the endline survey, we asked "Did you ever study with people outside your study group for this course?" We find that 32.8% engage with peers from only their assigned study group and respective seminar; 38.8% studied with their seminar group peers as well as with peers from other seminar groups within the same course; 17.9% studied exclusively with students from other seminar groups within the course; and 10.4% studied only by themselves. Overall, more than 70% of participants maintain their engagement within their assigned team or seminar group.

Our endline survey achieved a 60% response rate, substantially higher than the average response rate for online surveys (~44.4% according to Wu et al.,2022) and the typical response rate for student surveys in UK universities (*e.g.*, for teaching evaluations), which persistently averages 25-30%. Still, we adopt two strategies to account for survey nonresponse. First, we adjust for non-response using inverse probability weighting before running the regressions.[13] We estimate the probability of response conditional on observable characteristics and assign each respondent a weight corresponding to the inverse of their cell's response probability. Second, we calculate Oster (2019) bounds around the treatment effects, assuming that the relative importance of observed and unobserved omitted variables in generating selection bias is the same. See Appendix A.4 for estimation results and further details.

Finally, we adjust *p*-values following Barsbai et al. (2024) to account for the potential multiple hypothesis testing (MHT) problem.[14]

[13] For a detailed description of the inverse probability weighting method, see (Hernán & Robins, 2020). We present odd ratios of the logistic regression of endline response in Appendix OA.3.

[14] In empirical work, multiple outcomes raise concerns about multiple inference: significant coefficients may emerge by chance. A common approach to adjust for multiple hypothesis testing is to control for the Family Wise Error Rate (FWER). When a family of K hypotheses is tested, of which J are true, the FWER is the probability that at least one of the J true hypotheses is rejected. Within the FWER correction methods, the Bonferroni's correction is the most well-known. However, the Bonferroni's correction suffers from poor power, and its calculated upper bound *p*-values can sometimes exceed 1 (Anderson, 2008). Additionally, it does not account for dependence between outcomes, which is suboptimal when outcomes are correlated (Anderson, 2008), as in our study. List et al. (2019) propose a FWER correction that overcomes Bonferroni's correction issues with *p*-value size and outcomes' dependence, better detecting truly false null hypotheses. Barsbai et al. (2024) modify the List et al. (2019) approach to allow regression-based implementation with control variables. We use the Stata command -mhtreg- provided by Barsbai et al. (2024) to implement the regression-based multiple hypothesis testing correction.

5 Results

In this section, we present the results of our estimations concerning the effects of the proportion of women and the proportion of native English speakers within the context of collaborative team activities and seminar discussions.

5.1 Effects of group composition on individual performance

We study three performance outcomes: final exam marks, final programme grade (distinction), and dissertation grade (distinction). Our main explanatory variables are proportional measures: the proportion of women and the proportion of native English speakers in the team and the seminar group (excluding the individual's own team). The estimated coefficients in Equation 1 quantify the effect of moving from a group with 0% to 100% women (or native speakers) among peers outside their own team.

Table 5 presents estimates for exam performance. Table 6 presents results for obtaining distinction in the final programme grade (Panel (a)) and dissertation distinction (Panel (b)). In each table, Column (1) reports the estimates for the full sample, while Columns (2) to (5) report estimates for subsamples: women, men, native English speakers, and non-native English speakers, respectively.

Final Exam Marks

Table 5 presents the effects of group composition on final exam marks.[15] We find no statistically significant effects across all subsamples, consistent with prior work showing that the impact of peer group diversity often materializes in longer-term rather than immediate outcomes (Fisher 2017; Zölitz and Feld 2021).

Final Programme Grade

Table 6, Panel (a), shows the effects of group composition on the likelihood of graduating with a distinction in the programme (*i.e.*, the Master's degree). We find that a higher proportion of women in the seminar group increases the probability of distinction for both women and native English speakers. For women, the coefficient is 0.46 (Column 2, Row 4), with a *p*-value of 0.04 after the multiple hypothesis testing correction (Column 2, last row). For native English speakers, the coefficient is 0.60

[15] We also tested potential interaction effects and non-linear heterogeneous (Appendix A.3 (Tables A3.1, and A.3.5), as for the main specification, we find no statistically significant evidence of interactions or non-linearities in composition effects.

(Column 4, Row 4), with a corrected *p*-value of 0.07 (Column 3, last row). This means for a native English speaker in a seminar with two small teams of four, adding one woman to the other team (a 25% increase in the proportion of women among seminar peers) would increase their likelihood of graduating with distinction by 15% (0.6 × 25%).

The fact that native English speakers, but not non-native English speakers, benefit from a higher share of women in seminars is noteworthy. Seminar discussions are designed to encourage active engagement and feedback from instructors and peers, allowing participants to refine their ideas through deliberation. Although a greater proportion of women may ease participation for some, language barriers might continue to impede non-native English speakers from fully engaging and thus from reaping the potential benefits of these seminar dynamics. A similar pattern emerges in final programme grades, where both women and native English speakers benefit from a higher proportion of women in seminars. This might be due to the persistence of networks formed early in the academic year between seminar participants who could benefit from discussions with their classmates when preparing research-intensive elements of their coursework.

Dissertation Grade

Panel (b) of Table 6 presents the estimates for distinction in the dissertation as the dependent variable. The results closely mirror those for the final programme grade. Each additional woman in the seminar group significantly raises the likelihood of obtaining a dissertation distinction for both women and native English speakers. For women, the coefficient is 0.58 (Column 2, Row 4), indicating that for female students, each additional woman among seminar peers increases the likelihood of obtaining a distinction by 14.5% (0.58 × 25%) (corrected *p*-value = 0.01). For native English speakers (Column 4, Row 4), the coefficient is 0.66, suggesting an increase in likelihood of 16.5% (0.66 × 25%) per additional woman in the seminar group (corrected *p*-value = 0.04). These findings point to the importance of gender composition in more exposed exchanges, especially for women and for native English speakers.

Table 6 (b) also shows meaningful effects at the team level. For women, a higher proportion of native English speakers in their teams increases the likelihood of a dissertation distinction: the estimated coefficient is 0.22 (Column 2, Row 1) with a corrected *p*-value of 0.08 (Column 2, penultimate row). For men, introducing one woman into a male student's team (equivalent to a 0.33 increase in the proportion of women) would decrease his likelihood of dissertation distinction by approximately 0.3 points (0.33 × -0.92) (Column 3). The *p*-value after the multiple hypothesis testing correction is 0.04 (Column 3,

penultimate row). However, this finding must be interpreted with caution: it is based on a small number of male-majority teams, and all men who achieved distinction in these teams were native English speakers. Given the small sample size and limited variability in language background, we cannot definitively attribute the observed effect to gender composition alone; confounding by language proficiency or ability may also play a role. Put differently, while we cannot rule out gender as a contributing factor, we also cannot conclude that gender is the sole driver of this result.

We also tested potential interaction effects and non-linear heterogeneous effects at the seminar level for both final programme grade and dissertation outcomes. Specifically, Appendix A.3 (Tables A3.2, A.3.3, A.3.6, and A.3.7) report these additional specifications. However, we find no statistically significant evidence of further interactions or non-linearities in seminar composition effects.

Broader Patterns and Interpretation of Results

Comparing across levels of group interaction suggests that seminar composition—particularly the proportion of women—exerts a stronger influence on overall distinction outcomes, especially for women and native speakers. Team-level composition plays a more pronounced role in dissertation outcomes, with women benefiting from greater exposure to native English speakers.

The interaction effect between the proportion of women in the team and being female is positive and statistically significant for both programme grade and dissertation distinction outcomes (Appendix Tables A.3.2 and A.3.3), reinforcing the importance of female-majority team composition for women's academic success. Women placed in teams with a medium share of native speakers (35–75%) have significantly higher odds of earning dissertation distinction compared to those in teams with lower shares (<35%) (effect size = 0.20; corrected p-value = 0.01; see Appendix Table A.3.7).

Group composition's null effect on the course-specific exam contrasts with its significant effect on later outcomes; however, both affirm prior work that finds the early formation of university peer groups can produce delayed-onset influences on longer-term outcomes. For example, Fisher (2017) and Zölitz and Feld (2021) find that early group composition's significant effects on long-term course choice and programme trajectory were more observable than proximate course-specific outcomes. Zölitz and Feld (2021) attribute the delays to the greater relative influence of early-onset female-friendship formations than friendships formed later in one's study trajectory, even though that influence may take time to materialize. Further, in our setting, the individualized nature of the final exam may have muted peer

interaction effects, whereas diverse group compositions provided advantages in more complex, collaborative assessments such as the dissertation.

5.2 Effects of group composition on perception of voice

We begin by visualizing how gender and linguistic composition produced different perceptions of voice across team and seminar contexts using the full sample, we then present the result of our estimations.

Full sample

For the full sample, Figure 1 illustrates kernel density estimations of individuals' perceptions of being heard, differentiated by group composition in terms of gender and language. Perceptions of being heard are measured on a scale ranging from 0 to 10. The top panels compare groups based on gender composition, distinguishing between teams and seminars with low (<50%), medium (50%-99%), and high (100%) shares of women. The bottom panels similarly compare groups by the proportion of native English speakers (low: <35%, medium: 35%-75%, high: >75%).

In teams (top-left panel), the distributions indicate that groups composed entirely of women (high) report a higher density of feeling strongly heard (scores above 8), relative to mixed-gender (medium) and majority male groups (low). This suggests that women's perceived voice within teams is higher when the team is entirely female. In seminar settings (top-right panel), this pattern is even more pronounced, with a clear rightward shift indicating women in seminars where the other teams are composed entirely of women report significantly higher perceptions of being heard.

The lower panels explore the influence of native English speaker composition. In team contexts (bottom-left panel), teams with a low proportion (<35%) of native English speakers have greater densities at higher levels of perceived influence (above 9), while in seminar contexts (bottom-right panel), mixed composition (medium: 35%-75%) appears to yield the highest densities of perceived influence at higher scores, which points towards potentially different dynamics in more formal academic settings, where moderate linguistic diversity may be beneficial.

Taken together, the panels in Figure 1 provide a description of a pattern which underscores the possibility that different types of group diversity can produce differences in perceptions of voice across different contexts. On one hand, women’s inclusion may promote a group-wide perception that one’s voice is heard, irrespective of whether collaboration unfolds in team settings or in seminars; perhaps as a policy

priority, promoting women's inclusion seems a *prima facie* straightforward goal worth pursuing. On the other hand, however, the inclusion of non-native English speakers points in different directions depending on the setting. Although their inclusion in backstage team settings may promote the group-wide perception of voice, the effect reverses sign in frontstage seminar settings. This figure provides a *prima facie* support for testing diversity's effects in a dual-layered setting such as ours: marginality of one kind may well operate differently than marginality of another kind; by extension, the context in which groups work may implicate different effects across different dimensions of diversity.

We turn, therefore, to estimating the effects on voice at a more granular level.

Estimation

Table 7 presents the regression estimates for the item "My voice was heard during group discussions" (measured as level of agreement from 1-10). Results suggest that students benefited from a higher proportion of women in the group, as an increase in the proportion of women in one's own team increased the extent to which students agreed with the statement. Consistent with our expectations, the effect was especially pronounced among non-native English speakers, who may feel reluctant to speak in seminars.

In the full sample, moving from a team with no women to a fully female team increases the level of agreement by 1.32 points (Table 7, Row 3, Column 1), with a corrected *p*-value of 0.09 (Table 7, Column 1, last row). This effect is stronger among non-native English speakers: for this group, moving from a team with no women to a team consisting entirely of women increases the level of agreement by 2.41 points (Table 7 row 3, column 5; the *p*-value associated to this estimate when applying the multiple hypothesis testing correction is equal to 0.09, refer to Table 7, column 1, last row). Thus, in a team of four, one more woman in the group causes an average increase of 0.8 points in level of agreement with the statement. This finding is robust to the inclusion of survey weights (see Table A.3.10).

While the interacted specification (Table A.3.4 in Appendix 3) shows no statistically significant interaction effects, the estimated coefficients from the non-linear specification (Eq. 3) do reveal a complementary finding. For non-native English speakers, being in a team with a medium share of native English speakers (between 35% and 75%) is associated with a decrease of 1.08 points in the perception of being heard, relative to teams with low shares of native English speakers (Table A.3.8, Column 4, Row 1). This negative effect (corrected *p*-value equal to 0.06) suggests that non-native English speakers have the highest perception of being heard when in teams with a small proportion of native English

speakers. At the same time, women in seminar groups with a medium proportion of native English speakers experience a positive and significant increase of 1.26 points (Table A.3.8, Column 1, Row 3; the corrected p-value is equal to 0.03). However, the effect does not consistently increase with the percentage of native English speakers in the seminar, as the coefficient on the 75% dummy is not significant. Hence, a declining marginal benefit may accrue to an increase in the proportion of native English speakers in a seminar.

