Gender and Performance in Teams: Evidence From Students

Max Coveney [*] [†]   Teresa Bago d'Uva [*]   Pilar García-Gómez[*]

February 28, 2024

Abstract

Should firms take into account gender mix when forming work teams? This paper examines how the performance of teams completing tasks found in many white-collar occupations is influenced by their gender composition. Leveraging a first-year economics bachelor course in which students are randomly paired together and perform tasks such as document preparation, data analysis, and presentations, we document large differences in performance grades by gender composition. All-male teams are significantly outperformed by both mixed and all-female teams. These differences remain even when comprehensively controlling for the individual task aptitude of the group members, as well as other characteristics potentially relevant for teamwork that may vary by gender. In contrast, individuals in mixed-gender teams, especially women, report the worse outcomes along many subjective dimensions, including reported motivation, team atmosphere, and team unity.

[*] Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam 3062PA, the Netherlands

[†] Corresponding author. coveney@ese.eur.nl

[‡] The authors have no relevant or material financial interests that relate to the research described in this paper. This project is financed by the Erasmus Community for Learning & Innovation (CLI) Fellowship program. IRB approval for this study was obtained through Erasmus University (ETH2122-0522). The paper has also benefited from the comments and suggestions of participants at various seminars. All omissions and errors are our own.

1 Introduction

While progress has been slow and uneven, gender diversity at the workplace is historically high; women are increasingly found in occupations and roles previously dominated by men (Goldin, 2006, 2014). A raft of policy initiatives have been introduced to further bridge the gap in corporate boards, panels, and other areas where women's representation has been low (Hughes et al., 2017). At the same time, most firms now explicitly organise their employees into work teams for production (Lazear and Shaw, 2007), and the organisation of such teams is a critical firm decision.[1] What are the implications for this increase in gender diversity for work teams, and how can firms best take advantage of these demographic changes when assembling teams?

To shed light on these questions, this paper studies how the gender composition of work teams influences team performance. Using data on teams of university students engaged in generic tasks such as writing and document preparation, data processing and analysis, feedback-giving, and oral presentation tasks, we show a large and robust gender composition effect on team performance as measured by grades. Teams with more women tend to produce significantly better quality work, even controlling for the *individual* ability of each group member.

Leveraging the random allocation of 4 cohorts of roughly 2,600 students to 2,700 work teams, and using grades as a measure of performance, we are able to estimate the importance of a team's gender composition for its outcomes. Drawing on approximately 10,600 team-task observations, we reach several conclusions. First, there are sizable and significant differences in task performance, as measured by grades, depending on the gender composition of the team. Teams comprised of two women (one woman and one man), produce work that is graded on average 15% (20%) of a standard deviation better than teams comprised of two men.

Second, we find that these differences are not driven by individual task ability differences between women and men, or by other observable characteristics that may vary by gender. Teams with more women are found to perform better, even with the addition of compre-

[1] Figure 1 shows the prevalence of teamwork on the job across 10 large European economies and the US based on employee microdata. Across most occupational categories and countries, the majority of respondents report using teamwork on the job.

hensive individual ability controls for all members of the team, as well as when controlling for possible correlates of gender composition, such as the socio-economic status (SES), nationality, or ethnicity composition of teams.

Third, we show these results hold across different types of task and different group sizes. By distinguishing between the performance on various tasks in our data, we find the gender composition effect remains across all tasks: writing, data analysis, feedback giving, and presentations. Further, we find a qualitatively similar gender composition pattern in a small sample of larger groups, suggesting the effect is not isolated to pairs.

After establishing the existence of a gender composition effect on the quality of output, we turn to investigating differences in group processes and experience that may reveal mechanisms possibly driving the effect. Using data from a self-reported evaluation exercise, we group potential explanations following existing literature on gender and teamwork. Based on these explanations, we elicit individuals' group working experiences, the reported contributions of each member, the existence of particular team-working processes and leadership structures, and other differences between teams that may serve as potential mechanisms.

While we do not find strong evidence for one particular set of explanations, the subjective measures reveal a contrasting pattern to those found in the performance analysis; mixed-gender teams report worse outcomes along many dimensions – including conflict, trust, and motivation – compared to all-female and all-male teams. We find this is primarily driven by the responses of women within the mixed groups, and speculate that it may be driven by a mismatch in diligence and conscientiousness within these pairs with women taking on more menial and costly tasks.

The multi-disciplinary literature studying the effect of gender composition on team outcomes includes studies of research teams (Yang et al., 2022; Díaz-García et al., 2013; Hengel, 2020; Hengel and Moon, 2023), corporate teams (Green and Homroy, 2018), evaluative committees (Bagues and Esteve-Volart, 2010; Bagues et al., 2017), student business-game teams (Fenwick and Neal, 2001; Apesteguia et al., 2012; Hoogendoorn et al., 2013), political bodies (Hannagan and Larimer, 2010), and teams within the moving industry (Jehn et al., 1999).[2]

[2] Also see Bear and Woolley (2011) for an overview of the literature on gender and team performance, with a focus on research teams.

Three features of our context allow us to make novel contributions to this literature. First, the tasks performed by teams in our data are comparable with those performed in many white-collar occupations.[3] One drawback of existing research is that the findings from these specialized teams may not generalize well to many of the occupations where teamwork is highly prevalent; the tasks and characteristics of individuals involved in – for instance – evaluative committees may not easily generalize to teamwork in most firms.

Second, many existing papers study teams that have been endogenously formed (Apesteguia et al., 2012; Yang et al., 2022; Hengel and Moon, 2023). While analyses of such teams is informative, a potential caveat to these findings is that teams who chose to work in certain gender combinations may differ from other teams in important but unobserved ways. We avoid this problem by studying randomly allocated teams.

Third, we have rich data data on each individual in our sample, including their previously measured individual performance on similar tasks, ethnicity, and SES. By controlling for ability and other characteristics, we can rule out that the gender composition effect is being driven by these other variables that may correlate with gender in our sample. This provides further assurance that we are estimating a gender effect.

This paper contributes to the growing economic literature broadly studying gender and group work, often via experimental methods and samples of students (Karpowitz et al., 2023; Born et al., 2020; Sarsons et al., 2021; Keck and Tang, 2018).[4] In a similar spirit, ours is the first paper to show how gender composition impacts the performance of teams performing tasks comparable to those in many white-collar occupations. Though based on student data, the realism of our setting is enhanced by both the long team interaction period of multiple months and the nature of tasks performed having a large overlap with tasks performed in real occupations. Many of the students in our sample will go to work in such occupations.

Overall, our findings paint a more nuanced picture of the effect of gender composition in team performance. In line with earlier research, we document a positive impact of women

[3] We show this in Appendix A.1 by comparing the contents of the tasks performed by the student teams to those in a external taxonomy of US occupational tasks.

[4] Based on evidence from student teams, Karpowitz et al. (2023) show that minority women in majority-male teams participate less, and are less likely to be seen as influential or be chosen as group leader. Born et al. (2020) use experimental teams of students to show that women are significantly less willing to lead male-majority teams. Sarsons et al. (2021) use economists' CVs and experimental data to show that women tend to get less credit for group work than men. Using laboratory experiments, Keck and Tang (2018) find that group judgment quality is positively impacted by the presence of a female group member.

in teams on the quality of output (Woolley et al., 2010; De Paola et al., 2022; Keck and Tang, 2018; Fenwick and Neal, 2001; Hoogendoorn et al., 2013; Yang et al., 2022; Hengel, 2020; Hengel and Moon, 2023), although our subsequent analyses finds mixed-gender teams report worse outcome along many subjective dimensions, especially the women within these teams.

Although the low sample size and the self-reported nature of our data prevent a definitive statement on mechanisms, the pattern of results suggests that while the presence of women in teams may raise the quality of output through a boost in diligence and social-sensitivity, the potentially higher burden shouldered by women in these teams may lead to them experiencing worse team atmosphere, unity and motivation.

Respecting the potentially limited external validity of our student data, our findings lead to two main policy implications. Firstly, it appears that increasing gender diversity in traditionally male dominated teams will lead to average performance gains. Organisations could benefit significantly simply by ensuring work teams include at least one woman. Secondly, however, managers should be aware that performance gains may come at the cost of a larger burden of menial, costly, and potentially unrewarded tasks (Babcock et al., 2017) for women in these teams. Our results show that the reported team atmosphere, motivation, and unity experienced by women in these teams will suffer if this burden is unaddressed.

The remainder of the paper is organized as follows. Section 2 describes the setting and data. Section 3 outlines the regression methodology we use to identify the gender composition effect. Section 4 presents the baseline results of the analysis and various extensions. In Section 5 we test the robustness of our baseline results. Section 6 describes the group work self-reflection exercise and the analysis aimed at investigating the mechanisms of the gender composition effect. Finally, Section 7 concludes.

2 Context and Data

2.1. Setting Our context is the first year of the undergraduate Economics program at a top-ranked university in The Netherlands.[5] The program is offered in separate English and

[5] The university is continuously ranked as among the top universities in the Business and Economics category in the country.

a Dutch language version with otherwise identical contents and assessment, and admits roughly 600 students per year. Each academic year consists of consists of 5 blocks (semesters), each lasting eight weeks.

Course overview We leverage a compulsory course spanning blocks 2 to 5 of the first year of the Economics program. The course centers on fostering various important skills, with a focus on writing and document preparation, presentation skills, feedback-giving skills, and research and data analysis skills. The grade achieved in the course accounts for roughly 7% of the students' first year GPA.

Each block has a certain focus, with block 2 devoted to academic communication, block 3 broadly to writing, block 4 to data and analyses, and block 5 culminating in writing and presenting an entire research document. Alongside the tasks, students are required to attend tutorial sessions (four per block) in which course materials are explained.[6]

Course structure Figure 2 shows the structure of the course. The tasks in block 2 involve individual students researching and presenting on an economics subject, giving written feedback on their peers' presentations, and presenting an online pitch on an academic or business subject of their choice. The grades achieved in this block represent an *individual*-level measure of student's aptitude on tasks that are similar to the ones they will subsequently complete in teams.

In blocks 3, 4 and 5 students work in pairs to complete tasks. At the beginning of each block, students are randomized into new teams. Students are required to work in their allocated team for the whole block, and thus work together for 2 months. The teams are formed within classroom groups (sections), comprising of approximately 15 students.[7] These classroom groups meet several times per block for students to present their work and to discuss upcoming tasks. Each classroom is lead by a teaching assistant (TA), who is also in charge of grading the tasks of the students in their classroom.[8] A student's performance across all tasks in the three blocks forms their final grade for the course.

Students are randomly allocated to teams of two since the 2018 academic year. Thus,

[6] Students must attend at least three of the four tutorial sessions per block in order to pass their first year.

[7] In the case of an uneven number of students, one group of three is formed. We discard these groups in our main analyses but make use of them when examining larger teams.

[8] In Appendix A.3 we explore how TA gender influences grading patterns to investigate potential bias. While male graders tend to give higher grades, we find a gender composition effect for both TA genders.

our focus is the period covering the 2018, 2019, 2020 and 2021 academic years.[9] One exception to this is block 4 of the 2018 academic year, during which students were randomly allocated to teams of 4, 5, and 6 members, much larger than the groups of two used in all other blocks. We therefore exclude this block from our main analysis.[10] This leaves us with 11 academic blocks: two in the 2018 academic year, and three in each of the 2019, 2020, and 2021 academic years.

Task types and relevance Figure 2 gives the tasks that each team is required to complete per block. We identify four distinct types of tasks. Writing tasks include proposals writing, writing up smaller components of a research document, and writing up an entire research paper. Data tasks include identifying and using existing datasets, running surveys, and cleaning and analyzing data. Feedback involve giving feedback to other teams on their work, predominately on their writing work. Presentation tasks involves presenting their work in class. Each block consists of four graded research tasks.

How similar are these tasks to those done in everyday jobs, and which occupations have the largest overlap? We explore this question in Appendix A.1 by comparing the tasks done by the student teams with a taxonomy of US occupations and their required tasks on the job (ONET). We show that our tasks, especially writing tasks, have a large overlap with many white-collar type occupations.

2.2. Data Our data includes the random team allocation for the 11 blocks mentioned above. We match this with the grades achieved by the teams for the tasks within a block, and their overall grade for the block. Finally, we also include a wide range of student characteristics from the university's administrative database: gender, age, high school GPA (for Dutch students only), ethnicity (for Dutch students only), nationality, information on the educational attainment of students' parents, and the grades achieved by each students in all courses taken at the university.

Our aim is to measure how task performance differs by the gender composition of a team. If ability to perform a tasks differs by an individual's gender, then any gender composition effect may simply reflect these ability differences, rather than a pure gender effect.

[9] The Covid-19 pandemic and the associated lockdowns affected part of our sample period. In Table B.4 we show that our baseline results are unchanged during these Covid-19 blocks.

[10] We make use of this block in Section 4.3 when investigating groups of larger sizes.

Therefore, our baseline analysis includes ability controls for each individual in the team. Our main measure of individual ability is the Task Ability Measure. As described above, this is the average of the grades achieved by students on the *individual* tasks in block 2. This variable measures an individual's aptitude on the tasks they will subsequently perform in groups. Figure 2 gives the exact tasks involved in this measure. While this is our preferred measure of ability, given its similarity with the team-based tasks, we also use make use of both high school and university GPA in previous courses as ability measures in robustness checks. The latter is calculated based on the grades achieved in the first two blocks of students' first year. High school GPA is arguably a more comprehensive measure of general aptitude, given that it is the result of a full year of both course and exam results. However, it is available only for Dutch students. Appendix Figure B.3 shows the distribution of each of the three ability measures by gender.

Table 1 presents the summary statistics for each variable in our sample. Our data includes approximately 2,600 unique students over the 4 cohorts. Across the 11 blocks these students form approximately 2,700 work teams.[11] Around 45% of these teams are all-male teams, 45% are mixed-teams, and 10% are all-female teams. The various writing, data, feedback, and presentation tasks these research groups perform lead to 10,500 team-task grade observations.

2.3. Randomization tests As described above, students were randomized into research teams as the beginning of each block. Randomization of teams is vital to identifying any potential gender composition effect. The presence of sorting, or endogenous group formation, would make attributing differences in performance to gender impossible, as individuals favouring certain gender combinations may also differ in other (unobserved) dimensions.

We formally test for the successful randomization of students into groups using the randomization tests derived by Jochmans (2023). This test improves on previous randomization tests (e.g. those used in Sacerdote (2001) and Guryan et al. (2009)) by improving power and avoiding the so-called exclusion bias. Intuitively, the procedure tests the degree to which some characteristic of an individual (say gender) is systematically related to the characteristic of their assigned partner. In the case of random assignment, no systematic correla-

[11] Some students do not appear in all three blocks due to drop-out or missing data. Subsequent dropout is not influenced by the gender composition of a student's team.

tion should be present. We perform tests using the following characteristics: gender, ability (continuous), high ability (top 25% of ability distribution), low ability (bottom 25%), native Dutch, and non-Dutch. The results of these 6 randomization tests are shown in Appendix Table B.1, which shows that across all characteristics, there are no significant correlations between a student's characteristic and that of their partner. We therefore conclude that the randomization of teams was successful.

