Team Size and Diversity*

Brais Álvarez Pereira,† Shan Aman-Rana‡ and Alexia Delfino§

April 14, 2023

Abstract

We analyse the relationship between performance, team diversity and size. We first propose a model with knowledge spillovers in production, which predicts that the effect of having a person with a diverse knowledge set within a team increases with the size of the team. We experimentally test the model by randomly assigning students to solve knowledge questions in teams of different sizes, with or without a person with a diverse knowledge set. Our main finding is that the benefit of having a diverse rather than a same-skill colleague is greater in larger teams relative to small teams. We further show that such benefit is heterogeneous depending on the students' gender and the gender composition of teams. This has implications for how organizations can design their teams to maximize knowledge flows and performance.

JEL codes: J1, J15, J16, M50, O15
Keywords: Gender, diversity, team performance, information, communication

* We are grateful to seminar participants at multiple universities, conferences, and workshops, including the European University Institute, ASSET 2018, 43rd Spanish Economic Association Conference, Loyola Workshop on Economic Behavior, NOVAFRICA Conference 2022. We thank comments by Marco Casari, Andrea Mattozzi, Mathijs Janssen, Giulio Schinaia, Georgios Manalis, Matthias Schmidtblaicher, Ameet Morjaria, Andrea Ichino, David K Levine, Stefano Gagliarducci, Pablo Brañas-Garza, Alex Coutts. Lucia Solda, Jair Fernandes Martins, Moon Kim, Da Lin, Hengzhou Li, Weining Wang, Yanwen Wang, Laure Anique and Elizabeth Schroppe provided excellent research assistance. BELAB and Universidade Lusófona da Guiné-Bissau provided generous logistical support during the implementation of the lab experiment. All mistakes are our own. AEA Trial Registry RCT ID: AEARCTR-0007988.

†Nova School of Business and Economics (Nova SBE) NOVAFRICA, and BELAB. Email:brais.pereira@novasbe.pt
‡Department of Economics, University of Virginia. Email:sa8ey@virginia.edu
§Department of Economics, Bocconi University. Email:alexia.delfino@unibocconi.it

1 Introduction

The desire for organizational diversity is at an all-time high (Deloitte, 2021). Across sectors and geographies, the quest for diversity is shaping business strategy and challenging organizational design. On the one hand, organizations hope to reap the benefits of complementarities between diverse skill sets. On the other hand, diversity may lead to greater communication and coordination costs (Lazear, 1999; Prat, 2002; Hong and Page, 2004; Kahane et al., 2013; Garicano, 2000). The blooming or perishing of diverse teams ultimately depends on how diversity gets integrated within an organization (Shore et al., 2009).

This paper focuses on one particular organizational design problem: should diversity be introduced in small or large teams? As a motivating example, think about a hospital. Nurses mainly work with patients, but sometimes they are asked to solve management issues (e.g., optimal scheduling). To help with these tasks, the hospital hires an economist. Where would this new *diverse* hire be more useful, in a small or large team of nurses? Does this answer depend on gender? Our focus is on the policy relevant case of an organization who hires one diverse colleague, and has a choice to allocate them into teams of different sizes.[1]

Empirically, studying the effect of having one diverse member in teams of different sizes can be challenging. First, we need to observe the performance of a large number of teams which differ exclusively in size, keeping constant tasks and incentives, among other things. Second, we require variation in team composition which is as good as random, avoiding endogenous selection into teams. We overcome this challenge by means of a lab experiment, where we are able to randomly allocate participants to teams of different sizes who are facing the same tasks and incentives.

Going back to the hospital example, our main empirical object of interest is the marginal impact that the economist can have in a large compared to small team of nurses. However, relative to small mixed teams, the improvement in performance in large mixed teams may simply come from the larger number of nurses. To rule this out, we use experimental variation in a differences in differences framework: we compare the increase in performance in large versus small teams when an economist rather than a nurse is added.

Our main finding is that the benefit of a skill-diverse teammate rather than a skill-homogeneous one is increasing in team size. We also show that such benefit is heterogeneous depending on the participants' gender and the gender composition of teams. This suggests that organizations should take into account team size and gender composition when introducing skill diversity.

We first propose a theoretical framework which draws on Lazear (1999) to explore the effect of skill diversity on performance in teams of different size.[2] There are two types of tasks and two types of workers, each able to solve only one type of task. When a task is randomly drawn, each worker has to solve it individually, but can improve her performance by talking to colleagues who are potentially better than her at solving it. The main prediction of the model is that the contribution of one diverse worker to their colleagues' performance should be increasing in team size. The intuition is that adding workers with the same skill set to a team increases their colleagues' performance, but at a decreasing rate (since their skills become redundant as their number increases). Thus adding

[1]Another possibility is to compare teams of different sizes with the same exact share of diverse members. However, this would not take into account hiring constraints which may be binding in the real world. We discuss the implications of our design choice in the discussion section.

[2]We focus on the case in which diversity matters through differences in skills or knowledge. However, heterogeneous within-type and between-type costs of communication in the theoretical model make our findings also relevant for the study of discrimination (differences in inteam and outteam behavior), typically associated to ethnic, religious, gender or national diversity (Alesina and La Ferrara 2000, 2005; Bertrand and Duflo 2017).

a person with a diverse knowledge, instead of an extra worker with the same skill-set, can have a bigger impact in a larger team, where the returns to the same-type knowledge are declining.

We design our experiment to closely map the theoretical setting. Our participants are Nursing and Economics students —nurses and economists from now on. Each of them solves 25 multiple-choice questions, first in isolation and then in teams where communication is encouraged. To allow for gains from having a skill diverse teammate, each test contains questions in Nursing, Economics and General Knowledge. To get variation in team diversity and size, we assigned one economist or one nurse to teams with already one or three nurses. Nurses are the treated participants, while economists provide a source of exogenous variation in diversity within teams. We measure performance as the percentage of correct answers that changed between the team test and the one in isolation, thus controlling for individual baseline ability.

Our results are consistent with the model predictions. First, we find that nurses' average performance is the same in small and large teams made only of nurses. This reflects declining marginal returns from adding people with the same skill set in a team, as formalized in the model. Second, the marginal effects of having a skill diverse teammate depends on team size. In teams of size 2, relative to a homogeneous team, being paired with an economist increases average individual performance by 4 percentage points ($p < 0.05$). On the other hand, an economist rather than a nurse in teams of size 4, however, leads to an additional 5.8 percentage points higher performance ($p < 0.10$) for nurses. This shows that the benefit of having a skill diverse teammate is higher in larger teams.

While our framework focuses on skill diversity, skills often correlate with dimensions such as gender, ethnicity, socio-economic background. This complicates the organizational trade-offs involved in managing and coordinating a diverse workforce. For instance, communication frictions between colleagues with different educational backgrounds may be worsened by gender differences in communication styles and workplace behavior.

Theoretically, we explore the multi-dimensionality of diversity by considering the case in which the skill diverse teammate may also be of a different gender. We assume that there are higher communication costs for a man and a woman to communicate with each other compared to intra-gender communication.[3] The key insight in our theory is that inter-gender communication costs may reduce or enhance the advantages of working with a skill diverse teammate depending on whether these change with team size. If the frictions brought about by gender differences raise with team size, the advantages of skill diversity in large teams may be mitigated or even lost.

Following this theoretical extension, the experiment further manipulates an additional dimension of team diversity: whether the economist is a man or a woman. This implies that skill-diverse teams in our experiment may be gender homogeneous or not, a variation which allows us to ask whether the positive interaction between team size and skill diversity on performance is affected by gender diversity.

We find that the effects of having a skill diverse teammate differs by gender and by the team's gender composition. Male nurses reap the largest benefits from being in a larger team with an economist, independently of whether it's a female or male economist. In contrast, women do not experience any additional benefit from being in a larger team with a skill-diverse teammate when they are of a different gender. This means that the very presence of a man in the team can be disruptive for women, even when they are the majority. Interpreted through the lens of our

[3]Previous work has indeed highlighted the challenges that women face in interacting with men in team settings, especially ones that require sharing of ideas (Born et al., 2022; Coffman, 2014).

theoretical framework, these gender differences suggest that female nurses' cost of communication with male economists may be increasing in team size. Results remain robust to controlling for the average baseline performance of the entire team, suggesting that the heterogeneous effects by gender do not simply capture a gender gap in skills of economists.

For organizational design, the main takeaway from our paper is that both the size and composition of teams should be taken into account when introducing new skill sets within the firm. Previous work has highlighted the trade-off between enhanced productivity and communication costs in diverse teams. Our paper contributes by predicting where we expect the productivity effect to dominate and focuses on choice variables that firms are able to manipulate (size and gender composition).