A field experiment constrains our ability to make confident inferences about mechanisms that might explain these findings. However, correspondence with the instructors, who report that native English speakers were particularly vocal during seminars, hints at one possibility. Namely, a seminar discussion's vibrancy may have depended on some modicum level of engagement that native English speakers delivered, and which conferred especially pronounced benefits to women in this women-majority environment. However, beyond the point when enthusiastic participants saturate the discussion, a seminar's vibrancy can also deliver too much of a good thing, such that inclusion's positive effects on voice may disappear.

Finally, for a smaller sample in the second cohort, we also examined two follow-up items related to perceptions of voice: 'More confident in voicing my view' and 'My voice will be heard.' The estimated effects for these items are in the same direction as our main findings, though they do not keep statistical significance after multiple hypothesis testing correction (see Table A.3.9). Note that these effects are statistically significant prior to multiple hypothesis testing correction, and robust to selection on unobservables, as shown using Oster bounds in Appendix A.4 (see Table A.4.1).

7. Conclusions

This paper examined how gender and linguistic diversity within teams and seminars shape individual academic outcomes and perceptions of voice in a diverse, female-majority and international graduate course. Our findings show that diversity's effects depend not only on group composition—such as gender or language background—but also on the format of collaboration: whether individuals work in small, sustained teams or participate in larger, public-facing groups. These patterns highlight the importance of considering both demographic and interactional contexts when designing collaborative environments.

One of the advantages of our field experiment is that it captured real-world group dynamics with minimal intervention, a feature rarely achieved in related work. While this design prioritised a naturalistic setting

over isolating specific mechanisms, future research could complement our findings by using more controlled environments to test the channels through which gender and linguistic diversity interact — particularly in settings involving multiple levels of group interaction.

As classrooms and workplaces diversify, professionals and educators will need strategies to maximize diversity's benefits while minimizing potential costs. Further research could usefully clarify whether the findings we observe durably persist beyond female-majority settings, or in the completion of simpler collaborative tasks beyond the research-related ones tested here. More broadly, further research could explore the effects that arise from translating these findings into initiatives designed to promote inclusion. Our results suggest that assigning complex tasks to small teams with a critical mass of women — rather than male-dominated teams — may foster greater inclusion, particularly for non-native speakers. Similarly, larger groups with a higher share of women appear to support women's academic excellence and participation, in particular for female native English speakers. These insights are relevant for designing research teams, course structures, and collaborative projects in both educational and professional settings. Our results are likely relevant for several strands of research within several disciplines, such as industrial organization, organizational behavior, labor economics, economics of education and public policy.

References

Alan, S., Corekcioglu, G., Kaba, M. and Sutter, M., 2023. Female leadership and workplace climate. *Management Science*, forthcoming

Alan, S., Duysak, E., Kubilay, E. and Mumcu, I., 2023. Social Exclusion and Ethnic Segregation in Schools: The Role of Teachers' Ethnic Prejudice. *Review of Economics and Statistics*, *105*(5), pp.1039-1054.

Al-Ubaydli, O. and List, J.A., 2015. On the generalizability of experimental results in economics. *Handbook of experimental economic methodology*, pp.420-462.

Anderson, M.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

Anelli, M. & Peri, G. (2019), 'The effects of high school peers' gender on college major, college performance and income', The Economic Journal 129(618), 553–602.

Azmat, G. (2019). Gender diversity in teams. IZA World of Labor.

Barsbai, T., Licuanan, V., Steinmayr, A., Tiongson, E. and Yang, D., 2024. Information and immigrant settlement. Journal of Development Economics, 170, p.103305

Black, S. E., Devereux, P. J. & Salvanes, K. G. (2013), 'Under pressure? the effect of peers on outcomes of young adults', Journal of Labor Economics 31(1), 119–153.

Born, A., Ranehill, E. and Sandberg, A., 2022. Gender and willingness to lead: Does the gender composition of teams matter?. *Review of Economics and Statistics*, *104*(2), pp.259-275.

Braakmann, N. & McDonald, S. (2018), 'Student exposure to socio-economic diversity and students' university outcomes–evidence from English administrative data', MPRA Paper No. 90351 .

Chevalier, A., Isphording, I. E. & Lisauskaite, E. (2020), 'Peer diversity, college performance and educational choices', Labour Economics 64, 101833.

Chiswick, B. R. & Miller, P. W. (2005), 'Linguistic distance: A quantitative measure of the distance between english and other languages', Journal of Multilingual and Multicultural Development 26(1), 1–11.

Coffman, K., Flikkema, C.B. and Shurchkov, O., 2021. Gender stereotypes in deliberation and team decisions. *Games and Economic Behavior*, *129*, pp.329-349.

Dahl, G.B., Kotsadam, A. and Rooth, D.O., 2021. Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams. *The Quarterly Journal of Economics*, *136*(2), pp.987-1030.

Diette, T. M. & Uwaifo Oyelere, R. (2012), 'Do significant immigrant inflows create negative education impacts? lessons from the North Carolina public school system', IZA Discussion Paper

Fenoll, A.A. and Zaccagni, S., 2022. Gender mix and team performance: Differences between exogenously and endogenously formed teams. *Labour Economics*, *79*, p.102269.

Fischer, S., 2017. The downside of good peers: How classroom composition differentially affects men's and women's STEM persistence. *Labour Economics*, *46*, pp.211-226.

Folger, R. (1977). Distributive and procedural justice: Combined impact of voice and improvement on experienced inequity. Journal of personality and social psychology, 35(2), 108.

Geay, C., McNally, S. & Telhaj, S. (2013), 'Non-native speakers of English in the classroom: what are the effects on pupil performance?', The Economic Journal 123(570), F281–F307.

Gould, E. D., Lavy, V. & Daniele Paserman, M. (2009), 'Does immigration affect the long-term educational outcomes of natives? quasi-experimental evidence', The Economic Journal 119(540), 1243–1269.

Greenberger, D. B., & Strasser, S. (1986). Development and application of a model of personal control in organizations. Academy of Management Review, 11(1), 164-177.

Hardt, D., Mayer, L. and Rincke, J., 2023. Who does the talking here? The impact of gender composition on team interactions. Accepted, Management Science

Hernán, M. A. & Robins, J. M. (2020), Causal inference: What If, Boca Raton: Chapman & Hall/CRC.

Hoxby, C. (2000), Peer effects in the classroom: Learning from gender and race variation, Technical report, National Bureau of Economic Research.

Jamieson, K.H., 1995. *Beyond the double bind: Women and leadership*. Oxford University Press, USA.

Jensen, P. & Rasmussen, A. W. (2011), 'The effect of immigrant concentration in schools on native and immigrant children's reading and math skills', Economics of Education Review 30(6), 1503–1515.

Jochmans, K., 2023. Testing random assignment to peer groups. Journal of Applied Econometrics, 38(3), 321-333.

Karpowitz, C., O'Connell, S.D., Preece, J. and Stoddard, O., 2023. Strength in Numbers? Gender Composition, Leadership, and Women's Influence in Teams. Forthcoming, *Journal of Political Economy*

Knyazeva, A., Knyazeva, D., & Naveen, L. (2021). Diversity on corporate boards. Annual Review of Financial Economics, 13, 301-320.

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

Lazear, E. P. (1999), 'Culture and language', *Journal of political Economy* 107(S6), S95–S126.

List, J.A., Shaikh, A.M. and Xu, Y., 2019. Multiple hypothesis testing in experimental economics. Experimental Economics, 22, pp.773-793

Manne, K., 2017. *Down girl: The logic of misogyny*. Oxford University Press.

Manski, C. F. (1993), 'Identification of endogenous social effects: The reflection problem', *The review of economic studies* **60**(3), 531–542.

Morrison, E. W. (2014). Employee voice and silence. *Annual Review of Organizational Psychology and Organizational Behavior* ***1***(1), 173-197.

OECD (2020), International Migration Outlook 2020, OECD, Paris, France.

Oosterbeek, H. & Van Ewijk, R. (2014), 'Gender peer effects in university: Evidence from a randomized experiment', *Economics of Education Review* **38**, 51–63.

Oster, E. (2019) Unobservable Selection and Coefficient Stability: Theory and Evidence, *Journal of Business & Economic Statistics*, 37:2, 187-204,

Owan, H. (2014). How should teams be formed and managed?. *IZA World of Labor*.

Pinder, C. C., & Harlos, K. P. (2001). Employee silence: Quiescence and acquiescence as responses to perceived injustice. In *Research in personnel and human resources management* (Vol. 20, pp. 331-369). Emerald Group Publishing Limited.

Rodriguez, G. M., & Cruz, L. (2009). The transition to college of English learner and undocumented immigrant students: Resource and policy implications. *Teachers College Record*, *111*(10), 2385-2418.

Sacerdote, B. (2001), 'Peer effects with random assignment: Results for Dartmouth roommates', The Quarterly journal of economics **116**(2), 681–704.

Schneeweis, N. & Zweimüller, M. (2012), 'Girls, girls, girls: Gender composition and female school choice', Economics of Education review 31(4), 482–500.

Schneeweis, N. (2015), 'Immigrant concentration in schools: Consequences for native and migrant students', Labour Economics **35**, 63–76.

Shan, X. (2022), 'The minority trap: Minority status drives women out of male- dominated fields', Unpublished manuscript.

Stebleton, M. J., Huesman Jr, R. L., & Kuzhabekova, A. (2010). Do I belong here? Exploring immigrant college student responses on the SERU survey sense of belonging/satisfaction factor.

UNESCO (2022), Higher education global data report. UNESCO, Paris, France.

Weidmann, B., & Deming, D. J. (2021). Team players: How social skills improve team performance. *Econometrica*, *89*(6), 2637-2657.

World Bank (2022), 'Gender statistics', https://data.worldbank.org/indicator/ SL.TLF.CACT.FM.ZS?locations=O

Wu, M.J., Zhao, K. and Fils-Aime, F., 2022. Response rates of online surveys in published research: A meta-analysis. *Computers in Human Behavior Reports*, *7*, p.100206

Zölitz, U. and Feld, J., 2021. The effect of peer gender on major choice in business school. *Management Science*, *67*(11), pp.6963-6979.

Tables

Table 1: Summary statistics: Administrative data

		All	Men	Women	Non-native Speakers*	Native Speakers*
(a) Demographic characteristics						
Women	(%)	79.5	-	-	84.2	75.6
Native speakers	(%)	49	60	46.6	-	-
Age	Mean	24.6	25.4	24.4	25.3	24.1
	(st.dev)	(3.9)	(4.4)	(3.7)	(4.6)	(2.9)
(b) Prior studies						
Highest qualification	Bachelor (%)	81.9	83.1	81.9	75.4	89.0
	Master (%)	17.6	16.9	18.1	24.6	11.0
Studied in United Kingdom	(%)	33.4	37.6	32.1	22.2	42.2
(c) Academic outcomes						
Exam Mark	Mean	69.2	68.9	70.4	66.7	72.5
	(st.dev)	(13.0)	(13.2)	(12.6)	(12.8)	(12.1)
Dissertation Distinction	(%)	32.7	34.7	32.31	25.99	38.95
Programme Distinction	(%)	31.62	36.0	30.61	20.9	42.44
Total (N)		376	77	299	180	173

Notes: This table provides summary statistics based on administrative data from the LSE. "Non-native Speakers" and "Native Speakers" classification is derived from responses to the baseline survey. Percentages in the first row represent the proportion of students within each category. "Dissertation Distinction" and "Programme Distinction" denote the percentage of students achieving the highest-grade classification in those respective categories. Standard deviations (st.dev) in parentheses.