3 Methods

Our approach involves regressing the standardized $Grade_{agt}$ achieved in assignment a by group g in block b and classroom c on a set of dummy variables describing the gender composition of the group – $Mixed_g$ and $AllWomen_g$ – and both classroom-times-block ($ClassBlock$) and task fixed effects (A):

$$Grade_{agcb} = \beta_0 + \beta_1 Mixed_g + \beta_2 AllWomen_g + A_a + ClassBlock_{cb} + \epsilon_{agcb} \tag{1}$$

Coefficients β_1 and β_2 then give the difference in standardized grades for mixed-gender research teams and all-women teams, respectively, when compared to all-male teams. The fixed effects absorb any task- or classroom-block-level difference in grades. We cluster the error term at classroom level.

3.1. Specifications with ability controls Findings of significant gender compositions effects in Equation (1) could potentially be driven by ability differences between men and women in our sample.[12] We therefore control for ability differences between teams in some specifications. Doing so reveals the degree to which differences in grades by gender composition can be attributed to differences in gender composition, rather than differences in individual ability between man and female students in our sample. As discussed in Section 2, our measure of ability is the Task Ability Measure; the average grade of the individual tasks achieved by a student in block 2.[13] We use two approaches to control for ability.

[12] Appendix Figure B.3 shows some evidence of women having higher average ability, depending on the measure used.

[13] In robustness checks, we also present results using alternative measures of individual ability: highschool GPA and university GPA.

Best & worse ability controls Our first approach is to identify the “best” and “worst” member of each pair in terms of ability, based on our ability measure. We then compute $AbilityQuintile_g^{Best}$ ($AbilityQuintile_g^{Worst}$), a variable containing the quintile of the ability of the best (worst) member of the pair in group g of classroom t in block b.[14] We add dummies for each ability quintile of the best and worst member, resulting in the following modified version of Equation (1):

$$\begin{aligned} Grade_{agcb} = \beta_0 + \beta_1 Mixed_g + \beta_2 AllWomen_g + \\ \sum_{q=1}^{4} \theta_{1q} \mathbb{1}\left(AbilityQuintile_g^{Best} = q\right) + \sum_{q=1}^{4} \theta_{2q} \mathbb{1}\left(AbilityQuintile_g^{Worst} = q\right) \\ + A_a + ClassBlock_{cb} + \epsilon_{agcb} \end{aligned} \quad (2)$$

Ability combination controls Equation (2) controls separately for the individual researcher ability of both members of the group. However, it may be that interactions occur between the ability of the two group members; the effect of being in the top quintile of individual ability on $Grade_{agt}$ may depend on the ability quintile of the other member of the group. In total, there are 15 possible combinations of ability quintile categories for the best and worst member of the research team. Our second specification ensures that any potential ability interactions are controlled for by including dummies for each of these 15 categories:

$$\begin{aligned} Grade_{agcb} = \beta_0 + \beta_1 Mixed_g + \beta_2 AllWomen_g + \\ \sum_{q=1}^{5} \sum_{p=1}^{q} \theta_{q,p} \mathbb{1}\left(AbilityQuintile_g^{Best} = q, AbilityQuintile_g^{Worst} = p\right) \\ + A_a + ClassBlock_{cb} + \epsilon_{agcb} \end{aligned} \quad (3)$$

As well as Equation (1), our baseline results present estimates of β_1 and β_2 from Equation (2) and Equation (3). Due to the addition of these ability controls, any remaining differences in grade by gender composition cannot be attributed to underlying difference in the individual academic ability of the researchers.[15]

[14] The quintiles here and elsewhere in the paper are calculated by cohort and program.

[15] One may still worry about underlying gender difference in the individual academic ability if our measure contains too much noise. We present evidence to the contrary in Appendix A.2 by showing that our preferred ability controls are able to control for all differences in individual-based grades between men and women for other courses.

4 Results

We begin by simply plotting the density of (standardized) performance measure for teams with each of the three gender compositions. These distributions are shown in Figure 3 across all tasks, and separately by the Writing, Feedback, Presentation and Data tasks. The dashed lines show the average grade by gender composition. This figure reveals small systematic differences in task performance by gender composition; on average, all-male teams appear to do worse, mixed teams better than all-male teams, and all-female ones perform best.

4.1. Regression approach Table 2 presents our baseline results. Column (1) shows estimation results of running Equation (1) on the 10,600 group-task grade observations. The estimates for β_1 and β_2 are both large and highly significant. They imply that research pairs comprised of two women (one woman and one man), achieve grades 29% (20%) of a standard deviation higher than those comprised of two men.

How much of the differences by gender composition in column (1) can be explain by by differences in individual's task ability per gender? In columns (2) and (3) of Table 2 we present results from estimates of Equation (2) and Equation (3), respectively. The addition of ability controls reduces the magnitude of the estimated β_1 and β_2 coefficients, although both remain large in magnitude and statistically significant. These results imply that research pairs comprised of two women (one woman and one man), achieve grades 20% (15%) of a standard deviation higher than those comprised on two men.[16]

4.2. Results by task There are large difference in performance by gender composition across all tasks. Are these differences also present within the various types of tasks? To investigate this we estimate Equation (1), Equation (2), and Equation (3) on subsamples of each task type: data tasks, feedback tasks, presentation tasks, and writing tasks.

The results reveal that the pattern of all-male teams being outperformed by mixed and all-female teams is present across all types of task (Table 3). The effect of gender composition differ somewhat by task type; the largest differences are observed in presentation tasks, where mixed (all women) teams outperform all men teams by 20% (27%) in specifications

[16] In Appendix Table B.4 we interact the gender composition dummies with an indicator for blocks affected by Covid-19 lockdowns. We find statistically insignificant and small interaction terms, indicating that our results are not drive by periods affected or not affected by Covid-19.

including controls for the ability combinations. Smaller differences are observed in the data (14% and 15% for mixed and all-women teams, respectively) and feedback (11% and 18%) type tasks. However, across all specifications, the point estimates for each coefficient show that all men teams perform the worst, followed by mixed teams, with all women teams performing the best.

4.3. Larger groups The above results examine the gender composition effect for pairs. However, workers in firms and other contexts are obviously not restricted to pairs. Are the findings above present in other group sizes?

As described in Section 2, block 4 of the 2018 cohort was excluded from the main analysis as in this block teams were randomized into sizes of 4, 5, and 6, rather than 2. We also drop groups of size 3 from our main analysis that were formed in the remaining blocks when classrooms contain an odd number of students. In order to investigate whether the gender composition pattern above also exists in larger groups, we focus our analysis on this sample of groups of size 3 and above. In total, there are approximately 400 teams larger than 2, and 1,500 task observations of these teams. The average size of these groups is 3.6, with the average proportion women being 0.33. Summary statistics for these groups and task observations are given in Appendix Table B.3.

Figure 4 shows a binscatter plot of task performance and the proportion of females in these larger groups. We overlay the results of a non-parametric local-linear regression of the proportion of females on task performance, computing confidence intervals via a bootstrap procedure.[17] These non-parametic methods suggest a positive effect of the proportion of females in the group, except approximately between a proportion of 0.3 and 0.7, where the function is approximately flat. Keeping in mind the low sample size and limited support, we take this as suggestive evidence of a gender composition effect also in the larger groups.

We also investigate the gender composition effect in larger groups using regression specifications similar to those used above. Our first approach is to estimate Equation (1), Equation (2), and Equation (3) on the sample of larger groups. However, because the $MixedTeam$ dummy encompasses a wide range of teams with different proportions of female members, we also estimate specifications where $MixedTeam$ and $AllWomen$ are replaced with dum-

[17] A local-linear and local-constant kernel regression is used, using an Epanechnikov kernel function. Confidence intervals generated via 1,000 bootstraps.

mies representing quartiles of the proportion of women in the groups, where the 1^{st} quartile serves as the reference category. We also adjust our ability controls in order to account for the larger and variable team size using two different approaches: 1) control for the average ability of the group members; 2) control separately for ability quintile of the best and the worst member of the group (ignoring all other group members), as in Equation (2). In all specifications we add controls for group size. In order to maximize the number of observations in each regression, we don't restrict our sample to be the same across all specifications.[18]

The results of these regressions are shown in Table 4. Columns 1, 2, 3 give the result of Equation (1), Equation (2), and Equation (3), while columns 4, 5, and 6 repeat these specifications with the addition of the proportion of women quartile dummies. Columns 1 and 4 reveal no statistically significant effect of gender composition on team performance when ability controls are excluded. However, the point estimates suggest a positive effect on performance of more women in the team. The addition of average ability in column 5, and best/worst ability quintile controls in columns 3 and 6, however, suggest that groups with all women, or those in the 4^{th} quintile of the proportion of women in the sample, significantly outperform groups with no women, or those in the 1^{st} quintile. These findings are suggestive of a similar pattern to those found in Figure 4, whereby the performance gains are only experienced in groups in which the proportion of women reach a adequate proportion. While the results of this analysis in this subsection should be regarded as suggestive only due to the small sample size and low support, we nevertheless take these results as evidence that the gender composition patterns observed in pairs also appear to carry over – to some degree – to larger groups of sizes 3, 4 and 5.

4.4. Results by average group ability To what degree is the gender composition effect concentrated amongst groups of certain ability? To explore this question we divide all observations by the quintile of the team's average individual ability, and run Equation (3) for each of these subgroups. These reveal the degree to which the gender composition effect is present across the ability distribution.

Table 7 display the gender composition results for these subsamples. The effect appears

[18] Some groups have missing ability data. We do not drop these groups from the regressions excluding ability controls.

to be most concentrated in groups between the 2^{nd} and 4^{th} quintiles. The results for groups in the 1^{st} quintile are of similar magnitude, although the performance difference between all women and all men teams is not statistically significant. The results for groups in the 5^{th} quintile are positive but of much smaller magnitude and not statistically significant.

These results suggest that the gender composition effect is most concentrated in groups lying within the middle of the ability distribution. However, the insignificant results for highly-able groups may be due to power issues, and the fact that grade variation in these groups are limited due to the truncated nature of our performance measures.

5 Robustness

5.1. Extended Ability Regressions Our baseline results presented in Table 2 make use of students' individual task grades in block 2 of the academic skills course as an aptitude measure to control for potential ability differences between men and women. Controlling for these potential differences is important as their presence would lead to a gender composition effect even in the absence of any group-level dynamics or processes; they could simply reflect the fact that female students are better than male students in our sample.

In this section, we explore the robustness of our results to the use of alternative ability controls; namely highschool GPA and university GPA in previous courses.[19] Appendix A.2 gives corroborating evidence that the Task Ability Measure is the most suitable measure for capturing possible gender differences in ability between students. Here we repeat our baseline results with combinations of alternative ability measures and our preferred measure to test robustness.

Columns 1-6 of Table 5 show how the estimates of β_1 and β_2 change as the best and worst member quintiles are added for different combinations of the Task Ability Measure, university GPA, and highschool GPA.[20] For instance, column 6 shows the results of a specification that includes quintile dummies for the best and worst team member for all three ability measures, resulting in 15 separate dummies controlling for the ability composition of the group. While the point estimates vary somewhat, the results remain qualitatively similar

[19] See Section 2 for a precise definition of these variables.

[20] The sample size in Columns 4-5 and 10-12 are reduced as highschool GPA is only avaliable for Dutch students.

to our baseline results.

Columns 7-12 of Table 5 repeat this exercise using the combination ability control method (Equation (3)) with different combinations of the ability measures. The most demanding is shown in column 12, which includes 43 dummy variables controlling for the ability combinations of the best and worst member according the three ability controls. Again, these various combinations do not change our baseline results. We take these results as evidence that our results are not likely to be driven by unobserved task ability differences between women and men in our sample.

5.2. Other characteristics An alternative explanation for our gender composition finding is that the effect is driven not by gender, but some other characteristic that happens to correlate with gender in our sample. For instance, if most women in our sample are non-Dutch, it may be some nationality compositional effect driving our results, rather than gender.

While we are not able to rule out all potential unobserved correlates of gender, our student data allows us to control for many other important student characteristics. Namely, we have information on student's SES (measured by parental university attendance), nationality (Dutch and non-Dutch), and whether a Dutch student has a so-called immigration background. Should the gender composition effect remain with the inclusion of these controls, this would provide further evidence that our findings indeed capture a gender effect.

We repeat our Equation (3), the specification with the most demanding ability controls, while also controlling for the number of group members who possess the following characteristics: at least one parent attending university, have a non-Dutch nationality, or have an immigration background. For Dutch students, we have information on whether they have a "minority" immigrant background, a non-minority immigrant background, or no immigrant background.[21] Specifically, we control for each possible combination of Dutch nationality and Dutch ethnicity within a team.

Table 6 gives the results. Column 1 repeats our baseline specification for comparison purposes. Columns 2-4 add the above characteristics as controls separately, while column

[21] In The Netherlands, a "minority" immigrant background includes all students who are either first or second generation migrants with ties to any country in Africa, South America or Asia (excl. Indonesia and Japan) or from Turkey. In practice, more than 50% of the students in this group have Moroccan, Turkish, or Dutch Caribbean ancestry. All those with non-minority immigrant backgrounds are first or second generation migrants with ties to other countries.

5 controls all additional characteristics simultaneously. Across all regressions our baseline results of the gender composition effect remain virtually unchanged. We take this as evidence that suggests the gender effect is not driven by other demographic characteristics in our sample.

6 Why Do Teams With More Women Do Better?

The results above show that groups with more women tend to outperform those with men. This cannot be explained by task ability differences between the men and women in our sample, nor by other characteristics of these students that may vary by gender such as SES, nationality, or ethnicity. What then might be driving these differences?

To shed light on this question, we look to existing literature in economics, management, and small group research, and divide the potential explanations for the gender composition effect into five broad (though non-exhaustive and partly overlapping) categories: (1) *Team Work Preferences, Atmosphere & Friendship*, (2) *Contributions, Effort & Motivation*, (3) *Conflict, Unity & Trust*, (4) *Feedback, Monitoring & Decision-making*, and (5) *Leadership Style*.

Below we describe and motivate each category via supporting literature, and subsequently test for these explanations in our data.

Team Work Preferences, Atmosphere & Friendship Any gender differences skills and preferences for group work may lead to, broadly speaking, a better group atmosphere, levels of civility, and thus better outcomes in groups with more women. Previous research has proposed that so-called "interpersonal sensitivity" - the propensity to treat teammates with care and respect – is higher amongst women (Kennedy, 2003), and that men themselves may exhibit more of this trait when in mixed-gender teams (Williams and Polman, 2015). Several studies have also tried to quantify so-called "social-skills"; non-cognitive skills that allow an individual to boost team performance (Woolley et al., 2010; Weidmann and Deming, 2020), and it may be that these social skills are more concentrated amongst women than men. These gender differences may also partly explain the on-average larger preferences for cooperation over competition for women compared to men (Croson and Gneezy, 2009).