Our paper contributes to the vast Economics literature on the impact of diversity on productivity (Haltiwanger et al., 1999; Hoxby, 2000; Hansen et al., 2006; Gagliarducci and Paserman, 2012; Ghosh, 2022; Marx et al., 2021; Hjort, 2014; Hoogendoorn et al., 2012; Aman-Rana et al., 2021). Close to our skill-diversity variation, Hoogendoorn et al. (2012) find that cognitive ability dispersion has an inverted U-shape relationship with team performance. We contribute to this literature by highlighting how the same kind of diversity can have different effects on performance depending on the size of the team considered. Studies in management have also touched upon the relationship between team size and diversity (Krammer, 2021; Hu et al., 2021). We contribute by providing experimental estimates in a tightly controlled environment.

Recent studies have investigated empirically communication in teams (e.g. Bloom et al., 2014; Sandvik et al., 2020; Battiston et al., 2021; Menzel, 2021) and, in particular, uncovered frictions that arise when men and women work together (Sharma and Castagnetti, 2019; Coffman, 2014; Bordalo et al., 2016, 2019; Coffman et al., 2021; Shan, 2020). While most of this work has focused on the case in which women are the minority, we show that women's performance may be hindered in mixed-gender teams even when they are the majoritarian gender.

2 Theoretical framework

We draw on Lazear (1999) to explore the role that team size plays in the relationship between diversity and performance.

Each worker is assigned a task Z, a random variable that requires Nursing knowledge with probability p (task Z_1) and Economics knowledge with probability $1-p$ (task Z_2). Two types of workers with different skills can each solve one of the tasks, but not the other. Type-1 workers (nurses) have a stronger background in Health and type-2 workers (economists) in Economics, but they are otherwise identical (e.g., in motivation). There is also heterogeneity within each type of worker, as some are more skillful than others.

Both nurses and economists have the same probability of facing task Z_1 or Z_2. We consider the extreme case in which the knowledge of nurses and economists is fully disjoint: a nurse never knows the solution to an economics task Z_2 and an economist never knows the solution to a health task Z_1.

We depart from Lazear (1999) in two ways. First, we focus on how the effect of diversity on performance varies with team size. Second, our outcome is individual performance for a type-1 worker rather than team performance. Tasks are solved individually by each member, who may gain from being part of a team because of communication with others. We assume that agents become as good in solving a given task as the best colleague they have talked to. Communication is costly, and talking to a colleague with a different skill set is more costly than talking to a similarly-skilled colleague. We interpret communication costs as the opportunity cost of time (Garicano, 2000). As people with different technical language take longer to understand each other, we assume that there are higher communications costs between people holding different knowledge sets. Proofs can be found in Appendix D.

2.1 Performance in isolation

A worker's performance when working in isolation is determined by the task at hand, her type and ability. If the task is of class Z_1, a type-1 worker i will perform the job with quality $z_{1i} \in [0, \hat{Z}_1]$, drawn from the density of production possibilities $f(z_1)$. If the task is of class Z_2, a type-1 worker in isolation will produce 0. Similarly, a type-2 worker j will perform the job with quality $z_{2j} \in [0, \hat{Z}_2]$, drawn from $g(z_2)$ whenever the task is of class Z_2, and 0 otherwise. It follows that the expected production of workers i and j in isolation is respectively given by $E(z_i) = pE(z_{1i})$ and $E(z_j) = (1-p)E(z_{2j})$.

2.2 Performance in skill homogeneous teams

In a team setting, each individual might talk to her colleagues and improve her own performance. We assume that team size and communication costs are small enough for each worker to find it optimal to communicate with everyone else.

Consider type-1 worker i being part of a team with $m > 1$ workers with the same knowledge set. If i communicates with each of her $m-1$ colleagues at a cost c, her expected performance is:

$$E(z_i|m) = pE(z_{1i}|m) + (m-1)c = pZ_{1m} - (m-1)c, \tag{1}$$

where Z_{1m} is the expected highest order statistic from a sample of m draws from $f(z_1)$, which is a concave function of the sample size for any continuous distribution (de la Cal and Cárcamo, 2005). That is, each worker performs her task as well as the best worker she communicates with, minus the incurred communication costs.[4]

The concavity of the expected higher order statistic and linearity of costs imply that, for homogeneous teams composed of same-skill workers:

Proposition 1. *The effect of team size on expected performance is concave.*

Intuitively, this result is due to the redundancy of same-type workers' knowledge, which increases with the number of workers of this type, combined with linear costs of communication. The smaller gains in expected performance resulting from having additional workers of the same type in larger teams has implications for the effect of diversity.

2.3 Performance in skill diverse teams

A team is now composed of m type-1 workers (nurses) and k type-2 workers (economists). A type-1 worker can potentially improve her expected performance (with respect to working alone)

[4]Better workers may not want to share their superior knowledge once the task and abilities are revealed. There are incentives for workers to communicate their knowledge as long as part of individual payment depends on team performance, as in our experiment.

by communicating with both types. Communicating with a worker of a different type costs $c_o > 0$, and we normalize the cost of within-type communication to 0.

In a diverse team, the expected performance of a type-1 worker i is:

$$E(z_i|m,k) = pZ_{1m} + (1-p)Z_{2k} - kc_o. \tag{2}$$

When the team is homogeneous and only has $m+k$ type-1 workers, i's performance is instead:

$$E(z_i|m+k,0) = pZ_{1m+k}. \tag{3}$$

Denote as $\Delta_k(m)$ the difference between worker i's performance in the diverse and homogeneous team of a given size $m+k$:

$$\Delta_k(m) = E(z_i|m,k) - E(z_i|m+k,0) = p(Z_{1m} - Z_{1m+k}) + (1-p)Z_{2k} - kc_o. \tag{4}$$

For a given team size, $\Delta_k(m)$ is the additional contribution of k diverse workers to the expected performance of each of their m type-1 colleagues. The concavity of the expected highest order statistic, together with the knowledge of type-2 workers contributing the same in larger and smaller teams, produces the following result:

Proposition 2. *For a type-1 worker, the gain in expected performance from being placed in a diverse team with k type-2 workers - instead of a same-sized homogeneous team - is increasing in team size.*

The result follows from the concavity of the expected highest order statistic. Proposition 2 implies that skill diversity might have different effects in teams that differ in their size. As type-1 workers in a team increase, the knowledge of such workers becomes redundant, implying that a diverse worker will be relatively more valuable in large teams. Appendix Figure A1 illustrates the expected performance of type-1 workers, both in isolation and in teams of different size and diversity.

Appendix D.1 presents an extension of the model with skill and gender diversity. The main result is that additional costs of inter-gender communication may mitigate (or enhance) the relationship between team size and skill diversity of Proposition 2.

3 Context and research design

We carried out a lab experiment with university students studying Nursing and Economics in the largest university in Guinea Bissau. Students were drawn from their second, third and fourth (final) year in their respective degrees.

3.1 Experimental design

We designed the experiment to simultaneously study three factors that might affect individual performance in teams: size, skill diversity, and gender diversity. With this aim, participants were randomly assigned to teams of size 2 and 4, composed of only nurses or of nurses with one economist of either gender. Our treatments are illustrated in Figure 1. From a total of 248 students (61 economists and 187 nurses) we created 52 teams composed of two members and 36 teams composed of four members, for a total of 88 teams. Out of the teams with two members, 37 teams had one economist. Out of the teams composed of four members, 24 teams had one economist. In approximately half of the teams with an economist, the economist had the opposite gender from the rest of the team.

All the participants that started the experiment completed it to the end. Most participants took part in the study once, while there were 23 participants that took part in the study in two different sessions and one took part in three sessions. Since the tests were unique per session, we consider these participants as fresh draws from the pool of students and include them in the main results. However, the main results remain robust to including a control for repeated participants.[5]

3.2 Task and incentives

Each participant had to solve a knowledge test twice. The test consisted of 25 multiple-choice questions with five options and a unique correct answer. For a nurse (economist), ten questions were general-knowledge, ten questions were from Nursing (Economics) and five were questions from the other discipline.[6] Since each student faced 20% questions from the other discipline, the task required both Nursing and Economics skills. We thus embedded the need for skill complementarity within the design, as in the model.

Students had to first answer the test in isolation. This was followed by a second round of testing in teams, using the same exact questions that each participant had completed individually. This time participants were allowed to communicate and discuss their answers with the other members. Every participant in each session studying the same subject answered the exact same test, and had 30 minutes to complete it in each round and there was no penalty for incorrect answers in both rounds.

In the first round, students received 80 FCFA (0.12$) per right answer.[7] For the second round of testing, in order to incentivize participants to share their knowledge, one of the tests from the team was randomly chosen and all team-members were paid according to the number of correct answers in that test (80 FCFA for each correct answer).

