Coveney, Max; García Gómez, Pilar; Bago d'Uva, Teresa

Working Paper

Gender and performance in collaboration: Evidence from random student teams

Tinbergen Institute Discussion Paper, No. TI 2025-032/V

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Coveney, Max; García Gómez, Pilar; Bago d'Uva, Teresa (2025) : Gender and performance in collaboration: Evidence from random student teams, Tinbergen Institute Discussion Paper, No. TI 2025-032/V, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/316215

TI 2025-032/V
Tinbergen Institute Discussion Paper

Gender and Performance in Collaboration: Evidence from Random Student Teams

Max Coveney[1]
Pilar Garcia-Gomez[2]
Teresa Marreiros Bago d'Uva[3]

1 Erasmus University Rotterdam, Tinbergen Institute

2 Erasmus University Rotterdam, Tinbergen Institute

3 Erasmus University Rotterdam, Tinbergen Institute

Gender and Performance in Collaboration: Evidence from Random Student Teams

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

April 28, 2025

Abstract

Should gender composition be taken into account when forming teams? This paper examines how the output of teams completing tasks similar to those performed in many workplaces is influenced by their gender composition. Leveraging an economics bachelor course in which students are randomly paired together, we document large differences in performance grades by the gender make-up of the team. 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 each of the group members, as well as other characteristics potentially relevant for teamwork that may vary by gender. Exploring mechanisms, we find suggestive evidence that women have greater preferences for cooperation, and - even when controlling for individual ability - exert higher effort levels in teams compared to men. This asymmetry appears to lead to members of mixed-gender teams reporting the worst team experiences.

[*]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 School of Economics SDGs Research Programme and 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 changes when assembling teams?

To shed light on these questions, this paper studies how the gender composition of work teams influences their performance.[2] Our main data source consists of graded tasks performed by randomly allocated pairs of university students. These teams work together for approximately 2 months and perform tasks common to many work environments, such as writing and document preparation, data processing and analysis, feedback-giving, and oral presentations. We show a large and robust gender composition effect on performance. Teams with more women tend to produce significantly better quality work, even controlling for the *individual* ability of each team member.

Leveraging the random allocation of 4 cohorts of roughly 3,000 students to 3,200 work teams, and using grades on approximately 12,600 team-task observations, we estimate the importance of a team's gender composition. We find sizeable and significant differences in task performance 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 17% (15%) 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] Appendix Figure B.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.

[2] Although we recognise the distinction between sex and gender, this paper uses the concepts interchangeably.

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 task types, levels of task importance, and team sizes. We find the gender composition effect across all task types in our data: writing, data analysis, feedback giving, and presentations. Identifying low and high stakes tasks - based on their weight on the final grade - we show the gender composition effect is present for both. Further, we find a similar gender composition pattern in a sample of larger teams, showing the effect is not isolated to pairs.

After establishing the existence of a gender composition effect on the quality of team output, we turn to investigating differences in team processes and experience that may reveal mechanisms possibly driving the effect. Using data from a self-reported reflection exercise about this team work, we measure individuals' 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.

Based on this self-reflection exercise, we find suggestive evidence that the gender composition effect may be driven by the fact that women appear to be more conscientious and diligent team members. Women report a larger preference for team work, and report more hours spent on team work, than men. We also find that respondents from mixed-gender teams report worse outcomes along many dimensions of team work processes and experiences – including team atmosphere, unity, and motivation – compared to all-female and all-male teams. We speculate that this pattern may be driven by a mismatch in diligence and effort within these pairs, with women taking on a larger burden of 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).[3]

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

This paper contributes to the growing economic literature broadly studying gender and team work, often via experimental methods and samples of students (Keck and Tang, 2018; Born et al., 2020; Sarsons et al., 2021; Karpowitz et al., 2023; Hardt et al., 2024).[4] In a similar spirit, ours is the first paper to show how gender composition impacts the performance of teams completing tasks comparable to those in many white-collar occupations. Though also based on student data, our setting enhances external validity in two important respects: it contains a long team interaction period of multiple months, and the nature of tasks performed has a large overlap with those performed in real occupations.[5] Indeed, many of the students in our sample will go on to work in such occupations.

Beyond the generalisability of our results, three features of our context allow us to make novel contributions to this literature. First, 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 are 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, which hinders causal claims about the effect of gender. We avoid this problem by studying randomly allocated teams.

Second, we have rich administrative data on each individual in our sample, including their previously measured individual performance on similar tasks, high school and university GPA, ethnicity, and SES. We can therefore rule out that the gender composition effect is being driven by these other variables that may correlate with gender in our sample, as well as by individual ability. Our data also covers various task types and stakes, allowing investigation of the gender composition effect across these dimensions.

Third, we pair our administrative student and task performance data with a comprehensive self-reported reflection exercise, covering many aspects of individuals' team-working experiences and group dynamics. This grants a deeper exploration of potential mechanisms

[4]Using a sample of students and in a lab setting Hardt et al. (2024) find that all-women (all-men) teams communicate the least (most). 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 team 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 team work than men. Using laboratory experiments, Keck and Tang (2018) find that team judgement quality is positively impacted by the presence of a female team member.

[5]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.

driving the gender composition effect.

Overall, our findings paint a more nuanced picture of the effect of gender composition on team performance. In line with earlier research, we document a positive impact of women 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). On the other hand, our subsequent analyses finds that members of mixed-gender teams, especially the women therein, report worse outcomes along many dimensions of team processes and experiences. Although the restricted sample size available for these additional analyses and the self-reported nature of the data prevent a definitive statement on mechanisms, they provide suggestive evidence of the following: while the presence of women in teams may raise the quality of output through a boost in diligence and effort, the potentially higher burden shouldered by women in these teams may lead to all team members experiencing worse team atmosphere, unity and motivation.

Our findings, should they hold in other contexts, 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 costly, and potentially unrewarded (Babcock et al., 2017), tasks for women in these teams. Our results suggest that the reported team atmosphere, motivation, and unity experienced by members of these teams suffers if this burden is unaddressed.

The remainder of the paper is organised 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 results of the baseline analysis and various extensions. In Section 5 we test the robustness of our baseline results. Section 6 describes the team 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 economics bachelor at a top-ranked university in The Netherlands.[6] The bachelor admits an average of 750 students per year over our observation period. The bachelor program has various specialisations, some of which are offered in separate English and Dutch language versions.[7] The bachelor lasts 3 years in total, and each academic year consists of 5 blocks (semesters), each lasting eight weeks.

Course overview We leverage a course spanning blocks 2 to 5 of the first year of the bachelor, compulsory for all specialisations. The course centres 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 covering communication, block 3 covering writing skills, block 4 covering datawork and analyses, and block 5 covering 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.[8]

Course structure Figure 1 shows the structure of the course. The tasks in block 2 involve 3 *individual* tasks. Task 1 involves individual students researching and presenting on an economics subject, task 2 involves giving written feedback on their peers' presentations, task 3 involves 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 students' aptitude on tasks that are similar - both in type and in context - to the ones they will subsequently complete in teams.[9]

In blocks 3, 4 and 5 students work in pairs to complete tasks. At the beginning of each

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

[7] These specialisations include econometrics, economics and law, fiscal economics, and business and economics. During the first year of the program, students from all specialisations follow a shared curriculum, mostly taking the same courses together.

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

[9] While there is not a complete 1-to-1 overlap with these tasks and every tasks students will complete in the subsequent blocks, Appendix A.2 show that controlling for the average grade achieved by students across these three tasks cleans out any gender differences in individual grades in other courses. We also assess robustness to various alternative measures of individual ability in Section 5.1.

block, students are randomised 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.[10] 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 respective students.[11] Tasks are graded according to a detailed rubric that is provided to the TAs.[12] A student's performance across all tasks in the three blocks forms their final grade for the course.

Our sample period covers the 2018-2021 academic years.[13] 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, rather than pairs. We therefore exclude this block-year from our main analyses. We make use of it in Section 4.4 when investigating effects of gender composition in teams of larger sizes. This leaves us with 11 academic blocks for our main analysis: two in the 2018 academic year, and three in each of the 2019-2021 academic years.

Task types and relevance Figure 1 gives the tasks that each team is required to complete per block. We identify four distinct types of tasks. *Writing tasks* include writing research proposals, components of a research document, and ultimately an entire research paper. *Data tasks* include identifying and using existing datasets, running surveys, and cleaning and analyzing data. *Feedback tasks* involve giving feedback to other teams on their work, predominately on their writing work. *Presentation tasks* involve students presenting their work in class in-front of their peers. Each block consists of four of these graded 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.

[10] Classrooms consist of students within the same economics specialisation program, of which there are six. In the case of an uneven number of students within a classroom group, one team of three is formed. We discard these teams in our main analyses but make use of them when examining larger teams.

[11] In Section 5.4 we assess whether TA gender influences grading patterns to investigate potential bias. While male graders tend to give higher grades, we find the gender composition effect present across TA gender.

[12] An example of the instructions and grading rubric used for two tasks is given in Appendix Figure B.2.

[13] The COVID-19 pandemic and the associated lock-downs occurred during our observation period. We show that our baseline results are unchanged during these COVID-19 blocks.

2.2. Data We use data on performance grades achieved by the randomly allocated teams for the tasks within a block, and their overall grade for the block. We merge in a range of student characteristics from the university's administrative database: gender, age, high school GPA (available for Dutch students), ethnicity (idem), nationality, information on the educational attainment of students' parents, and the grades achieved by each student in all other courses taken at the university.

Our aim is to measure how task performance differs by the gender composition of a team. If the ability to perform a task differs by an individual's gender, then any gender composition effect will (at least partly) reflect these average ability differences. Therefore, our baseline analysis includes ability controls for each individual in the team.

Our main measure of individual ability is the block 2 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 teams. Figure 1 gives the exact tasks involved in this measure. This is our preferred measure of ability, given its similarity with the team-based tasks. 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. University GPA in previous courses (based on the grades achieved in the first two blocks of students' first year) is available for all students in our sample. Appendix Figure B.5 shows the distribution of each of the three ability measures by gender. We use university and high school GPA as alternative measures in robustness checks.

Table 1 presents the summary statistics for each variable in our sample. Our data includes approximately 3,000 unique students over the 4 cohorts. Across the 11 blocks these students form approximately 3,200 work teams.[14] Around 48% of these teams are all-male teams, 42% are mixed-teams, and 10% are all-female teams. The various writing, data, feedback, and presentation tasks these teams perform lead to 12,600 team-task grade observations.

2.3. Randomisation tests Students are randomised into teams at the beginning of each block. Randomisation of teams is vital to identifying any potential gender composition

[14] 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.

effect. The presence of sorting, or endogenous team 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 randomisation of students into teams using the randomisation tests derived by Jochmans (2023). This test improves on previous randomisation 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 correlation should be present.

We perform randomisation tests on the following characteristics: gender, ability (continuous), high ability (top 25% of ability distribution), low ability (bottom 25%), non-Dutch background, non-immigrant Dutch background, immigrant Dutch background, and parental university attendance. The results of these 8 randomisation tests are shown in Appendix Table B.3. As expected given the randomisation procedure, there are no significant correlations between a student's characteristic and that of their partner for any of these student characteristics. We therefore conclude that the randomisation of teams was successful.

3 Methods

We regress the standardised $Performance_{agt}$ on task a achieved by team g in block b and classroom c on a set of dummy variables describing the gender composition of the team – $Mixed_g$ and $AllWomen_g$ – and both classroom-times-block ($ClassBlock_{cb}$) and task fixed effects (A_a):

$$Performance_{agcb} = \beta_0 + \beta_1 Mixed_g + \beta_2 AllWomen_g + A_a + ClassBlock_{cb} + \epsilon_{agcb} \quad (1)$$

Coefficients β_1 and β_2 then give the difference in standardised performance 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 performance. We cluster the error term at classroom the 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.[15] We therefore also estimate model specifications controlling for ability differences between teams. As discussed in Section 2, our preferred measure of ability is the Task Ability Measure - the average grade of the individual tasks achieved by a student in block 2. We use two approaches to control for this measure in the main analysis and also show robustness to alternative measures of individual ability - high school GPA and university GPA in blocks 1 and 2 - in Section 5.1.