Contributions, Effort & Motivation A straightforward explanation of the better outcomes in groups with more women may simply reflect that woman dedicated more time and effort on the tasks than men. In a laboratory experiment using student pairs, Babcock et al. (2017) find that women in mixed-gender teams are far more likely than men to volunteer to perform menial and costly work. Showing this volunteering gap disappears when teams are single-sex, they argue that this pattern is driven the the belief that women will eventually volunteer for menial and costly tasks in mixed teams, rather than differences in preferences by gender. Further laboratory evidence suggests that men tend to free-ride in teams more than women (Cadsby and Maynes, 1998), and that women are more likely to cooperate in public good games (Furtner et al., 2021). If women tend to contribute more in group settings, this would lead to differences in performance at the group level by gender composition.

Conflict, Unity & Trust Group conflict has been identified as an important component driving group outcomes, although there remains some debate about the direction of its effect and about the importance of different types of conflict (Jehn, 1995). The relationship between gender composition of teams and conflict has also long been a topic of interest within the management literature, with early papers generally a positive correlation between gender diversity and conflict levels (Pelled, 1996; Hope Pelled, 1996; Jehn, 1995). However, evidence from board rooms (Nielsen and Huse, 2010) and legislative groups (Rosenthal, 2000) suggests a positive effect of the presence of women in such groups on performance through decreased conflict levels.

Feedback, Monitoring & Decision-making Differences in internal organization and processes of groups, depending on gender composition, may explain the gender composition effect. We consider three different possible dimensions of team processes: decision making, mutual monitoring, and feedback processes.

Decision-making processes may differ between groups, leading to differences in group performances. Research from the lab (Hannagan and Larimer, 2010), student groups (Fenwick and Neal, 2001), and political legislators (Rosenthal, 2000) suggests that groups with more women tend to employ more cooperative strategies when making decisions. Mutual monitoring – the practice of group members monitoring the effort and work of their teammates – has been studied as means of addressing incentive and skirting problems in teams

(Carpenter et al., 2006), and it's presence in board rooms has been shown to correlate with a firm's future value (Li, 2014). Also in the context of board rooms, Adams and Ferreira (2009) find that boards with more women allocate more effort to monitoring practices. Management literature points to feedback within groups as an important determinant of group performance, with both experimental (Barr and Conlon, 1994) and theoretical work (Robinson and Weldon, 1993) pointing to group feedback playing an important role in group performance. Other literature argues that female-majority groups may be more receptive to feedback that male-majority groups (Karakowsky and Miller, 2002).

Leadership Style Leadership structures may differ between groups depending on gender composition. If leadership influences performance, this in turn may lead to a gender composition effect. Research has shown that women are less likely to appear in leadership roles within groups with more men, which may stem from the fact that women tend to get less support Born et al. (2020), credit (Sarsons et al., 2021), and more menial tasks (Babcock et al., 2017) in such groups. Moreover, some evidence points to different average leadership styles between men and women when they are leaders, with women tending to adopt more democratic leadership styles in contrast to more autocratic styles (Eagly and Johnson, 1990).

6.1. Self-Reflection Exercise To investigate these possible explanations for the gender effect we use data from a comprehensive self-reflection exercise introduced in blocks 3 and 5 of the 2021 cohort.[22] This exercise was designed to help students reflect on their team work experience in that particular block, and prompts students with questions relating to the explanations above.[23]

We use the self-reflection exercise data to investigate explanations for the gender-composition effect on quality of team output. However, we note that the categories above may both be mechanisms and outcomes of the gender composition effect, and our data does not allow us to disentangle the two. We therefore interpret the proceeding results as a speculative exploration of potential explanations, rather than clear-cut evidence.

Summary statistics for the 22 outcomes resulting from the self-reflection exercise are given in Table 8, collected under the headings of the potential explanations above. In to-

[22] These blocks and the self-reflection exercise took place in 2022, after the end of Covid-19 related lockdowns.

[23] This exercise was completed by individual students, rather than at the team level.

tal, we record approximately 1,000 responses by students across blocks 3 and 5 of the 2021 cohort.[24] The outcomes include both directly elicited questions, as well as outcomes that are the result of combining multiple items through principal component analyses (PCA) with the aim of measuring a particular underlying construct. A full description of the self-reflection exercise, the elicited questions, and the construction of the PCAs is given in Appendix A.4.

Regression Approach for Teams Self-Reflection Data The results in Section 4 show that both women and men appear to benefit, performance-wise, from having a female partner compared to a male partner. Using the self-reflection exercise data, we therefore explore the effect of being allocated a female partner – rather than a male partner – on the outcomes in Table 8 for women and men.

In contrast to the group-performance outcomes, the self-reflection exercise is elicited at the level of the individual across blocks. This allows us to run individual-level regressions that permit the inclusion of individual fixed effects.[25] For each outcome in Table 8 we run regressions of the following form:

$$\begin{aligned} Outcome_{ijb} &= \gamma_0 + \gamma_1 FemaleTeammate_j + \gamma_2 FemaleTeammate_j \times Woman_i \\ &+ Tut_{ij} + Block_b + S_i + \sum_{p=1}^{5} \theta_p \mathbb{1}\Big(AbilityQuintile_j = p\Big) + \epsilon_{ijb} \end{aligned} \tag{4}$$

Where $Outcome_{ijb}$ refers to some outcome of the self-reflection exercise for respondent i, allocated partner j, in block b. Coefficient γ_1 ($\gamma_1 + \gamma_2$) then gives the average difference in $Outcome$ for a man (woman) allocated a female teammate compared to those allocated a male partner. Hence, we measure the within-student effect of being allocated a female partner, rather than a male partner. The specification also flexibly controls for the ability quintile of teammate j.

Findings From Self-Reflection Data Coefficient estimates of γ_1 (effect for men) and $\gamma_1 + \gamma_2$ (effect for women) from Equation (4) for each outcome of Table 8 are given in Figure 5. Those estimates significant at $\alpha = 0.10$ are dashed. The regression results underlying these plots

[24] Appendix A.5 shows that the qualitative pattern of our group performance results holds in this smaller sample.

[25] For comparison purposes, we also run analyses at the group level, in a similar fashion to Equation (3). The results of this analysis are shown in Appendix A.5.

are shown in Appendix Table B.11.

The effect of being allocated a female teammate has opposite signs for men and women for many outcomes. Women allocated a female working partner, rather than a male, report more familiarity, better group atmosphere, larger motivation for themselves and their partner, higher group contributions, less conflict, more unity, and less group hierarchy. On the other hand, despite the boost in performance, men allocated a female working partner tend to report less previous and current familiarity and worse group atmosphere. These results reveal that members of mixed teams appear to report "worse" experiences, largely driven by women in these teams. What can this tell us about the gender composition effect? Keeping in mind the low sample size and the self-reported nature of these outcomes, we speculate on some potential explanations below.

One explanation for these patterns of results is that women are more conscientious or diligent group members than men. Although we control for each individual's task ability – which should capture both cognitive and non-cognitive skills – these may not include individual's "group-work" ability. It may be that there exists a group conscientiousness or diligence factor that differs on average per gender which leads to improved group performances. This is consistent with several findings from the self-reflection exercise. Firstly, women allocated a male report significantly lower *Team Contribution* PCA scores. Second, women are more likely to report being the leader when matched with a man. Notably, while not significant, Figure 5 shows that both men and women allocated a female partner report higher values for the feedback, monitory, and decision-making principal components. Taken at face value, this suggests that teams with women are better functioning, which may in turn partly explain the boost in performance.

Such an explanation is also consistent with patterns from laboratory experiments, usually involving students. Woolley et al. (2010) document higher levels of a factor predicting success in group work in groups with a higher fraction of women. They attribute this finding to higher levels of "social-sensibility". Using data on student teams with randomly allocated leaders, De Paola et al. (2022) find that teams lead by women tend to outperform those headed by men. They speculate that traits like conscientiousness and readiness to collaborate leads to higher levels of group performance. Keck and Tang (2018) show that groups with at least one woman are more effective at sharing information with each other, possibly due to better interpersonal sensitivity, and that this leads to better-calibrated group decisions. This pattern is also found in observation data in similar contexts; the work of research teams in science has a strong overlap with the tasks studied in this paper. Based on publications in the medical science field, Yang et al. (2022) show that research papers that are more impactful and with citation patterns indicative of innovative research are more likely to come from research teams with more female team members. In economics, papers with more women co-authors tend to have a higher citation count upon publication and have higher readability scores (Hengel, 2020).

The self-reflection findings, especially the lower levels of reported team atmosphere, self-motivation, and unity by women in mixed teams, also square with previous literature on the allocation of menial tasks in teams. Using pairs of students, Babcock et al. (2017) show in a laboratory setting that women in mix-gender pairs are far more likely than men to volunteer to perform costly yet menial tasks, but that this gender volunteering gap disappears when teams are single-sex; men expect women to volunteer more in teams, and therefore contribute less to menial tasks in mix-gender groups. In our setting, the poorer group experiences reported by women in mixed groups may be the result of them being burdened with more menial, low-recognition tasks.

One explanation of both the performance and self-reflection results is therefore that the presence of women boosts team performance due their average higher levels of social-sensibility and conscientiousness, similar to the traits identified in Woolley et al. (2010); De Paola et al. (2022) and Keck and Tang (2018), while at the same time, the miss-match in diligence and burden of tasks in mixed teams – in a similar fashion to Babcock et al. (2017) – leads to worse reported team atmosphere, motivation, unity, and conflict levels in mixed teams, especially experienced by women.

7 Conclusion

Using data on randomly formed student teams performing tasks comparable to those in many while-collar occupations, this paper investigates how the gender composition of such teams influences their performance, using task grades as a performance measure. Using 10,600 task-grade observations, we document a substantial gender composition effect; mixed-gender (all-female) pairs outperform all-male pairs by 15% (20%) of a standard deviation. This gender composition effect is robust to the inclusion of many alternative measures of individual ability for each member of the team, and thus does not reflect differences in *individual* ability between men and women in our sample. The effect is also robust to controlling for other characteristics that may vary by gender in our sample, such as ethnicity. The gender composition performance gap exists in all task types (writing, feedback, data and presentation tasks), and also appears to be present in groups larger than two. These findings are in line with earlier research documenting a positive impact of women in self-selected teams (Woolley et al., 2010; De Paola et al., 2022; Keck and Tang, 2018; Fenwick and Neal, 2001; Hoogendoorn et al., 2013; Yang et al., 2022; Hengel, 2020; Hengel and Moon, 2023).

The data from a self-reflection exercise shed further light on possible mechanisms behind the performance findings. In contrast to the ranking of groups by performance, where mixed and all-women teams do better than all-male teams, the more subjective measures show that – along many dimensions – mixed-teams tended to report worse outcomes. This pattern fits a scenario where women's higher group-diligence, conscientiousness, or social-sensitivity (Woolley et al., 2010) are effective in boosting a team's performance, while the increased mismatch of these traits within the team, and an uneven burden of tasks, leads to a worse team atmosphere, motivation, and unity.

These results highlight the potential trade-offs between the subjective experiences of group members and their objective performance. While mixed-gender groups produce significantly better quality work than all-male groups, members of these groups reported the "worst" subjective outcomes on, for instance, group atmosphere. While the performance of such groups is higher, their long-term sustainability may be questionable.

Should these results on student data hold in other contexts, it appears that the increased gender diversity in traditionally male firms, boards, panels, and other work teams will lead to performance gains, as more women break the glass ceiling in these domains. At first glance, organisations could make significant gains simply by ensuring work teams include at least one woman. However, our results also highlight that policy makers and managers should be aware that performance gains may come at the cost of a larger burden of menial, costly, and potentially non-promotable tasks for women in these teams (Babcock et al., 2017), especially as other evidence suggests women are not given the same credit for team work as men (Sarsons et al., 2021). This uneven distribution of tasks may lead to more dysfunction along harder to measure dimensions, such as group atmosphere, and threaten the long-term sustainability of these teams.

References

Adams, R. B. and D. Ferreira (2009). Women in the boardroom and their impact on governance and performance. *Journal of financial economics 94*(2), 291–309.

Apesteguia, J., G. Azmat, and N. Iriberri (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science 58*(1), 78–93.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review 107*(3), 714–747.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017). Does the gender composition of scientific committees matter? *American Economic Review 107*(4), 1207–1238.

Bagues, M. F. and B. Esteve-Volart (2010). Can gender parity break the glass ceiling? evidence from a repeated randomized experiment. *The Review of Economic Studies 77*(4), 1301–1328.

Barr, S. H. and E. J. Conlon (1994). Effects of distribution of feedback in work groups. *Academy of Management Journal 37*(3), 641–655.

Bear, J. B. and A. W. Woolley (2011). The role of gender in team collaboration and performance. *Interdisciplinary science reviews 36*(2), 146–153.

Born, A., E. Ranehill, and A. Sandberg (2020). Gender and willingness to lead: Does the gender composition of teams matter? *The Review of Economics and Statistics*, 1–46.

Cadsby, C. B. and E. Maynes (1998). Gender and free riding in a threshold public goods game: Experimental evidence. *Journal of economic behavior & organization 34*(4), 603–620.

Carpenter, J. P., S. Bowles, and H. Gintis (2006). Mutual monitoring in teams: Theory and experimental evidence on the importance of reciprocity.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic literature 47*(2), 448–74.

De Paola, M., F. Gioia, and V. Scoppa (2022). Female leadership: Effectiveness and perception. *Journal of Economic Behavior & Organization 201*, 134–162.

Díaz-García, C., A. González-Moreno, and F. Jose Sáez-Martínez (2013). Gender diversity within r&d teams: Its impact on radicalness of innovation. *Innovation 15*(2), 149–160.

Eagly, A. H. and B. T. Johnson (1990). Gender and leadership style: A meta-analysis. *Psychological bulletin 108*(2), 233.

Feld, J., N. Salamanca, and D. S. Hamermesh (2016). Endophilia or exophobia: Beyond discrimination. *The Economic Journal 126*(594), 1503–1527.

Fenwick, G. D. and D. J. Neal (2001). Effect of gender composition on group performance. *Gender, Work & Organization 8*(2), 205–225.

Furtner, N. C., M. G. Kocher, P. Martinsson, D. Matzat, and C. Wollbrant (2021). Gender and cooperative preferences. *Journal of Economic Behavior & Organization 181*, 39–48.

Goldin, C. (2006). The quiet revolution that transformed women's employment, education, and family. *American economic review 96*(2), 1–21.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American economic review 104*(4), 1091–1119.

Green, C. P. and S. Homroy (2018). Female directors, board committees and firm performance. *European Economic Review 102*, 19–38.

Guryan, J., K. Kroft, and M. J. Notowidigdo (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics 1*(4), 34–68.

Hannagan, R. J. and C. W. Larimer (2010). Does gender composition affect group decision outcomes? evidence from a laboratory experiment. *Political Behavior 32*, 51–67.

Hengel, E. (2020). Publishing while female: Are women held to higher standards? evidence from peer review.

Hengel, E. and E. Moon (2023). Gender and equality at top economics journals.

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

Hope Pelled, L. (1996). Relational demography and perceptions of group conflict and performance: A field investigation. *International Journal of Conflict Management 7*(3), 230–246.

Hughes, M. M., P. Paxton, and M. L. Krook (2017). Gender quotas for legislatures and corporate boards. *Annual Review of Sociology 43*, 331–352.

Jehn, K. A. (1995). A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative science quarterly*, 256–282.