3.3 Balance

Nurses assigned to different types of teams were very similar on a variety of observable characteristics (Figure A2). Appendix Table A5 shows that the statistically significant differences (moved to capital for college, age and self-reported altruism) are not associated with performance in the baseline (individual) test. These checks support balanced assignment of the four treatment groups in pre-treatment covariates.[8]

3.4 Baseline differences in skills between nurses and economists

To understand the skill differences that exist across nurses and economists of either gender we present the baseline skill gap in Figure 2. This figure shows the average percentage difference in the score of nurses (female on the left, male on the right) with either a female (gray bar) or a male (orange bar) economist.

[5] Appendix B details recruitment and logistics.
[6] See the Online Appendix for more details on the source and characteristics of the questions.
[7] Students were paid between 500 and 5000 FCFA (between 0.76 and 7.58$), depending on performance. These incentives were sizeable: the minimum daily wage is around 600 FCFA.
[8] Appendix Figure A3 presents an equivalent graph for economists. There are only 3 significant coefficients out of 36 tested.

As expected, students perform better in the subject in which they specialize. Economists perform 25% higher than nurses in Economics questions, while nurses perform almost 20% better than economists in Nursing questions. This is true for both male and female nurses, and reassures that our empirical variation captures differences in skills.

However, we find some gender differences in general knowledge. Female nurses perform 22% worse than male economists in this category, while there is no statistically significant difference between female nurses and female economists. These findings suggest that being placed in the same team with a male economist can benefit female nurses both in Economics and general knowledge. Therefore, a skill diverse and gender heterogeneous team should improve the performance of female nurses more than a skill diverse but gender homogeneous team, as long as the costs of communication across genders are negligible. In contrast, male nurses should benefit from being matched with an economist only in Economics, as they perform weakly better than economists in general knowledge questions.

4 Main results

4.1 Empirical strategy

We report estimates restricting attention to the sample of nurses, and briefly discuss results on economists in Appendix C. To test the effect of a skill diverse teammate in teams of different size we use the following estimation. For nursing student i we estimate:

$$\Delta Score_i = \alpha SkillDiversity_i + \gamma TeamSize4_i + \beta SkillDiversity_i \times TeamSize4_i + \theta Male_i + u_i \quad (5)$$

where $\Delta Score_i$ is the difference in the percentage of correct answers between each individual i and their team test (the same test is used in both rounds).[9] $SkillDiversity_i$ is a dummy variable equal to 1 if a skill diverse teammate i.e., an economist is allocated to nursing student i, $TeamSize4_i$ is a dummy variable that takes value 1 if the nursing student is allocated to a team of size four. $Male_i$ is a dummy variable for being a male. The reference category comprises participants in skill homogeneous teams of size 2. Standard errors u_i are clustered at the team level (at which the treatments are allocated).[10] Coefficient α is the effect on performance of a nurse working with one economist, relative to their performance when working with another nurse. γ is the effect of allocating a nurse to a homogeneous-skill team of size 4 relative to size 2. The main coefficient of interest is β, which captures the effect of being allocated to work with an economist relative to a nurse in a team of size 4 relative to size 2. Following Proposition 2, we expect $\beta > 0$.

4.2 When is skill diversity most beneficial?

We first present evidence that our treatments affect the distribution of changes in score (%) between the first round (answering alone) and the second (answering in teams), in line with the

[9]Following the model, we focus on individual performance rather than team performance. We think this is relevant as many organizations use individual incentives even when employees are organized in teams.

[10]Specifications with the inclusion of session fixed effects (α_s) and a dummy for students participating in more than one session are reported in the Appendix (Tables A1 and A2). The effects are similar to the main results, but lower variation within sessions reduces our statistical power.

presence of intra-team communication. Next we present results estimating Equation 5.

Figure 3 shows the distribution of the average test-score change between the individual and the team test by diversity (with an economist or not) and size (2 or 4). The distribution of the average change in test-scores is concentrated on positive values irrespective of team size or diversity suggesting that people are communicating within their teams.[11] Second, Figure 3 shows that the gains from team communication are greater when an economist is allocated to the team. For both team sizes, the distribution of performance is shifted to the right in skill-diverse teams (red dashed lines). A third pattern emerging from Figure 3 is the difference in performance improvements between small and large teams. Working with an economist improves the score of nurses, but this improvement is much larger in a team of size 4 than size 2.

Table 1 shows the average gains from working with an economist accruing in large versus small teams, overall and by nurses' gender, estimating Equation 5. Column (1) shows that there are positive and statistically significant gains from working with an economist in small teams: the average gain is 4 percentage points (8% of average individual test scores) in skill-diverse pairs with respect to skill-homogeneous pairs. In contrast, working in a large homogeneous team with respect to a small one does not significantly improve scores. This is consistent with Proposition 1 from the model. The coefficient on the interaction between skill diversity and team size 4 (β) is positive, significant and large, representing 12% of the average individual test scores. This result is consistent with our main theoretical prediction (Proposition 2). A skill diverse teammate is most beneficial in large teams, where the addition of a person with diverse knowledge has the largest marginal impact on each member's performance.

Columns (2) and (3) of Table 1 split the sample by nurses' gender and show that diversity in large teams benefit both men and women, but to different degrees.

First, in skill homogeneous teams women see larger benefits from team size than men. The coefficient on the dummy for team size 4 is 2.1 for women and -1.7 for men, even if the difference is not statistically significant ($p = 0.41$). This suggests that interactions between female nurses may be more cooperative than among male nurses, as highlighted also in previous work on collaborations in female-dominated fields (Charness and Rustichini, 2011). Another reason could be that there is larger room for size to improve individual performance among female nurses, as they have worse general-knowledge.

Second, an economist in a large team benefits men more than twice with respect to women. The differential effect of working with an economist in large teams is 8.3 pp for men ($p < 0.05$) and only 3.5 pp for women ($p > 0.10$), even if we cannot reject the equality of the interaction coefficients between the two genders ($p = 0.43$). In other words, three male nurses matched with an economist seem to be able to improve their performance more than three female nurses with one economist.

To what extent are these differences coming from the challenges of communicating in teams which are not only skill-diverse, but also gender diverse? Previous work has highlighted the challenges that women face in interacting with men in team settings, especially in sharing of ideas (Coffman, 2014; Born et al., 2018; Heikensten and Isaksson, 2019). While most work has focused on studying women's behavior as a minority, we are interested in knowing the impact of diversity on the majority's outcomes. For instance, consider an all-women team in a given organization.

[11] The change in scores is the joint effect of doing the test twice and intra-team communication. We maintain the assumption that the change due to repetition is the same across treatments. Empirically, this effect seems small: when assigned to working with only one economist, nurses do not improve their performance in the 5 Nursing questions that are not present in the test for economists.

Does introducing a man with different skills changes incumbents' outcomes, and how? How does adding a woman affect all-men teams instead?

We address these questions in the next section.

4.3 Does gender diversity inhibit the effect of skill diversity?

This section investigates whether gender diversity inhibits the effect of skill diversity by interacting the $Skill\ Diversity$ regressor in Equation 5 with a dummy for whether the diverse teammate is of the same or of a different gender to the rest of the team.[12]

We first focus on the effects of working with an economist of the same gender as nurses (coefficients α and γ). Table 2 Column (1) shows almost identical results to the ones discussed in the previous section on the average effects of working with a skill diverse teammate. We see that skill diversity has an attenuated impact in small teams and significantly improves the average scores in larger teams. Columns (2) and (3) show that this is true for *both* men and women, but splitting the sample by gender increases the noisiness of the estimates. Therefore, while the average estimate for coefficient γ in Column (1) is significant, we cannot reject the null hypothesis that the interaction between skill diversity and large team size is equal to zero for either men or women, even if the coefficients' magnitude is similar (Columns (2) and (3) respectively). Overall, the effects of introducing a gender homogeneous skill diverse teammate seems to go in the direction suggested by the model.

Coefficients β and δ explore the joint effect of multi-dimensional diversity, in skills and gender, on performance. In the overall sample (Column (1)), a nurse paired with one economist experiences a substantial benefit when the economist is of the opposite gender (5.92 pp increase in score). This benefit is even larger than being matched with an economist of the same gender (coefficients $\alpha = \beta$, $p = 0.06$).

When economists are of a different gender relative to nurses, moving from pairs to teams of four people does not disproportionately affect performance for nurses. When differences in knowledge are brought by participants of different genders, diversity seems as beneficial in small and large teams. Thus there is no "size premium" for skill diversity when accompanied by gender diversity. Below we unpack this result and show that it is driven mainly by female nurses. Columns (2) and (3) uncover important gender differences in teams with both gender and skill diversity. Column (2) shows that male nurses working one-to-one with a female economist increase their scores as much as male nurses working with male economists ($\alpha = \beta$, p=0.52). Similarly, teams of three male nurses working with a female economist are able to take advantage of skill diversity as much as when working with a male economist (δ=γ, $p = 0.69$).