Best & worse ability controls We 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 team g of classroom t in block b.[16] Our first approach is to include indicators for each ability quintile of the best and worst member, resulting in the following extended version of Equation (1):

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

Ability combination controls Equation (2) controls separately for the individual ability of both members of the team. However, there may be interaction effects between the ability of the two team members; the effect of being in the top quintile of individual ability on $Performance_{agt}$ may depend on the ability quintile of the other member. In total, there are 15 possible combinations of ability quintile categories of the best and worst member of the team. Our second specification ensures that any potential ability interactions are

[15] Appendix Figure B.5 shows some evidence of women having higher average ability, depending on the measure used.

[16] The quintiles here and elsewhere in the paper are calculated by specialisation program as classrooms are grouped by specialisation.

controlled for by including indicators of each of these 15 categories:

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

As well as Equation (1), our baseline results include 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 differences in the individual academic ability of the individuals. One may still worry about underlying gender differences in the individual academic ability if our measure contains too much noise. We present evidence in Appendix A.2 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

Figure 2 shows the density of the (standardised) performance measure for teams with each of the three gender compositions 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 and systematic differences in raw task performance by gender composition; on average, all-male teams do worse, mixed teams better than all-male teams, and all-female teams perform best.

4.1. Regression approach Table 2 presents our baseline results. Column (1) shows results of estimating Equation (1) with the 12,600 team-task grade observations. The estimates for β_1 and β_2 are both large and highly significant. They imply that pairs comprised of two women (one woman and one man), achieve grades 25% (19%) of a standard deviation higher than those comprised of two men.

How much of the differences by gender composition in column (1) can be explained by differences in individual's task ability per gender? In columns (2) and (3) of Table 2 we present estimation results of Equation (2) and Equation (3), respectively. The addition of

ability controls in these two ways reduces the magnitude of the estimated β_1 and β_2 coefficients, although both remain large in magnitude and statistically significant.

The results in column (3), which we take as our preferred specification due to accounting for ability combinations between team members, imply that pairs comprised of two women (one woman and one man), achieve grades 17% (15%) of a standard deviation higher than those comprised of two men. Also presented in Table 2 are the results of a test of the equality of two gender composition coefficients. In none of the three specifications is the difference between them significant, indicating no statistical difference between the performance of mixed gender pairs and all-women pairs.

4.2. Results by task type and importance There are large differences in performance by gender composition on the average grade of all tasks. Are these differences also present within the various types of tasks? For instance, previous research documents gender differences in preference for giving presentations (De Paola et al., 2021), which could manifest in differing performances for these tasks.

To investigate this we estimate our preferred specification, Equation (3), on sub-samples of each task type: data tasks, feedback tasks, presentation tasks, and writing tasks. It may also be the case that the gender effect is only present in tasks of certain importance, as defined by their weight in student's final grades. We define high (low) importance tasks as those that count for at least (less than) 50% of the grade in a particular block. We estimate Equation (3) on both of these sub-samples.

The results of these sub-group analyses are shown in Table 3. They reveal that the pattern of all-male teams being outperformed by mixed and all-female teams is present across all types of task, with the exception of all-women teams in data tasks, where the coefficient is positive but insignificant.

The effect of gender composition effect differ somewhat by task type; the largest differences are observed in presentation tasks, where mixed (all women) teams outperform all men teams by 18% (20%). 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. The gender composition effect is present in tasks of both low and high importance, with the point estimates of the latter being slightly higher than the former.

4.3. Results by team ability To what extent is the gender composition effect concentrated in certain parts of the team ability distribution? Finding that the effect is only present in - for instance - teams with lower ability has implications not only for the external validity of these findings, but may also point to mechanisms at play. We divide all observations by the quintile of the team's average individual ability, and estimate Equation (3) for each of these subgroups.[17]

Table 4 displays the gender composition results for these sub-samples. The grade advantage of mixed teams compared to all-male teams is present across the ability distribution, except in the highest ability teams (5^{th} quintile). Statistically significant differences between all-men and all-women teams are found in the 3^{rd} and 4^{th} ability quintile. Notably, the gender composition effect does not appear to be present for the most able teams. This may partly reflect the lack of grade variation in these teams due to the truncated nature of our performance measures.

4.4. Larger teams The above results examine the gender composition effect for pairs. However, workers in firms and other contexts may obviously also collaborate in larger teams. Are the findings above present in other team 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 teams of size 3 from our main analysis that were formed in the remaining blocks in classrooms with an odd number of students. In order to investigate whether the gender composition pattern above also exists in larger teams, we analyse it here for the sample of teams of size 3 and above. In total, there are approximately 470 teams larger than 2, and 1,800 task observations of these teams. The average size of these teams is 3.5, with the average proportion of women being 0.3. Summary statistics for these teams and task observations are given in Appendix Table B.5.

Figure 3 shows a binscatter plot of task performance and the proportion of females in these larger teams. 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

[17] Qualitatively similar results are found from other methods of dividing teams into ability categories, such as using the ability quintile of the lowest-ability member of the team.

procedure.[18] These non-parametic methods suggest a positive effect of the proportion of females in the team, except approximately between a proportion of 0.3 and 0.7, where the function is approximately flat. Keeping in mind the low sample size relatively discrete nature of the underlying proportion values, we take this as suggestive evidence of a gender composition effect also in the larger teams.

We also investigate the gender composition effect in larger teams 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 teams. However, because the $MixedTeam$ indicator encompasses a wide range of teams with different proportions of female members, we also estimate specifications where $MixedTeam$ and $AllWomen$ are replaced with indicators of quartiles of the proportion of women in the teams, with 1^{st} quartile as reference category.[19] We also adjust our ability controls in order to account for the larger and variable team size using two different approaches: 1) we control for the average ability of all team members; 2) we control separately for ability quintile of the best and the worst member of the team (ignoring all other team members), as in Equation (2). In all specifications we control for team size. In order to maximize the number of observations in each regression, we do not restrict the sample to be the same across all specifications.[20]

The results of these regressions are shown in Table 5. Columns 1, 2, 3 give the results of Equation (1), Equation (2), and Equation (3), while columns 4, 5, and 6 repeat these specifications with the addition of indicators of quartiles of the proportion of women in the team. Columns 1 and 3 reveal a similar pattern in larger teams to those found in the pairs results; gender mixed and all-women teams tend to do significantly better than all-male teams. This is confirmed by the results in columns 4 to 6, which show that teams in the 3^{rd} and 4^{th} quintile of the proportion of women in the sample significantly outperform teams with no women - those in the 1^{st} quintile.

These findings are suggestive of a similar pattern to that found in Figure 3, whereby the performance gains are only experienced in teams in which the proportion of women reaches

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

[19] The average proportion of women in each quintile category are 0.00 (i.e. all-male groups), 0.21, 0.33, 0.60, and 0.91 respectively.

[20] Some teams have missing ability data. In order to maximize sample size per regression we do not drop these teams from the regressions excluding ability controls.

an adequate proportion. While the results of this analysis in this subsection should be regarded as suggestive only due to the small sample size, we nevertheless take them as evidence that the gender composition patterns observed in pairs also appear to carry over to larger teams of sizes 3 to 6.

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 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 on performance even in the absence of any such effect on team-level dynamics or processes; they could simply reflect the fact that female students are better than male students in our sample.

Although Appendix A.2 gives corroborating evidence that our preferred individual task ability measure is the most suitable measure for capturing possible gender differences in ability between students, we now explore the robustness of our results to the use of alternative ability controls; namely highschool GPA and university GPA in previous courses.[21] We do so by estimating our baseline specifications with (combinations of) these alternative ability measures and our preferred measure.

Columns 1-6 of Table 6 show estimates of β_1 and β_2 as the best and worst member quintiles are added for the individual task ability measure, university GPA, and highschool GPA.[22] For instance, column 6 shows the results of a specification that includes quintiles for the best and worst team member for all three ability measures, resulting in 15 separate indicators of ability composition of the team. While the point estimates vary somewhat, the results remain qualitatively similar to the baseline results.

Columns 7-12 of Table 6 repeat this exercise using the combination ability control method (Equation (3)) with different combinations of the ability measures. The most complete is shown in column 12, which includes 43 dummy variables controlling for the ability com-

[21] See Section 2 for definitions of these variables.

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

binations of the best and worst member according the three ability controls. Again, these various combinations do not change our baseline results. We take this 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 team composition effects is that they are driven, not (fully) by gender, but by some other characteristic that happens to correlate with gender in our sample. For instance, if most women in our sample are non-Dutch, then there may be some nationality compositional effect driving our results.

While we are not able to rule out all potential unobserved correlates of gender, our student data allows us to control for many 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 effect remain with the inclusion of these controls, this would provide further evidence that it indeed captures a gender composition effect.

We repeat Equation (3), the specification with the most demanding ability controls, while also controlling for the number of team 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 are first- or second-generation immigrants, or instead have no immigrant background. Specifically, we control for each possible combination of Dutch nationality and Dutch immigration background within a team.

Table 7 gives the results. Column 1 repeats our baseline specification for comparison purposes. Columns 2-4 add the above characteristics as controls separately, while column 5 controls for all of them 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.

5.3. COVID-19 period Our sample period includes periods affected by the COVID-19 pandemic. Namely, education was moved online for blocks 4-5 of the 2019 cohort, all blocks of the 2020 cohort, and block 3 of the 2021 cohort. Given that the manner in which students conducted team work during these periods may have differed, we explore the change in the

gender composition effect during the COVID-19 affected blocks.

To do so, we estimated an extended version of Equation (3) including interaction effects between the gender composition variables and an indicator of whether or not the task was completed in a COVID-19 block. These interaction coefficients then give the change in gender composition coefficients during the COVID-19 periods. The results of these regressions, both with and without ability controls, are shown in Table B.6. They show no significant difference between the gender composition effect for team work done during COVID-19 blocks and otherwise.

5.4. Grader bias We make use of grades as a measure of performance in team work. The tasks performed by teams are graded by TAs under the guidance of rubrics provided by a senior lecturer. If the graders 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 students' work they are grading.

In a similar context, Feld et al. (2016) show 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 of the grader of each task in our sample, and estimate Equation (3) separately for male and female graders. Appendix Table B.8 gives the results of these regressions, where column 1 repeats the main specification for comparison, column 2 gives the results for assignments graded by a female tutor, and column 3 for a male tutor. 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. Hence, our results are not driven by the same-gender bias of graders.

6 Why Do Teams With More Women Do Better?

The results above show that teams 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, and is present across task type and importance. What then might be driving these differences?

To shed light on this question, we look to existing literature in economics, management, and small team 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 with our data.

Team Work Preferences, Atmosphere & Friendship Gender differences in skills and preferences for team work may lead to, broadly speaking, a better team atmosphere, levels of civility, and thus better outcomes in teams with more women. Previous research has proposed that so-called "interpersonal sensitivity" - the propensity to treat teammates with care and respect – is higher among 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 find mixed evidence regarding higher concentration of them among women than men. Such 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 teams with more women may be that they 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 by 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 team settings, this would lead to differences in performance at the team level by gender composition.

Conflict, Unity & Trust Team conflict has been identified as an important component driving team outcomes, although there remains some debate about the direction of its effect and 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 showing 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 teams (Rosenthal, 2000) suggests a positive effect of the presence of women in such teams on performance through decreased conflict levels.

Feedback, Monitoring & Decision-making Differences in internal organization and processes of teams, 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 teams, leading to differences in team performance. Research from the lab (Hannagan and Larimer, 2010), student teams (Fenwick and Neal, 2001), and political legislators (Rosenthal, 2000) suggests that teams with more women tend to employ more cooperative strategies when making decisions. Mutual monitoring – the practice of team members monitoring the effort and work of their teammates – has been studied as a means of addressing incentive and skirting problems in teams (Carpenter et al., 2006), and its 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 teams as an important determinant of team performance, with both experimental (Barr and Conlon, 1994) and theoretical work

(Robinson and Weldon, 1993) pointing to team feedback playing an important role. Other literature argues that female-majority teams may be more receptive to feedback than male-majority teams (Karakowsky and Miller, 2002). Experimental research on student teams also shows significant differences in the levels and types of within-team communication by gender (Hardt et al., 2024); all-women teams appear to communicate the least, with all-male teams communicating the most. This may signal different internal team processes across gender compositions.

Leadership Style Leadership structures may differ between teams depending on gender composition. Research suggests certain individuals possess managerial or leadership qualities that can boost team performance (Weidmann et al., 2024). If these traits are unequally concentrated by gender, this in turn may lead to a gender composition effect. Women are less likely to appear in leadership roles within teams 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 teams. Moreover, some evidence points to different leadership styles between men and women when they are leaders, with women tending to adopt more democratic leadership styles as opposed to more autocratic styles (Eagly and Johnson, 1990).