Table 2: Summary statistics: Surveys

Variable		All	Men	Women	Non-native Speakers	Native Speakers
(1) **Baseline Survey**						
Language Score (1-3)	Mean	-	1.9	1.9	1.9	-
	(st.dev)	-	(0.6)	(0.5)	(0.5)	-
Familiarity with qualitative research methods (0-10)	Mean	5.9	5.6	5.9	5.8	6.0
	(st.dev)	(3.4)	(2.1)	(2.5)	(2.4)	(2.5)
Familiarity with quantitative research methods (0-10)	Mean	4.1	4.2	4.1	4.1	4.2
	(st.dev)	(2.2)	(2.2)	(2.2)	(2.3)	(2.2)
Expected Mark (0-100)	Mean	72.6	72.0	72.8	73.7	71.4
	(st.dev)	(12.6)	(13.1)	(12.4)	(12.7)	(12.3)
Average response rate	%	93.9	92.2	94.3	-	-
N		355	71	284	182	173
(2) **End of term survey**						
My voice was heard during group discussions (Agreement 0-10)	Mean	8.7	8.3	8.8	8.6	8.9
	(st.dev)	(1.8)	(2.1)	(1.7)	(1.7)	(1.6)
More confident in voicing my view in future interactions (Agreement 0-10)	Mean	5.1	4.8	5.2	5.3	4.9
	(st.dev)	(3.1)	(2.9)	(3.1)	(3.1)	(3.1)
More confident that my view will be heard in future interactions (Agreement 0-10)	Mean	5.2	5.1	5.3	5.5	5.1
	(st.dev)	(2.9)	(2.5)	(3.1)	(2.9)	(3.0)
Average response rate (%)	%	59.8	58.4	60.2	63.9	60.7
N		85	22	63	33	47

Notes: This table reports summary statistics from baseline and end-of-term surveys. For the baseline survey, we present mean and standard deviation (st. dev) for language score (derived using Chiswick & Miller (2005) measures), self-reported familiarity with qualitative and quantitative research methods (on a scale of 0-10), and expected marks (on a scale of 0–100). The end-of-term survey, we present mean and standard deviation (st. dev) for three measures of perceptions of voice. The items "More confident in voicing my view in future interactions" and "More confident that my view will be heard in future interactions" are specific to the 2021–2022 cohort. Response rates by gender are calculated using administrative records available for all enrolled students. Response rates by native language are calculated only among students who completed the baseline survey, as native language information was collected in that survey and is not available for the full sample.

Table 3: Identification: Random assignment test statistics and p-values

		Proportion of native speakers in seminar group	Proportion of native speakers in team
Native English	Test statistic	1.02	-.097
	p-value	(0.31)	(0.92)
	Urns (N)	2	33
		Proportion of women in seminar group	**Proportion of women in team**
Female	Test statistic	-1.41	1.53
	p-value	(0.16)	(0.13)
	Urns (N)	2	33

Note: This table reports test statistics and p-values from a regression-based test assessing random assignment to seminars and teams, following Jochmans (2023). This test examines the within-group correlation between each individual's characteristics (native language and gender) and the average characteristic of their peers. The null hypothesis of the test is random assignment, the large p-values suggest that seminar and team allocations are exogenous to student gender and native language. The number of urns represents the groups from which peers are drawn for each test.

Table 4: Balancing tests: Effect of individual characteristics on seminar composition

Variables	(1) Share of women in seminar	(2) Share of native speakers in seminar
English as first language	-0.016	
	(0.011)	
Gender		-0.019
		(0.023)
Age	0.001	0.001
	(0.001)	(0.002)
Previous UK studies	-0.003	0.012
	(0.011)	(0.020)
Highest Level of Education	0.015	0.005
	(0.013)	(0.021)
Experience with	-0.002	0.001
quantitative methods	(0.003)	(0.004)
Experience with	-0.002	-0.001
qualitative methods	(0.002)	(0.004)
Expected mark	-0.000	0.001*
	(0.000)	(0.001)
Identify as leader	-0.010	0.014
	(0.008)	(0.019)
Observations	376	
N tests	16	
N tests significant at 1%	0	
N tests significant at 5%	0	
N tests significant at 10%	1	

Notes: This table reports the results of 16 balancing tests assessing the association between individual characteristics and seminar group composition. Each row shows the estimated coefficient from a separate linear regression, where the dependent variables are (1) the share of women in the individual's seminar group and (2) the share of native English speakers in the seminar group. Independent variables include individual characteristics such as English as a first language, gender, age, prior UK studies, highest level of education, experience with quantitative and qualitative methods, expected mark, and an indicator for usual role in the team (leader). *** p<0.01, ** p<0.05, * p<0.1, standard errors in parentheses. The results indicate no significant correlations at the 5% level, and only one test shows significance at the 10% level. This suggests that variations in the gender and language composition of seminar groups are unlikely to be systematically influenced by these individual characteristics, supporting the random assignment to seminar groups. The number of tests and the levels of significance are provided for reference.

Table 5: Regression coefficients for Exam Marks

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	-0.39	0.09	-4.47	0.85	-2.87
	(2.39)	(2.84)	(4.82)	(3.50)	(3.74)
Proportion Native English Speakers in Seminar (β_2)	2.40	2.02	6.06	0.40	4.14
	(3.39)	(3.91)	(7.97)	(5.16)	(5.21)
Proportion of Women in Team (β_3)	1.46	1.98	-4.19	-2.92	6.93
	(3.11)	(3.61)	(7.09)	(4.41)	(4.97)
Proportion of Women in Seminar (β_4)	7.47	8.11	-2.02	11.97	0.20
	(4.72)	(5.38)	(12.11)	(6.90)	(7.66)
N	337	268	69	166	150
N Seminars	32	32	29	31	32
N Teams	85	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.90	0.93	0.82	0.96	0.70
$\beta_2 = 0$ (p-value MHT)	0.85	0.83	0.76	0.93	0.81
$\beta_3 = 0$ (p-value MHT)	0.80	0.88	0.87	0.84	0.22
$\beta_4 = 0$ (p-value MHT)	0.27	0.28	0.89	0.26	0.97

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team)

Table 6: Regression coefficients: Individual Performance

(a) Classification: Final Grade Distinction

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	0.06	0.15	-0.45	0.07	0.03
	(0.09)	(0.10)	(0.20)	(0.14)	(0.12)
Proportion Native English Speakers in Seminar (β_2)	0.05	0.01	0.05	-0.18	0.35
	(0.12)	(0.14)	(0.33)	(0.21)	(0.17)
Proportion of Women in Team (β_3)	0.05	0.18	-0.73	0.13	0.01
	(0.11)	(0.13)	(0.30)	(0.18)	(0.16)
Proportion of Women in Seminar (β_4)	0.36*	0.46**	-0.79	0.60*	-0.06
	(0.17)	(0.19)	(0.51)	(0.28)	(0.24)
N	339	270	69	167	151
N Seminars	32	32	29	31	32
N Teams	85	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.87	0.32	0.13	0.63	0.98
$\beta_2 = 0$ (p-value MHT)	0.68	0.94	0.81	0.79	0.10
$\beta_3 = 0$ (p-value MHT)	0.82	0.27	0.11	0.71	0.84
$\beta_4 = 0$ (p-value MHT)	0.10	0.04	0.26	0.07	0.97

(b) Classification: Dissertation Distinction

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	0.15	0.22*	-0.31	0.16	0.00
	(0.09)	(0.10)	(0.21)	(0.14)	(0.13)
Proportion Native English Speakers in Seminar (β_2)	0.04	0.08	-0.50	-0.01	-0.06
	(0.13)	(0.14)	(0.34)	(0.21)	(0.18)
Proportion of Women in Team (β_3)	0.09	0.22	-0.91**	-0.06	0.25
	(0.12)	(0.13)	(0.30)	(0.18)	(0.18)
Proportion of Women in Seminar (β_4)	0.48**	0.58**	-0.47	0.66*	0.16
	(0.18)	(0.19)	(0.52)	(0.27)	(0.27)
N	339	270	69	167	151
N Seminars	32	32	29	31	32
N Teams	85	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.24	0.08	0.40	0.55	0.98
$\beta_2 = 0$ (p-value MHT)	0.76	0.56	0.39	0.96	0.96
$\beta_3 = 0$ (p-value MHT)	0.68	0.16	0.04	0.94	0.26
$\beta_4 = 0$ (p-value MHT)	0.02	0.01	0.43	0.04	0.92

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, and whether the advisor was native English speaker.Language score is included for the non-native speaker sample (specification 5). Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team).

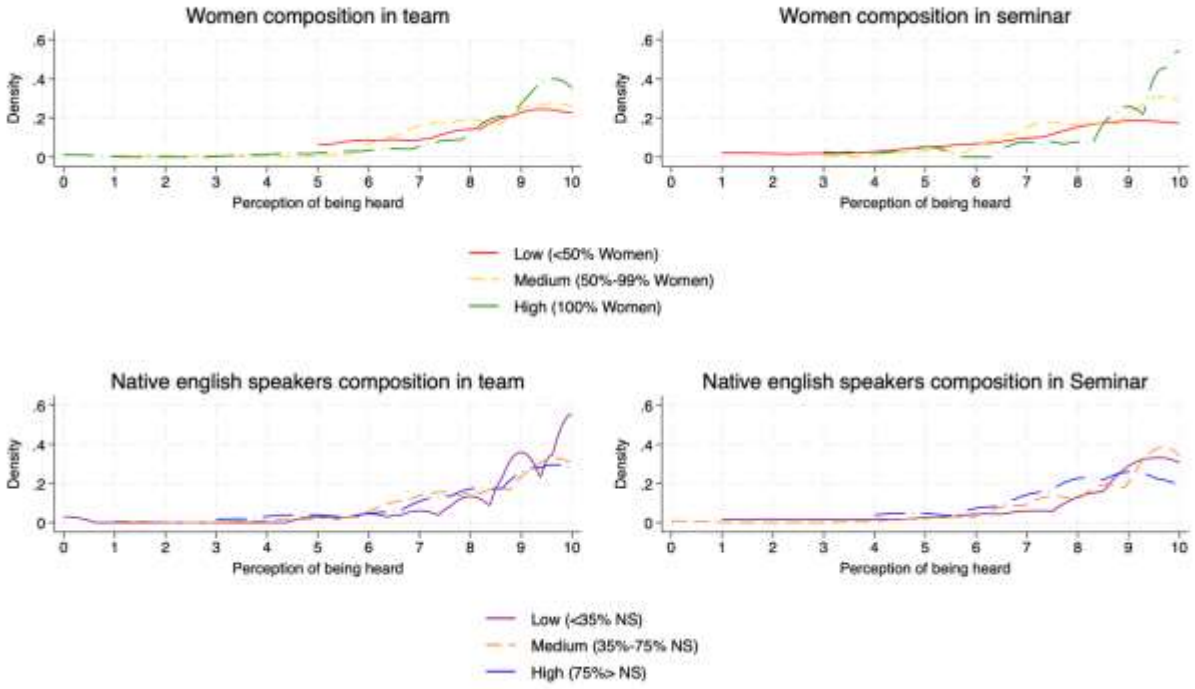
Table 7: Regression coefficients: "My voice was heard during group discussions"

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	-0.23	-0.49	0.25	0.48	-1.52
	(0.42)	(0.45)	(1.59)	(0.56)	(0.69)
Proportion Native English Speakers in Seminar (β_2)	0.02	0.20	-1.40	-0.48	0.70
	(0.61)	(0.61)	(2.50)	(0.89)	(0.89)
Proportion of Women in Team (β_3)	1.32*	1.40	0.47	1.38	2.41*
	(0.54)	(0.55)	(2.08)	(0.77)	(0.82)
Proportion of Women in Seminar (β_4)	1.61	1.99	-1.57	1.44	2.56
	(0.89)	(0.87)	(4.74)	(1.30)	(1.40)
N	215	171	44	104	98
N Seminars	32	32	23	28	31
N Teams	79	77	36	61	61
$\beta_1 = 0$ (p-value MHT)	0.86	0.53	0.88	0.76	0.15
$\beta_2 = 0$ (p-value MHT)	0.98	0.83	0.93	0.57	0.58
$\beta_3 = 0$ (p-value MHT)	0.09	0.15	0.96	0.33	0.09
$\beta_4 = 0$ (p-value MHT)	0.41	0.30	0.97	0.40	0.47

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, and whether the advisor was native English speaker. Language score is included for non-native speakers (specification 5). Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. The variable "My voice was heard during group discussions" measures the level of agreement with the statement in a scale of 0-10. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team). Data source: Endline survey, 2020-2021, and 2021-2022 cohorts.

Figures

Figure 1: Density plot: Perception of being heard



Notes: kernel density estimates for participants' self-reported perception of being heard in group discussions, measured on a scale from 0 to 10. Panels differentiate the composition of groups by the percentage of women (top row) and native English speakers (bottom row), separately for team and seminar settings. Group composition categories are defined as follows: low (less than 50% women or less than 35% native speakers), medium (50–99% women or 35–75% native speakers), and high (100% women or more than 75% native speakers). Kernel densities are estimated using an Epanechnikov kernel function.