On the other hand, results in Column (3) show that female nurses miss a comparable gain in performance from being allocated to large gender and skill diverse teams. The increase in performance for female nurses is the same in small gender diverse teams (β) or large gender homogeneous teams (γ). When it comes to interacting with a male economist in a large team, however, women see no benefit compared to working with another nurse. In other words, female nurses benefit equally from diversity in small teams, when they work only with one male economist, and in large teams, where they work with other two female nurses in addition to a male economist.

This implies that our main prediction that introducing a skill diverse teammate is most beneficial in large teams holds generally true for men, while for women it is empirically satisfied only in gender

[12]In our experiment gender diversity is only introduced in skill-diverse teams.

homogeneous teams. Strikingly, the magnitude of the δ coefficient for women is indeed one-tenth of the coefficient for men, even if we cannot reject the equality between the two coefficients under conventional significance levels ($\delta_{male} = \delta_{female}$, $p = 0.22$),

For female nurses this result is consistent with gender differences in the costs of communicating with the opposite gender which are a function of team size. On the other hand results suggest that this cost is independent of team size for men.

We also present results in Appendix Table A3 and Table A4 controlling for the average baseline performance of the entire team. This is done to rule out the possibility that heterogeneous effects by gender simply capture differences in the baseline skills of men and women economists. We find that not only the result remain robust, but the coefficient γ is also more precise and marginally significant in the case of female nurses. This suggests that, conditional on baseline group knowledge, we still see that female nurses stand to benefit more from the presence of a female rather than a male economist in a large team.

Overall, comparing coefficients δ_{male} and δ_{female} with β_{male} and β_{female}, we can say that an organization of female nurses would reap the highest benefits of skill diversity from placing a male economist in small teams and placing a female economist in large teams. In contrast, an organization of male nurses may place either a female or male economist in large teams to maximise the benefit from their diverse knowledge.

Why do women and men benefit differently from the combination of gender and skill diversity in large teams? While we do not have experimental variation to further tell apart different mechanisms, below we discuss a few potential channels.

4.4 Discussion

Social image as a potential explanation for gender differences. In order to explore why men and women perform differently in large gender diverse teams, we look at results by knowledge area. Appendix Figure A4 shows the difference in the percentage of answers which improved in General Knowledge, Economics and Nursing questions between teams of size 4 and 2, which can be gender homogeneous (light bars) or gender diverse (dark bars). We distinguish between female (top figure) and male nurses (bottom figure).

First, for female nurses, we observe that the improvement in performance is larger in gender homogeneous teams across all subjects. In Nursing, 13.9% of answers are improved when working with two female nurses and one female economist with respect to working with only one female economist. However, this improvement drops to 7.4% by simply replacing the female economist with a male economist, despite keeping everything else constant including the number of female nurses that one can work with in the larger team. This is despite the fact that there are no skill differences across male and female economists in Nursing questions (see Figure 2). Similarly, the increase in performance in General Knowledge and Economics is at least twice as big in large gender homogeneous teams compared to gender diverse teams. This is also hard to explain through skill gaps, since female nurses should have gained *more* when paired with male economists as compared with female economists (see Figure 2).

This evidence has two implications. First, in large teams with a male economist, women are unable to take advantage of the economist's diverse knowledge. Second, women also modify the way they interact with each other to solve Nursing questions. The lower improvement in Nursing indeed suggests that women cannot effectively interact with each other in the presence of a man.

While we cannot pin down the exact channel, social image concerns can accommodate this evidence. If women fear being perceived as incompetent by the other gender, they would refrain from asking in any subject, either to each other or directly to the male economist. Such image concerns, instead, seem to be attenuated in teams with only women.[13]

The bottom panel of Appendix Figure A4 shows that male nurses behave in the opposite way of female nurses, and that the presence of a female economist in large teams improves the overall team dynamics. This is shown by the fact that men's Nursing scores get even better when there is a female economist in the team (improvement of 19.4%), compared to a situation in which there is a male economist and the same exact number of other male nurses (improvement of 8.1%). This is again hard to explain through skill differences since there are no differences across male and female economists in Nursing skills (see Figure 2). Social image gains can rationalize men's behavior. If men get utility from signaling their competence to a person of a different gender, then they would put more effort in communicating their Nursing knowledge in the presence of a woman rather than a man. For instance, knowledgeable men may be showing-off to the opposite gender by helping others.[14]

Diversity share as a potential mechanism. Our experiment considers a policy-relevant situation in which one employee has to be allocated in teams that differ in their sizes. This design implies that the "share of diversity" is not held constant in small versus large teams, and this can be a potential mechanism behind the results. To see this, compute the probability that any two team members picked randomly have different skill sets. This probability is 1 in teams with one economist and one nurse, but it is 1/2 in teams with one economist and three nurses. Thus in our experiment the share of diversity falls as the size of skill-diverse teams increases. If communication costs increase with the share of diversity, rather than the absolute number of skill-diverse teammates, then our results for gender-homogeneous teams may be explained by these differences in treatment intensity. Theoretically, this could be obtained in our framework by having the costs of communication increase as a function of the team diversity share. While our framework focused on the relative returns from adding a diverse skill member to teams of different size, such changes in relative communication costs might be another way of explaining our results. However, we believe this alternative formulation is less appropriate in our setting because communication costs are kept very low (e.g., at the extreme, people could just swap their exam papers without even talking).

Conclusion

The jobs of the future increasingly require workers to solve complex cognitive tasks within global teams (Lazear, 1999, 2000; Antràs et al., 2006, 2008; Edmondson, 2012). These trends make knowledge spillovers within diverse teams increasingly important determinants of firms'

[13]Another explanation is mansplaining: the male economist makes the team spend most of the time discussing Economics and General Knowledge questions at the expenses of Nursing. However, this explanation predicts that female nurses should weakly increase their Economics or General Knowledge scores with male economists, which is not observed. There can also be supply-side explanations, like men not wanting to help or wanting to show off without effectively communicating to women, which we cannot fully rule out. However, this does not explain the change in women's performance in Nursing.

[14]An alternative explanation is that three male nurses are better able to extract information from a female economist compared to a single male nurse (e.g., by pressuring her). However, male nurses' score in Economics improves similarly when in pairs or in large teams with a woman (-0.04% improvement, $p > 0.10$), suggesting that male peer pressure is not the main mechanism at play.

comparative advantage and competitiveness. How should organizations design their teams to maximise knowledge flows between members?

In this paper, we study three aspects of this design choice: skill diversity, team size and gender. In particular, we are interested in whether including one person with a diverse skills set is most beneficial for performance in small or large teams. We find support for a positive relationship between gains from diversity and team size.

We also find that women and men benefit differently from the combination of gender and skill diversity in large teams. This result is consistent with gender differences in the costs of communicating with the opposite gender which are an increasing function of team size for women. This result speaks to emerging evidence on possible frictions to learning between genders (Conlon et al., 2021) and highlights that environmental features (e.g, size) may play a key role in limiting knowledge diffusion. A promising direction for future research is studying how firms and institutions should design their workplaces to make the most of diversity, while mitigating its costs and avoiding lack of inclusion for minorities.

References

Alesina, Alberto and Eliana La Ferrara, "Participation in heterogeneous communities," *The Quarterly Journal of Economics*, 2000, *115* (3), 847–904.

— **and** — , "Ethnic diversity and economic performance," *Journal of Economic Literature*, 2005, *43* (3), 762–800.

Aman-Rana, Shan, Alexia Delfino, Clement Minaudier, Shamyla Chaudry, and Brais Álvarez Pereira, "Gender and Choice: Experimental Evidence," *Available at SSRN 3958386*, 2021.

Antràs, Pol, Luis Garicano, and Esteban Rossi-Hansberg, "Offshoring in a knowledge economy," *The Quarterly Journal of Economics*, 2006, *121* (1), 31–77.

— **,** — **, and** — , "Organizing Offshoring: Middle Managers and Communication Costs," *Harvard University Press*, 2008, pp. 311–39.

Battiston, Diego, Jordi Blanes i Vidal, and Tom Kirchmaier, "Face-to-face communication in organizations," *The Review of Economic Studies*, 2021, *88* (2), 574–609.

Bertrand, Marianne and Esther Duflo, "Field Experiments on Discrimination," in "Handbook of Economic Field Experiments," Vol. 1, North-Holland, 2017, pp. 309–393.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, and John Van Reenen, "The Distinct Effects of Information Technology and Communication Technology on Firm Organization," *Management Science*, 2014, *60* (12), 2859–2885.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer, "Stereotypes," *The Quarterly Journal of Economics*, 2016, *131* (4), 1753–1794.