6.1. Self-reflection task To investigate these possible explanations for the gender effect we exploit data from a comprehensive self-reflection exercise introduced at the end of blocks 3 and 5 of the 2021 cohort. This exercise was designed to help students reflect on their team work experience in the respective block (throughout which they worked on the tasks analysed above, within the same team). It prompts students with questions relating to the explanations above. This exercise was completed by individual students, rather than at the team level.

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 of them.

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 total, we record approximately 1,600 student responses across blocks 3 and 5 of the 2021 cohort. This equates to 83% of the relevant potential sample. Panel A of Appendix Table B.9 shows that neither own gender, nor partner gender, nor the interaction of own and partner gender, significantly affect the probability of answering the self-reflection exercise. Panel B of Appendix Table B.9 shows the gender composition effect on performance for this sample. Columns 3 and 4 show the gender composition effect for all assignments completed in blocks 3 and 5 by the 2021 cohort, while columns 5 and 6 give the effect for teams in which both members responded to the self-reflection exercise. With the addition of ability controls the effect is not significant for this latter team-task sample, which we believe reflects the fact the regression is based on only 15% of our main sample. Despite this, we believe that patterns in the self-reflection exercise may provide insights into the potential drivers of the gender composition effect.

The outcomes are either measured directly through single questions, or 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, questions and the construction of the components is given in Appendix A.3.

6.2. Team-level analysis of self-reflection exercise outcomes We begin by analysing the self-reflection exercise outcomes listed in Table 8 at the team level, in a similar manner to Equation (3). To explore differences at the team level we first calculate the team-average of each outcome. We then regress each team-average outcome on dummy variables reflecting the gender composition of the team, as well as controls for classroom-times-block effects and the ability composition of the team:

$$\begin{aligned} AvgOutcome_{rgb} &= \delta_0 + \delta_1 Mixed_r + \delta_2 AllWomen_r + ClassBlock_{gb} \\ &+ \sum_{q=1}^{5} \sum_{p=1}^{q} \theta_{q,p} \mathbb{1}\Big(AbilityQuintile_r^{Best} = q, AbilityQuintile_r^{Worst} = p \Big) + \epsilon_{rgb} \end{aligned} \tag{4}$$

Where $AvgOutcome_{rgb}$ is the average of some outcome for team r, which is a member of classroom g in block b. Coefficients δ_1 and δ_2 then give the average difference in the (team-average) response for their respective variables, compared to all-male teams. The regression also flexibly controls for combinations of the ability quintile of both the best and

worst individual in the pair, and any classroom-times-block effects.

Team-level results The estimates for the δ_1 (*Mixed Team*) and δ_2 (*All Women*) coefficients from Equation (4) for each of the 22 outcomes across the 5 categories are shown visually in Figure 4.[23] The *Mixed Team* and *All Women* coefficients show the average difference, compared to all-male, in mixed and all-female teams respectively, in the team-level average per outcome.

The results in Figure 4 show no statistically significant difference between all-men and all-women teams for any of the team-level averaged outcomes. Rather, the largest differences are found in mixed teams. Compared to all-male teams, mixed teams report less familiarity (both before and after working together), worse team atmosphere (measured directly and via multiple items), lower levels of own and teammate motivation, lower levels of team contributions, and lower levels of unity. Thus, despite the increase in performance shown in Section 4, mixed teams appear to do worse compared to all-male teams along many dimensions of team work experience, such as atmosphere. We now turn to the individual-level responses to identify the source of these differences.

6.3. Individual-level analysis of self-reflection exercise outcomes The self-reflection exercise outcomes are elicited at the level of the individual. This allows us to estimate individual-level models, where we distinguish between the gender of both the respondent and their partner. For each outcome in Table 8 we estimate the following specification:

$$Outcome_{ijb} = \gamma_0 + \gamma_1 FemaleTeammate_j + \gamma_2 Woman_i + \gamma_3 \Big(FemaleTeammate_j \times Woman_i \Big) + \sum_{t=1}^{5} \sum_{p=1}^{5} \theta_{t,p} \mathbb{1} \Big(AbilityQuintile_i = t, AbilityQuintile_j = p \Big) + ClassBlock_{ijb} + \epsilon_{ijb} \tag{5}$$

Where $Outcome_{ijb}$ refers to some outcome of the self-reflection exercise for respondent i, who is allocated partner j, in block b. Coefficient γ_1 (γ_2) then gives the average difference in $Outcome$ for a man allocated a female teammate (woman allocated a male teammate) compared to the reference category of men allocated a male partner. The sum of γ_1, γ_2, and γ_3 gives the estimated difference between this reference category and women allocated

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

a female partner. Thus, using these estimates we can calculate the estimated difference between the reference category respondent type (men allocated a male partner) and the three remaining respondent types: men allocated a female partner, women allocated a male partner, and women allocated a female partner. The specification also flexibly controls for combinations of the ability quintile of the respondent and their teammate, as well as any class-block fixed effects.

Individual-level results Coefficient estimates from Equation (5) for the three different respondent types, compared to the reference category (man working with another man), are shown in Figure 5 for each outcome of Table 8. The regression results underlying these plots are shown in Appendix Table B.11.

Consistent with the previous results, both men and women in mixed teams report lower levels of familiarity, team atmosphere, own motivation, and team contributions.[24] Women within these teams are notable in several respects; compared to the reference category, these women report having less motivated partners, lower levels of feedback, and higher levels of monitoring within their team. Significant results on these outcomes for men within mixed teams are absent, although these men report significantly less team trust. Women, regardless of allocated partner gender, report higher levels of team work preferences, as well as more hours per week of work.

One explanation for these patterns is that women may be more conscientious and diligent team members than men (conditional on each individual's task ability, comprised of both individuals' cognitive and non-cognitive skills), leading to improved team performance. This is consistent with several findings from the self-reflection exercise. Women have higher preferences for teamwork and cooperation, rather than working alone, and spend more hours on the teamwork assignments than men, regardless of the gender of their assigned partner.

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 team work in teams with a higher fraction of women. They attribute this find-

[24] It should be noted however that, while members of mixed teams do report significantly worse outcomes along some dimensions, the absolute scores remain relatively high. For instance, the (unadjusted) average rating of team atmosphere among women and men in mixed teams is 3.97 and 4.08 respectively, out of a maximum of 5. The same values for women and men in non-mixed teams are 4.07 and 4.14 respectively.

ing to higher levels of "social-sensibility". Using data on student teams with randomly allocated leaders, De Paola et al. (2022) find that teams led by women tend to outperform those headed by men. They speculate that traits like conscientiousness and readiness to collaborate may lead to higher levels of team performance. Keck and Tang (2018) show that teams 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 team decisions.

A mismatch between team work preferences and diligence may also explain the poor outcomes in mixed teams regarding team atmosphere, motivation, and unity, and is consistent with women allocated a male partner reporting worse *Team Contribution* PCA scores, lowers levels of team feedback, and higher levels of teammate monitoring. In other words, women reporting having to monitor male partners more closely, and receiving less feedback on their work from male partner. This suggests that teams with women perform better as they are more diligent and harder-working team members, but that this comes at a cost along other dimensions when this diligence is not reciprocated.

This finding also squares with previous literature on the gender allocation of menial tasks in teams. Using pairs of students, Babcock et al. (2017) show in a laboratory setting that women in mixed-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 mixed-gender teams. In our setting, the poorer team experiences reported by women in mixed teams may be partly the result of them being burdened with more tasks.

While these assertions are speculative, 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 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 mismatch 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, and unity in mixed teams.

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, as measured by task grades. Using 12,600 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 (and combinations) 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. 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 of higher and lower stakes, and is also present in teams larger than two.

The findings on performance are in line with earlier research documenting a positive impact of women in teams in laboratory settings (Woolley et al., 2010; De Paola et al., 2022; Keck and Tang, 2018; Fenwick and Neal, 2001; Hoogendoorn et al., 2013). Although our use of student teams producing work for an academic course limits our ability to speak confidently about work teams in other settings, we note that the patterns of results we find is similar to those found using workplace and research teams (Yang et al., 2022; Hengel, 2020; Hengel and Moon, 2023), and that – as shown in Appendix A.1 – the types of tasks performed in our context have a large overlap with many workplaces.

Using data from a self-reflection exercise on the team work, we are able to shed further light on possible mechanisms behind the performance findings. We find a higher preference for cooperation for women and that they exert higher effort levels in team work than men. However, in contrast to the ranking of teams by performance, where mixed and all-women teams do better than all-male teams, the self-reflection exercise shows that – along many dimensions, such as team contributions, atmosphere, motivation, and unity – mixed-teams tended to report worse outcomes than homogeneous ones. The findings fit a scenario where women's higher team-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 atmo-

sphere, motivation, and unity.

These results highlight potential trade-offs between team work experiences and team performance. While mixed-gender teams produce significantly better quality work than all-male teams, members of those teams reported the worst team experience outcomes on, for instance, team atmosphere. Consequently, while the performance of such teams is higher, their long-term sustainability may be impaired.

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 team atmosphere, and challenge 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.

De Paola, M., R. Lombardo, V. Pupo, and V. Scoppa (2021). Do women shy away from public speaking? a field experiment. *Labour Economics 70*, 102001.

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.

Hardt, D., L. Mayer, and J. Rincke (2024). Who does the talking here? the impact of gender composition on team interactions. *Management Science*.

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.

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.

Weidmann, B., J. Vecci, F. Said, D. J. Deming, and S. R. Bhalotra (2024). How do you find a good manager? Technical report, National Bureau of Economic Research.

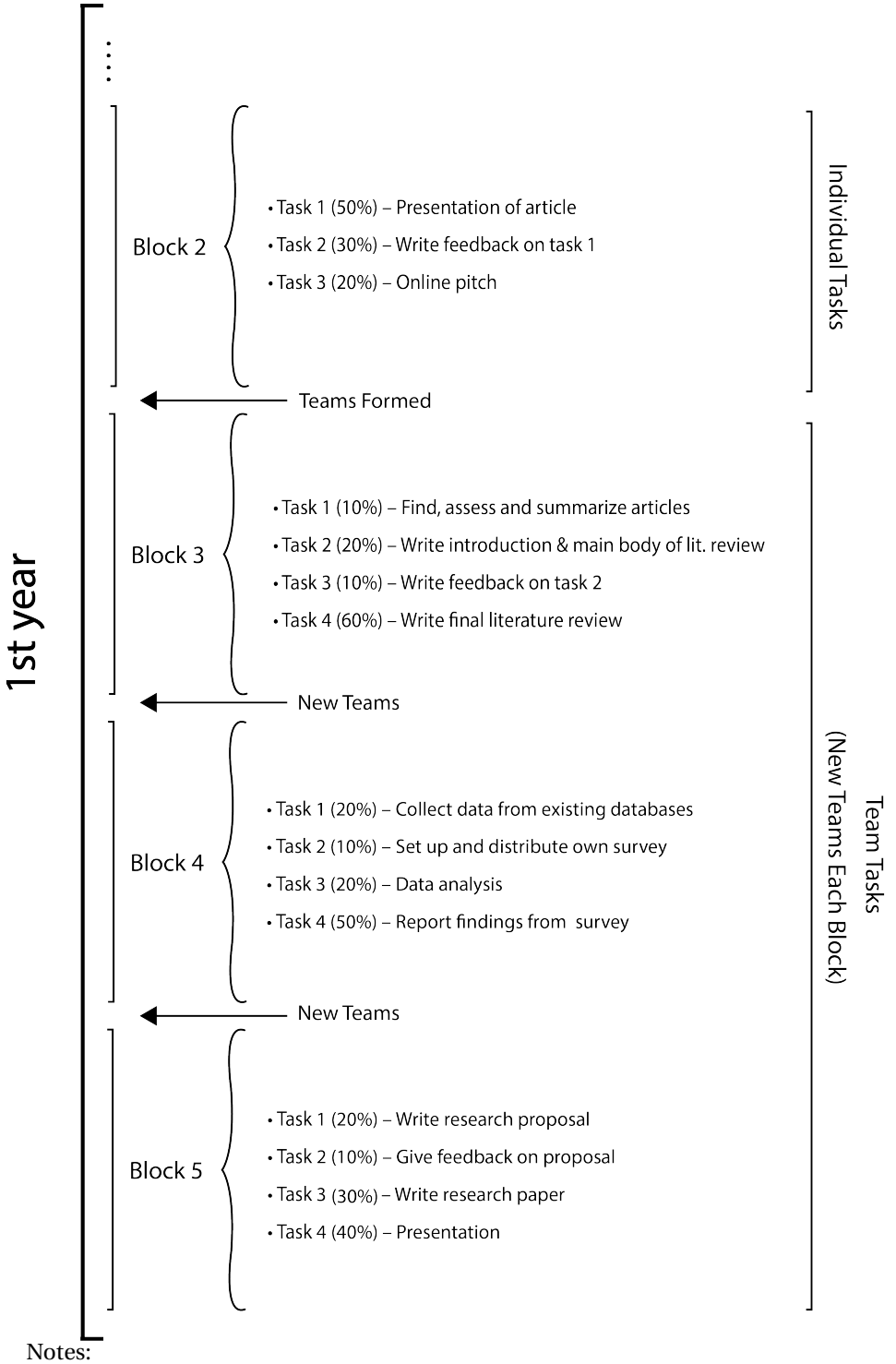
Williams, M. and E. Polman (2015). Is it me or her? how gender composition evokes interpersonally sensitive behavior on collaborative cross-boundary projects. *Organization Science 26*(2), 334–355.