Jehn, K. A., G. B. Northcraft, and M. A. Neale (1999). Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly 44*(4), 741–763.

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

Karakowsky, L. and D. Miller (2002). Teams that listen and teams that do not: exploring the role of gender in group responsiveness to negative feeback. *Team Performance Management: An International Journal 8*(7/8), 146–156.

Karpowitz, C., S. D. O'Connell, J. Preece, and O. Stoddard (2023). Strength in numbers? gender composition, leadership, and women's influence in teams. *Journal of Political Economy 0*(ja), null.

Keck, S. and W. Tang (2018). Gender composition and group confidence judgment: The perils of all-male groups. *Management Science 64*(12), 5877–5898.

Kennedy, C. (2003). Gender differences in committee decision-making: Process and outputs in an experimental setting. *Women & Politics 25*(3), 27–45.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of public Economics 92*(10-11), 2083–2105.

Lazear, E. P. and K. L. Shaw (2007). Personnel economics: The economist's view of human resources. *Journal of economic perspectives 21*(4), 91–114.

Li, Z. F. (2014). Mutual monitoring and corporate governance. *Journal of Banking & Finance 45*, 255–269.

Nielsen, S. and M. Huse (2010). The contribution of women on boards of directors: Going beyond the surface. *Corporate governance: An international review 18*(2), 136–148.

Pelled, L. H. (1996). Demographic diversity, conflict, and work group outcomes: An intervening process theory. *Organization science 7*(6), 615–631.

Robinson, S. and E. Weldon (1993). Feedback seeking in groups: A theoretical perspective. *British Journal of Social Psychology 32*(1), 71–86.

Rosenthal, C. S. (2000). Gender styles in state legislative committees: Raising their voices in resolving conflict. *Women & Politics 21*(2), 21–45.

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

Sarsons, H., K. Gërxhani, E. Reuben, and A. Schram (2021). Gender differences in recognition for group work. *Journal of Political Economy 129*(1), 101–147.

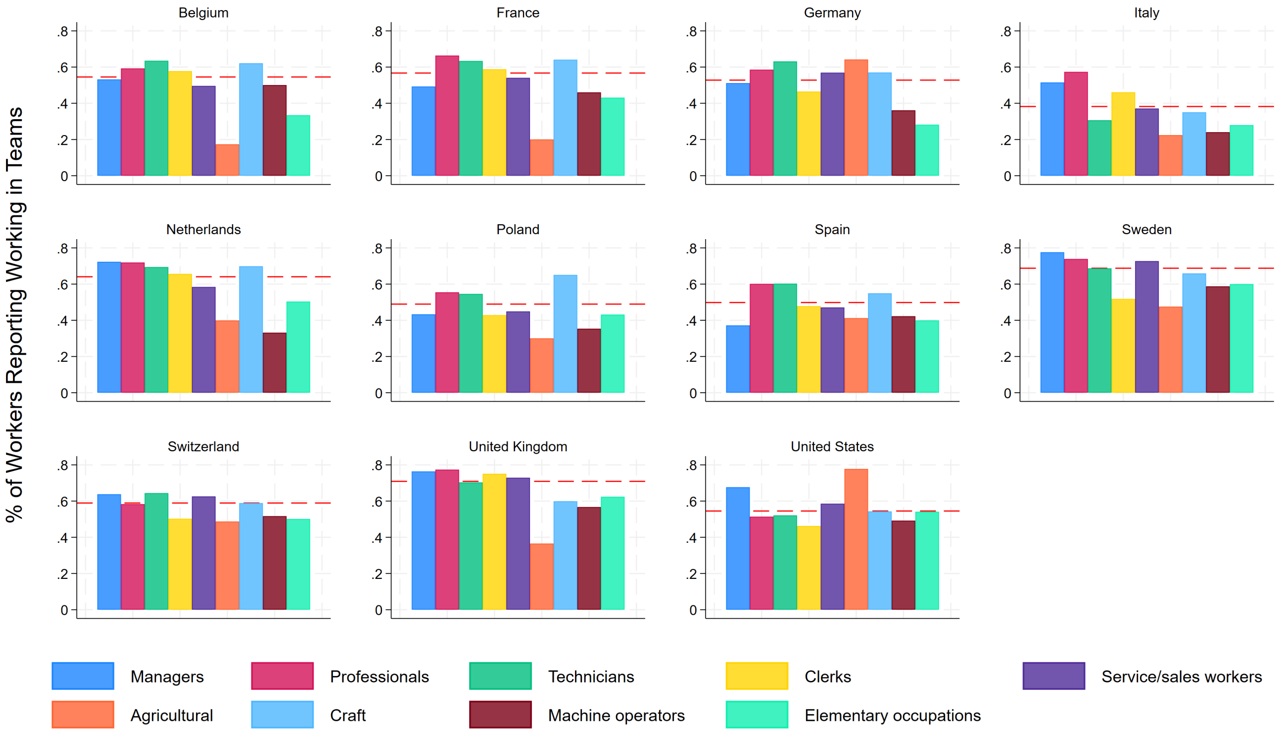
Weidmann, B. and D. J. Deming (2020). Team players: how social skills improve group performance. Technical report, National Bureau of Economic Research.

Williams, M. and E. Polman (2015). Is it me or her? how gender composition evokes interpersonally sensitive behavior on collaborative cross-boundary projects. *Organization Science 26*(2), 334–355.

Woolley, A. W., C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone (2010). Evidence for a collective intelligence factor in the performance of human groups. *science 330*(6004), 686–688.

Yang, Y., T. Y. Tian, T. K. Woodruff, B. F. Jones, and B. Uzzi (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences 119*(36), e2200841119.

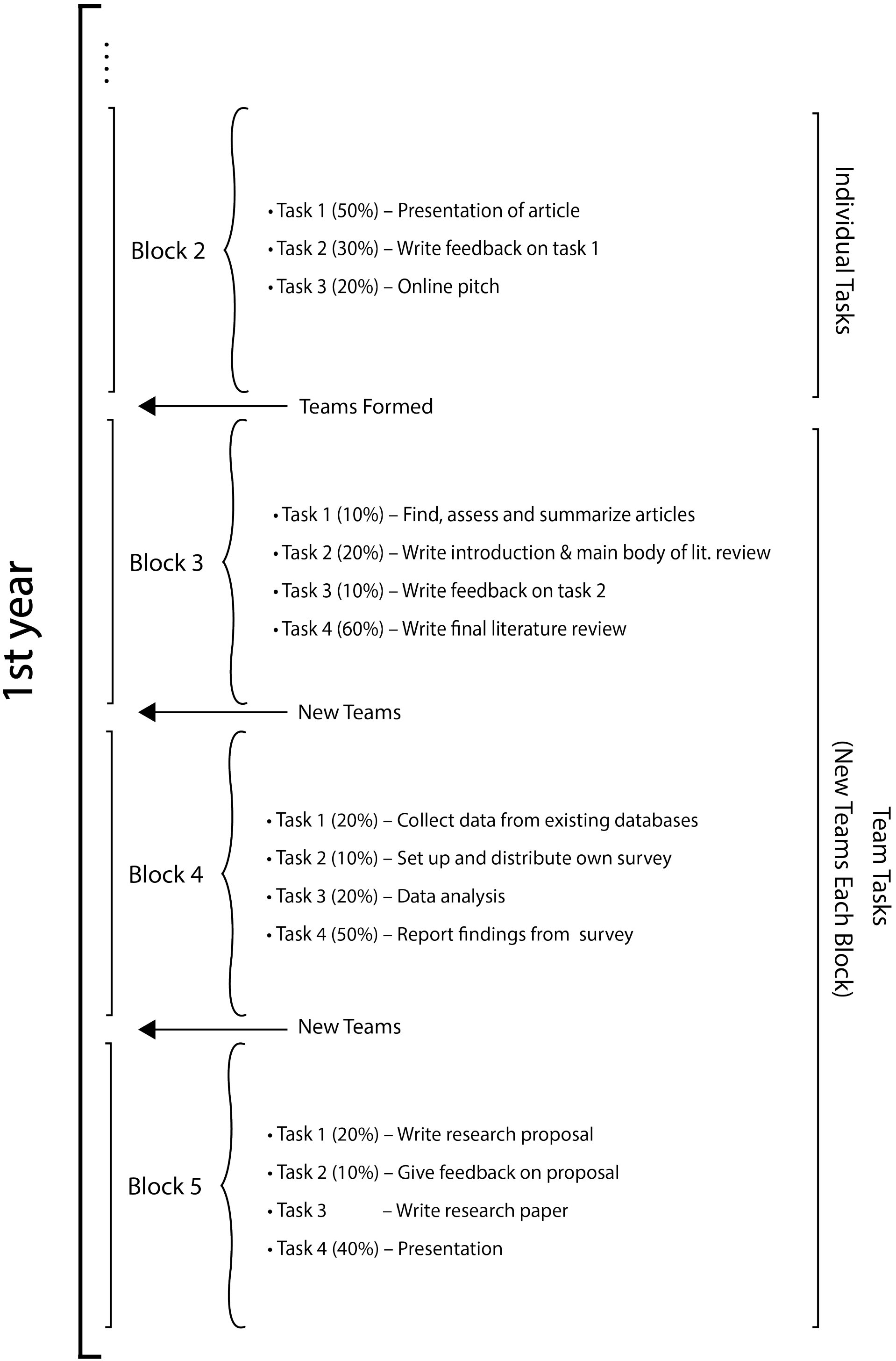
Figure 1: Percentage of Workers in Teams



Notes:

1. Figure shows the percentage of workers who report working in teams in 10 European countries and the US. The dashed red line shows the average prevalence of reported employee teamwork per country.
2. Data comes from the 2015 wave of the European Working Conditions Survey (EWCS) for Europe and from the 2018 wave of the General Social Survey (GSS). The relevant EWCS question asks: *"Do you work in a group or team that has common tasks and can plan its work?"*. The GSS question asks: *"In your job, do you normally work as part of a team, or do you mostly work on your own?"*.
3. All statistics are calculated using representative survey weights.

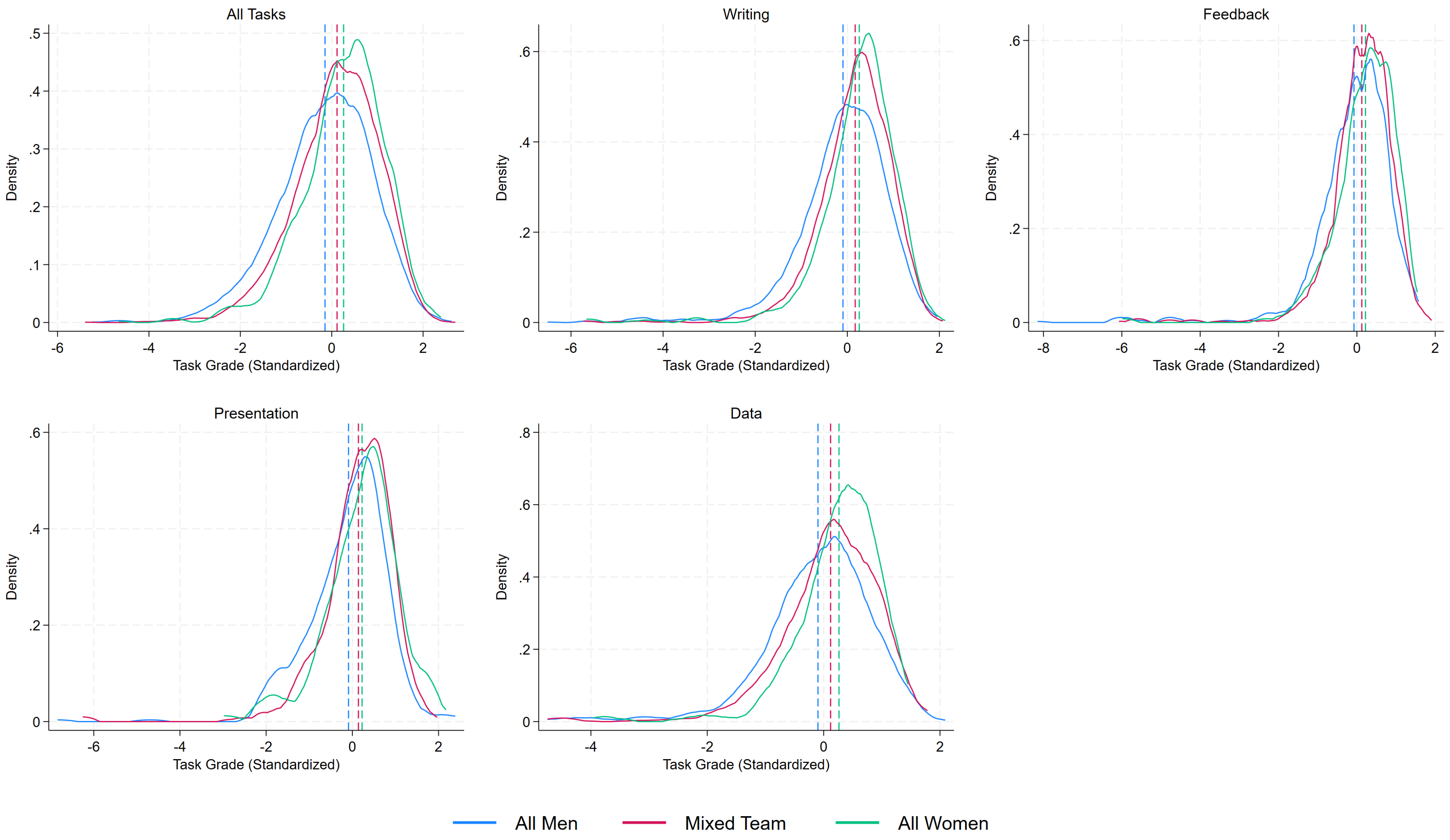
Figure 2: Structure of Data Collection and Outcomes


1st year
Block 2
• Task 1 (50%) – Presentation of article
• Task 2 (30%) – Write feedback on task 1
• Task 3 (20%) – Online pitch
Individual Tasks
Teams Formed
Block 3
• Task 1 (10%) – Find, assess and summarize articles
• Task 2 (20%) – Write introduction & main body of lit. review
• Task 3 (10%) – Write feedback on task 2
• Task 4 (60%) – Write final literature review
New Teams
Block 4
• Task 1 (20%) – Collect data from existing databases
• Task 2 (10%) – Set up and distribute own survey
• Task 3 (20%) – Data analysis
• Task 4 (50%) – Report findings from survey
New Teams
Block 5
• Task 1 (20%) – Write research proposal
• Task 2 (10%) – Give feedback on proposal
• Task 3 – Write research paper
• Task 4 (40%) – Presentation
Team Tasks
(New Teams Each Block)

Notes:

1. Figure shows the structure of the course from which the task data come from.

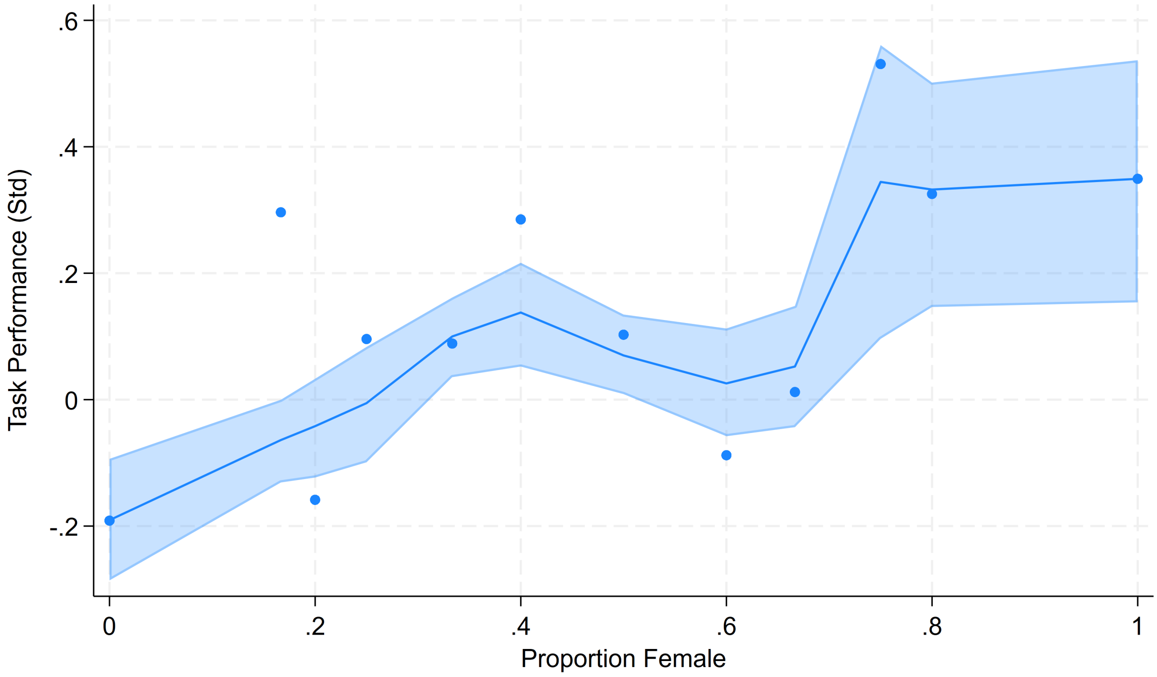
Figure 3: Task Histograms



Notes:

1. Figures show histograms of the standardized grades achieved in tasks per group type (All Men, Mixed Team, All Women). Dashed lines show averages per group type.
2. Histograms show distribution across all tasks types, and per task type.

Figure 4: Results for Larger Groups



Notes:

1. Figure shows binscatter of the proportion of female team members and standardized task grade for groups larger than 3.
2. Line shows the results of a local-linear and local-constant kernel regression, using an Epanechnikov kernel function. Confidence intervals generated via 1,000 bootstraps.

Table 1: Descriptive Statistics

	Mean	SD	Observations
		Student Data	
Number of students			2,583
Female	0.317	(0.465)	2,583
Task Ability Measure	74.05	(18.85)	2,583
High school GPA	6.920	(0.600)	1,468
University GPA	6.798	(1.000)	2,583
Non-Dutch	0.408	(0.491)	2,583
Native Dutch	0.450	(0.498)	2,583
Immigrant Dutch (Non-West)	0.102	(0.303)	2,583
Immigrant Dutch (West)	0.040	(0.196)	2,583
Both parents university	0.489	(0.500)	2,205
		Team Data	
Number of teams			2,710
Number of teams 2018			534
Number of teams 2019			805
Number of teams 2020			898
Number of teams 2021			473
All men	0.4590	(0.498)	2,710
Mixed	0.4390	(0.496)	2,710
All women	0.1030	(0.303)	2,710
		Task Data	
Average task grade	73.08	(14.28)	10,675
Average task grade Writing	71.92	(13.68)	5,577
Average task grade Data	67.38	(14.44)	2,177
Average task grade Presentation	76.14	(10.35)	940
Average task grade Feedback	81.19	(13.52)	1,981

1. Table shows the summary statistics of the student, team, and task data.
2. Student data comes from the internal administrative data of the university.
3. Team and task data come from the course outlined in Section 2.

Table 2: Baseline Results

	(1)	(2)	(3)
	Task Results (Std)		
Mixed Team	0.194***	0.147***	0.157***
	(0.0309)	(0.0291)	(0.0295)
All Women	0.292***	0.211***	0.216***
	(0.0536)	(0.0531)	(0.0531)
Ability Controls			
Best/Worst Ability	No	Yes	No
Ability Combinations	No	No	Yes
Mixed Team=All Women			
F-statistic	3.989	1.817	1.495
p-value	0.0468	0.179	0.222
Observations	10,675	10,675	10,675

1. Standard errors in parentheses, clustered on the classroom group level.

2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3. Table shows the baseline results of estimating Equation (1), Equation (2), Equation (3) on the team-task data in Table 1.

Table 3: Results Per Task

	(1)	(2)	(3)	(4)	(5)	(6)
		Data Tasks			Feedback Tasks	
Mixed Team	0.201^{***}	0.135^{**}	0.142^{***}	0.141^{***}	0.105^{**}	0.116^{***}
	(0.0556)	(0.0537)	(0.0539)	(0.0425)	(0.0429)	(0.0432)
All Women	0.242^{***}	0.141	0.152^{*}	0.245^{***}	0.183^{**}	0.175^{**}
	(0.0892)	(0.0864)	(0.0882)	(0.0729)	(0.0754)	(0.0754)
Ability Controls						
Best/Worst Ability	No	Yes	No	No	Yes	No
Ability Combinations	No	No	Yes	No	No	Yes
Observations	2,177	2,177	2,177	1,931	1,931	1,931
	(1)	(2)	(3)	(4)	(5)	(6)
		Presentation Tasks			Writing Tasks	
Mixed Team	0.242^{***}	0.206^{***}	0.215^{***}	0.202^{***}	0.154^{***}	0.165^{***}
	(0.0664)	(0.0674)	(0.0670)	(0.0366)	(0.0353)	(0.0361)
All Women	0.369^{***}	0.276^{**}	0.277^{**}	0.315^{***}	0.234^{***}	0.239^{***}
	(0.111)	(0.116)	(0.116)	(0.0637)	(0.0640)	(0.0636)
Ability Controls						
Best/Worst Ability	No	Yes	No	No	Yes	No
Ability Combinations	No	No	Yes	No	No	Yes
Observations	918	918	918	5,537	5,537	5,537

1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows the results of estimating Equation (1), Equation (2), Equation (3) on each type of task.

Table 4: Larger Groups

	(1)	(2)	(3)	(4)	(5)	(6)
Mixed Team	0.166	0.132	0.179			
	(0.113)	(0.114)	(0.134)			
All Women	0.356	0.358	0.888***			
	(0.287)	(0.279)	(0.244)			
2^{nd} Quartile Female Prop.				0.0847	0.0338	0.0655
				(0.120)	(0.123)	(0.152)
3^{rd} Quartile Female Prop.				0.240	0.207	0.235
				(0.152)	(0.159)	(0.149)
4^{th} Quartile Female Prop.				0.335	0.336*	0.473**
				(0.205)	(0.192)	(0.226)
Ability Controls						
Group Ability Average	No	Yes	No	No	Yes	No
Best/Worst Ability Quintiles	No	No	Yes	No	No	Yes
Observations	1,570	1,492	1,441	1,570	1,492	1,441

1. Standard errors in parentheses, clustered on the classroom group level.

2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3. Table shows the results of estimating Equation (1), Equation (2), Equation (3) and an altered specification using quartiles of proportion on teams larger than 2. The summary statistics of this data are shown in Table B.3.

Table 5: Extended Ability Regressions

	(1)	(2)	(3)	(4)	(5)	(6)
	Best/Worst Ability Quintiles					
Mixed Team	0.194^{***}	0.188^{***}	0.155^{***}	0.210^{***}	0.172^{***}	0.194^{***}
	(0.0309)	(0.0279)	(0.0267)	(0.0439)	(0.0436)	(0.0410)
All Women	0.292^{***}	0.267^{***}	0.211^{***}	0.308^{***}	0.257^{***}	0.258^{***}
	(0.0536)	(0.0499)	(0.0502)	(0.0874)	(0.0896)	(0.0893)
Best/Worst Uni. GPA Quint.	No	Yes	Yes	No	No	Yes
Best/Worst Task Ability Quint.	No	No	Yes	No	Yes	Yes
Best/Worst HS GPA Quint.	No	No	No	Yes	Yes	Yes
F-test	3.99	2.93	1.46	1.51	1.18	.602
p-value	0.0468	0.0883	0.228	0.22	0.28	0.439
Observations	10,675	10,675	10,675	5,158	5,158	5,158
	(7)	(8)	(9)	(10)	(11)	(12)
	Best/Worst Ability Quintile Combinations					
Mixed Team	0.194^{***}	0.182^{***}	0.157^{***}	0.212^{***}	0.185^{***}	0.195^{***}
	(0.0309)	(0.0281)	(0.0272)	(0.0430)	(0.0428)	(0.0404)
All Women	0.292^{***}	0.267^{***}	0.217^{***}	0.307^{***}	0.262^{***}	0.263^{***}
	(0.0536)	(0.0496)	(0.0501)	(0.0884)	(0.0863)	(0.0863)
Uni. GPA Quint. Comb.	No	Yes	Yes	No	No	Yes
Task Ability Comb.	No	No	Yes	No	Yes	Yes
HS GPA Quint. Comb.	No	No	No	Yes	Yes	Yes
F-test	3.99	3.41	1.71	1.4	.937	.717
p-value	0.047	0.066	0.192	0.239	0.334	0.398
Observations	10,675	10,675	10675	5,158	5,158	5,158

1. Standard errors in parentheses, clustered on the classroom group level.

2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3. Table shows results of estimating Equation (1), Equation (2), Equation (3) with different types of ability controls; Task Ability Measure, highschool GPA, and university GPA.

4. The reduced number of observations when using highschool GPA are due to the fact that this variable is only avaliable for Dutch students.

Table 6: Regression Results - Other Characteristics

	(1)	(2)	(3)	(4)	(5)
			Task Results (Std)		
Mixed Team	0.157***	0.154***	0.159***	0.158***	0.157***
	(0.0295)	(0.0296)	(0.0294)	(0.0295)	(0.0296)
All Women	0.216***	0.211***	0.213***	0.209***	0.205***
	(0.0531)	(0.0539)	(0.0537)	(0.0538)	(0.0545)
Ability Combinations	Yes	Yes	Yes	Yes	Yes
Parent Uni. Count Controls	No	Yes	No	No	Yes
Non-Dutch Count Controls	No	No	Yes	No	No
Ethnicity × Dutch Controls	No	No	No	Yes	Yes
Observations	10,675	10,388	10,675	10,675	10,388

1. Standard errors in parentheses, clustered on the classroom group level.

2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3. Table shows results of estimating Equation (3) with the addition of controls for other groups characteristics: parental university attendance, non-Dutch nationality, and Dutch ethnicity.

Table 7: Results by Team Ability Quintile

	(1)	(2)	(3)	(4)	(5)
	Average Group Ability Quintile				
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
Mixed Team	0.260**	0.228**	0.294***	0.158*	0.0799
	(0.117)	(0.108)	(0.0877)	(0.0853)	(0.0846)
All Women	0.282	0.298**	0.371***	0.344***	0.0998
	(0.243)	(0.135)	(0.118)	(0.120)	(0.120)
Ability Combinations	Yes	Yes	Yes	Yes	Yes
Observations	2,160	2,185	2,087	2,146	2,095

1. Standard errors in parentheses, clustered on the classroom level.

2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3. Table shows results of estimating Equation (3) for different subsets on the data depending on the quintile of the average ability of the team members.

Table 8: Summary Statistics Team Self-reflection Exercise Outcomes

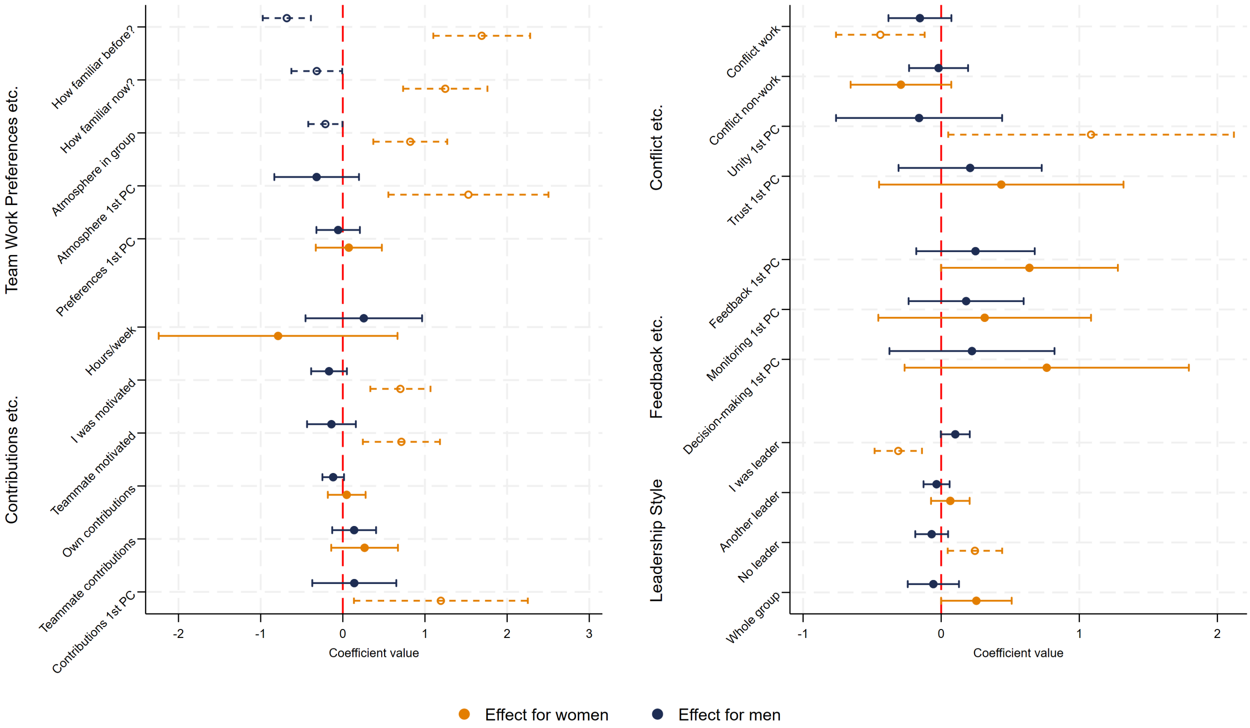
	Mean	SD	Min	Max	Observations
Team Atmosphere, Friendship, & Work Preferences					
How familiar before?	2.135	(1.048)	1	5	1,062
How familiar now?	2.926	(0.924)	1	5	1,062
Atmosphere within group?	4.077	(0.813)	1	5	1,062
Team Atmosphere PCA	0.000	(1.744)	-5.95	2.42	1,062
Team Work Preferences PCA	0.000	(1.263)	-5.20	2.96	1,062
Contributions, Effort, & Motivation					
Hours/week spent on course?	6.210	(3.264)	0	20	1,062
Own motivation to work with team	3.731	(0.925)	1	5	1,062
Partner's motivation to work with team	3.751	(0.950)	1	5	1,062
Rating of own contribution to group	4.116	(0.686)	1	5	1,062
Rating partner's contribution to group	4.007	(0.847)	1	5	1,062
Team Contributions PCA	0.000	(1.684)	-5.04	2.58	1,062
Conflicts, Unity & Trust					
Extent of conflict about group work?	1.694	(0.834)	1	5	1,062
Extent of conflict about other matters?	1.425	(0.771)	1	5	1,062
Team Unity PCA	0.000	(1.897)	-8.19	2.54	1,062
Team Trust PCA	0.000	(1.629)	-5.89	2.75	1,062
Feedback, Monitoring & Decision Making					
Team Feedback PCA	0.000	(1.603)	-5.38	2.84	1,062
Team Monitoring PCA	0.000	(1.662)	-6.97	2.77	1,062
Team Decision Making PCA	0.000	(1.794)	-7.84	3.50	1,062
Leadership Styles					
I was leader	0.219	(0.414)	0	1	1,062
Another member was leader	0.107	(0.310)	0	1	1,062
No leader	0.673	(0.469)	0	1	1,062
Worked as group rather than as individuals	0.419	(0.494)	0	1	1,062

1. Table shows the summary statistics of the outcomes derived from the self-reflection exercise.

2. Outcomes are organised by various headers describing possible explanations for the gender composition effect.

3. Some outcomes are directly elicited questions. Others are derived from PCA on a larger set of instruments. See Appendix A.4 for a full description of how these variables were constructed.