— **,** — **,** — **, and** — , "Beliefs about Gender," *American Economic Review*, 2019, *109* (3), 739–73.

Born, Andreas, Eva Ranehill, and Anna Sandberg, "A man´ s world?–The impact of a male dominated environment on female leadership," 2018.

— **,** — **, and** — , "Gender and willingness to lead: Does the gender composition of teams matter?," *Review of Economics and Statistics*, 2022, *104* (2), 259–275.

Charness, Gary and Aldo Rustichini, "Gender differences in cooperation with group membership," *Games and Economic Behavior*, 2011, *72* (1), 77–85.

Coffman, Katherine Baldiga, "Evidence on self-stereotyping and the contribution of ideas," *The Quarterly Journal of Economics*, 2014, *129* (4), 1625–1660.

Coffman, Katherine, Clio Bryant Flikkema, and Olga Shurchkov, "Gender Stereotypes in Deliberation and team decisions," *Games and Economic Behavior*, 2021.

Conlon, John J, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach, "Learning in the Household," Technical Report, National Bureau of Economic Research 2021.

de la Cal, J. and J. Cárcamo, "Inequalities for expected extreme order statistics," *Statistics and Probability Letters*, 2005, *73* (3), 219–231.

Deloitte, "First Steps into the Labour Market survey results," 2021.

Edmondson, Amy C, *Teaming: How organizations learn, innovate, and compete in the knowledge economy*, John Wiley & Sons, 2012.

Folke, Olle and Johanna Rickne, "Sexual Harassment and Gender Inequality in the Labor Market," *The Quarterly Journal of Economics*, 2022, *137* (4), 2163–2212.

Gagliarducci, Stefano and M Daniele Paserman, "Gender Interactions within Hierarchies: Evidence from the Political Arena," *The Review of Economic Studies*, 2012, *79* (3), 1021–1052.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 2000, *108* (5), 874–904.

Ghosh, Arkadev, "Religious Divisions and Production Technology: Experimental Evidence from India," Technical Report 2022.

Haltiwanger, John C, Julia I Lane, and James Spletzer, "Productivity differences across employers: The roles of employer size, age, and human capital," *American Economic Review*, 1999, *89* (2), 94–98.

Hansen, Zeynep, Hideo Owan, and Jie Pan, "The Impact of Group Diversity on Performance and Knowledge Spillover–An Experiment in a College Classroom," 2006.

Heikensten, Emma and Siri Isaksson, "Simon Says: Examining Gender Differences in Advice Seeking and Influence in the Lab," *Available at SSRN 3273186*, 2019.

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

Hong, Lu and Scott E Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences*, 2004, *101* (46), 16385–16389.

Hoogendoorn, Sander, Simon C Parker, and Mirjam van Praag, "Ability dispersion and team performance: A field experiment," 2012.

Hoxby, Caroline, "Peer effects in the classroom: Learning from gender and race variation," 2000.

Hu, Die, Maoyan She, Lingfeng Ye, and Zhiwei Wang, "The more the merrier? Inventor team size, diversity, and innovation quality," *Science and Public Policy*, 2021, *48* (4), 508–520.

Kahane, Leo, Neil Longley, and Robert Simmons, "The effects of coworker heterogeneity on firm-level output: assessing the impacts of cultural and language diversity in the National Hockey League," *Review of Economics and Statistics*, 2013, *95* (1), 302–314.

Krammer, Sorin, "Research performance in Business and Management: The impact of team size and diversity," *Academy of Management Proceedings*, 2021, *2021* (1), 10981.

Lazear, Edward P, "Globalisation and the market for team-mates," *The Economic Journal*, 1999, *109* (454), 15–40.

— , "Diversity and immigration," in "Issues in the Economics of Immigration," University of Chicago Press, 2000, pp. 117–142.

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

Menzel, Andreas, "Knowledge exchange and productivity spill-overs in Bangladeshi garment factories," *Journal of Economic Behavior & Organization*, 2021, *185*, 721–746.

Prat, Andrea, "Should a team be homogeneous?," *European Economic Review*, 2002, *46* (7), 1187–1207.

Sandvik, Jason J, Richard E Saouma, Nathan T Seegert, and Christopher T Stanton, "Workplace knowledge flows," *The Quarterly Journal of Economics*, 2020, *135* (3), 1635–1680.

Shan, Xiaoyue, "Does minority status drive women out of male-dominated fields?," *Unpublished manuscript*, 2020.

Sharma, Karmini and Alessandro Castagnetti, "Demand for Information by Gender: An Experimental Study," Technical Report, Working Paper 2019.

Shore, Lynn M., Beth G. Chung-Herrera, Michelle A. Dean, Karen Holcombe Ehrhart, Don I. Jung, Amy E. Randel, and Gangaram Singh, "Diversity in organizations: Where are we now and where are we going?," *Human Resource Management Review*, 2009, *19* (2), 117–133.

Tables

Table 1: When is skill diversity most beneficial?

Dependent Variable:	**Change in score (%)**		
Sample:		*Overall*	
	All Nurses (1)	Male Nurses (2)	Female Nurses (3)
Skill diverse team (α)	4.075**	3.267	4.601
	(1.994)	(2.511)	(3.019)
Skill diverse team × Team size 4 (β)	5.762*	8.256**	3.523
	(3.047)	(3.975)	(4.513)
Team size 4 (γ)	0.361	-1.667	2.111
	(2.369)	(2.991)	(3.517)
Male	-4.278***		
	(1.599)		
Constant	10.778***	7.333***	10.222***
	(1.747)	(1.979)	(2.272)
P val: Diversity + Diversity × Team 4 = 0	0.000	0.001	0.020
P val: Team 4 + Diversity × Team 4 = 0	0.002	0.016	0.053
P val: $\alpha = \gamma$	0.087	0.077	0.460
P val: $\alpha_{male} = \alpha_{female}$			0.734
P val: $\beta_{male} = \beta_{female}$			0.432
P val: $\gamma_{male} = \gamma_{female}$			0.414
Mean of Individual Score (%)	47.658	51.570	43.787
Observations	187	93	94
Cluster	88	45	44

* p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at study team levels. Change in score is score in the test with team minus score in the test taken individually. Test remains the same for each person across the two rounds. Skill diverse team is a dummy variable that equals to one when there is an economics student in the team. Team size 4 is a dummy variable that equals one when there are four students in the team. Male is a dummy variable that equals one when the participant is a male student.

Table 2: Does gender diversity inhibit the effect of skill diversity?

Dependent Variable:	**Change in score (%)**		
	All Nurses	Male Nurses	Female Nurses
	(1)	(2)	(3)
Skill diverse × gender homogeneous (α)	1.623	2.267	-0.222
	(2.084)	(2.409)	(3.431)
Skill diverse × gender diverse (β)	5.918**	4.267	7.232**
	(2.393)	(3.431)	(3.328)
Skill diverse × gender homogeneous × team size 4 (γ)	7.155*	7.178	8.333
	(3.679)	(4.894)	(5.487)
Skill diverse × gender diverse × team size 4 (δ)	4.976	9.225*	0.905
	(3.373)	(4.599)	(4.868)
Team size 4	0.351	-1.667	2.111
	(2.383)	(3.025)	(3.557)
Male	-4.176**		
	(1.603)		
Constant	10.737***	7.333***	10.222***
	(1.756)	(2.001)	(2.298)
P val: $\alpha = \beta$	0.062	0.521	0.039
P val: $\gamma = \delta$	0.556	0.694	0.171
P val: $\alpha_{male} = \alpha_{female}$			0.553
P val: $\beta_{male} = \beta_{female}$			0.535
P val: $\gamma_{male} = \gamma_{female}$			0.875
P val: $\delta_{male} = \delta_{female}$			0.217
Mean of Individual Score (%)	47.658	51.570	43.787
Observations	187	93	94
Cluster	88	45	44

* p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at study team levels. Change in score is score in the test with team minus score in the test taken individually. Test remains the same for each person across the two rounds. Skill diverse is a dummy variable that equals one when there is an economics student in the team. Team size 4 is a dummy variable that equals one when there are four students in the team. Gender homogeneous is a dummy variable that equals one when all the team members have the same sex. Gender diverse is a dummy variable that equals one when the economist is of a different sex from the nurses.