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.

Tables and Figures

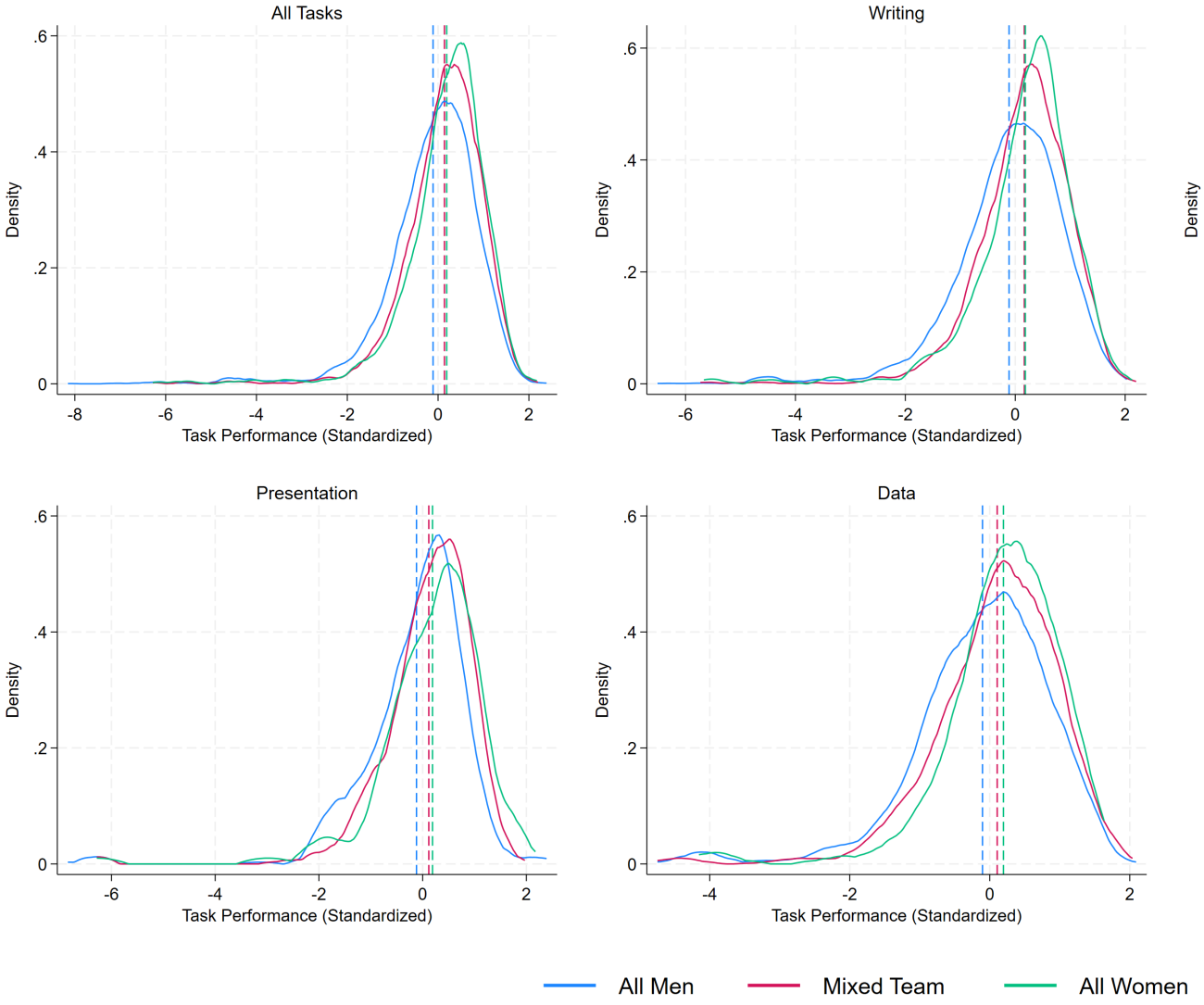
Figure 1: Structure of Data Collection and Outcomes



Notes:

1. Figure shows the structure of the course and the weight of each task per block.
2. Note that team work starts from block 3 onwards, and new teams are formed each block.

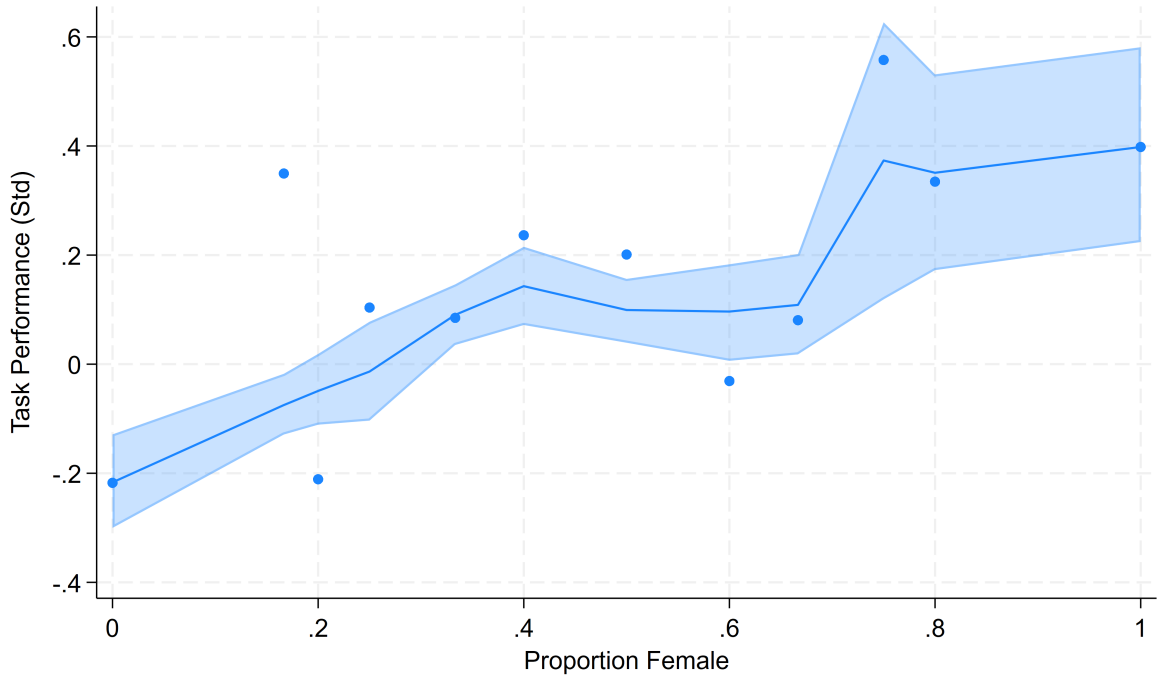
Figure 2: Task Histograms



Notes:

1. Figures show histograms of the standardised performance 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 3: Results for Larger Teams



Notes:

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

Table 1: Descriptive Statistics

	Mean	SD	Observations
Student Data			
Number of students			2,984
Female	0.307	(0.461)	2,984
Task ability measure	74.86	(18.323)	2,984
High school GPA	6.930	(0.596)	1,746
University GPA	6.806	(0.996)	2,984
Non-Dutch	0.390	(0.488)	2,984
Dutch (Non-Immigrant)	0.465	(0.499)	2,984
Dutch (Immigrant)	0.144	(0.351)	2,984
Both parents university	0.483	(0.500)	2,605
Team Data			
Number of teams			3,247
Number of teams in 2018 cohort			529
Number of teams in 2019 cohort			818
Number of teams in 2020 cohort			866
Number of teams in 2021 cohort			1,034
All men	0.476	(0.499)	3,247
Mixed	0.424	(0.494)	3,247
All women	0.100	(0.300)	3,247
Task Data			
Average task performance	72.76	(14.76)	12,631
Average task performance Writing	71.27	(14.17)	6,454
Average task performance Data	67.28	(14.68)	2,731
Average task performance Presentation	76.32	(11.40)	1,113
Average task performance Feedback	81.61	(13.57)	2,333

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: Estimated Effects of Gender Composition on Team Performance - Baseline Results

	(1)	(2)	(3)
Mixed Team	0.193***	0.153***	0.154***
	(0.0295)	(0.0277)	(0.0273)
All Women	0.253***	0.167***	0.165***
	(0.0498)	(0.0481)	(0.0479)
Best/Worst Ability Controls		✓	
Ability Combinations Controls			✓
Mixed Team=All Women			
F-statistic	1.724	0.101	0.059
p-value	0.190	0.751	0.808
Observations	12,631	12,631	12,631
R^2	0.231	0.258	0.257

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), and Equation (3) on the team-task data in Table 1.

Table 3: Estimated Effects of Gender Composition on Team Performance - Results Per Task Type and Importance

	Data	Feedback	Presentation
	(1)	(2)	(3)
Mixed Teams	0.162***	0.118***	0.183***
	(0.0448)	(0.0393)	(0.0601)
All Women	0.104	0.177**	0.202*
	(0.0782)	(0.0701)	(0.118)
Ability Combination Controls	✓	✓	✓
Observations	2,731	2,316	1,108
R^2	0.282	0.417	0.394
	Writing	Low Importance	High Importance
	(4)	(5)	(6)
Mixed Teams	0.160***	0.150***	0.174***
	(0.0337)	(0.0267)	(0.0432)
All Women	0.162***	0.156***	0.176**
	(0.0587)	(0.0461)	(0.0714)
Ability Combination Controls	✓	✓	✓
Observations	6,451	10,501	2,116
R^2	0.322	0.260	0.402

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 (3) on the team-task data in Table 1 per type of task and grade importance of task. High importance tasks are those that count for at least 50% of the grade in a given block.

Table 4: Estimated Effects of Gender Composition on Team Performance - Results by Team Ability Quintile

	Average Group Ability Quintile				
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
	(1)	(2)	(3)	(4)	(5)
Mixed Team	0.317***	0.269***	0.214**	0.177**	0.0535
	(0.111)	(0.0988)	(0.0839)	(0.0756)	(0.0711)
All Women	0.350	0.258	0.247**	0.356***	0.0584
	(0.267)	(0.160)	(0.122)	(0.125)	(0.0898)
Ability Combinations Controls	✓	✓	✓	✓	✓
Observations	2,574	2,531	2,523	2,536	2,461
R^2	0.440	0.392	0.433	0.384	0.324

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 5: Estimated Effects of Gender Composition on Team Performance - Effects for Larger Groups

	(1)	(2)	(3)	(4)	(5)	(6)
Mixed Team	0.215*	0.177	0.221*			
	(0.114)	(0.114)	(0.129)			
All Women	0.740***	0.723***	1.036***			
	(0.244)	(0.242)	(0.338)			
2^{nd} Quartile Female Prop.				0.114	0.0531	0.102
				(0.119)	(0.124)	(0.146)
3^{rd} Quartile Female Prop.				0.354**	0.320**	0.320**
				(0.139)	(0.143)	(0.131)
4^{th} Quartile Female Prop.				0.311*	0.291*	0.340*
				(0.168)	(0.165)	(0.186)
Group Ability Average		✓			✓	
Best/Worst Ability Quintiles			✓			✓
Observations	1,849	1,775	1,719	1,849	1,775	1,719
R^2	0.400	0.402	0.427	0.402	0.404	0.427

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 a more detailed specification using quartiles of proportion on teams larger than 2. The summary statistics of this data are shown in Table B.5.