Figure 5: Effect of Female Partner on Self-Reflection Exercise Outcomes



Notes:

1. Figures shows the estimated effect of being allocated a female partner - as opposed to a male partner - on the outcomes from the self-reflection exercise, shown in Table B.7, separately for women and men.
2. Results are derived from Equation (4). The effects for men are estimates of γ_1, while the effect for women are estimates of $\gamma_1 + \gamma_2$.
3. 90% confidence intervals are shown. Dashed intervals reflect estimates significant at $\alpha = 0.10$.

Gender and Performance in Teams: Evidence From Students

Online Appendix

Max Coveney Teresa Bago d’Uva Pilar García-Gómez

February 28, 2024

A Appendix

A.1. Relevance of Tasks How relevant are these tasks to those performed in actual occupations? To get a sense of the overlap between the tasks performed by the student groups and those in actual occupations we make use of the Occupational Information Network (ONET) database, maintained by the U.S Department of Labour. The database contains a complete taxonomy of approximately 900 occupations, with detailed descriptions of key tasks for each occupation sourced from job incumbent surveys and occupational experts. For instance, one occupation in the database is *Sewing Machine Operators*. One key task listed for such workers is to "*Tape or twist together thread or cord to repair breaks*." Given the results in Figure 1, we assume many of these tasks are done in teams. However, such information is not available on the ONET database.

For each task type (writing, data, feedback, and presentation tasks), we perform a string search through the ONET occupation-task database of certain keywords that would indicate an occupational task shares an overlap with one of our task types. The keywords were developed with assistance from ChatGPT, and are shown for each category in Appendix Table B.2.

Appendix Figure B.1 shows the percentage of occupations per International Standard Classification of Occupations (ISCO) group that share some overlap with the tasks given their description in the ONET occupation-task. Writing tasks share the biggest overlap with actual occupational tasks, with 80% of occupations having a writing-based task. Data and presentation-type tasks appear in 40% of occupations, while only 10% of occupations have some type of feedback-based tasks.

These results also reveal that occupations with the largest overlap to the tasks in our data appear most in so-called white-collar occupations. These are those defined as managers, professionals, technicians, and clerks by the ISCO classification system. Appendix Figure B.2 gives a breakdown of the prevalence of each keyword across each occupation category.

A.2. Choice of Ability Control In this section we provide evidence that our ability controls are able to successfully remove any differences in performance between men and women. To test the degree to which different ability controls remove any such differences we make use of our extensive student course data. We look at all other courses taken by students in blocks 3, 4 and 5 of their first year (i.e. all courses taken in that period except the course from which the task data comes from), and show how differences in achievement in these courses by gender changes with the addition of different ability control variables. To be precise, we run regressions of the following form:

$$CourseGrade_{iscb} = \theta_0 + \theta_1 Female_i + f(Ability_i) + ClassBlock_{cb} + \epsilon_{iscb} \tag{5}$$

Where a student i's (standardized) grade in course s, observed in classroom classroom c of block b, is regressed on a $Female$ dummy, and some function of student i's $Ability$, as well as classroom-times-block fixed effects. Intuitively, the degree to which the function of $Ability$ is able to remove any observed gender difference in individual course results θ_1 gives an indication of the degree to which it may successfully control for any underlying differences in ability by the gender composition of a group. In practice, we flexible control for ability via separately included quintile dummies of the following ability measures: (pre-intervention) university GPA, highschool GPA (Dutch students only), and our preferred Task Ability Measure. See Section 2 for further explanation of these variables.

Appendix Table B.5 shows the results of running regression Equation (5) on approximately 12,000 student-course grades observed in blocks 3-5 of the first year. Column 1 gives the estimate of θ_1 without the addition of any ability controls. This indicates that there does appear to be differences in individual student ability in our sample; female students outperform male students in course grades by approximately 9% of a standard deviation. In column 2, we add dummies reflecting the student's university GPA quintile (calculated based on courses in blocks 1 and 2). This reduces the estimates differences by approximately 2 percentage points of a standard deviation, but the differences between male and female students remains significant. Column 3 adds quintiles controlling for the student's highschool GPA. This is available only for Dutch students, reducing the sample size by approximately half. Controls for highschool ability further reduce the observed gender differences to approximately 4% of a standard deviation, resulting in the difference no longer being statistically significant. Finally, column 4 gives the estimate for θ_1 using our preferred ability control, available for virtually all students in the sample. The Task Ability Measure shrinks the observed difference between male and female students to only 1.4% of a standard deviation. This difference is highly statistically insignificant.

The results of Appendix Table B.5 gives further rationale for our use of the Task Ability Measure quintiles as an ability control in the group performance regressions. This measure is the most successful in removing individual performances differences between male and female students in all other first year courses; the addition of these quintiles virtually removes all observable differences between male and female achievement in our sample.

A.3. Grader Gender Bias We make use of grades as a performance measure when trying to identify a gender composition effect in group performance. The tasks performed by teams are graded by TAs under the guidance of a senior lecturer. If the graders themselves exhibit a gender bias in grading, this would lead to a problematic bias in our performance measure. TAs typically grade the tasks of the teams containing students in their classroom. As the names of the students are visible on the assignments, it is plausible that TAs are aware of which student's work they are grading. This plausibly leaves open the possibility of grading bias.

The strongest evidence on the existence of gender grading bias in a similar context comes from Feld et al. (2016). This paper shows that graders of exams at a large Dutch university tend to give higher grades on student's exams when they know the student has similar characteristics in terms of gender and ethnic background to themselves. Assuming a similar pattern in our case, a potential explanation for our results would be that female graders tend to give teams with more women higher grades, even in absence of any performance difference on the tasks. To check for this possibility, we hand collect the gender the grader of each task in our sample, and run Equation (1), Equation (2), and Equation (3) separately for both male and female graders. Appendix Table B.6 gives the results of these regressions. These results show that our baseline results hold for both types of grader. Mixed teams and all women teams significantly outperform male teams when their tasks are graded by both male and female graders, although the magnitude of the gender composition effect is smaller for female graders and only significant with the addition of ability controls at the 10% level,

Hence, we can show that our results are not driven by the same-gender grading bias of female graders for female students. While we cannot rule out bias stemming from a bias towards female students from both, we think this is unlikely to explain our results. Firstly, such bias would not be in line with the same-gender bias pattern observed in very similar settings (Feld et al., 2016), or elsewhere in the literature where the opposite pattern is observed (Lavy, 2008). Second, the magnitude of our results in Table 2 is larger than could plausibly be explained by bias.

A.4. Self-Reflection Exercise Outcomes We use both directly elicited questions, as well as instruments consisting of multiple items taken from existing papers that aim to measure a particular underlying construct. For each instrument, we combine the various measures through a Principal Component Analysis (PCA), from which we extract the first principal component. The full self-reflection exercise is shown Table B.8, where the different instruments are shown in bold, with the items measuring the construct beneath. The results of the PCA for each instrument, including the loadings for each item, the Eigenvalue, and the proportion of explained variance for the first principal component is given in Table B.9. [1]

After construction on the principal components, we are left we 22 different outcomes, measured for 1,062 students across the two blocks the exercise was completed. Summary statistics for the 22 outcomes are given in Table 8. Both members completed the exercise in 235 teams, resulting in 814 team-task observations for which we have answers for all members. Appendix Table B.7 shows the results of running our baseline specifications on this sample. Despite the reduced number of observations, we still find that teams with two women significantly outperform those with two men across all specifications. The results for mixed teams are similar in magnitude to our baseline results, but are not statistically significant when ability controls are added. Given the similar magnitude and the fact that the sample is approximately 8% of the original sample, we believe this stems from power issues, rather than a true null effect.

Below, for each category of explanation for the gender effect, we describe the measures elicited from students, and how these measures are transformed to arrive at those show in Table 8.

Team Work Preferences, Atmosphere & Friendship We use several holistic instruments to measure preferences for group work, the levels of friendship, and the overall group atmosphere within a group. To gauge the degree of friendship we ask respondents to rate how familiar they were with their teammate on a scale between 1 ("Strangers") and 5 ("Best

[1] Although our main results use PCA as a data reduction technique, results are virtually identical when combining the items as simple averages with items signed intuitively.

friends") both before and after the group work. To measure the general atmosphere within the team, respondent are asked to rate the atmosphere within their group between 1 ("Very bad") and 5 ("Very good"). Team atmosphere is also elicited indirectly through an instrument that combines four items, making up *Team Work Atmosphere*. These items relate to individual's agreement with statements relating to their satisfaction and enjoyment working in the group, and willingness to do so again.[2] In order to elicit a measure of individual's preferences for teamwork, we construct the *Team-Work Preferences* principal component, combining four items of individual's reported level of agreement with statements relating to enjoyment of working with others and preference for cooperation over competition.

Contributions, Effort & Motivation We ask respondents how many hours per week, on average, they spent on the course that the tasks made up. We also ask them to rate both their own and their teammate's contributions to the team (1 "Very bad" - 5 "Very good"), as well as the frequency that they themselves felt motivated to work with their teammate, and the frequency that their teammate appeared motivated to work with them (1 "Never" - 5 "Always"). Lastly, we construct a measure of *Team Contributions*, extracting the first principal component of four items measuring agreement with statements regarding whether work was fairly shared, there was equal effort provisions, and the degree of free-riding.

Conflict, Unity & Trust Respondents are asked about the frequency of both work and non-work related conflicts within the team (1 "Never" - 5 "Always"). *Unity* is the first principal component of five items measuring team loyalty, responsibility taking, and shared assistance. *Team Trust* is the first principal component of five items measuring trust and confidence in, and willingness to take on board, the input of team mates.

Feedback, Monitoring & Decision-making We construct three outcomes to measure these three distinct group processes. *Group Feedback* is the first principal component if a four items measuring the degree of feedback and revisions given by and to team members. *Team Monitoring* is the first principal component of four items measuring the degree to which members of the group checked the progress of their team members and held them to deadlines. *Decision-making* is the first principal component of seven items measuring the degree to which decision were made in a collaborative, constructive, and safe environment.

[2] See Table B.9 for a list of the exact items used in each PCA.

Leadership Style We ask respondents whether they themselves were the leader, their teammate was the leader, or there was no leader in the group, and whether the team worked as individuals or as a group.

A.5. Group-Level Analysis of Self-Reflection Exercise Outcomes In this section, we analyze the self-reflection exercise outcomes at the group level, in a similar manner to Equation (3). Intuitively, we explore differences in the group-level averages of the outcomes listed in Table 8 by the gender composition of the group. We first calculate the group-average of each outcome. We then regress each group-average outcome on dummies reflecting the gender composition of the group, as well as controls for tutorial group, block and the ability composition of the group:

$$\begin{aligned} AvgOutcome_{rgb} &= \delta_0 + \delta_1 Mixed_r + \delta_2 AllWomen_r + Tut_g + Block_b \\ &+ \sum_{q=1}^{5} \sum_{p=1}^{q} \theta_{q,p} \mathbb{1}\Big(AbilityQuintile_r^{Best} = q, AbilityQuintile_r^{Worst} = p\Big) + \epsilon_{rgb} \end{aligned} \tag{6}$$

Where $AvgOutcome_{rgb}$ is the average of some outcome for group r, which is a member of classroom g in block b. Coefficients δ_1 and δ_2 then give the average difference in the (group-average) response. The regression also flexibly controls for the ability quintile of both the best and worst individual in the pair.

Group-Level Results The estimates for the δ_1 and δ_2 coefficients from Equation (6) for each of the 22 outcomes across the 5 categories are shown visually in Appendix Figure B.4.[3] The larger confidence intervals for female groups are likely due to their small number. The *Mixed Team* and *All Women* coefficients show the average difference, compared to all-male, in mixed and all-women teams respectively, in the group-level average per outcome.

Despite the monotonic relationship between the number of women in a group and performance found in Section 4, the results in Figure B.4 show that, for most of the group-level averaged outcomes, there is no statistically significant difference between all-men and all-women teams. The largest differences are found in mixed teams, with one member from each gender. Compared to all-male teams, mixed teams report less familiarity (both before and after working together), worse group atmosphere (measured directly and via multiple items), lower levels of own and teammate motivation, lower levels of unity, and lower prob-

[3] The regression results underlying these plots are shown in Appendix Table B.10

ability of working together as a team. However, as shown in Figure 5, these results hide significant heterogeneity by gender of the respondent.

B Appendix Tables and Figures

Table B.1: Balancing Tests

	(1) Female	(2) Ability (Continuous)	(3) High Ability (Dummy)
T-statistic	-0.6671	0.9974	-1.5758
p-value	0.5047	0.3186	0.1151
Observations	5,420	5,420	5,420
	(4) Low Ability (Dummy)	(5) Dutch (Native)	(6) Non-Dutch
T-statistic	0.6881	0.3981	1.1216
p-value	0.4914	0.6905	0.2620
Observations	5,420	5,420	5,420

1. Table shows the results of 6 balancing tests, testing the random allocation of students to groups.
2. Test from Jochmans (2023) used, where the characteristic of each student is compared to that of their allocated partner. Conditional on the pool of potential partners, there should be no significant relationship between a student and their allocated partner.
3. Table shows no significant relationship for any of the 6 characteristics.

Table B.2: Tasks Keyword Table

Writing Tasks	Feedback Tasks	Data Tasks	Presentation Tasks
write	feedback	data entry	present
draft	audit	calculate	speak
edit	appraise	graph	communicate
format	proofread	chart	address
compile		statistics	announce
document		collect data	lecture
author		interpret data	speech
		data analysis	brief
		database	
		survey	

1. Table shows the variable keywords per task type used to search for overlap in the ONET occupation database.
2. Keywords per type drafted by authors and via ChatGPT prompts.