A.1 Appendix to Background
Example of Seminar Activity

Week Four: Causation and its role in policy research

Group Exercise

Study the famous "Cochrane hierarchy of scientific evidence":

Level 1: Correlation between an intervention programme and an outcome measure at one point in time.
Level 2: Outcome measures before and after the programme, with no comparable control condition.
Level 3: Outcome measures before and after the programme in experimental and comparable control units, controlling for other variables that influence the outcome
Level 4: Outcome measures before and after the programme in multiple experimental and control units, controlling for other variables that influence the outcome.
Level 5: Random assignment of programme and control conditions to units

> *Source*: Adapted from Guyatt, G. H., Sackett, D. L., Sinclair, J. C., et al. (1995). A method for grading health care recommendations. *JAMA*, *274*(22): 1800-1804.

In your group, use the Library, Google Scholar, and whatever other resources at your disposal to sort the following fifteen studies into the five 'Levels'. If your group disagrees over how to code a study, *record the disagreement for further discussion during seminar*.

A.2 Appendix to Data

Table A.2.1: Number (N) of students by country and cohort

Country of birth	2020-2021 (N)	2021-2022 (N)	Total (N)	Country of birth	2020-2021 (N)	2021-2022 (N)	Total (N)
Albania	1	0	1	**Mauritania**	1	0	1
Argentina	2	2	4	**Mexico**	3	2	5
Armenia	0	1	1	**Nepal**	1	0	1
Australia	2	1	3	**Netherlands**	2	0	2
Bahrain	0	1	1	**Nigeria**	2	0	2
Bangladesh	1	2	3	**Norway**	0	3	3
Belgium	0	1	1	**Pakistan**	5	5	10
Brazil	2	1	3	**Panama**	1	0	1
Bulgaria	2	0	2	**Paraguay**	1	0	1
Burma (Myanmar)	0	1	1	**Peru**	0	1	1
Canada	5	7	12	**Philippines**	2	1	3
Chile	0	3	3	**Poland**	2	2	4
China	26	21	47	**Qatar**	0	1	1
Colombia	1	5	6	**Romania**	1	1	2
Dominican Rep.	1	0	1	**Russia**	0	1	1
England	38	38	76	**Saudi Arabia**	0	1	1
Eritrea	0	1	1	**Scotland**	2	0	2
FYR Macedonia	1	0	1	**Singapore**	0	3	3
Finland	1	0	1	**South Africa**	0	1	1
France	9	5	14	**South Korea**	4	5	9
Germany	1	5	6	**Spain**	2	1	3
Ghana	0	1	1	**Sri Lanka**	0	1	1
Greece	2	0	2	**Sudan**	1	0	1
Hong Kong	3	6	9	**Sweden**	0	1	1
Hungary	1	0	1	**Taiwan**	0	1	1
India	12	13	25	**Thailand**	0	2	2
Indonesia	2	0	2	**Turkey**	3	1	4
Ireland	0	2	2	**USA**	18	20	38
Italy	7	10	17	**Ukraine**	2	0	2
Japan	2	2	4	**Uruguay**	1	1	2
Jordan	1	0	1	**Utd Arab Emts.**	1	1	2
Kazakhstan	1	0	1	**Venezuela**	1	1	2
Kenya	1	1	2	**Vietnam**	1	0	1
Lebanon	1	0	1	**Wales**	1	1	2
Lithuania	1	1	2	**Zimbabwe**	0	1	1
Luxembourg	0	1	1				
Malaysia	1	0	1	**Total**	186	191	377

Notes: Table present the number (N) of students by country of birth and cohort. Source: University administrative data records for cohorts 2020-2021 and 2021-2022.

Table A.2.2: Non-native English speakers Language scores (Linguistic Distance)

Native Language	N	%	Score	Native Language	N	%	Score
Arabic	4	2%	1.5	Nepali	1	1%	1.75
Bengali	1	1%	1.75	Norwegian	1	1%	3
Bulgarian	1	1%	2	Polish	5	3%	2
Burmese	1	1%	1.75	Portuguese	3	2%	2.5
Cantonese	5	3%	1.25	Punjabi	1	1%	1.75
Chinese	45	27%	1.5	Romanian	2	1%	3
Dutch	2	1%	2.75	Russian	1	1%	2.25
French	14	8%	2.5	Spanish	23	14%	2.25
German	5	3%	2.25	Swedish	3	2%	3
Greek	1	1%	1.75	Tagalog	1	1%	2
Gujarati	1	1%	1.75	Tamil	3	2%	1.75
Hindi	13	8%	1.75	Telugu	1	1%	1.75
Hungarian	1	1%	2	Thai	1	1%	2
Indonesian	1	1%	2	Turkish	3	2%	2
Italian	14	8%	2.5	Vietnamese	1	1%	1.5
Japanese	4	2%	1				
Korean	7	4%	1				
Malayalam	2	1%	1.75	Total	166		1.92

Notes: N represent the total number of students who reported each language as their Native language. Score is the Chiswick & Miller (2005) measure of linguistic distance from each language to English. The measure ranges from 1 to 3, with three being the most similar to English. Distance scores are reported for 166 individuals. Eleven additional non-native speakers are not shown because their native languages have no corresponding Chiswick & Miller (2005) measure.

Table A2.3: Proportion of women and native speakers per team

	Gender composition		
	25% to 50%	50% to 85%	100%
Mean native English speakers composition	64.7%	50.9%	44.8%
Mean team size	4.3	4.4	4.1
N teams	4	47	37
N Individuals	17	206	146
	Native English speakers composition		
	0 to 33%	34% to 66%	67% to 100%
Mean gender composition	82.7%	82.2%	71.4%
Mean team size	4.0	4.5	4.2
N teams	25	41	22
N Individuals	98	180	91

Notes: This table presents summary statistics on team composition by gender and native English speaker status. The top panel groups teams by the proportion of women, and reports the average proportion of native English speakers, mean team size, and number of teams and individuals in each category. The bottom panel groups teams by the proportion of native English speakers, and reports the average proportion of women, mean team size, and the number of teams and individuals in each category. *"N teams"* refers to the number of unique teams in each composition category, and *"N individuals"* refers to the total number of students in those teams.

A.3 Appendix to Results

Table A.3.1: Regression coefficients: Marks (interacted model)

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Prop. Native English (Team)	-2.88	-3.28	3.12	-3.90	4.65
	(5.78)	(3.79)	(7.83)	(6.59)	(8.43)
Prop. Native English (Seminar)	9.37	6.59	-8.80	10.21	-5.05
	(8.91)	(5.11)	(13.61)	(10.53)	(17.66)
Prop. Women (Team)	5.74	6.97	-6.55	-2.28	4.08
	(8.53)	(5.03)	(14.35)	(9.32)	(14.45)
Prop. Women (Seminar)	-0.81	4.17	-5.75	5.00	11.48
	(13.29)	(7.51)	(17.53)	(19.05)	(20.87)
Prop. Native English (Team) x English	1.69	6.41	-10.40	-	-
	(4.78)	(5.64)	(9.42)		
Prop. Native English (Seminar) x English	-6.40	-10.51	20.36	-	-
	(6.80)	(7.74)	(16.01)		
Prop. Women (Team) x English	-9.72	-9.58	3.01	-	-
	(6.34)	(7.13)	(17.42)		
Prop. Women (Seminar) x English	6.93	6.48	10.61	-	-
	(9.37)	(10.61)	(24.30)		
Prop. Native English (Team) x Female	1.38	-	-	7.06	-8.81
	(5.72)			(7.72)	(9.22)
Prop. Native English (Seminar) x Female	-4.77	-	-	-13.14	10.07
	(8.70)			(11.87)	(18.57)
Prop. Women (Team) x Female	1.24	-	-	-0.31	3.03
	(8.23)			(10.53)	(15.43)
Prop. Women (Seminar) x Female	5.11	-	-	7.77	-13.88
	(13.58)			(20.53)	(22.75)
Native English Speaker = Yes	8.77	8.99	-12.38	-	-
	(12.04)	(13.27)	(33.70)		
Gender = Female	-5.15	-	-	-4.39	7.87
	(17.72)			(26.46)	(31.44)
N	339	270	69	167	151
N Seminars	32	32	29	31	32
N Teams	85	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.63	0.53	0.78	0.66	0.58
$\beta_2 = 0$ (p-value MHT)	0.65	0.57	0.89	0.49	0.88
$\beta_3 = 0$ (p-value MHT)	0.35	0.16	0.87	0.76	0.70
$\beta_4 = 0$ (p-value MHT)	0.95	0.76	0.95	0.76	0.90
$\beta_5 = 0$ (p-value MHT)	0.73	0.57	0.74	-	-
$\beta_6 = 0$ (p-value MHT)	0.70	0.37	0.59	-	-
$\beta_7 = 0$ (p-value MHT)	0.24	0.36	1.00	-	-
$\beta_8 = 0$ (p-value MHT)	0.68	0.51	0.89	-	-
$\beta_9 = 0$ (p-value MHT)	0.95	-	-	0.44	0.32
$\beta_{10} = 0$ (p-value MHT)	0.80	-	-	0.44	0.88
$\beta_{11} = 0$ (p-value MHT)	0.85	-	-	0.97	0.95
$\beta_{12} = 0$ (p-value MHT)	0.70	-	-	0.66	0.90
$\beta_{13} = 0$ (p-value MHT)	0.85	0.65	0.92	-	-
$\beta_{14} = 0$ (p-value MHT)	0.76	-	-	0.84	0.93

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Panel (a) reports main effects of team- and seminar-level proportions of native English speakers and women. Panel (b) interacts those proportions with students' native-English status; Panel (c) interacts them with gender (female). Specification (1) shows the full sample with both sets of interactions; (2)–(3) focus on interactions with English-native status; (4)–(5) on interactions with gender. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is Prop. Native English (Team), β_2 is Prop. Native English (Seminar), etc.)

Table A.3.2: Regression coefficients: Final Programme Grades (interacted model)

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Prop. Native English (Team)	-0.43	0.07	-0.67	-0.32	-0.35
	(0.21)	(0.13)	(0.33)	(0.27)	(0.27)
Prop. Native English (Seminar)	0.42	0.20	0.66	-0.22	0.68
	(0.32)	(0.18)	(0.57)	(0.42)	(0.56)
Prop. Women (Team)	-0.70*	0.05	-0.77	-0.61	-0.38
	(0.31)	(0.17)	(0.60)	(0.38)	(0.45)
Prop. Women (Seminar)	-0.73	0.09	-0.53	-0.95	-0.06
	(0.48)	(0.26)	(0.74)	(0.77)	(0.66)
Prop. Native English (Team) x English	0.14	0.17	0.27	-	-
	(0.17)	(0.20)	(0.40)		
Prop. Native English (Seminar) x English	-0.50	-0.43	-0.91	-	-
	(0.24)	(0.27)	(0.67)		
Prop. Women (Team) x English	0.21	0.27	0.05	-	-
	(0.23)	(0.25)	(0.73)		
Prop. Women (Seminar) x English	0.45	0.66	-0.67	-	-
	(0.34)	(0.37)	(1.02)		
Prop. Native English (Team) x Female	0.50*	-	-	0.51	0.45
	(0.21)			(0.31)	(0.29)
Prop. Native English (Seminar) x Female	-0.19	-	-	-0.00	-0.37
	(0.31)			(0.48)	(0.59)
Prop. Women (Team) x Female	0.79**	-	-	0.94	0.45
	(0.30)			(0.42)	(0.49)
Prop. Women (Seminar) x Female	0.95	-	-	1.76*	-0.01
	(0.49)			(0.82)	(0.72)
Native English Speaker = Yes	-0.17	-0.42	1.00	-	-
	(0.43)	(0.46)	(1.42)		
Gender = Female	-1.58*	-	-	-2.46*	-0.49
	(0.64)			(1.06)	(0.99)
N	339	270	69	167	151
N Seminars	32	32	29	31	32
N Teams	85	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.15	0.50	0.34	0.63	0.47
$\beta_2 = 0$ (p-value MHT)	0.49	0.49	0.61	0.65	0.48
$\beta_3 = 0$ (p-value MHT)	0.10	0.74	0.69	0.33	0.70
$\beta_4 = 0$ (p-value MHT)	0.33	0.69	0.90	0.60	0.99
$\beta_5 = 0$ (p-value MHT)	0.77	0.74	0.76	-	-
$\beta_6 = 0$ (p-value MHT)	0.12	0.32	0.51	-	-
$\beta_7 = 0$ (p-value MHT)	0.35	0.49	1.00	-	-
$\beta_8 = 0$ (p-value MHT)	0.48	0.18	0.94	-	-
$\beta_9 = 0$ (p-value MHT)	0.06	-	-	0.34	0.32
$\beta_{10} = 0$ (p-value MHT)	0.89	-	-	0.99	0.82
$\beta_{11} = 0$ (p-value MHT)	0.02	-	-	0.09	0.81
$\beta_{12} = 0$ (p-value MHT)	0.15	-	-	0.11	0.99
$\beta_{13} = 0$ (p-value MHT)	0.88	0.79	0.88	-	-
$\beta_{14} = 0$ (p-value MHT)	0.03	-	-	0.06	0.94

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Panel (a) reports main effects of team- and seminar-level proportions of native English speakers and women. Panel (b) interacts those proportions with students' native-English status; Panel (c) interacts them with gender (female). Specification (1) shows the full sample with both sets of interactions; (2)–(3) focus on interactions with English-native status; (4)–(5) on interactions with gender. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is Prop. Native English (Team), β_2 is Prop. Native English (Seminar), etc.)