Table 6: Estimated Effects of Gender Composition on Team Performance - Extended Ability Specifications

	(1)	(2)	(3)	(4)	(5)	(6)
			Best/Worst Ability Quintiles			
Mixed Team	0.193***	0.198***	0.169***	0.211***	0.164***	0.193***
	(0.0295)	(0.0265)	(0.0255)	(0.0427)	(0.0412)	(0.0385)
All Women	0.253***	0.251***	0.191***	0.239***	0.173**	0.210**
	(0.0498)	(0.0451)	(0.0450)	(0.0833)	(0.0835)	(0.0790)
Best/Worst Uni. GPA Quint.		✓	✓			✓
Best/Worst Task Ability Quint.			✓		✓	✓
Best/Worst HS GPA Quint.				✓	✓	✓
Observations	12,631	12,631	12,631	6,238	6,238	6,238
R^2	0.231	0.277	0.289	0.253	0.272	0.287
	(7)	(8)	(9)	(10)	(11)	(12)
			Best/Worst Ability Quintile Combinations			
Mixed Team	0.193***	0.195***	0.166***	0.208***	0.166***	0.195***
	(0.0295)	(0.0266)	(0.0250)	(0.0417)	(0.0406)	(0.0379)
All Women	0.253***	0.249***	0.186***	0.230***	0.144*	0.186**
	(0.0498)	(0.0447)	(0.0442)	(0.0836)	(0.0838)	(0.0795)
Uni. GPA Quint. Comb.		✓	✓			✓
Task Ability Comb.			✓		✓	✓
HS GPA Quint. Comb.				✓	✓	✓
Observations	12,631	12,631	12,631	6,238	6,238	6,238
R^2	0.231	0.277	0.291	0.255	0.274	0.290

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 available for Dutch students.

Table 7: Estimated Effects of Gender Composition on Team Performance - Controlling for Other Characteristics

	(1)	(2)	(3)	(4)	(5)
Mixed Team	0.154^{***}	0.151^{***}	0.156^{***}	0.156^{***}	0.153^{***}
	(0.0273)	(0.0276)	(0.0271)	(0.0272)	(0.0276)
All Women	0.165^{***}	0.159^{***}	0.161^{***}	0.161^{***}	0.156^{***}
	(0.0479)	(0.0486)	(0.0484)	(0.0483)	(0.0490)
Ability Combinations Control	✓	✓	✓	✓	✓
Parent Uni. Count Controls		✓			✓
Non-Dutch Count Controls			✓		
Ethnicity × Nationality Controls				✓	✓
Observations	12,631	12,350	12,631	12,631	12,350
R^2	0.257	0.258	0.258	0.258	0.259

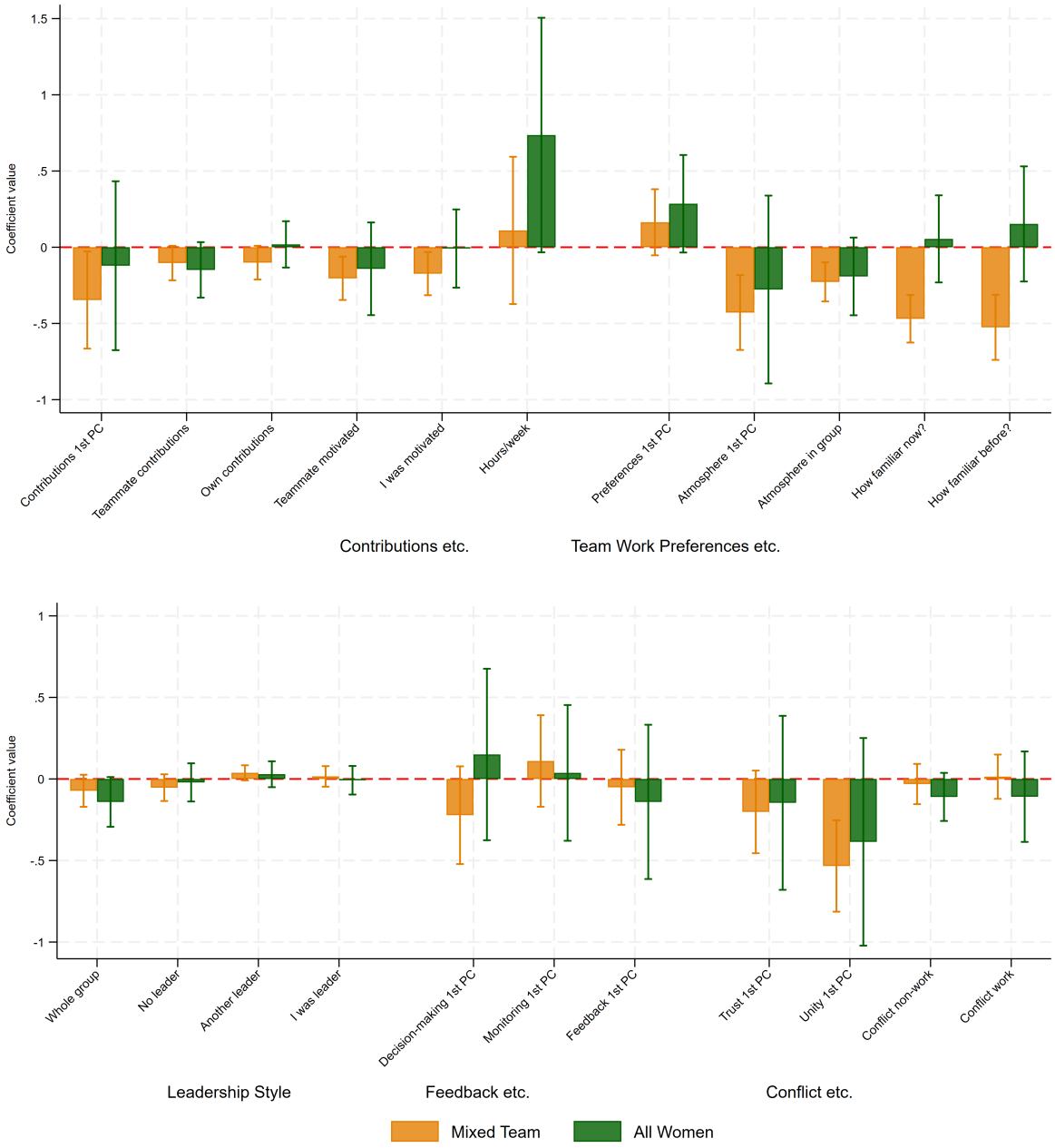
1. Standard errors in parentheses, clustered on the classroom group level.
2. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
3. Table shows results of estimating Equation (3) with the addition of controls for other group characteristics: parental university attendance, non-Dutch nationality, Dutch ethnicity, and a combination of Dutch ethnicity and nationality.

Table 8: Summary Statistics Team Self-Reflection Exercise Outcomes

	Mean	SD	Min	Max	Observations
Team Atmosphere, Friendship, & Work Preferences					
How familiar before?	2.132	(1.054)	1	5	1,665
How familiar now?	2.914	(0.925)	1	5	1,665
Atmosphere within group?	4.070	(0.801)	1	5	1,665
Team Atmosphere PCA	0.000	(1.727)	-6.05	2.47	1,665
Team Work Preferences PCA	0.000	(1.277)	-5.22	3.08	1,665
Contributions, Effort, & Motivation					
Hours/week spent on course?	5.968	(3.193)	0	20	1,665
Own motivation to work with team	3.702	(0.901)	1	5	1,662
Partner's motivation to work with team	3.708	(0.942)	1	5	1,662
Rating of own contribution to group	4.091	(0.669)	1	5	1,665
Rating partner's contribution to group	3.985	(0.837)	1	5	1,665
Team Contributions PCA	0.000	(1.683)	-5.06	2.59	1,665
Conflicts, Unity & Trust					
Extent of conflict about group work?	1.674	(0.820)	1	5	1,665
Extent of conflict about other matters?	1.417	(0.753)	1	5	1,665
Team Unity PCA	0.000	(1.878)	-8.38	2.66	1,665
Team Trust PCA	0.000	(1.634)	-7.79	2.79	1,665
Feedback, Monitoring & Decision Making					
Team Feedback PCA	0.000	(1.607)	-5.32	2.95	1,665
Team Monitoring PCA	0.000	(1.634)	-6.85	2.90	1,665
Team Decision Making PCA	0.000	(1.784)	-8.01	3.58	1,665
Leadership Styles					
I was leader	0.220	(0.414)	0	1	1,665
Another member was leader	0.102	(0.303)	0	1	1,665
No leader	0.678	(0.467)	0	1	1,665
Worked as group rather than as individuals	0.411	(0.492)	0	1	1,665
Worked as individuals	0.589	(0.492)	0	1	1,665

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

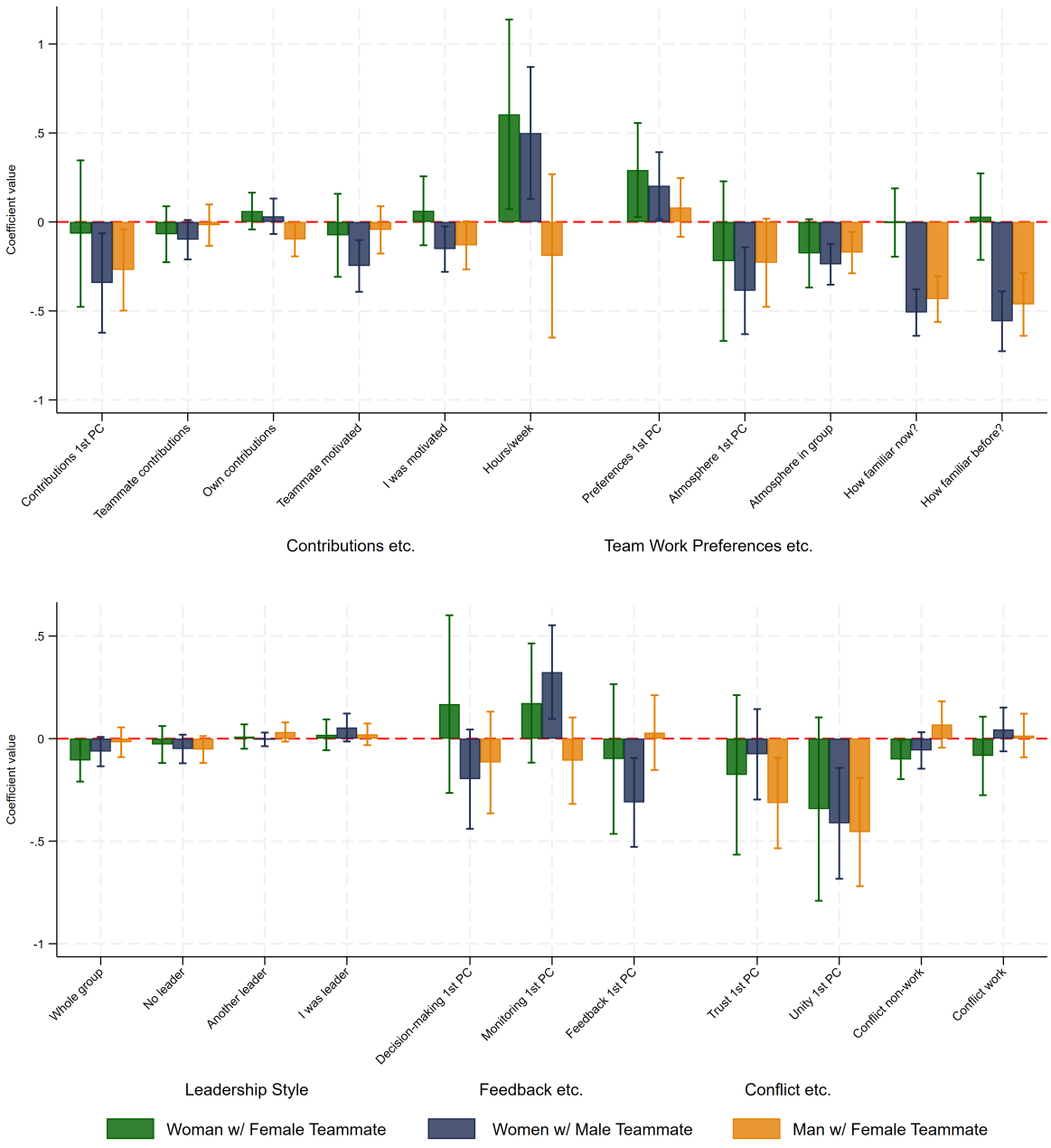
Figure 4: Group Level Differences in Self-Reflection Exercise Outcomes



Notes:

1. Figures shows the estimated effect for different group types, compared to the reference category of all-male groups, on the outcomes shown in Table 8.
2. Results are derived from Equation (4), where δ_1 (δ_2) gives the estimated difference in the average group response for mixed (all women) teams, compared to the reference category of all men teams.
3. 90% confidence intervals are shown.