Figures

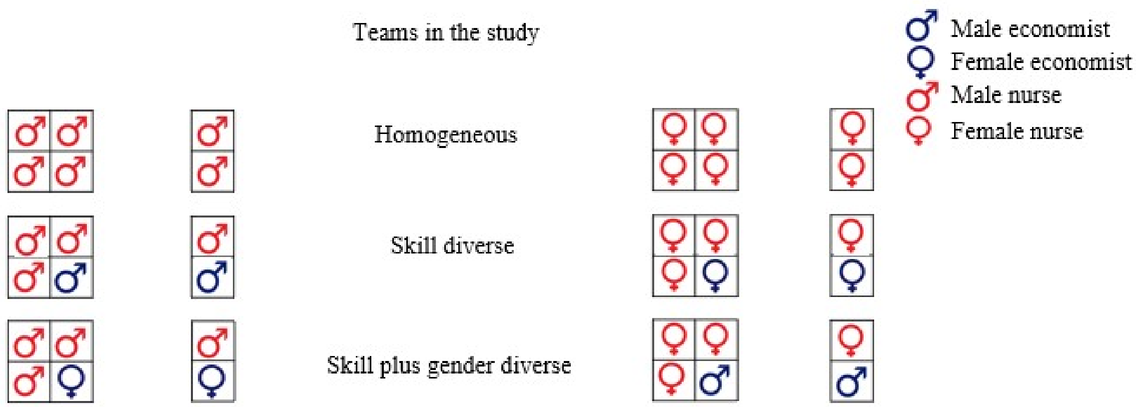

Figure 1: This figure shows the teams in the experiment. Notice that gender diversity only occurs in skill-diverse teams.

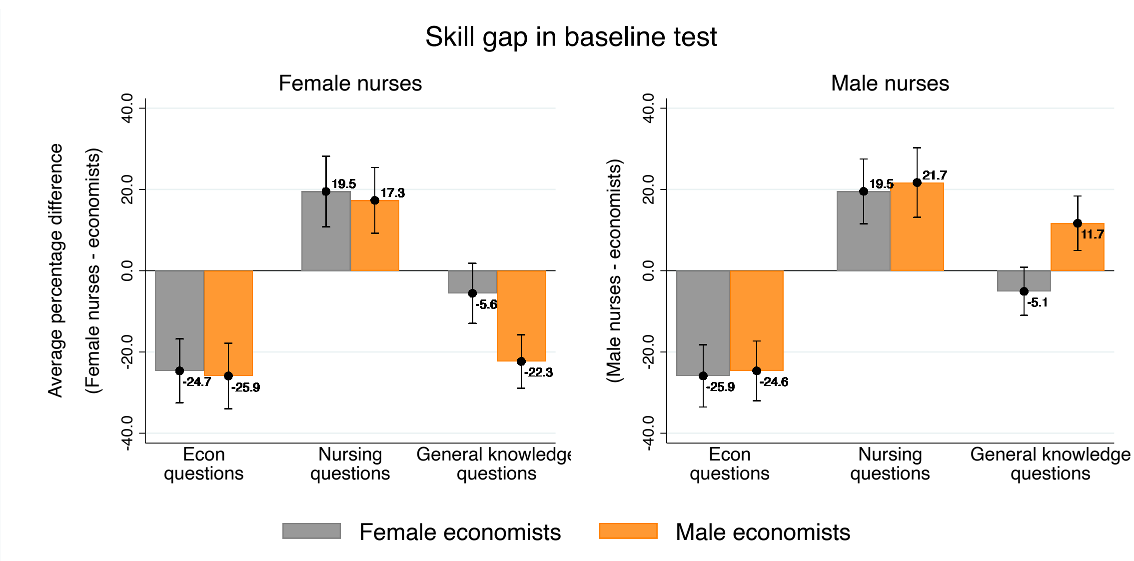

Figure 2: Skill gap in baseline test between Nurses and Economists

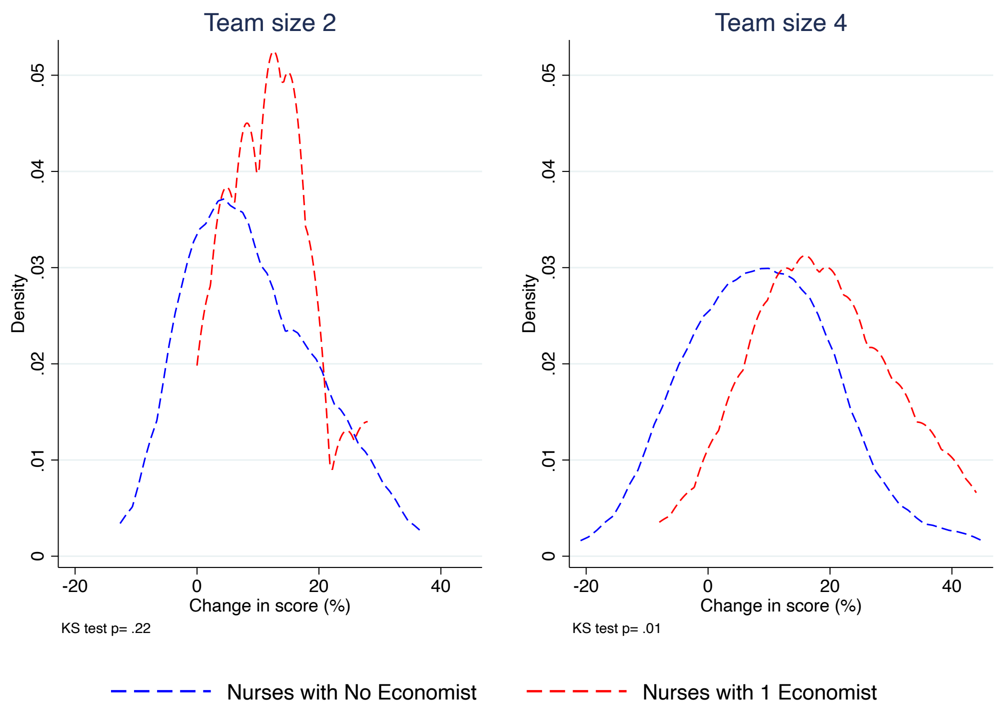

Figure 3: The effect of skill diversity on the percentage change in performance of Nursing students in small teams of 2 (left-hand side) & large teams of 4 (right-hand side). A Kolmogorov–Smirnov test of the equality of the two distributions has a p-value of 0.22 for teams of size 2, and a p-value of 0.01 for teams of size 4.

A Appendix Figures and Tables

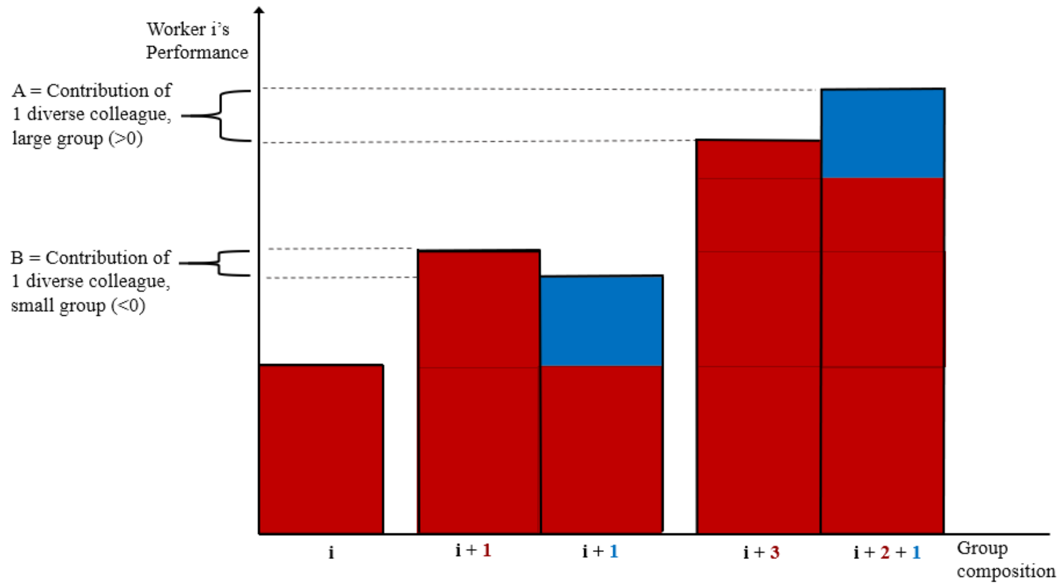

Figure A1: Illustration of the theoretical model for small teams. Each bar in this figure indicates the expected performance (net of communication costs) for an isolated type-1 worker *i* and for the same worker in different team compositions. The number after the plus (+) sign below each column indicate the number of i's teammates, with red indicating same-type (type-1) and blue other-type (type-2) teammates.