Table A.3.3: Regression coefficients for Dissertation distinction (interacted model)

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Panel (a): Main effects					
Prop. Native English (Team)	-0.30	0.16	-0.47	-0.18	-0.44
	(0.22)	(0.14)	(0.34)	(0.26)	(0.29)
Prop. Native English (Seminar)	-0.25	0.08	0.06	-0.67	0.12
	(0.34)	(0.19)	(0.60)	(0.41)	(0.62)
Prop. Women (Team)	-0.51	0.36*	-0.66	-0.79	-0.42
	(0.32)	(0.18)	(0.64)	(0.37)	(0.52)
Prop. Women (Seminar)	-0.47	0.37	-0.13	-0.81	-0.16
	(0.50)	(0.27)	(0.76)	(0.75)	(0.73)
Panel (b): Interaction with English as Native Language					
Prop. Native English (Team) x English	0.10	0.09	0.21	-	-
	(0.18)	(0.20)	(0.41)		
Prop. Native English (Seminar) x English	-0.15	0.02	-0.74	-	-
	(0.26)	(0.28)	(0.70)		
Prop. Women (Team) x English	-0.25	-0.30	-0.35	-	-
	(0.24)	(0.26)	(0.77)		
Prop. Women (Seminar) x English	0.28	0.43	-0.62	-	-
	(0.35)	(0.38)	(1.06)		
Panel (c): Interaction with Gender (female)					
Prop. Native English (Team) x Female	0.47*	-	-	0.42	0.51
	(0.22)			(0.30)	(0.32)
Prop. Native English (Seminar) x Female	0.41	-	-	0.81	-0.19
	(0.33)			(0.47)	(0.65)
Prop. Women (Team) x Female	0.88**	-	-	0.90*	0.78
	(0.31)			(0.41)	(0.55)
Prop. Women (Seminar) x Female	0.91	-	-	1.68	0.35
	(0.51)			(0.81)	(0.79)
Native English Speaker = Yes	0.14	-0.03	1.31	-	-
	(0.45)	(0.48)	(1.48)		
Gender = Female	-1.85**			-2.73**	-1.10
	(0.67)	-	-	(1.04)	(1.11)
N	339	270	69	167	151
N Seminars	32	32	29	31	32
N Teams	85	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.33	0.52	0.49	0.48	0.31
$\beta_2 = 0$ (p-value MHT)	0.73	0.69	0.96	0.34	0.93
$\beta_3 = 0$ (p-value MHT)	0.27	0.09	0.72	0.11	0.81
$\beta_4 = 0$ (p-value MHT)	0.70	0.49	0.99	0.56	0.98
$\beta_5 = 0$ (p-value MHT)	0.82	0.65	0.64	-	-
$\beta_6 = 0$ (p-value MHT)	0.58	0.93	0.63	-	-
$\beta_7 = 0$ (p-value MHT)	0.61	0.57	1.00	-	-
$\beta_8 = 0$ (p-value MHT)	0.77	0.49	0.94	-	-
$\beta_9 = 0$ (p-value MHT)	0.09	-	-	0.40	0.35
$\beta_{10} = 0$ (p-value MHT)	0.60	-	-	0.28	0.82
$\beta_{11} = 0$ (p-value MHT)	0.02	-	-	0.10	0.46
$\beta_{12} = 0$ (p-value MHT)	0.22	-	-	0.11	0.90
$\beta_{13} = 0$ (p-value MHT)	0.75	0.94	0.88	-	-
$\beta_{14} = 0$ (p-value MHT)	0.02	-	-	0.04	0.62

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Panel (a) reports main effects of team- and seminar-level proportions of native English speakers and women. Panel (b) interacts those proportions with students' native-English status; Panel (c) interacts them with gender (female). Specification (1) shows the full sample with both sets of interactions; (2)–(3) focus on interactions with English-native status; (4)–(5) on interactions with gender. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is Prop. Native English (Team), β_2 is Prop. Native English (Seminar), etc.)

Table A.3.4: Regression coefficients: Voice (interacted model)

	(1) All	(2) Women	(3) Men	(4) Native Speakers	(5) Non-Native Speakers
Panel (a): Main effects					
Prop. Native English (Team)	-1.73	-0.88	-5.42	1.03	-4.37
	(1.03)	(0.64)	(2.35)	(1.00)	(1.64)
Prop. Native English (Seminar)	0.40	0.61	4.17	-1.66	3.71
	(1.74)	(0.82)	(4.56)	(1.62)	(3.37)
Prop. Women (Team)	1.57	1.90	0.02	1.05	2.26
	(1.51)	(0.76)	(3.78)	(1.40)	(2.58)
Prop. Women (Seminar)	-1.71	2.70	-0.80	-2.25	-0.57
	(3.12)	(1.18)	(7.42)	(3.30)	(5.23)
Panel (b): Interaction with English as Native Language					
Prop. Native English (Team) x English	1.88*	0.69	6.23	-	-
	(0.84)	(0.91)	(2.60)		
Prop. Native English (Seminar) x English	-1.52	-1.18	-5.75	-	-
	(1.23)	(1.25)	(5.15)		
Prop. Women (Team) x English	-1.16	-1.07	0.49	-	-
	(1.09)	(1.15)	(4.52)		
Prop. Women (Seminar) x English	-1.26	-1.75	-3.05	-	-
	(1.71)	(1.77)	(8.90)		
Panel (c): Interaction with Gender (female)					
Prop. Native English (Team) x Female	0.19	-	-	-1.57	3.58
	(1.00)			(1.18)	(1.78)
Prop. Native English (Seminar) x Female	0.26	-	-	1.37	-3.09
	(1.64)			(1.86)	(3.52)
Prop. Women (Team) x Female	0.32	-	-	0.30	-0.09
	(1.43)			(1.64)	(2.69)
Prop. Women (Seminar) x Female	4.13	-	-	4.25	3.45
	(3.06)			(3.60)	(5.41)
Native English Speaker = Yes	2.28	2.87	3.14	-	-
	(2.13)	(2.11)	(11.31)		
Gender = Female	-3.54	-	-	-3.38	-2.21
	(3.86)			(4.42)	(6.96)
N	215	171	44	104	98
N Seminars	32	32	23	28	31
N Teams	79	77	36	61	61
$\beta_1 = 0$ (p-value MHT)	0.41	0.51	0.50	0.76	0.12
$\beta_2 = 0$ (p-value MHT)	0.83	0.89	0.85	0.46	0.76
$\beta_3 = 0$ (p-value MHT)	0.56	0.19	1.00	0.84	0.81
$\beta_4 = 0$ (p-value MHT)	0.89	0.51	0.96	0.81	0.93
$\beta_5 = 0$ (p-value MHT)	0.19	0.73	0.62	-	-
$\beta_6 = 0$ (p-value MHT)	0.60	0.78	0.55	-	-
$\beta_7 = 0$ (p-value MHT)	0.56	0.43	0.95	-	-
$\beta_8 = 0$ (p-value MHT)	0.60	0.74	0.86	-	-
$\beta_9 = 0$ (p-value MHT)	0.86	-	-	0.27	0.31
$\beta_{10} = 0$ (p-value MHT)	0.88	-	-	0.68	0.76
$\beta_{11} = 0$ (p-value MHT)	0.97	-	-	0.99	0.98
$\beta_{12} = 0$ (p-value MHT)	0.41	-	-	0.52	0.87
$\beta_{13} = 0$ (p-value MHT)	0.84	0.77	0.88	-	-
$\beta_{14} = 0$ (p-value MHT)	0.64	-	-	0.75	0.80

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Panel (a) reports main effects of team- and seminar-level proportions of native English speakers and women. Panel (b) interacts those proportions with students' native-English status; Panel (c) interacts them with gender (female). Specification (1) shows the full sample with both sets of interactions; (2)–(3) focus on interactions with English-native status; (4)–(5) on interactions with gender. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is Prop. Native English (Team), β_2 is Prop. Native English (Seminar), etc.)

Table A.3.5: Non-linear Effects on "Exam Marks" ,

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team				
Medium proportion	3.03	1.43	1.89	2.10
	(1.81)	(3.30)	(2.62)	(2.39)
High proportion	-1.75	-5.21	-0.79	-5.89
	(2.30)	(3.75)	(2.68)	(3.25)
Proportion Native English Speakers in Seminar				
Medium proportion	-1.30	5.03	-1.86	1.49
	(2.52)	(4.44)	(3.29)	(3.09)
High proportion	-0.87	9.48	-1.22	2.87
	(3.20)	(5.70)	(3.81)	(4.79)
Proportion of Women in Team				
Medium proportion	-0.67	1.71	-1.26	2.59
	(2.47)	(4.79)	(3.10)	(3.26)
High proportion	2.36	18.53	1.60	11.60
	(5.36)	(8.39)	(5.89)	(8.08)
Proportion of Women in Seminar				
Medium proportion	2.62	17.64	4.93	7.70
	(5.29)	(8.12)	(5.77)	(7.76)
High proportion	1.11	0.57	0.76	-0.19
	(2.46)	(4.57)	(3.29)	(2.93)
N	270	69	167	151
N Seminars	32	29	31	32
N Teams	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.20	0.74	0.80	0.63
$\beta_2 = 0$ (p-value MHT)	0.72	0.49	0.96	0.32
$\beta_3 = 0$ (p-value MHT)	0.90	0.84	0.79	0.83
$\beta_4 = 0$ (p-value MHT)	0.81	0.52	0.82	0.87
$\beta_5 = 0$ (p-value MHT)	0.99	0.95	0.98	0.72
$\beta_6 = 0$ (p-value MHT)	0.60	0.21	0.97	0.38
$\beta_7 = 0$ (p-value MHT)	0.59	0.28	0.70	0.62
$\beta_8 = 0$ (p-value MHT)	0.84	0.92	0.63	0.97

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) and (2) control for native speaker status, and (3) and (4) for gender. Group composition categories are based on the share of women or native speakers *in the team excluding the individual*. The categories are defined as low (< 50% women or < 35% native speakers), medium (50–99% women or 35–75% native speakers), and high (100% women or > 75% native speakers), with the low category as the omitted reference. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is medium proportion of native English speakers in team, β_2 is high proportion of native English speakers in team, etc.)

Table A.3.6: Non-linear Effects on "Final programme grade"

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team				
Medium proportion	0.12	-0.26	0.01	0.08
	(0.06)	(0.14)	(0.11)	(0.08)
High proportion	0.08	-0.23	0.02	0.02
	(0.08)	(0.16)	(0.11)	(0.10)
Proportion Native English Speakers in Seminar				
Medium proportion	-0.04	-0.04	-0.14	0.15
	(0.09)	(0.19)	(0.13)	(0.10)
High proportion	-0.14	0.11	-0.25	0.26
	(0.11)	(0.24)	(0.16)	(0.15)
Proportion of Women in Team				
Medium proportion	-0.01	-0.29	-0.01	-0.01
	(0.09)	(0.21)	(0.13)	(0.10)
High proportion	0.19	-0.21	-0.04	0.15
	(0.19)	(0.36)	(0.24)	(0.25)
Proportion of Women in Seminar				
Medium proportion	0.17	0.24	-0.05	0.19
	(0.19)	(0.35)	(0.24)	(0.24)
High proportion	0.03	-0.28	0.18	-0.12
	(0.08)	(0.20)	(0.13)	(0.09)
N	270	69	167	151
N Seminars	32	29	31	32
N Teams	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.19	0.35	0.94	0.62
$\beta_2 = 0$ (p-value MHT)	0.75	0.45	0.84	0.96
$\beta_3 = 0$ (p-value MHT)	0.84	0.85	0.60	0.36
$\beta_4 = 0$ (p-value MHT)	0.62	0.72	0.29	0.40
$\beta_5 = 0$ (p-value MHT)	0.94	0.61	0.90	0.94
$\beta_6 = 0$ (p-value MHT)	0.46	0.75	0.97	0.46
$\beta_7 = 0$ (p-value MHT)	0.47	0.82	0.94	0.58
$\beta_8 = 0$ (p-value MHT)	0.76	0.49	0.47	0.54

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) and (2) control for native speaker status, and (3) and (4) for gender. Group composition categories are based on the share of women or native speakers *in the team excluding the individual*. The categories are defined as low (< 50% women or < 35% native speakers), medium (50–99% women or 35–75% native speakers), and high (100% women or > 75% native speakers), with the low category as the omitted reference. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is medium proportion of native English speakers in team, β_2 is high proportion of native English speakers in team, etc.)