Figure 5: Individual Differences in Self-Reflection Exercise Outcomes



Notes:

1. Figures shows the estimated effect for different respondents, compared to the reference category of men allocated a male partner, on the outcomes shown in Table 8.
2. Results are derived from Equation (5). The effects for men allocated a female partner are estimates of γ_1, the effect for women allocated a male partner are estimates of γ_2, and the effects of women allocated a female partner are are estimates of $\gamma_1 + \gamma_2 + \gamma_3$.
3. 90% confidence intervals are shown.

Gender and Performance in Teams: Evidence From Students

Online Appendix

Max Coveney Teresa Bago d'Uva Pilar García-Gómez

April 28, 2025

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 teams 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 *Accountants and Auditors*. One key task listed for such workers is to "*Prepare detailed reports on audit findings.*" Given the results in Figure B.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 in our data (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.4.

Appendix Figure B.3 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.4 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:

$$CourseGrade_{iscb} = \theta_0 + \theta_1 Female_i + f(Ability_i) + ClassBlock_{cb} + \epsilon_{iscb} \quad (6)$$

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 team. 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.7 shows the results of running regression Equation (6) 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 8% 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 2% 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 0.45% of a standard deviation. This difference is highly statistically insignificant.

The results of Appendix Table B.7 gives further rationale for our use of the Task Ability Measure quintiles as an ability control in the team 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. Self-reflection exercise outcomes We use both items derived directly from single questions, as well as instruments consisting of multiple items to measure a particular underlying construct. For each instrument, we combine the various items through a Principal Component Analysis (PCA), from which we extract the first principal component. The full self-reflection exercise is shown in Table B.1, 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 are given in Table B.2. [1]

After construction of the principal components, we are left with 22 different outcomes. Across the two blocks we have data on the self-reflection exercise for 1,314 (83% of) students. For 96% of groups across blocks there is data for at least one group member.[2] Summary statistics for the 22 outcomes are given in Table 8.

Below, for each category of explanation for the gender effect, we describe the measures shown in Table 8 and how they were obtained from the self-reflection exercise data.

Team Work Preferences, Atmosphere & Friendship We use several holistic measures of preferences for group work, the levels of friendship, and the overall group atmosphere within a group. Students are asked to rate how familiar they were with their teammate on a scale between 1 ("Strangers") and 5 ("Best friends") both before and after the group work. To measure the general atmosphere within the team, they rate the atmosphere within their group between 1 ("Very bad") and 5 ("Very good"). Team atmosphere can also be measured indirectly through the combination of four items, making up *Team Work Atmosphere*. These

[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.

[2] Appendix Table B.9 shows that the treatment - partner gender - has no impact on the probability of completing the self-reflection exercise for the pooled sample nor for men or women separately.

items relate to individual's agreement with statements relating to their satisfaction and enjoyment working in the group, and willingness to do so again.[3] In order to measure 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 Students report how many hours per week, on average, they spent on the group work. They also 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 with which 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 The self-reflection exercise also asks 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 on team loyalty, responsibility taking, and shared assistance. *Team Trust* is the first principal component of five items regarding 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 of four items on the degree of feedback and revisions given by and to team members. *Team Monitoring* is the first principal component of four items on 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 regarding the degree to which decision were made in a collaborative, constructive, and safe environment.

Leadership Style Students report 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.

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

Table B.1: 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.2: Team Self-Reflection Assignment Principal Component Results

	Loading	Eigen-value	Proportion
Team Work Preferences		1.63105	0.4078
I like to work with other people.	0.5916		
Cooperation is preferable to competition.	0.5544		
I consider myself to be a competitive person.	-0.2491		
Work assignments are better when I do them myself.	-0.5297		
Team Atmosphere		2.98351	0.7459
In general, I am satisfied with the work of my group.	0.4838		
I enjoyed working with my group.	0.5184		
Working in this group was frustrating.	-0.4866		
I want to work with this group in the future.	0.5103		
Team Trust		2.67017	0.5340
I did not have difficulties accepting suggestions from my fellow group member	0.3709		
I trusted the knowledge of my fellow group member about the group work was sufficient	0.5343		
I trusted the information that my fellow group member brought to the discussion	0.5264		
When my fellow group member gave information, I wanted to double-check this information	-0.2903		
I did not have much confidence in the expertise of my fellow group member.	-0.4644		
Team Unity		3.52672	0.7053
Our group was united in trying to reach its goals for performance.	0.4493		
In this group, we all took our responsibility for setbacks or poor group performance	0.4259		
We helped each other to complete group tasks.	0.4394		
We worked well together.	0.4695		
We were loyal to each other.	0.4507		
Team Feedback		2.5816	0.6454
I gave feedback on the work of my fellow group member	0.4937		
I made revisions to the work of my fellow group member	0.4622		
My fellow group member(s) gave feedback on my work.	0.5260		
My fellow group member(s) made revisions to my work.	0.5157		
Team Monitoring		2.67004	0.6675
We checked to make sure that everyone in the group continued to work on the assi	0.5055		
We monitored each other's progress on the assignments.	0.5191		
We checked whether everybody was meeting their obligations to the group.	0.5346		
We made sure that everyone in the group met their deadlines.	0.4350		
Team Decision Making		3.18193	0.4546
Decisions were mainly taken by one group member.	-0.3650		
Decisions were worked out together in this group.	0.4328		
Some members contributed less to decision-making than others.	-0.3670		
When deciding on the strategies, the opinion of all group members was actively a	0.3929		
Some group members pushed their opinion through without much regard of what the	-0.3589		
I felt safe sharing my opinion and ideas with the other group members.	0.3684		
We adhered to any assignment-related decisions we made together.	0.3549		
Team Contributions		2.8332	0.7083
All group members contributed to the assignments equally.	0.5283		
I had to do more than my fair share of work for the assignments.	-0.4702		
All group members put in the same effort for the assignments.	0.5189		
I experienced free-riding problems in my group.	-0.4801		

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.

B Online Appendix Tables and Figures

Table B.3: Balancing Tests

	(1) Female	(2) Ability (Continuous)	(3) High Ability (Dummy)	(4) Low Ability (Dummy)
T-statistic	0.141	0.170	0.003	-0.672
p-value	0.888	0.865	0.998	0.501
Observations	6,445	6,445	6,445	6,445
	(5) Non-Dutch	(6) Native Dutch	(7) Non-Native Dutch	(8) Both Parents University
T-statistic	0.694	0.787	0.691	0.547
p-value	0.488	0.431	0.489	0.584
Observations	6,445	6,445	6,445	5,688

1. Table shows the results of 8 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 the characteristic of a student and that of their allocated partner.
3. The observations are at the student-teammate-block level. The lower number of observations in column 8 is due to missing data for parental university attendance for some students.

Table B.4: 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.5: Descriptive Statistics - Larger Teams

	Mean	SD	Count
Team Data			
Number of teams			474
Number of teams 2018			206
Number of teams 2019			31
Number of teams 2020			110
Number of teams 2021			127
Team size	3.540	(0.852)	474
Proportion of females	0.313	(0.260)	474
All men	0.285	(0.452)	474
Mixed	0.690	(0.463)	474
All women	0.025	(0.157)	474
Task Data			
Average task grade	73.794	(12.622)	1,849
Average task grade Writing	72.301	(12.721)	740
Average task grade Data	72.318	(13.110)	708
Average task grade Presentation	76.285	(8.004)	163
Average task grade Feedback	81.119	(10.470)	238

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 4.4.
2. These teams consist of team sizes 3-6, either from block 4 of 2018 where larger sizes of teams were created, or teams of 3 created from leftover students within classrooms in the other blocks-years.
3. The smaller number of larger groups in 2019 is due to the fact that left-over students in the majority of blocks in the cohort were made to work alone, rather than form groups of 3.

Table B.6: Baseline Results with COVID-19 Interaction

	(1)	(2)
Mixed Group	0.184***	0.141***
	(0.0393)	(0.0384)
All Women	0.255***	0.161**
	(0.0626)	(0.0653)
Mixed Team	0.0168	0.0240
× COVID-19 Block	(0.0543)	(0.0528)
All Women	-0.00463	0.00713
× COVID-19 Block	(0.0970)	(0.0955)
Ability Combination Controls		✓
Observations	12,631	12,631
R^2	0.231	0.257

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 interaction dummies for whether the block was affected by COVID-19 measures.
4. COVID-19 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.7: Individual Course Results

	(1)	(2)	(3)	(4)
Female Student	0.0828***	0.0654***	0.0207	0.0045
	(0.0301)	(0.0215)	(0.0362)	(0.0289)
University GPA Quint.		✓		
Highschool GPA Quint.			✓	
Task Ability Measure Quint.				✓
Observations	14,296	14,296	8,517	14,296
R^2	0.142	0.336	0.261	0.174

1. Standard errors in parentheses.
2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.
3. Table shows results of estimating Equation (6) on the sample of individual grades including quintile dummy variables of the various ability measures. See Appendix A.2.

Table B.8: Gender Tutor Effects

	All Tutors	Female Tutor	Male Tutor
	(1)	(2)	(3)
Mixed Team	0.154***	0.136***	0.181***
	(0.0273)	(0.0342)	(0.0424)
All Women	0.165***	0.130**	0.209***
	(0.0479)	(0.0654)	(0.0678)
Ability Combination Controls	✓	✓	✓
Observations	12,631	7,412	5,107
R^2	0.257	0.257	0.258

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 results split by tutor gender.

Table B.9: Self-Reflection Exercise Response Student and Team Results

	Panel A: Student Data			
	(1)	(2)		
Female Teammate	-0.0204	-0.0050		
	(0.0253)	(0.0299)		
Woman	-0.0004	0.0150		
	(0.0253)	(0.0247)		
Woman × Female Teammate		-0.0451		
		(0.0425)		
Ability Combination Controls	✓	✓		
Observations	1,565	1,565		
R^2	0.172	0.173		
	Panel B: Team-Assignment Data			
	(3)	(4)	(5)	(6)
Mixed Team	0.188***	0.133**	0.157**	0.0772
	(0.0573)	(0.0539)	(0.0644)	(0.0598)
All Women	0.175*	0.110	0.112	-0.0211
	(0.0946)	(0.101)	(0.111)	(0.114)
Ability Combination Controls		✓		✓
Observations	2,409	2,409	1,792	1,792
R^2	0.184	0.213	0.207	0.252

1. Columns 1 and 2 show how the probability of answering the self-reflection exercise depends on own and partner gender. Columns 3 and 4 show the gender composition effect for the entire 2021 cohort for assignments in blocks 3 and 5, and columns 5 and 6 show the gender composition effect only for team-assignment observations where both team members responded to the self-reflection exercise.
2. Standard errors in parentheses.
3. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.10: Team Self-reflection Assignment - Team Level 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
Mixed Team	-0.521***	-0.463***	-0.234***	-0.435***	0.185	0.129	-0.178**	-0.213**	-0.106	-0.110	-0.352*
	(0.133)	(0.0947)	(0.0804)	(0.153)	(0.126)	(0.297)	(0.0876)	(0.0870)	(0.0668)	(0.0701)	(0.196)
All Women	0.175	0.0778	-0.190	-0.262	0.275	0.670	-0.0161	-0.128	0.0124	-0.150	-0.0665
	(0.232)	(0.170)	(0.156)	(0.369)	(0.184)	(0.473)	(0.158)	(0.185)	(0.0863)	(0.111)	(0.327)
Observations	515	515	515	515	515	515	513	513	515	515	515
R^2	0.377	0.413	0.331	0.350	0.366	0.406	0.382	0.374	0.417	0.395	0.314
	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
Mixed Team	0.0362	-0.0265	-0.533***	-0.212	-0.0662	0.120	-0.213	0.0121	0.0370	-0.0490	-0.0815
	(0.0841)	(0.0758)	(0.175)	(0.157)	(0.145)	(0.170)	(0.187)	(0.0380)	(0.0281)	(0.0494)	(0.0593)
All Women	-0.0885	-0.102	-0.332	-0.173	-0.120	0.118	0.127	-0.000995	0.0265	-0.0255	-0.131
	(0.171)	(0.0922)	(0.386)	(0.322)	(0.285)	(0.245)	(0.315)	(0.0541)	(0.0476)	(0.0716)	(0.0910)
Observations	515	515	515	515	515	515	515	515	515	515	515
R^2	0.343	0.306	0.359	0.345	0.370	0.391	0.327	0.300	0.329	0.312	0.313

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 (4) on the various self-reflection exercise outcomes, shown in Table 8.