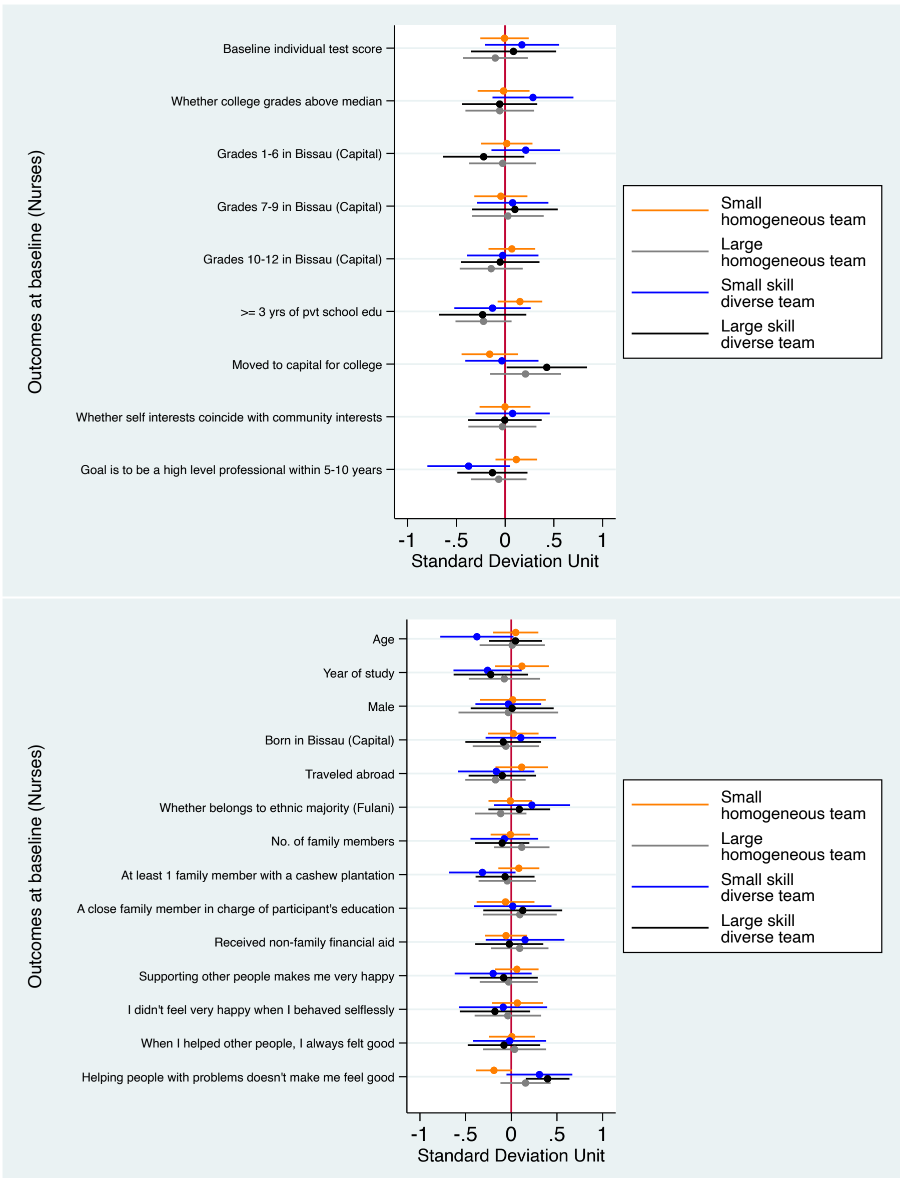

Figure A2: Balance tests for nurses: baseline variables by different types of teams. Reporting coefficients and 95% confidence intervals.

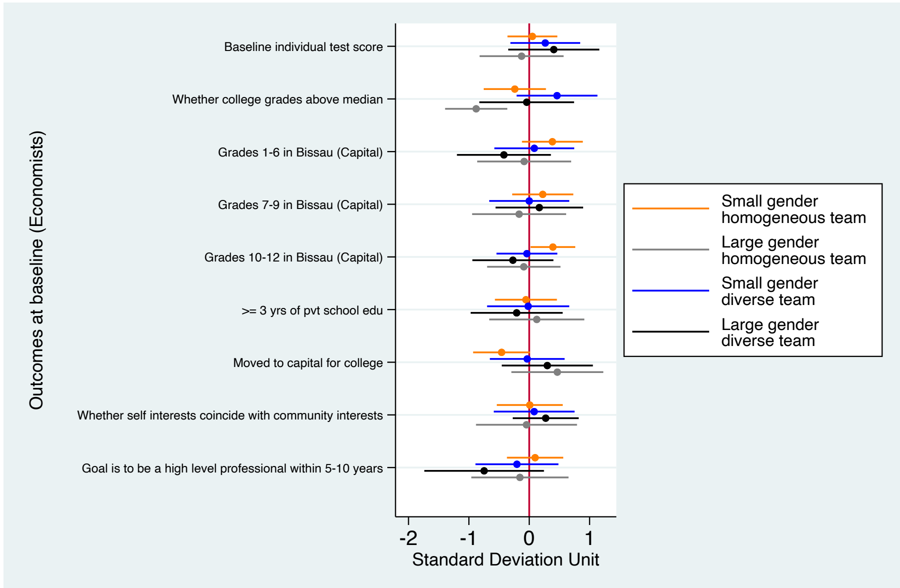

Figure A3: Balance tests for economists: baseline variables by different types of teams. Reporting coefficients and 95% confidence intervals.

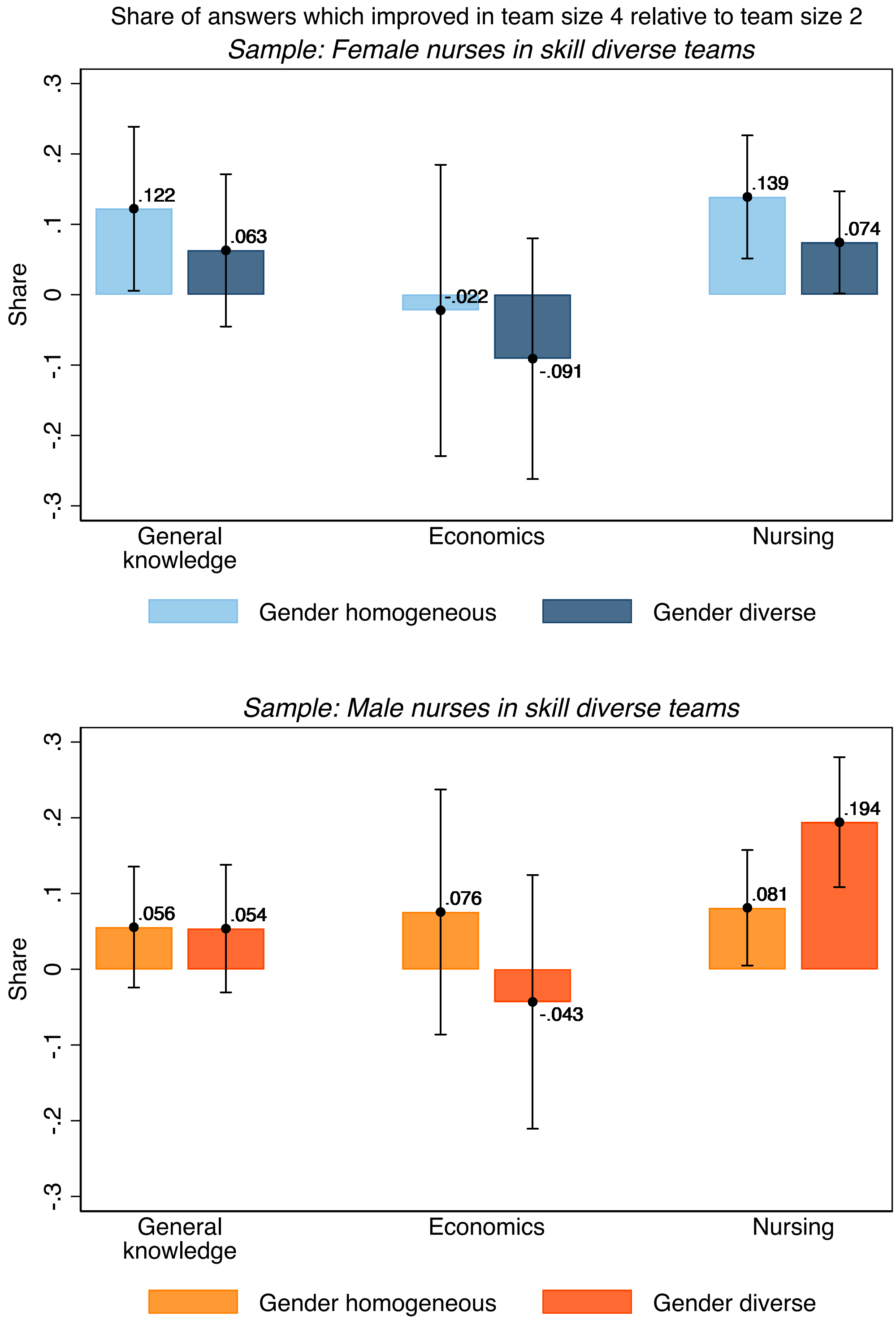

Figure A4: Share of answers that improved with team size

Table A1: When is skill diversity most beneficial? Including session fixed effects