Table A.3.7: Non-linear Effects on "Dissertation distinction"

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team				
Medium proportion	0.20**	-0.09	0.15	0.08
	(0.07)	(0.15)	(0.10)	(0.08)
High proportion	0.07	-0.06	0.09	-0.01
	(0.08)	(0.17)	(0.10)	(0.11)
Proportion Native English Speakers in Seminar				
Medium proportion	-0.00	-0.08	-0.02	0.04
	(0.09)	(0.20)	(0.13)	(0.11)
High proportion	-0.09	-0.17	-0.24	0.08
	(0.12)	(0.26)	(0.15)	(0.17)
Proportion of Women in Team				
Medium proportion	0.11	-0.20	0.05	0.20
	(0.09)	(0.22)	(0.12)	(0.11)
High proportion	0.27	-0.19	-0.02	0.24
	(0.19)	(0.38)	(0.23)	(0.28)
Proportion of Women in Seminar				
Medium proportion	0.20	0.26	0.04	0.07
	(0.19)	(0.37)	(0.23)	(0.27)
High proportion	0.12	-0.09	0.16	0.04
	(0.09)	(0.21)	(0.13)	(0.10)
N	270	69	167	151
N Seminars	32	29	31	32
N Teams	85	48	75	73
$\beta_1 = 0$ (p-value MHT)	0.01	0.93	0.46	0.41
$\beta_2 = 0$ (p-value MHT)	0.74	0.95	0.74	0.92
$\beta_3 = 0$ (p-value MHT)	1.00	0.96	0.85	0.70
$\beta_4 = 0$ (p-value MHT)	0.79	0.90	0.34	0.67
$\beta_5 = 0$ (p-value MHT)	0.53	0.87	0.99	0.18
$\beta_6 = 0$ (p-value MHT)	0.35	0.68	0.93	0.40
$\beta_7 = 0$ (p-value MHT)	0.46	0.87	0.88	0.77
$\beta_8 = 0$ (p-value MHT)	0.49	0.97	0.51	0.87

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) and (2) control for native speaker status, and (3) and (4) for gender. Group composition categories are based on the share of women or native speakers *in the team excluding the individual*. The categories are defined as low (< 50% women or < 35% native speakers), medium (50–99% women or 35–75% native speakers), and high (100% women or > 75% native speakers), with the low category as the omitted reference. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is medium proportion of native English speakers in team, β_2 is high proportion of native English speakers in team, etc.)

Table A.3.8: Non-linear Effects on "Voice"

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team				
Medium proportion	-0.40	0.59	0.33	-1.08*
	(0.25)	(1.08)	(0.46)	(0.37)
High proportion	0.05	-0.31	0.52	-0.84
	(0.34)	(1.17)	(0.44)	(0.51)
Proportion Native English Speakers in Seminar				
Medium proportion	1.26**	0.60	0.68	0.84
	(0.41)	(1.42)	(0.59)	(0.54)
High proportion	-0.21	0.80	-0.26	-1.00
	(0.46)	(1.82)	(0.65)	(0.71)
Proportion of Women in Team				
Medium proportion	-0.10	0.13	0.10	-0.08
	(0.37)	(1.54)	(0.56)	(0.49)
High proportion	2.00	0.45	1.74	7.23
	(0.74)	(1.48)	(0.99)	(1.65)
Proportion of Women in Seminar				
Medium proportion	1.92	0.00	1.28	6.78
	(0.74)	(.)	(1.02)	(1.62)
High proportion	-0.47	-0.71	-0.23	-0.52
	(0.37)	(1.51)	(0.63)	(0.43)
N	171	44	104	98
N Seminars	32	23	28	31
N Teams	77	36	61	61
$\beta_1 = 0$ (p-value MHT)	0.13	0.90	0.78	0.06
$\beta_2 = 0$ (p-value MHT)	0.87	0.81	0.72	0.40
$\beta_3 = 0$ (p-value MHT)	0.03	0.94	0.47	0.44
$\beta_4 = 0$ (p-value MHT)	0.91	0.92	0.87	0.61
$\beta_5 = 0$ (p-value MHT)	0.97	0.95	0.95	0.99
$\beta_6 = 0$ (p-value MHT)	0.44	0.73	0.35	0.43
$\beta_7 = 0$ (p-value MHT)	0.56	0.89	0.67	0.40
$\beta_8 = 0$ (p-value MHT)	0.54	0.92	0.68	0.58

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2020) (four hypotheses included). All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) and (2) control for native speaker status, and (3) and (4) for gender. Group composition categories are based on the share of women or native speakers *in the team excluding the individual*. The categories are defined as low (< 50% women or < 35% native speakers), medium (50–99% women or 35–75% native speakers), and high (100% women or > 75% native speakers), with the low category as the omitted reference. The β_k coefficients tested in the bottom rows correspond to the order in which variables appear in the main coefficient panel above (e.g., β_1 is medium proportion of native English speakers in team, β_2 is high proportion of native English speakers in team, etc.

Table A.3.9 Regression coefficients: Perceptions of voice (2)

	(a) "More confident in voicing my view"			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	-0.25	3.76	-0.96	0.38
	(1.89)	(5.18)	(2.17)	(3.36)
Proportion Native English Speakers in Seminar (β_2)	5.21	0.11	3.44	6.43
	(3.87)	(6.31)	(4.23)	(6.44)
Proportion of Women in Team (β_3)	-1.27	1.13	-5.32	10.14
	(3.09)	(10.21)	(3.47)	(5.21)
Proportion of Women in Seminar (β_4)	-6.03	19.02	-13.10	10.36
	(8.89)	(28.90)	(10.64)	(14.02)
N	58	21	47	29
N Seminars	22	17	21	17
N Teams	35	21	29	26
$\beta_1 = 0$ (p-value MHT)	0.89	0.99	0.71	0.95
$\beta_2 = 0$ (p-value MHT)	0.65	1.00	0.80	0.91
$\beta_3 = 0$ (p-value MHT)	0.87	1.00	0.35	0.65
$\beta_4 = 0$ (p-value MHT)	0.83	0.99	0.61	0.91
	(b) "My voice will be heard"			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	-0.15	1.35	-2.27	-0.47
	(1.94)	(5.99)	(2.34)	(3.19)
Proportion Native English Speakers in Seminar (β_2)	3.85	1.56	6.65	4.23
	(3.88)	(6.42)	(4.46)	(6.11)
Proportion of Women in Team (β_3)	-0.49	0.61	-5.39	9.52
	(3.18)	(9.67)	(3.67)	(4.94)
Proportion of Women in Seminar (β_4)	-5.71	11.70	-16.63	5.10
	(9.30)	(29.17)	(11.49)	(13.30)
N	57	20	45	30
N Seminars	21	17	19	17
N Teams	35	21	27	27
$\beta_1 = 0$ (p-value MHT)	0.95	1.00	0.39	0.94
$\beta_2 = 0$ (p-value MHT)	0.84	1.00	0.42	0.96
$\beta_3 = 0$ (p-value MHT)	0.98	1.00	0.42	0.65
$\beta_4 = 0$ (p-value MHT)	0.87	1.00	0.48	0.97

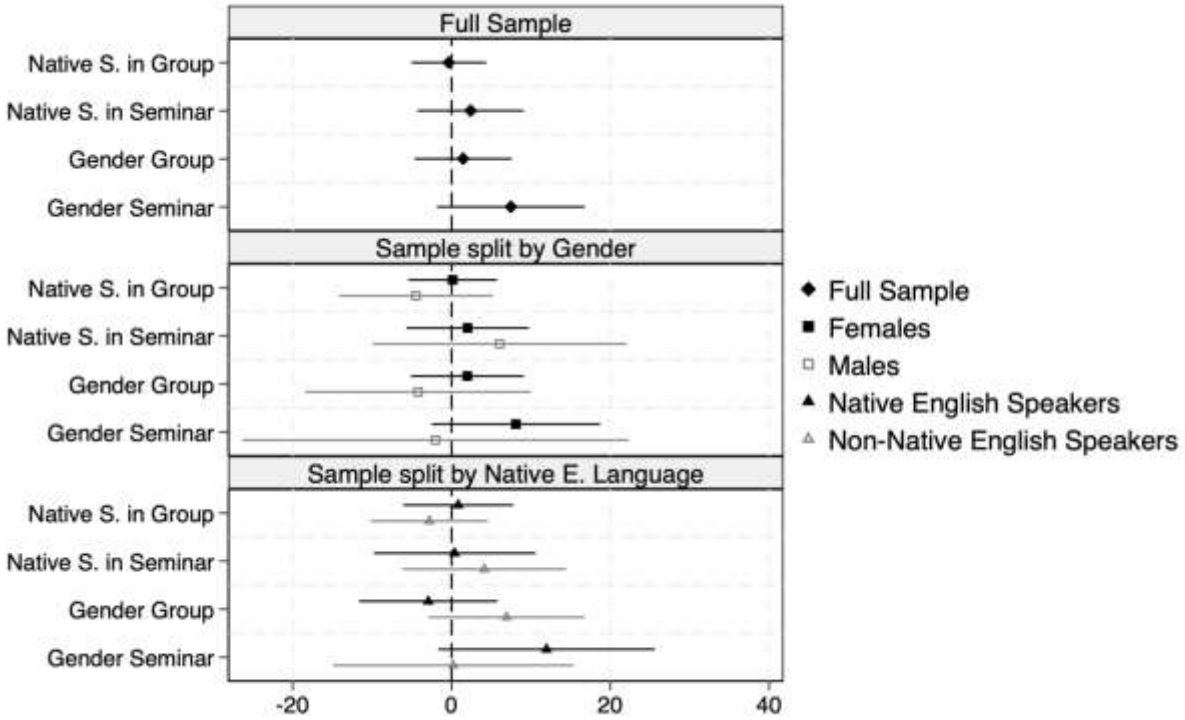
Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01. p-values adjusted for multiple hypothesis testing (MHT) as in Barsbai et al. (2024) (four hypotheses included). All models include controls for the student's age, education level, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, and whether the advisor was native English speaker. The variable "More confident in voicing my view" measures the level of agreement (from 0 to 10) with the statement "Working in teams for SP401 made more confident than before in voicing my view in future interactions." Data source: End year survey 2021-2022. The dependent variable "My voice will be heard" indicates the level of agreement (from 0-10) with the statement "Working in teams for SP401 made more confident than before that my view will be heard in future interactions." Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team). Data source: Endline survey, 2021-2022 cohort

Table A.3.10: Regression coefficients: "My voice was heard during group discussions" – Adjusted survey weights

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	-0.50	0.23	0.50	-1.59*
	(0.45)	(0.99)	(0.62)	(0.64)
Proportion Native English Speakers in Seminar (β_2)	0.36	-1.28	-0.29	0.79
	(0.83)	(1.31)	(0.68)	(1.07)
Proportion of Women in Team (β_3)	1.36*	0.73	1.38	2.34*
	(0.64)	(1.29)	(0.70)	(0.95)
Proportion of Women in Seminar (β_4)	2.11	-1.08	1.50	2.62
	(1.18)	(2.81)	(0.87)	(2.04)
N	171	44	104	98
N Seminars	43	29	38	40
N Teams	77	36	61	61