Table B.3: Descriptive Statistics - Larger Groups

	Mean	SD	Count
Team Data			
Number of teams			397
Number of teams 2018			214
Number of teams 2019			30
Number of teams 2020			108
Number of teams 2021			45
Group size	3.680	(0.905)	397
Proportion of females	0.326	(0.268)	397
All men	0.270	(0.444)	397
Mixed	0.698	(0.460)	397
All women	0.033	(0.178)	397
Task Data			
Average task grade	74.478	(12.276)	1570
Average task grade Writing	73.432	(12.446)	601
Average task grade Data	73.285	(12.639)	660
Average task grade Presentation	76.296	(8.340)	132
Average task grade Feedback	81.122	(10.469)	177

1. Table shows summary statistics of both team and task data of groups larger than 2, dropped in the main analysis, but used in Section 4.3.
2. These teams consist of groups size 3-6, either from block 3 of 2018 where larger sizes of teams were created, or teams of 3 created from leftover students within classrooms.

Table B.4: Baseline Results by with Covid Interaction

	(1)	(2)	(3)	(4)
		Course Results (Std)		
Mixed Group	0.194^{***}	0.184^{***}	0.157^{***}	0.144^{***}
	(0.0309)	(0.0430)	(0.0295)	(0.0430)
All Women	0.292^{***}	0.297^{***}	0.216^{***}	0.208^{***}
	(0.0536)	(0.0607)	(0.0531)	(0.0652)
Mixed Group		0.0159		0.0211
× Covid Block		(0.0589)		(0.0557)
All Women		-0.00869		0.0124
× Covid Block		(0.0979)		(0.0954)
Ability Combinations	No	No	Yes	Yes
Observations	10,675	10,675	10,675	10,675

1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows results of estimating Equation (1) and Equation (3) with and without interaction dummies for whether the block was affected by Covid-19 measures.
4. Covid Blocks were those affected by Covid-19 lockdowns: blocks 4-5 of the 2019 cohort, blocks 3-5 of the 2020 cohort, and block 3 of the 2021 cohort.

Table B.5: Individual Course Results

	(1)	(2)	(3)	(4)
	Course Results (Std)			
Female	0.0956***	0.0694***	0.0380	0.0140
	(0.0334)	(0.0243)	(0.0404)	(0.0327)
Ability Controls				
University GPA Quint.	No	Yes	No	No
Highschool GPA Quint.	No	No	Yes	No
Task Ability Measure Quint.	No	No	No	Yes
Observations	12,220	12,166	6,901	12,214

1. Standard errors in parentheses.

2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

3. Table shows results of estimating Equation (5) on the sample of individual grades including quintile dummy variables of the various ability measures. See Appendix A.2.

Table B.6: Regression Results with Gender Tutor Effect

	Female Grader			Male Grader		
	(1)	(2)	(3)	(4)	(5)	(6)
Mixed Team	0.171^{***}	0.120^{***}	0.129^{***}	0.230^{***}	0.189^{***}	0.198^{***}
	(0.0392)	(0.0371)	(0.0365)	(0.0482)	(0.0464)	(0.0490)
All Women	0.229^{***}	0.125^{*}	0.134^{*}	0.384^{***}	0.336^{***}	0.323^{***}
	(0.0667)	(0.0698)	(0.0686)	(0.0818)	(0.0784)	(0.0781)
Ability Controls						
Best/Worst Ability	No	Yes	No	No	Yes	No
Ability Combinations	No	No	Yes	No	No	Yes
Observations	6,478	6,478	6,478	4,089	4,089	4,089

1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows results of estimating Equation (1), Equation (2), and Equation (3) separately by TA gender. See Appendix A.3.

Table B.7: Baseline Results With Survey Sample

	(1)	(2)	(3)
	Task Results (Std)		
Mixed Team	0.170**	0.136	0.129
	(0.0837)	(0.0844)	(0.0847)
All Women	0.452***	0.383***	0.359**
	(0.111)	(0.136)	(0.137)
Ability Controls			
Best/Worst University GPA Quint.	No	Yes	No
Skills B1&2 Quint. Comb.	No	No	Yes
F-statistic	8.627	5.287	3.526
p-value	0.00462	0.0248	0.0651
Observations	814	814	814

1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows results of estimating Equation (1), Equation (2), and Equation (3) on the small sample of students who took part in the self-reflection exercise.

Table B.8: Team Work Self-reflection Assignment Questions

Question	Scale
How familiar were you with your fellow group member(s) before working together in this course?	Strangers (1) - Best friends (5)
How familiar are you with your fellow group members now, after working together in this course?	Strangers (1) - Best friends (5)
How many hours per week on average did you spend on this course?	0-20+
Overall, how would you describe the atmosphere within your group?	Very bad (1) - Very good (5)
I felt motivated to work with my fellow group member.	Never (1) - Always (5)
My fellow group member appeared motivated to work with me.	Never (1) - Always (5)
Worked as group	Yes/No
Worked as individuals	Yes/No
I was the leader	Yes/No
Another member was the leader	Yes/No
Mostly shared leadership or no defined leader(s).	Yes/No
Extent of conflict/disagreement about group work?	Never (1) - Always (5)
Extent of conflict/disagreement about other matters?	Never (1) - Always (5)
Were the conflicts managed/resolved constructively and effectively?	Never (1) - Always (5)
How would you rate your own contributions to the work of your group?	Very bad (1) - Very good (5)
How would you rate the average contributions of your fellow group member?	Very bad (1) - Very good (5)
Team Work Preferences	
I like to work with other people.	Strongly disagree (1) - Strongly agree (5)
Cooperation is preferable to competition.	Strongly disagree (1) - Strongly agree (5)
I consider myself to be a competitive person.	Strongly disagree (1) - Strongly agree (5)
Work assignments are better when I do them myself.	Strongly disagree (1) - Strongly agree (5)
Team Work Atmosphere	
In general, I am satisfied with the work of my group.	Strongly disagree (1) - Strongly agree (5)
I enjoyed working with my group.	Strongly disagree (1) - Strongly agree (5)
Working in this group was frustrating.	Strongly disagree (1) - Strongly agree (5)
I want to work with this group in the future.	Strongly disagree (1) - Strongly agree (5)
Team Unity	
Our group was united in trying to reach its goals for performance.	Strongly disagree (1) - Strongly agree (5)
In this group, we all took our responsibility for setbacks or poor group perform	Strongly disagree (1) - Strongly agree (5)
We helped each other to complete group tasks.	Strongly disagree (1) - Strongly agree (5)
We worked well together.	Strongly disagree (1) - Strongly agree (5)
We were loyal to each other.	Strongly disagree (1) - Strongly agree (5)
Team Feedback	
I gave feedback on the work of my fellow group member.	Never (1) - Always (5)
I made revisions to the work of my fellow group member.	Never (1) - Always (5)
My fellow group member gave feedback on my work.	Never (1) - Always (5)
My fellow group member made revisions to my work.	Never (1) - Always (5)
Team Trust	
I did not have difficulties accepting suggestions from my fellow group member	Strongly disagree (1) - Strongly agree (5)
I trusted the knowledge of my fellow group member about the group work was sufficient.	Strongly disagree (1) - Strongly agree (5)
I trusted the information that my fellow group member brought to the discussion.	Strongly disagree (1) - Strongly agree (5)
When my fellow group member gave information, I wanted to double-check this information.	Strongly disagree (1) - Strongly agree (5)
I did not have much confidence in the expertise of my fellow group member.	Strongly disagree (1) - Strongly agree (5)
Team Monitoring	
We checked to make sure that everyone in the group continued to work on the assignments.	Strongly disagree (1) - Strongly agree (5)
We monitored each other's progress on the assignments.	Strongly disagree (1) - Strongly agree (5)
We checked whether everybody was meeting their obligations to the group.	Strongly disagree (1) - Strongly agree (5)
We made sure that everyone in the group met their deadlines.	Strongly disagree (1) - Strongly agree (5)
Team Decision Making	
Decisions were mainly taken by one group member.	Strongly disagree (1) - Strongly agree (5)
Decisions were worked out together in this group.	Strongly disagree (1) - Strongly agree (5)
Some members contributed less to decision-making than others.	Strongly disagree (1) - Strongly agree (5)
When deciding on the strategies, the opinion of all group members was actively asked for.	Strongly disagree (1) - Strongly agree (5)
Some group members pushed their opinion through without much regard.	Strongly disagree (1) - Strongly agree (5)
I felt safe sharing my opinion and ideas with the other group members.	Strongly disagree (1) - Strongly agree (5)
We adhered to any assignment-related decisions we made together.	Strongly disagree (1) - Strongly agree (5)
Team Contributions	
All group members contributed to the assignments equally.	Strongly disagree (1) - Strongly agree (5)
I had to do more than my fair share of work for the assignments.	Strongly disagree (1) - Strongly agree (5)
All group members put in the same effort for the assignments.	Strongly disagree (1) - Strongly agree (5)
I experienced free-riding problems in my group.	Strongly disagree (1) - Strongly agree (5)

1. Table shows the full self-reflection exercise that students completed in blocks 3 and 5 of the 2021 cohort.
2. The questions are organised by possible explanations of the gender composition effect.

Table B.9: Team Self-Reflection Assignment Principal Component Results

	Loading	Eigen-value	Proportion
Team Work Preferences		1.59443	0.3986
I like to work with other people.	0.6047		
Cooperation is preferable to competition.	0.5653		
I consider myself to be a competitive person.	-0.2161		
Work assignments are better when I do them myself.	-0.5178		
Team Atmosphere		3.04291	0.7607
In general, I am satisfied with the work of my group.	0.4865		
I enjoyed working with my group.	0.5182		
Working in this group was frustrating.	-0.4861		
I want to work with this group in the future.	0.5084		
Team Trust		2.654	0.5308
I did not have difficulties accepting suggestions from my fellow group member	0.3670		
I trusted the knowledge of my fellow group member about the group work was sufficient	0.5356		
I trusted the information that my fellow group member brought to the discussion	0.5315		
When my fellow group member gave information, I wanted to double-check this information	-0.2802		
I did not have much confidence in the expertise of my fellow group member.	-0.4663		
Team Unity		3.59712	0.7194
Our group was united in trying to reach its goals for performance.	0.4563		
In this group, we all took our responsibility for setbacks or poor group performance	0.4274		
We helped each other to complete group tasks.	0.4392		
We worked well together.	0.4663		
We were loyal to each other.	0.4458		
Team Feedback		2.57058	0.6426
I gave feedback on the work of my fellow group member	0.4988		
I made revisions to the work of my fellow group member	0.4548		
My fellow group member(s) gave feedback on my work.	0.5267		
My fellow group member(s) made revisions to my work.	0.5167		
Team Monitoring		2.76067	0.6902
We checked to make sure that everyone in the group continued to work on the assi	0.5046		
We monitored each other's progress on the assignments.	0.5096		
We checked whether everybody was meeting their obligations to the group.	0.5311		
We made sure that everyone in the group met their deadlines.	0.4513		
Team Decision Making		3.21802	0.4597
Decisions were mainly taken by one group member.	-0.3594		
Decisions were worked out together in this group.	0.4312		
Some members contributed less to decision-making than others.	-0.3666		
When deciding on the strategies, the opinion of all group members was actively a	0.3960		
Some group members pushed their opinion through without much regard of what the	-0.3617		
I felt safe sharing my opinion and ideas with the other group members.	0.3568		
We adhered to any assignment-related decisions we made together.	0.3680		
Team Contributions		2.83549	0.7089
All group members contributed to the assignments equally.	0.5287		
I had to do more than my fair share of work for the assignments.	-0.4701		
All group members put in the same effort for the assignments.	0.5174		
I experienced free-riding problems in my group.	-0.4814		

1. Table shows the results of the principal component analysis of some questions included in the self-reflection exercise.
2. Per PCA, the loadings per question, Eigen-value, and proportion of explained variance are shown for the first principal component.

Table B.10: Team Self-reflection Assignment - Team Level Results

	Team Work Preferences, Atmosphere & Friendship					**Contributions, Effort & Motivation**					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	How familiar before	How familiar now	Atmosphere in group	Atmosphere 1st PC	Preferences 1st PC	Hours /week	I was motivated	Teammate motivated	Own contributions	Teammate contributions	Contributions 1st PC
Mixed Team	-0.760***	-0.582***	-0.275*	-0.440*	-0.149	0.383	-0.335***	-0.363***	-0.0764	-0.0947	-0.287
	(0.172)	(0.138)	(0.138)	(0.230)	(0.187)	(0.403)	(0.124)	(0.133)	(0.110)	(0.138)	(0.294)
All Women	0.0561	-0.0273	-0.0774	0.0201	0.111	0.869	-0.130	-0.171	0.0120	-0.0862	0.158
	(0.338)	(0.269)	(0.238)	(0.492)	(0.272)	(0.571)	(0.189)	(0.226)	(0.139)	(0.201)	(0.408)
Observations	235	235	235	235	235	235	234	234	235	235	235

	Conflict, Unity & Trust				**Feedback, Monitoring & Decision-making**			**Leadership Style**			
	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
	Conflict work	Conflict non-work	Unity 1st PC	Trust 1st PC	Feedback 1st PC	Monitoring 1st PC	Decision-making 1st PC	I was leader	Another leader	No leader	Whole group
Mixed Team	0.0507	0.0156	-0.634**	-0.0703	-0.211	0.180	-0.191	0.0768	0.0281	-0.105	-0.115*
	(0.138)	(0.125)	(0.313)	(0.210)	(0.260)	(0.266)	(0.343)	(0.0567)	(0.0501)	(0.0798)	(0.0669)
All Women	-0.124	-0.0851	-0.270	0.111	-0.0932	0.141	0.493	-0.102*	0.0478	0.0542	-0.0682
	(0.170)	(0.126)	(0.506)	(0.487)	(0.384)	(0.314)	(0.447)	(0.0569)	(0.0609)	(0.0868)	(0.146)
Observations	235	235	235	235	235	235	235	235	235	235	235

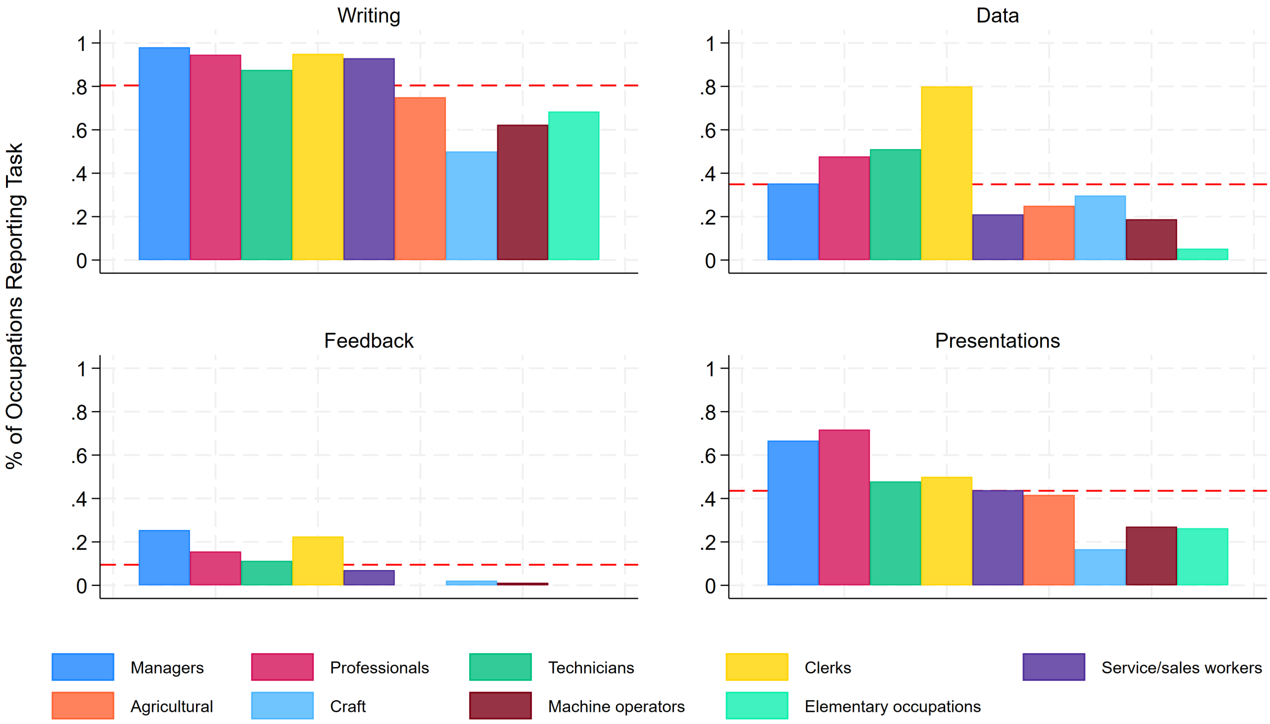
1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows results of estimating Equation (5) on the various self-reflection exercise outcomes, shown in Table B.7.