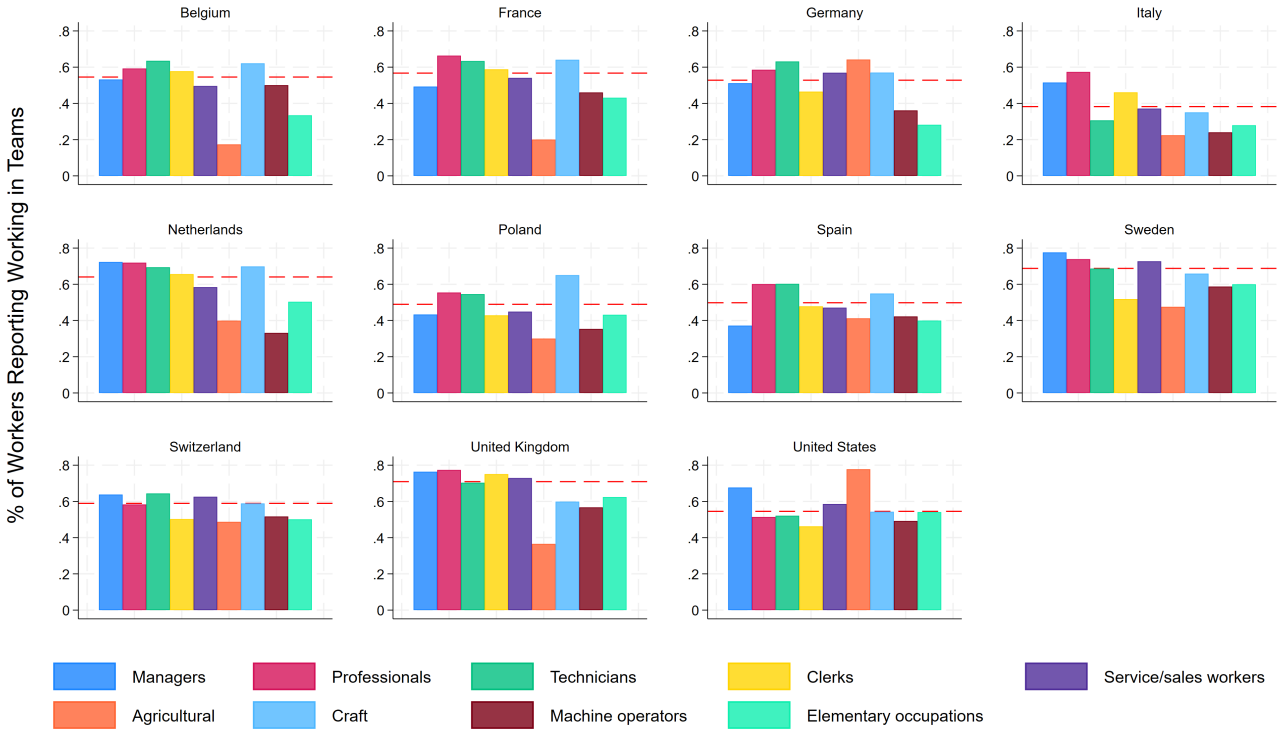
Table B.11: Team Self-reflection Exercise Regression Results

	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	-0.463***	-0.433***	-0.172**	-0.230	0.0814	-0.191	-0.131	-0.0446	-0.0983*	-0.0179	-0.270*
	(0.106)	(0.0773)	(0.0700)	(0.149)	(0.0991)	(0.276)	(0.0812)	(0.0799)	(0.0572)	(0.0698)	(0.137)
Woman	-0.558***	-0.509***	-0.238***	-0.387***	0.204*	0.500**	-0.153*	-0.247***	0.0318	-0.0999	-0.343**
	(0.101)	(0.0785)	(0.0688)	(0.146)	(0.113)	(0.223)	(0.0768)	(0.0873)	(0.0598)	(0.0666)	(0.168)
Woman x Female Teammate	1.052***	0.939***	0.234*	0.396	0.00685	0.296	0.347**	0.217	0.128	0.0485	0.548**
	(0.195)	(0.158)	(0.124)	(0.288)	(0.200)	(0.403)	(0.132)	(0.158)	(0.0902)	(0.112)	(0.238)
Women w/ Female Effect	0.0299	-0.00332	-0.177	-0.220	0.292	0.605	0.0627	-0.0753	0.0612	-0.0694	-0.0657
F-test	0.0421	0.000829	2.351	0.669	3.382	3.575	0.290	0.289	0.974	0.540	0.0707
p-value	0.838	0.977	0.129	0.416	0.0698	0.0625	0.592	0.593	0.327	0.465	0.791
Observations	1,314	1,314	1,314	1,314	1,314	1,314	1,311	1,311	1,314	1,314	1,314
R^2	0.268	0.259	0.187	0.197	0.180	0.225	0.195	0.193	0.204	0.189	0.176

	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	0.0151	0.0688	-0.456***	-0.314**	0.0292	-0.108	-0.116	0.0209	0.0322	-0.0532	-0.0177
	(0.0638)	(0.0679)	(0.159)	(0.132)	(0.110)	(0.127)	(0.149)	(0.0317)	(0.0281)	(0.0394)	(0.0435)
Woman	0.0447	-0.0577	-0.413**	-0.0765	-0.311**	0.324**	-0.198	0.0543	-0.00379	-0.0505	-0.0633
	(0.0641)	(0.0534)	(0.162)	(0.132)	(0.130)	(0.137)	(0.145)	(0.0410)	(0.0200)	(0.0419)	(0.0429)
Woman x Female Teammate	-0.144	-0.113	0.525	0.214	0.182	-0.0432	0.482*	-0.0566	-0.0184	0.0750	-0.0254
	(0.141)	(0.104)	(0.318)	(0.270)	(0.234)	(0.238)	(0.286)	(0.0647)	(0.0420)	(0.0669)	(0.0831)
Women w/ Female Effect	-0.0843	-0.102	-0.344	-0.177	-0.0994	0.173	0.168	0.0186	0.0100	-0.0286	-0.106
F-test	0.538	3.140	1.637	0.573	0.206	0.988	0.419	0.169	0.0786	0.280	2.945
p-value	0.465	0.0804	0.205	0.451	0.651	0.323	0.520	0.682	0.780	0.598	0.0902
Observations	1,314	1,314	1,314	1,314	1,314	1,314	1,314	1,314	1,314	1,314	1,314
R^2	0.174	0.155	0.203	0.178	0.200	0.171	0.187	0.158	0.167	0.162	0.184

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 8.

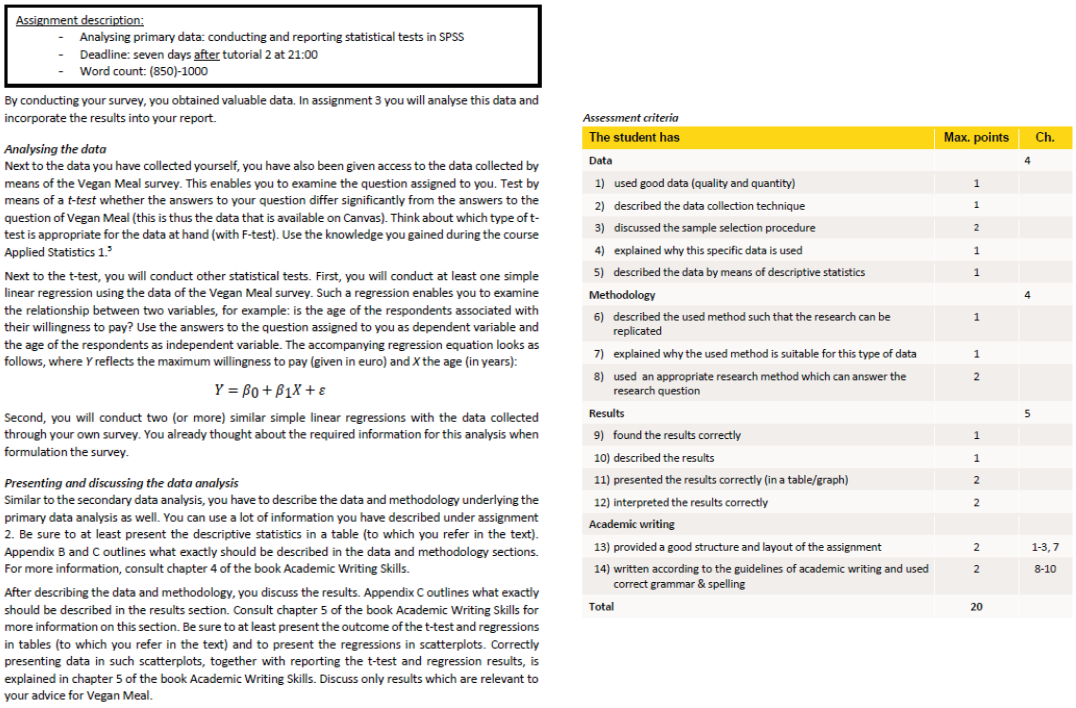
Figure B.1: Percentage of Workers in Teams



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) for the US. 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 B.2: Task Examples



Assignment description:

- Analysing primary data: conducting and reporting statistical tests in SPSS
- Deadline: seven days after tutorial 2 at 21:00
- Word count: (850)-1000

By conducting your survey, you obtained valuable data. In assignment 3 you will analyse this data and incorporate the results into your report.

Analysing the data

Next to the data you have collected yourself, you have also been given access to the data collected by means of the Vegan Meal survey. This enables you to examine the question assigned to you. Test by means of a *t-test* whether the answers to your question differ significantly from the answers to the question of Vegan Meal (this is thus the data that is available on Canvas). Think about which type of t-test is appropriate for the data at hand (with F-test). Use the knowledge you gained during the course Applied Statistics 1.[3]

Next to the t-test, you will conduct other statistical tests. First, you will conduct at least one simple linear regression using the data of the Vegan Meal survey. Such a regression enables you to examine the relationship between two variables, for example: is the age of the respondents associated with their willingness to pay? Use the answers to the question assigned to you as dependent variable and the age of the respondents as independent variable. The accompanying regression equation looks as follows, where *Y* reflects the maximum willingness to pay (given in euro) and *X* the age (in years):

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Second, you will conduct two (or more) similar simple linear regressions with the data collected through your own survey. You already thought about the required information for this analysis when formulation the survey.

Presenting and discussing the data analysis

Similar to the secondary data analysis, you have to describe the data and methodology underlying the primary data analysis as well. You can use a lot of information you have described under assignment 2. Be sure to at least present the descriptive statistics in a table (to which you refer in the text). Appendix B and C outlines what exactly should be described in the data and methodology sections. For more information, consult chapter 4 of the book Academic Writing Skills.

After describing the data and methodology, you discuss the results. Appendix C outlines what exactly should be described in the results section. Consult chapter 5 of the book Academic Writing Skills for more information on this section. Be sure to at least present the outcome of the t-test and regressions in tables (to which you refer in the text) and to present the regressions in scatterplots. Correctly presenting data in such scatterplots, together with reporting the t-test and regression results, is explained in chapter 5 of the book Academic Writing Skills. Discuss only results which are relevant to your advice for Vegan Meal.