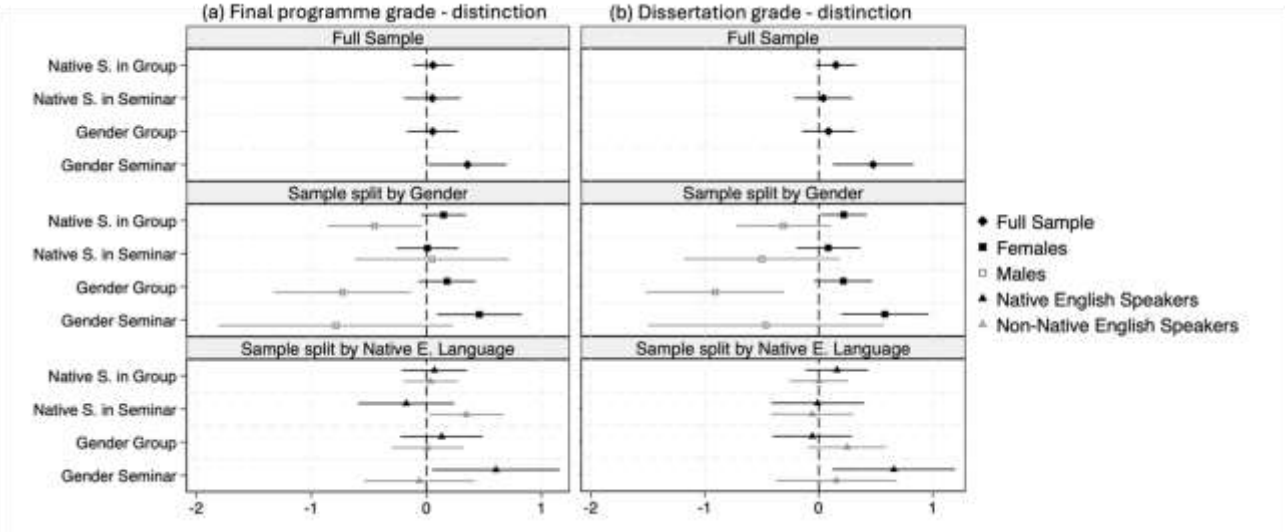
Notes: Dependent variable is agreement (0–10) with "My voice was heard during group discussions." Regressions are weighted using adjusted survey weights to correct for non-response via inverse-probability weighting: we first estimate each student's probability of responding to the endline survey using a logistic regression on observable characteristics: cohort, birth year, gender, native-English status, prior UK studies, and seminar group (odds ratios reported in Table OA.3), then assign each respondent a weight equal to the inverse of their predicted response probability. Standard errors in parentheses . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team). Data source: Endline survey, 2020-2021, and 2021-2022

Figure A.3.1 Regression coefficients plot: Exam grades – (95% CI, Fisher p-values)



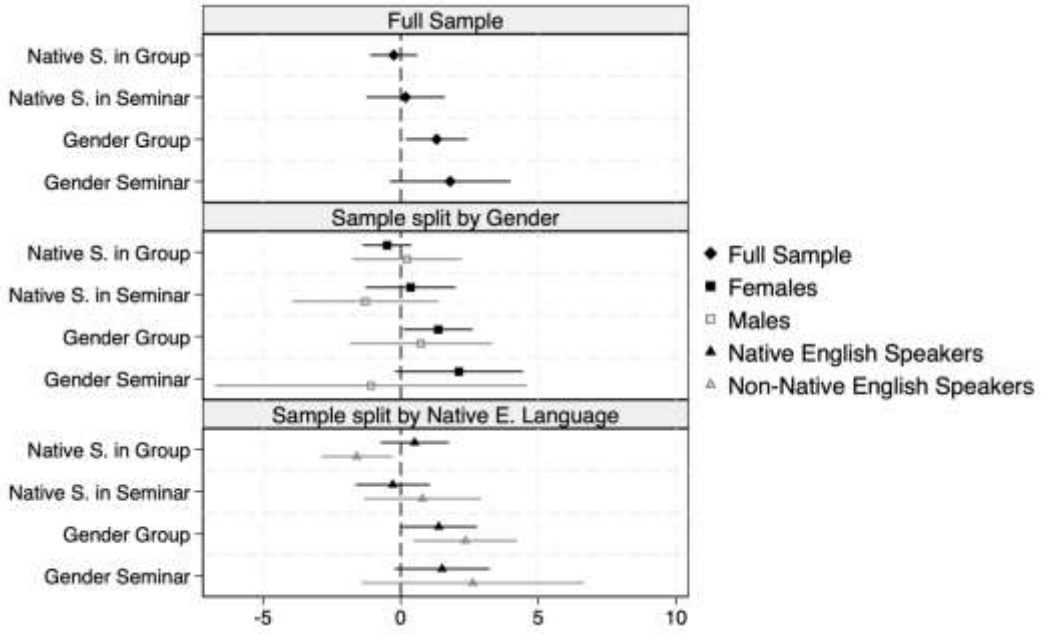
Notes: This figure plots estimated coefficients (with 95 % confidence intervals) where the dependent variable is Exam Grades (scale from 0-100). All models include controls for the student's age, education level, experience with quantitative methods, experience with qualitative methods, proportion of native speakers in the team and in the seminar group, and proportion of women in the team and seminar group.. Data source: Administrative records for cohorts 2020-2021, and 2021-2022.

Figure A.3.2 Regression coefficients plot: Final programme grade distinction and dissertation distinction – (95% CI, Fisher p-values)



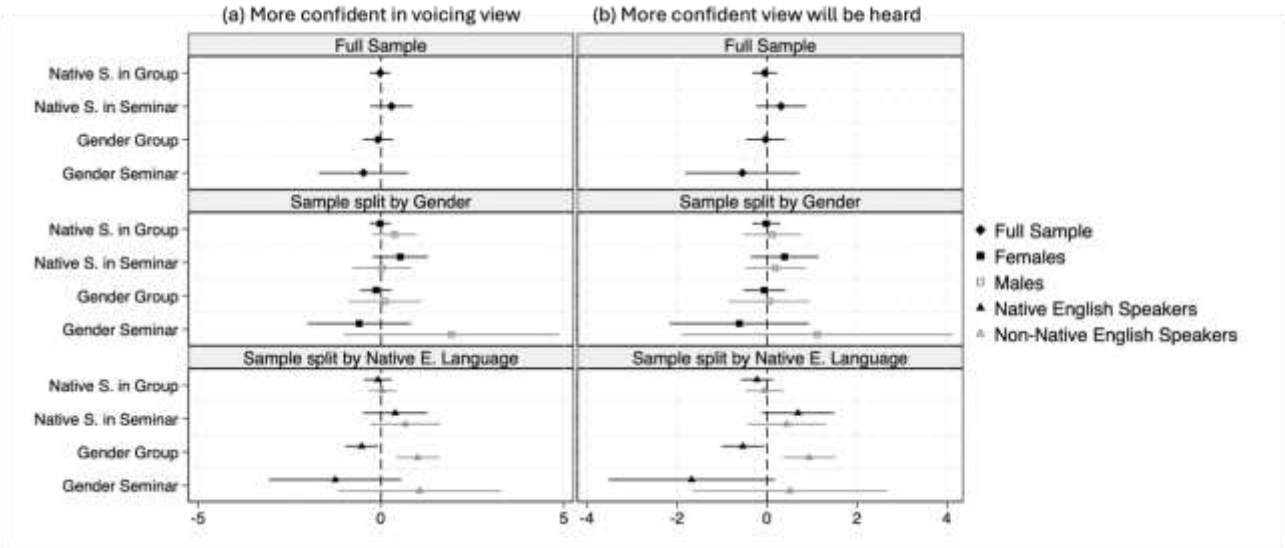
Notes: Panels (a) and (b) plot estimated coefficients (with 95 % confidence intervals) from two binary outcomes: Panel (a): Final Programme Grade – Distinction, coded 1 if a student's overall postgraduate programme mark is ≥ 70, and 0 otherwise. Panel (b): Dissertation Grade – Distinction, coded 1 if a student's dissertation mark is ≥ 70, and 0 otherwise. Each row of each panel shows estimates for the full sample (top), samples split by gender (middle), and samples split by native-English status (bottom). All models include controls for the student's age, education level, experience with quantitative methods, experience with qualitative methods, proportion of native speakers in the team and in the seminar group, and proportion of women in the team and seminar group. Data source: administrative records for 2020–21 and 2021–22 cohorts.

Figure A.3.3 Regression coefficients plot: Voice – (95% CI, Fisher p-values)



Notes: This figure plots estimated coefficients (with 95 % confidence intervals) where the dependent variable is the student's response to "My voice was heard during group discussions" (scale: 0–10). Each row shows results for: Full sample (top panel), sample split by gender (middle panel), sample split by native-English status (bottom panel). All models include controls for the student's age, education level, experience with quantitative methods, experience with qualitative methods, proportion of native speakers in the team and in the seminar group, and proportion of women in the team and seminar group. Data source: Endline survey

Figure A.3.4 Regression coefficients plot: confidence in voicing view and view will be heard – (95% CI, Fisher p-values)



Notes: Panels (a) and (b) plot estimated coefficients (with 95 % confidence intervals) from two outcomes: Panel (a): More confident in voicing view– measures the level of agreement (from 0 to 10) with the statement "Working in teams for SP401 made more confident than before in voicing my view in future interactions.". Panel (b): more confident view will be heard indicates the level of agreement (from 0-10) with the statement "Working in teams for SP401 made more confident than before that my view will be heard in future interactions.". All models include controls for the student's age, education level, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, proportion of native speakers in the team and in the seminar group, and proportion of women in the team and seminar group. Data source: Endline survey, 2021-2022 cohort.

A.4 Oster Bounds Estimation

In this section, to assess the robustness of our results to potential omitted variable bias (OVB) we calculate the bounds proposed by Oster (2019) for the treatment coefficients that were statistically significant (before MHT). We estimate such bounds under the assumption that the relative importance of observed versus unobserved omitted variables in generating selection bias is the same. Additionally, to calculate the bounds, there is need to make an additional assumption about the size of the of a hypothetical regression that controls for all relevant observed and unobserved factors. Oster (2019) suggests a possible value for this hypothetical R^2, referred to as R_{max} to be equal to the R^2 from the regression controlling for all observable factors multiplied by a factor of 1.3, and we use this value in our calculations as well.

Table A.4.1 presents the estimated bounds for the various model specifications in our analysis. For each specification, we provide the R-squared from the controlled models "R", the estimated coefficient, and the calculated bound.

Table A.4.1: Oster Bounds for Selected Coefficients

Independent Variable	R	Estimated Coefficient	Bound	Delta
A: My voice was heard during group discussions (Figure A.3.3)				
Full sample				
Proportion of Women (Team)	0.102	1.30	[1.30,1.46]	1.37
Women subsample				
Proportion of Women (Team)	0.127	2.11	[2.11, 2.22]	1.34
Non-Native Speakers subsample				
Proportion of Native Speakers (Team)	0.289	-1.60	[-1.60, -1.60]	1.39
Proportion of Women (Team)	0.289	2.35	[2.35, 2.74]	1.30
B: More confident in voicing my view (Figure A.3.4)				
Non-Native Speakers subsample				
Proportion of Women (Team)	0.486	1.00	[1.00, 1.83]	1.36
C: My voice will be heard (Figure A.3.4)				
Non-Native Speakers subsample				
Proportion of Women (Team)	0.469	0.94	[0.94,1.74]	1.38

Note: Bounds are calculated using the method proposed by Oster (2019) to assess the sensitivity of our results to omitted variable bias. To calculate the bounds, we assume that the relative importance of observed versus unobserved omitted variables in generating selection bias is the same. The third column of the table presents the estimated coefficient, while the last one presents the Oster bound. If the Oster bound is close to the estimated coefficient, this suggests that our results are less sensitive to potential omitted variable bias. If it is far from the estimated coefficient, our results may be more sensitive to OVB.

Online Appendix

Table OA.1: Description of dependent variables

Variable	Description	Range	Data Source	Questionnaire Item
Academic outcomes				
Exam Grades	Grade on the course's final exam. This is the only assessment in the course.	0-100	Administrative records	
Dissertation - Distinction	Binary variable that indicates if a student obtains distinction (value 1) or not (value 0). A student is awarded distinction if their dissertation mark is equal or above 70.	0,1		
Final Programme Grade - Distinction	Binary variable that indicates if a student is awarded distinction (value 1) or not (value 0) as their overall mark in the postgraduate programme.	0,1		
Team dynamics				
"My voice was heard during group discussions"	The variable indicates the level of agreement with the statement "My voice was heard during group discussions"	0-10	Endline survey	Level of agreement from 0 to 10 with the following statements: (1) "My voice was heard during group discussions" (2) "Working in teams for SP401 made more confident than before in voicing my view in future interactions" *(2021-2022 only)* (3) "Working in teams for SP401 made more confident than before that my view will be heard in future interactions. *(2021-2022 only)*
"More confident in voicing my view"	The variable indicates the level of agreement with the statement "Working in teams for SP401 made more confident than before in voicing my view in future interactions"			
"My voice will be heard"	The variable indicates the level of agreement with the statement "Working in teams for SP401 made more confident than before that my view will be heard in future interactions"			

Survey Attrition

Although most students answered the baseline survey, there is some attrition at the endline survey. Table OA.2 presents descriptive statistics for the endline survey respondents (Response=1), and non-respondents (Response=0). Table OA.3 presents the odd ratios of the logistic regression used for the inverse probability weighting.