Table B.11: Team Self-reflection Exercise Regression Results

	Team Work Preferences, Atmosphere & Friendship					**Contributions, Effort & Motivation**					
	(1) How familiar before	(2) How familiar now	(3) Atmosphere in group	(4) Atmosphere 1st PC	(5) Preferences 1st PC	(6) Hours /week	(7) I was motivated	(8) Teammate motivated	(9) Own contributions	(10) Teammate contributions	(11) Contributions 1st PC
Female teammate	-0.681***	-0.317*	-0.212*	-0.318	-0.0566	0.255	-0.168	-0.138	-0.117	0.138	0.140
	(0.175)	(0.184)	(0.124)	(0.307)	(0.158)	(0.423)	(0.130)	(0.177)	(0.0784)	(0.159)	(0.305)
Female teammate ×	1.691***	1.247***	0.821***	1.529**	0.0732	-0.787	0.701***	0.713**	0.0476	0.265	1.194*
Woman	(0.351)	(0.306)	(0.268)	(0.581)	(0.240)	(0.867)	(0.218)	(0.280)	(0.137)	(0.242)	(0.631)
Full effect	1.01	.93	.609	1.21	.0166	-.532	.533	.575	-.0693	.403	1.33
F-test	10.5	14.4	7.59	4.61	.0107	.502	9.59	7.18	.486	4.61	6.37
p-value	.00217	.000417	.00826	.0368	.918	.482	.00327	.0101	.489	.0368	.0149
Observations	348	348	348	348	348	348	346	346	348	348	348
	Conflict, Unity & Trust				**Feedback, Monitoring & Decision-making**			**Leadership Style**			
	(12) Conflict work	(13) Conflict non-work	(14) Unity 1st PC	(15) Trust 1st PC	(16) Feedback 1st PC	(17) Monitoring 1st PC	(18) Decision-making 1st PC	(19) I was leader	(20) Another leader	(21) No leader	(22) Whole group
Female teammate	-0.154	-0.0189	-0.160	0.210	0.248	0.181	0.223	0.102	-0.0333	-0.0688	-0.0566
	(0.136)	(0.127)	(0.359)	(0.309)	(0.256)	(0.248)	(0.356)	(0.0626)	(0.0562)	(0.0708)	(0.111)
Female teammate ×	-0.441**	-0.291	1.085*	0.435	0.639	0.315	0.764	-0.311***	0.0663	0.244**	0.255
Woman	(0.191)	(0.217)	(0.617)	(0.528)	(0.382)	(0.459)	(0.614)	(0.102)	(0.0834)	(0.118)	(0.152)
Full effect	-.595	-.31	.925	.645	.887	.496	.988	-.209	.0329	.176	.198
F-test	16.1	2.87	3.06	2.26	7.83	1.28	4.47	6.09	.375	3.16	4.18
p-value	.000207	.097	.0868	.139	.00738	.264	.0397	.0172	.543	.0817	.0465
Observations	348	348	348	348	348	348	348	348	348	348	348

1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows results of estimating Equation (6) on the various self-reflection exercise outcomes, shown in Table B.7.

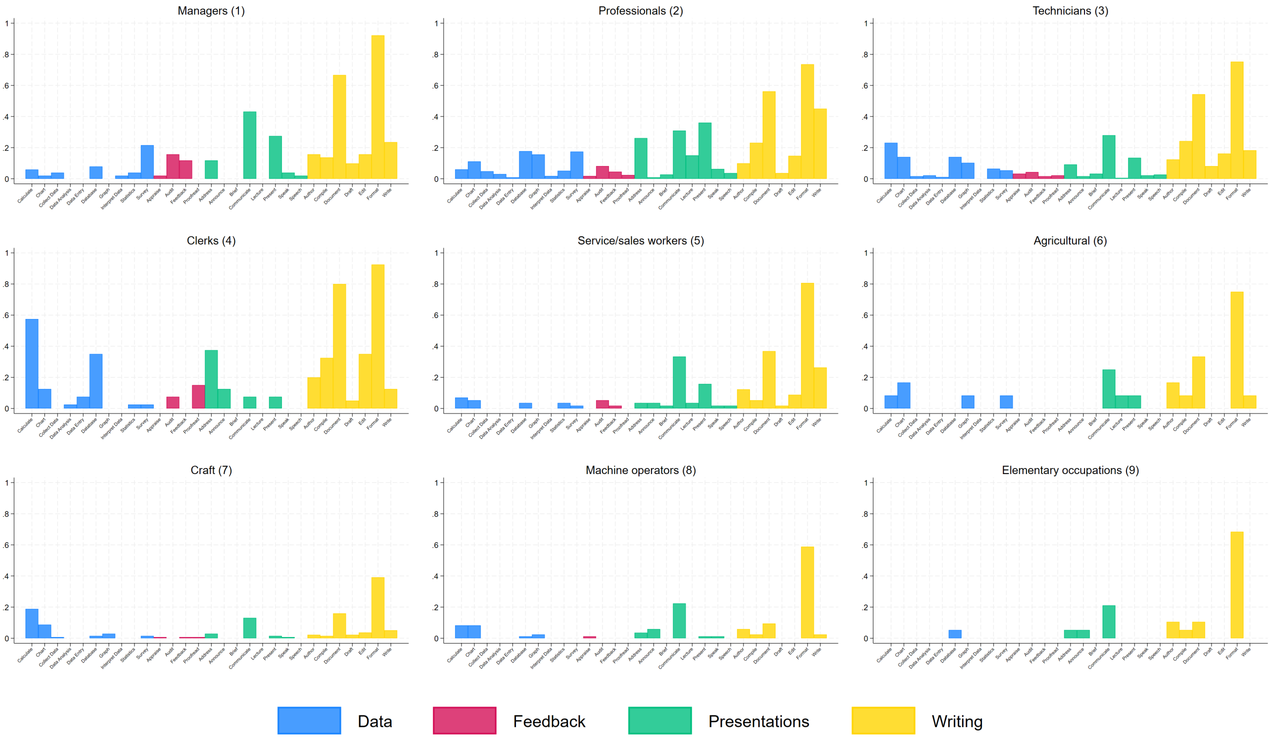
Figure B.1: Overlap of Tasks With Occupations



Notes:

1. Figure shows the proportion of occupations per ISCO category with tasks that share a keyword with those shown in Appendix Table B.2.
2. Data comes from the ONET occupation-task database 28.0.
3. The dashed red line shows the proportion of overlapping occupations across all occupations per task type.

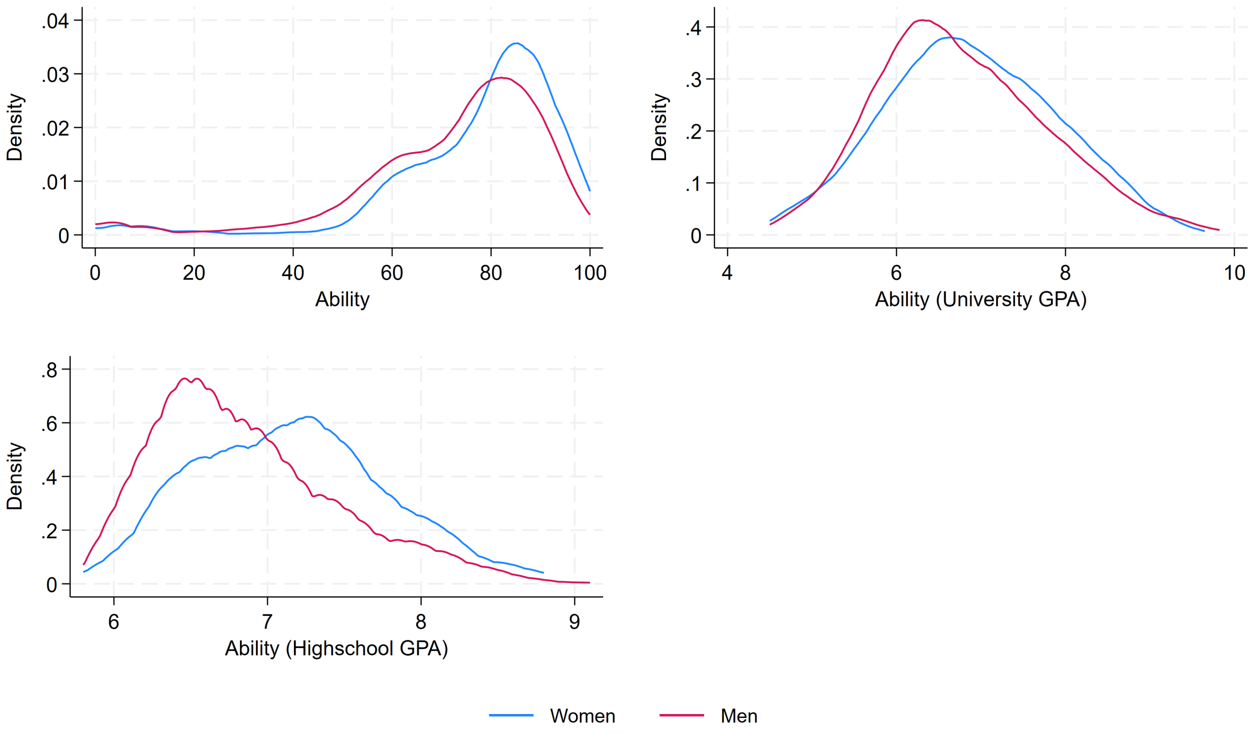
Figure B.2: Overlap of Keywords With Occupations



Notes:

1. Figure shows the proportion of overlap per task keyword, shown in Appendix Table B.2, by each ISCO occupation category.
2. Data comes from the ONET occupation-task database 28.0.

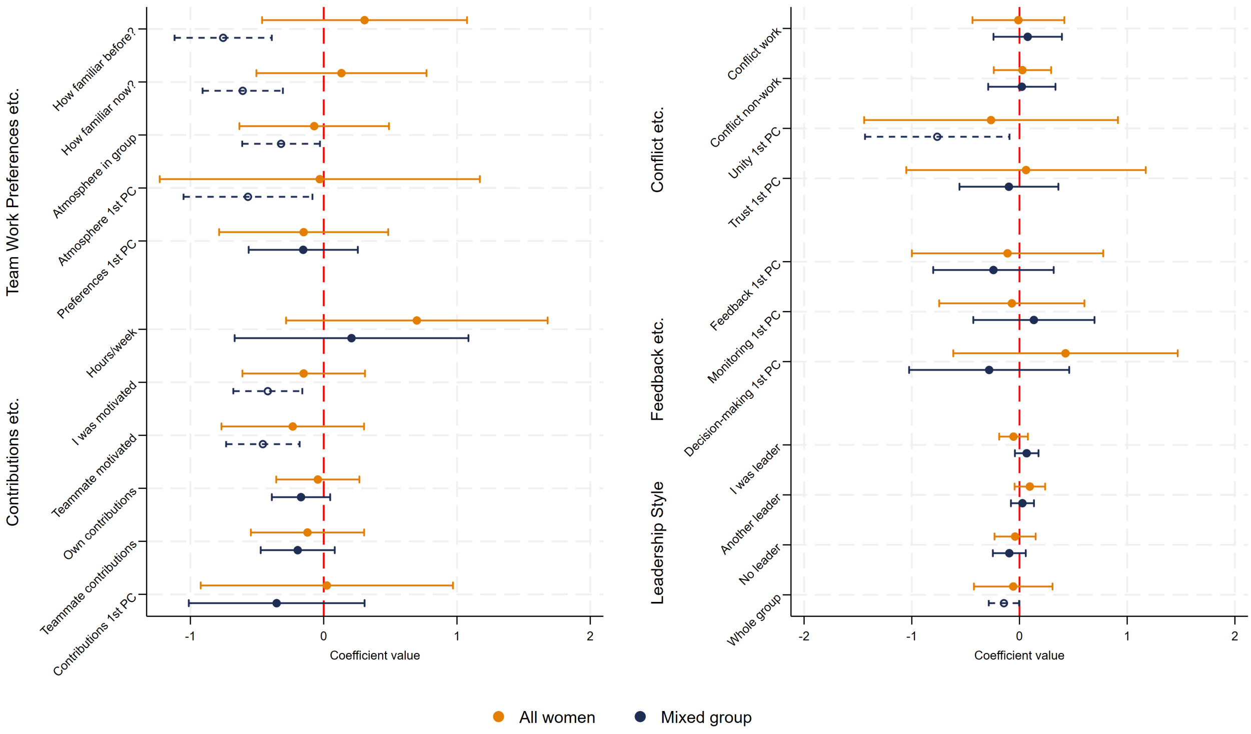
Figure B.3: Ability Histograms



Notes:

1. Figures shows histograms of three ability measures of students, separately for men and women in our sample.
2. The Ability measures is the Task Ability Measures, our preferred measure of individual task ability throughout the paper.
3. See Section 2 for a detailed description of these variables.

Figure B.4: Differences in Self-reflection Assignment Across Team Types



Notes:

1. Figures shows the estimated differences in the average response per group type per self-reflection exercise outcome, shown in Table B.7.
2. Results are derived from Equation (6). The effects for mixed teams are estimates of δ_1, while the effect for all women teams are estimates of δ_2.
3. 90% confidence intervals are shown. Dashed intervals reflect estimates significant at $\alpha = 0.10$.