Assessment criteria

The student has	Max. points	Ch.
Data		4
1) used good data (quality and quantity)	1	
2) described the data collection technique	1	
3) discussed the sample selection procedure	2	
4) explained why this specific data is used	1	
5) described the data by means of descriptive statistics	1	
Methodology		4
6) described the used method such that the research can be replicated	1	
7) explained why the used method is suitable for this type of data	1	
8) used an appropriate research method which can answer the research question	2	
Results		5
9) found the results correctly	1	
10) described the results	1	
11) presented the results correctly (in a table/graph)	2	
12) interpreted the results correctly	2	
Academic writing		
13) provided a good structure and layout of the assignment	2	1-3, 7
14) written according to the guidelines of academic writing and used correct grammar & spelling	2	8-10
Total	20	

Assignment 8: Presentation

General information

In block 2, you have practiced your presentation skills. During the sixth and last tutorial of this block, you will present the final version of your paper in class. This presentation should last ten minutes at most and must be made in PowerPoint. In your presentation, you should discuss all of the following questions clearly and 'to the point':

- Introduction:
  - What is your topic and research question?
  - Why is your research socially relevant?
  - Why is your research scientifically relevant?
- Theoretical framework:
  - How do you define (the most) important concepts? Tip: Only discuss the relevant concepts and be very concise.
  - What is the main economic literature that fits your research (that is: which articles did you use and why did you choose these)?
  - What are your hypotheses?
  - How did you embed the hypotheses in the literature?
- Data and methods:
  - Which data and methods did you use and why?
- Results:
  - Which results did you find?
  - Did you represent your results in a clear manner (in a table/graph)?
  - Did you describe explicitly whether the hypotheses are rejected/not rejected?
- Discussion/conclusion:
  - What was your research question?
  - Which conclusion did you draw (from the results)?
  - What is the link between the results and the theory you discussed?
  - What are the limitations of your research?
  - What are your suggestions for future research?

These are many questions for a ten-minute presentation. Therefore, you should be as brief and concise as possible. It helps to include limited (and only the most relevant) information on the slides. A clear and uncluttered table/graph with the most important results according to the guidelines of Academic writing is very much recommended! You should practice your presentation in advance, as you will be stopped after ten minutes. For the parts that you have not presented yet, you will not get any points.

After your presentation, there is time for some questions. Other teams and your teaching assistant will ask some questions. You should answer these questions in a thorough and concise manner, as your answers will influence your grade. On Tuesday June 28, 2022 (i.e., when the preliminary grades are published), your teaching assistant will provide feedback on the presentation (via Canvas).

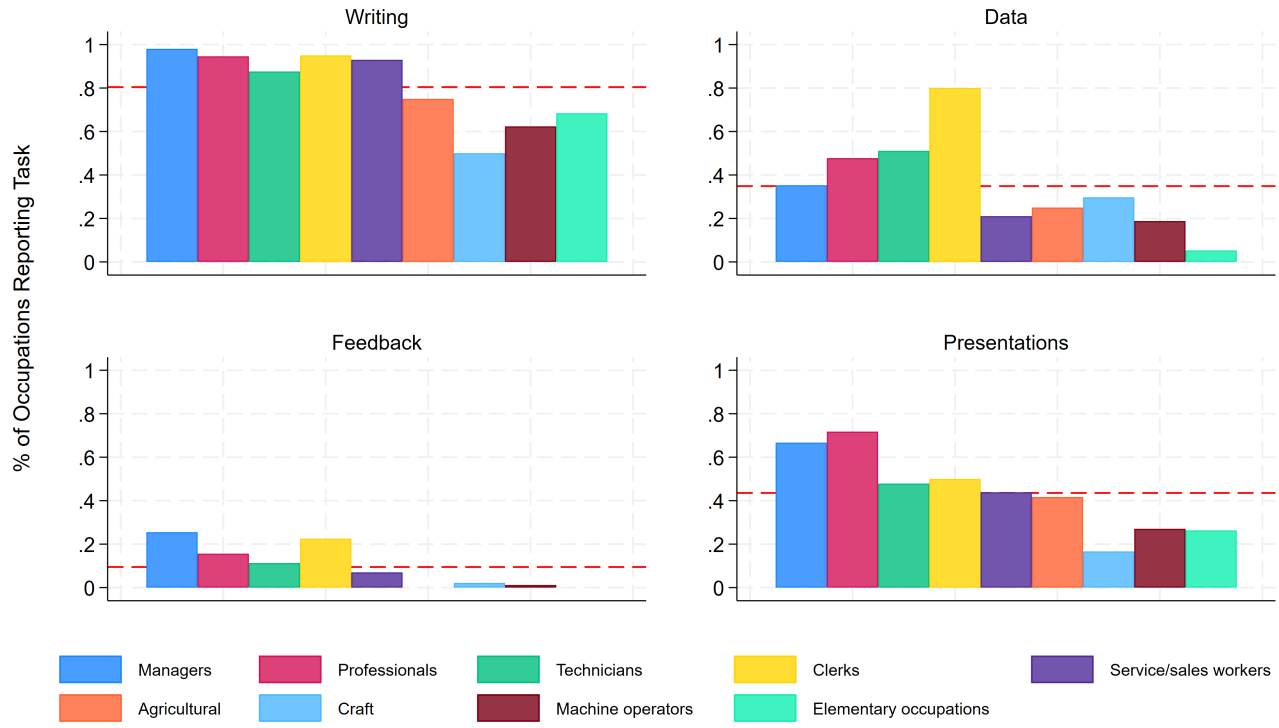
Criteria for assignment 8: Presentation

Assessment criteria	Max. points
The students have:	
PowerPoint presentation	
1) made a structured, clear and attractive presentation	3
Presenting	
2) presented well (open posture, spoken in a clear manner, enthusiastic)	3
Introduction	
3) introduced the topic in a catchy manner	2
4) discussed the research question	3
5) demonstrated the social relevance	1
6) demonstrated the scientific relevance	2
Theoretical framework	
7) defined (the most) important concepts	2
8) discussed the hypotheses	2
9) discussed how the hypotheses are embedded in the literature	3
Data and methods	
10) discussed which data and methods were used	2
11) explained why these data and methods were used	2
Results	
12) discussed the results that were found	2
13) displayed the results in a clear manner (in a graph or table)	3
14) indicated explicitly whether the hypotheses are rejected/not rejected	1
Discussion/conclusion	
15) drawn a clear conclusion (from the results)	2
16) linked the results and the theory they discussed	2
17) mentioned the limitations of the research	2
18) given suggestions for future research	1
Answering questions	
19) answered the questions well	2
Total	40

Notes:

1. Figure shows examples of two team task - the top is task 3 of block 4, the bottom is task 4 of block 5.
2. On the left hand side are the descriptions and instructions of given to the student teams. On the right hand side are the marking rubrics that the assignment is graded on.

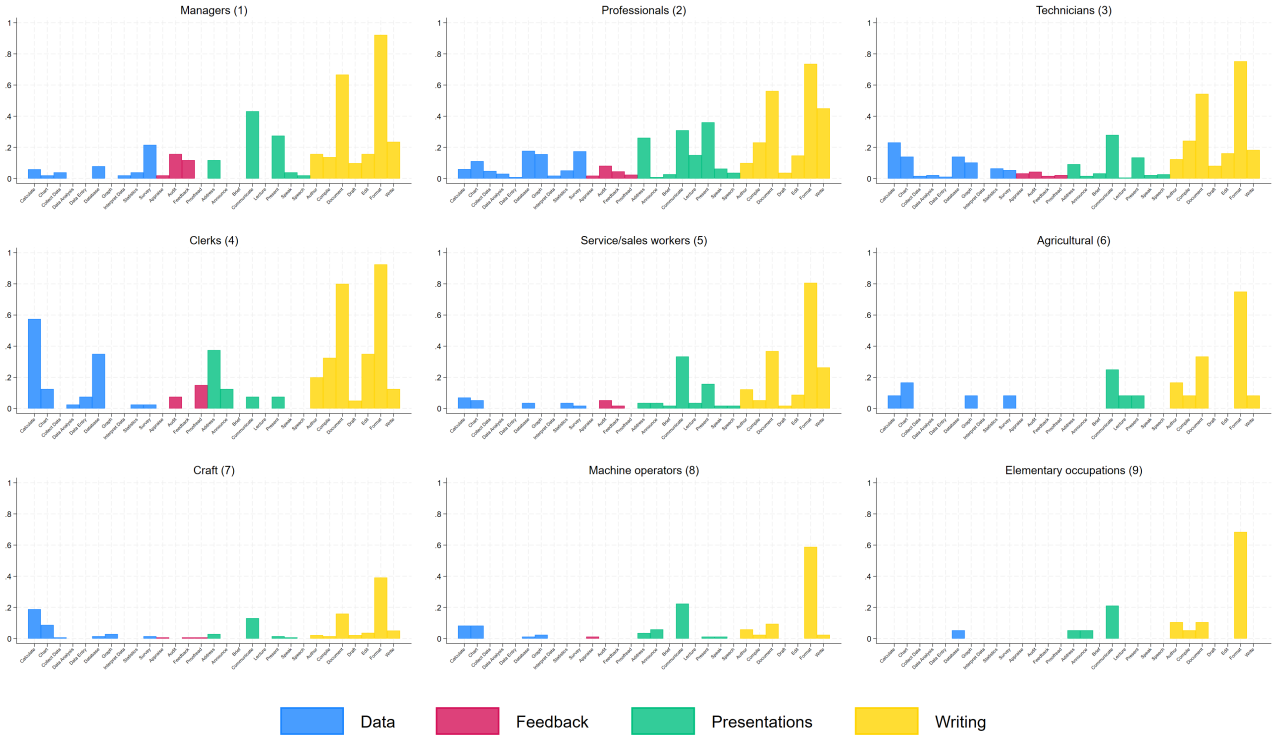
Figure B.3: Overlap of Tasks With Occupations



Notes:

1. Figure shows the proportion of occupations per ISCO category with tasks that share a keyword with those shown in Appendix Table B.4.
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.

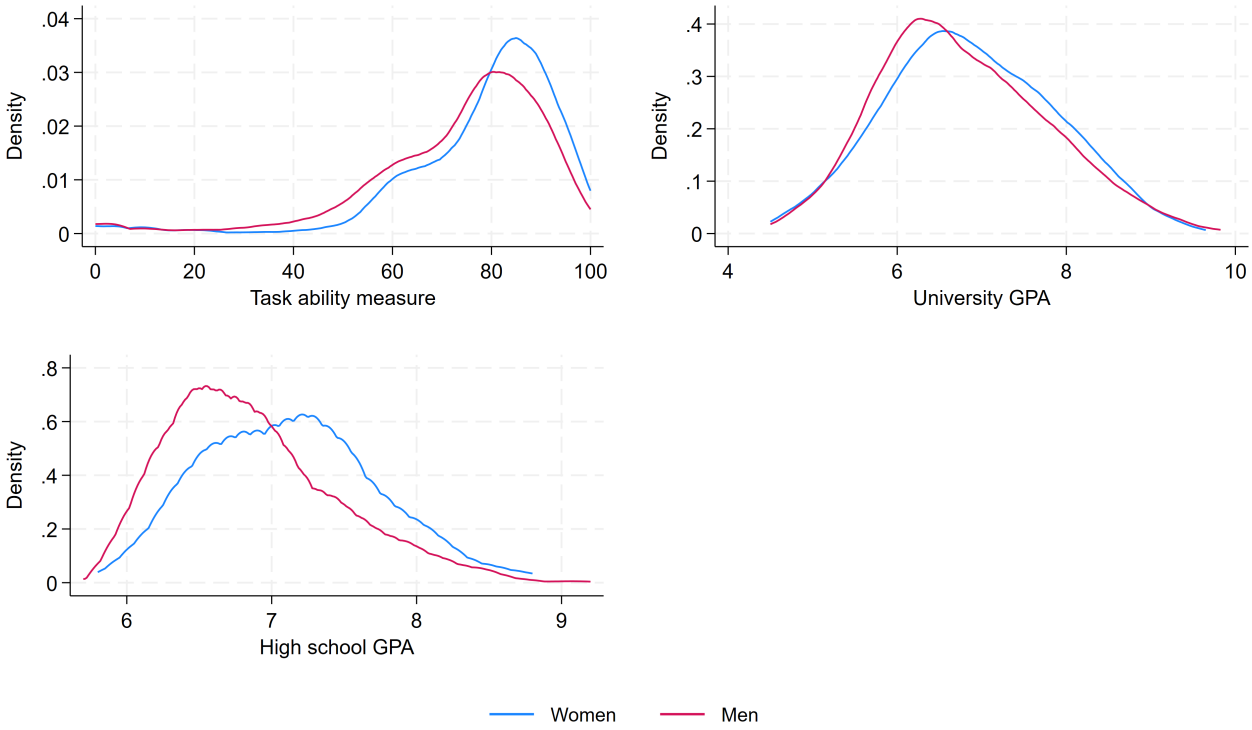
Figure B.4: Overlap of Keywords With Occupations



Notes:

1. Figure shows the proportion of overlap per task keyword, shown in Appendix Table B.4, by each ISCO occupation category.
2. Data comes from the ONET occupation-task database 28.0.

Figure B.5: 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.