Table OA.2: Covariates' mean value for respondents and non-respondents

Covariates	Means	
	Response=0	Response=1
Age	24.7	24.5
Female=1	0.79	0.80
Native English=1	.50	0.47
UK Studies=1	0.35	0.32
Highest level of education (Master) = 1	0.20	0.19
Proportion Native Speakers (Team)	45.8%	51.4%
Proportion Native Speakers (Seminar -O)	42.1%	47.8%
Proportion of Women (Team)	78.2%	80.5%
Proportion of Women (Seminar -O)	78.2%	81.6%
Expected Grade	71.5	73.1
Experience with Quantitative methods	4.0	4.2
Experience with Qualitative methods	5.7	5.9
Adviser Gender (female=1)	0.54	0.46
Adviser Native Language (English=1)	0.64	0 .55
Cohort (2021==1)	0.37	0.60
Total (N)	153	226

Notes: Expected Grade refers to students' self-reported anticipated performance on a scale from 0 to 100. Experience with quantitative methods and qualitative methods are self-reported at baseline, each rated from 0 (no familiarity) to 10 (high familiarity). Team composition variables—Proportion Native Speakers and Proportion of Women—are calculated as percentages of native English speakers and women within the individual's team and seminar group (excluding the individual's own team). Data sources include baseline survey responses and administrative records.

Table OA.3: Odds ratios for end survey response

	Response
Cohort	0.82***
	(0.24)
Year Born	0.01
	(0.03)
Female =1	-0.05
	(0.28)
Native English =1	-0.18
	(0.23)
UK studies =1	-0.24
	(0.25)
Seminar Group	0.03
	(0.02)
N	355

Notes: This table reports odds ratios from a logistic regression predicting whether a student completed the endline survey (1 = response). The estimated response probabilities from this model are used to construct inverse probability weights—each respondent's weight equals the inverse of their fitted probability of response—which we apply to correct for survey nonresponse. *p < 0.05; **p < 0.01; ***p < 0.001.

OB.1 Clustered errors at seminar level

Table OB.1.1: Regression coefficients: "Exam Marks" - errors clustered at seminar level

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	0.09	-4.47	0.85	-2.87
	(2.48)	(4.28)	(3.36)	(3.11)
Proportion Native English Speakers in Seminar (β_2)	2.02	6.06	0.40	4.14
	(2.94)	(7.70)	(4.02)	(3.54)
Proportion of Women in Team (β_3)	1.98	-4.19	-2.92	6.93
	(3.36)	(4.64)	(3.62)	(3.59)
Proportion of Women in Seminar (β_4)	8.11	-2.02	11.97*	0.20
	(4.32)	(14.70)	(4.65)	(6.84)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73

Notes: Standard errors clustered at the seminar level in parentheses . * p < 0.05, ** p < 0.01, *** p < 0.001. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team)

Table OB.1.2: Regression coefficients: Final Programme Grades - errors clustered at seminar level

	(c) Classification: Final Grade Distinction			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	0.15	-0.45*	0.07	0.03
	(0.09)	(0.18)	(0.14)	(0.10)
Proportion Native English Speakers in Seminar (β_2)	0.01	0.05	-0.18	0.35*
	(0.15)	(0.32)	(0.20)	(0.15)
Proportion of Women in Team (β_3)	0.18	-0.73**	0.13	0.01
	(0.12)	(0.26)	(0.15)	(0.16)
Proportion of Women in Seminar (β_4)	0.46*	-0.79	0.60*	-0.06
	(0.19)	(0.42)	(0.24)	(0.23)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73
	(d) Classification: Dissertation Distinction			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	0.22*	-0.31	0.16	0.00
	(0.09)	(0.17)	(0.11)	(0.12)
Proportion Native English Speakers in Seminar (β_2)	0.08	-0.50	-0.01	-0.06
	(0.12)	(0.35)	(0.16)	(0.15)
Proportion of Women in Team (β_3)	0.22	-0.91***	-0.06	0.25
	(0.12)	(0.22)	(0.16)	(0.15)
Proportion of Women in Seminar (β_4)	0.58***	-0.47	0.66*	0.16
	(0.15)	(0.40)	(0.26)	(0.22)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73

Notes: Standard errors clustered at the seminar level in parentheses . * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team).

Table OB.1.3: Regression coefficients: "My voice was heard during group discussions" - errors clustered at seminar level

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	-0.49	0.25	0.48	-1.52*
	(0.43)	(1.26)	(0.75)	(0.60)
Proportion Native English Speakers in Seminar (β_2)	0.20	-1.40	-0.48	0.70
	(0.98)	(1.50)	(0.68)	(1.15)
Proportion of Women in Team (β_3)	1.40	0.47	1.38	2.41*
	(0.74)	(1.56)	(0.69)	(1.04)
Proportion of Women in Seminar (β_4)	1.99	-1.57	1.44	2.56
	(1.49)	(3.30)	(0.94)	(2.24)
N	171	44	104	98
N Seminars	43	29	38	40
N Teams	77	36	61	61

Notes: Standard errors clustered at the seminar level in parentheses . * p < 0.05, ** p < 0.01, *** p < 0.001. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. The variable "My voice was heard during group discussions" measures the level of agreement with the statement in a scale of 0-10. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team). Data source: Endline survey, 2020-2021, and 2021-2022 cohorts.

OB.2 Clustered errors at team level

Table OB.2.1: Regression coefficients: "Exam Marks" - errors clustered at team level

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	0.09	-4.47	0.85	-2.87
	(2.61)	(4.19)	(3.25)	(3.34)
Proportion Native English Speakers in Seminar (β_2)	2.02	6.06	0.40	4.14
	(3.31)	(7.63)	(4.14)	(3.90)
Proportion of Women in Team (β_3)	1.98	-4.19	-2.92	6.93
	(3.37)	(4.33)	(4.09)	(3.57)
Proportion of Women in Seminar (β_4)	8.11	-2.02	11.97*	0.20
	(4.85)	(12.79)	(5.79)	(6.68)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73

Notes: Standard errors clustered at the team level in parentheses . * p < 0.05, ** p < 0.01, *** p < 0.001. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own tea

Table OB.2.2: Regression coefficients: Final Programme Grades - errors clustered at team level

	(a) Classification: Final Grade Distinction			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	0.15	-0.45*	0.07	0.03
	(0.09)	(0.17)	(0.14)	(0.11)
Proportion Native English Speakers in Seminar (β_2)	0.01	0.05	-0.18	0.35*
	(0.14)	(0.35)	(0.19)	(0.14)
Proportion of Women in Team (β_3)	0.18	-0.73**	0.13	0.01
	(0.13)	(0.27)	(0.17)	(0.16)
Proportion of Women in Seminar (β_4)	0.46*	-0.79	0.60*	-0.06
	(0.21)	(0.43)	(0.26)	(0.27)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73
	(b) Classification: Dissertation Distinction			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	0.22*	-0.31	0.16	0.00
	(0.10)	(0.17)	(0.12)	(0.13)
Proportion Native English Speakers in Seminar (β_2)	0.08	-0.50	-0.01	-0.06
	(0.11)	(0.34)	(0.17)	(0.15)
Proportion of Women in Team (β_3)	0.22	-0.91***	-0.06	0.25
	(0.12)	(0.25)	(0.16)	(0.14)
Proportion of Women in Seminar (β_4)	0.58***	-0.47	0.66**	0.16
	(0.17)	(0.39)	(0.23)	(0.24)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73

Notes: Standard errors clustered at the team level in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team).

Table OB.2.3: Regression coefficients: "My voice was heard during group discussions" - errors clustered at team level

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	-0.49	0.25	0.48	-1.52*
	(0.47)	(1.23)	(0.76)	(0.65)
Proportion Native English Speakers in Seminar (β_2)	0.20	-1.40	-0.48	0.70
	(0.99)	(1.68)	(0.72)	(1.14)
Proportion of Women in Team (β_3)	1.40	0.47	1.38*	2.41*
	(0.72)	(1.54)	(0.66)	(1.05)
Proportion of Women in Seminar (β_4)	1.99	-1.57	1.44	2.56
	(1.47)	(3.78)	(0.94)	(2.24)
N	171	44	104	98
N Seminars	43	29	38	40
N Teams	77	36	61	61

Notes: Standard errors clustered at the team level in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. The variable "My voice was heard during group discussions" measures the level of agreement with the statement in a scale of 0-10. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team). Data source: Endline survey, 2020-2021, and 2021-2022 cohorts

OB.3 Wild cluster bootstrap errors at team level

Table OB.3.1: Regression coefficients: "Exam Marks" - Wild cluster bootstrap errors clustered at team level

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	0.09	-4.47	0.85	-2.87
	(2.61)	(4.19)	(3.25)	(3.34)
Proportion Native English Speakers in Seminar (β_2)	2.02	6.06	0.40	4.14
	(3.31)	(7.63)	(4.14)	(3.90)
Proportion of Women in Team (β_3)	1.98	-4.19	-2.92	6.93
	(3.37)	(4.33)	(4.09)	(3.57)
Proportion of Women in Seminar (β_4)	8.11	-2.02	11.97*	0.20
	(4.85)	(12.79)	(5.79)	(6.68)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73

Notes: Standard errors clustered at the team level in parentheses . * p < 0.05, ** p < 0.01, *** p < 0.001. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team)

Table OB.3.2: Regression coefficients: Final Programme Grades - Wild cluster bootstrap errors clustered at team level

	(a) Classification: Final Grade Distinction			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	0.15	-0.45*	0.07	0.03
	(0.09)	(0.17)	(0.14)	(0.11)
Proportion Native English Speakers in Seminar (β_2)	0.01	0.05	-0.18	0.35*
	(0.14)	(0.35)	(0.19)	(0.14)
Proportion of Women in Team (β_3)	0.18	-0.73**	0.13	0.01
	(0.13)	(0.27)	(0.17)	(0.16)
Proportion of Women in Seminar (β_4)	0.46*	-0.79	0.60*	-0.06
	(0.21)	(0.43)	(0.26)	(0.27)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73
	(b) Classification: Dissertation Distinction			
	(1) Women	**(2) Men**	**(3) Native Speakers**	**(4) Non-Native Speakers**
Proportion Native English Speakers in Team (β_1)	0.22*	-0.31	0.16	0.00
	(0.10)	(0.17)	(0.12)	(0.13)
Proportion Native English Speakers in Seminar (β_2)	0.08	-0.50	-0.01	-0.06
	(0.11)	(0.34)	(0.17)	(0.15)
Proportion of Women in Team (β_3)	0.22	-0.91***	-0.06	0.25
	(0.12)	(0.25)	(0.16)	(0.14)
Proportion of Women in Seminar (β_4)	0.58***	-0.47	0.66**	0.16
	(0.17)	(0.39)	(0.23)	(0.24)
N	268	69	166	150
N Seminars	43	35	42	42
N Teams	85	48	75	73

Notes: Standard errors clustered at the team level in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team).

Table OB.3.3: Regression coefficients: "My voice was heard during group discussions" - Wild cluster bootstrap errors clustered at team level

	(1) Women	(2) Men	(3) Native Speakers	(4) Non-Native Speakers
Proportion Native English Speakers in Team (β_1)	-0.49	0.25	0.48	-1.52*
	(0.47)	(1.23)	(0.76)	(0.65)
Proportion Native English Speakers in Seminar (β_2)	0.20	-1.40	-0.48	0.70
	(0.99)	(1.68)	(0.72)	(1.14)
Proportion of Women in Team (β_3)	1.40	0.47	1.38*	2.41*
	(0.72)	(1.54)	(0.66)	(1.05)
Proportion of Women in Seminar (β_4)	1.99	-1.57	1.44	2.56
	(1.47)	(3.78)	(0.94)	(2.24)
N	171	44	104	98
N Seminars	43	29	38	40
N Teams	77	36	61	61

Notes: Standard errors clustered at the team level in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001. All models include controls for the student's age, education level, previous UK studies, experience with quantitative methods, experience with qualitative methods, the seminar leader's gender, whether the seminar leader was native English speaker, the advisor's gender, whether the advisor was native English speaker, and Language score for non-native speakers. Specification (1) controls for native English speaker status and gender, (2) and (3) control for native speaker status, and (4) and (5) for gender. The variable "My voice was heard during group discussions" measures the level of agreement with the statement in a scale of 0-10. Coefficients β_2 and β_4 correspond to the estimated effect of the group composition of other members in the seminar group (excluding the individual's own team). Data source: Endline survey, 2020-2021, and 2021-2022 cohort