Dependent Variable:	**Change in score (%)**		
Sample:	*Overall*		
	All Nurses (1)	Male Nurses (2)	Female Nurses (3)
Skill diverse team (α)	3.778*	3.771	3.998
	(1.933)	(2.636)	(2.689)
Skill diverse team × Team size 4 (β)	4.851	6.419*	2.678
	(2.966)	(3.794)	(4.530)
Team size 4 (γ)	1.333	0.038	3.349
	(2.134)	(2.904)	(3.075)
Male	-4.393***		
	(1.459)		
Constant	15.708***	11.575***	15.186***
	(2.302)	(3.473)	(2.957)
P val: Diversity + Diversity × Team 4 = 0	0.000	0.000	0.046
P val: Team 4 + Diversity × Team 4 = 0	0.002	0.013	0.056
P val: $\alpha = \gamma$	0.219	0.150	0.833
P val: $\alpha_{male} = \alpha_{female}$			0.999
P val: $\beta_{male} = \beta_{female}$			0.649
P val: $\gamma_{male} = \gamma_{female}$			0.597
Mean of Individual Score (%)	47.658	51.570	43.787
Observations	187	93	94
Cluster	88	45	44
Session fixed effect	Yes	Yes	Yes

* p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at study team levels. Change in score is score in the test with team minus score in the test taken individually. Test remains the same for each person across the two rounds. Skill diverse team is a dummy variable that equals to one when there is an economics student in the team. Team size 4 is a dummy variable that equals one when there are four students in the team. Male is a dummy variable that equals one when the participant is a male student. Control for participants who attended multiple sessions included.

Table A2: Does gender diversity inhibit the effect of skill diversity? Including session fixed effects

Dependent Variable:	**Change in score (%)**		
	All Nurses	Male Nurses	Female Nurses
	(1)	(2)	(3)
Skill diverse × gender homogeneous (α)	0.370	2.210	-2.225
	(2.096)	(2.500)	(3.637)
Skill diverse × gender diverse (β)	6.267***	5.443	7.257***
	(2.343)	(3.706)	(2.660)
Skill diverse × gender homogeneous × team size 4 (γ)	6.638*	5.146	8.545
	(3.487)	(4.456)	(5.329)
Skill diverse × gender diverse × team size 4 (δ)	3.812	7.106	-0.333
	(3.322)	(4.516)	(5.141)
Team size 4	1.393	0.270	3.388
	(2.119)	(2.892)	(3.079)
Male	-4.265***		
	(1.450)		
Constant	15.906***	11.709***	15.491***
	(2.313)	(3.706)	(2.522)
P val: $\alpha = \beta$	0.021	0.358	0.019
P val: $\gamma = \delta$	0.444	0.683	0.120
P val: $\alpha_{male} = \alpha_{female}$			0.359
P val: $\beta_{male} = \beta_{female}$			0.811
P val: $\gamma_{male} = \gamma_{female}$			0.617
P val: $\delta_{male} = \delta_{female}$			0.407
Mean of Individual Score (%)	47.658	51.570	43.787
Observations	187	93	94
Cluster	88	45	44
Session fixed effect	Yes	Yes	Yes

* p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at study team levels. Change in score is score in the test with team minus score in the test taken individually. Test remains the same for each person across the two rounds. Skill diverse is a dummy variable that equals one when there is an economics student in the team. Team size 4 is a dummy variable that equals one when there are four students in the team. Gender homogeneous is a dummy variable that equals one when the all the team members have the same sex. Gender diverse is a dummy variable that equals one when the economist is of a different sex from the nurses. Control for participants who attended multiple sessions included.

Table A3: When is skill diversity most beneficial? Including a control for each team's average baseline performance

Dependent Variable:	**Change in score (%)**		
Sample:	*Overall*		
	All Nurses	Male Nurses	Female Nurses
	(1)	(2)	(3)
Skill diverse team (α)	3.943*	2.809	3.076
	(2.029)	(2.400)	(3.330)
Skill diverse team $\times$ Team size 4 (β)	5.968*	8.174**	5.066
	(3.056)	(3.907)	(4.548)
Team size 4 (γ)	0.318	-2.074	1.505
	(2.363)	(3.018)	(3.467)
Male	-4.564***		
	(1.721)		
Constant	8.569**	17.547**	0.769
	(4.309)	(7.700)	(4.381)
P val: Diversity + Diversity $\times$ Team 4 = 0	0.000	0.001	0.011
P val: Team 4 + Diversity $\times$ Team 4 = 0	0.002	0.028	0.023
P val: $\alpha = \gamma$	0.094	0.081	0.614
P val: $\alpha_{male} = \alpha_{female}$			0.839
P val: $\beta_{male} = \beta_{female}$			0.475
P val: $\gamma_{male} = \gamma_{female}$			0.461
Mean of Individual Score (%)	47.658	51.570	43.787
Observations	187	93	94
Cluster	88	45	44

* p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at study team levels. Change in score is score in the test with team minus score in the test taken individually. Test remains the same for each person across the two rounds. Skill diverse team is a dummy variable that equals to one when there is an economics student in the team. Team size 4 is a dummy variable that equals one when there are four students in the team. Male is a dummy variable that equals one when the participant is a male student. All regressions include a control for the group's average baseline performance in the individual test.

Table A4: Does gender diversity inhibit the effect of skill diversity? Including a control for each team's average baseline performance

Dependent Variable:	**Change in score (%)**		
	All Nurses (1)	Male Nurses (2)	Female Nurses (3)
Skill diverse × gender homogeneous (α)	1.569	2.059	-0.902
	(2.096)	(2.383)	(3.734)
Skill diverse × gender diverse (β)	5.832**	3.390	5.369
	(2.427)	(3.273)	(3.767)
Skill diverse × gender homogeneous × team size 4 (γ)	7.313*	6.345	9.682*
	(3.729)	(4.626)	(5.537)
Skill diverse × gender diverse × team size 4 (δ)	5.037	9.841**	2.096
	(3.390)	(4.644)	(4.968)
Team size 4	0.327	-2.150	1.536
	(2.380)	(3.060)	(3.519)
Male	-4.330**		
	(1.741)		
P val: $\alpha = \beta$	0.066	0.661	0.095
P val: $\gamma = \delta$	0.541	0.491	0.163
P val: $\alpha_{male} = \alpha_{female}$			0.548
P val: $\beta_{male} = \beta_{female}$			0.605
P val: $\gamma_{male} = \gamma_{female}$			0.869
P val: $\delta_{male} = \delta_{female}$			0.243
Mean of Individual Score (%)	47.658	51.570	43.787
Observations	187	93	94
Cluster	88	45	44

* p<0.1, ** p<0.05, *** p<0.01. Standard errors clustered at study team levels. Change in score is score in the test with team minus score in the test taken individually. Test remains the same for each person across the two rounds. Skill diverse is a dummy variable that equals one when there is an economics student in the team. Team size 4 is a dummy variable that equals one when there are four students in the team. Gender homogeneous is a dummy variable that equals one when the all the team members have the same sex. Gender diverse is a dummy variable that equals one when the economist is of a different sex from the nurses. All regressions include a control for the group's average baseline performance in the individual test.

Table A5: Whether variables that are not balanced across different types of teams are correlated with baseline performance (individual test)

	Baseline performance (individual test)		
	Overall	Nurse	Economist
	(1)	(2)	(3)
Moved to capital for college	-0.078	0.313	-0.455
	(0.439)	(0.509)	(1.017)
Age	-0.059	-0.076	0.070
	(0.050)	(0.053)	(0.141)
Helping people with problems doesn't make me feel good	-0.904	-1.187	-0.273
	(0.609)	(0.692)	(1.338)
Constant	14.032***	14.083***	11.449**
	(1.280)	(1.373)	(3.544)
Observations	195	146	49

* p<0.1, ** p<0.05, *** p<0.01. Moved to capital for college is a dummy variable that turns on 1 if a student moved from inland / islands to the capital to complete their higher education. Age of student is measured in years. "Helping people with problems doesn't make me feel good" is measured on a Likert scale with a score of 1 for strongly disagree and a score of 5 for strongly agree.

B Appendix: Recruitment and randomization

Recruitment All the students were recruited from the Bachelor's degrees in Economics and Nursing at the Universidade Lusófona da Guiné-Bissau (ULG). We excluded first year students because of the expected lower familiarity with their respective discipline. We chose students from these degrees for two reasons. First, Nursing was the degree with most students at the university, and the only one with enough female and male students to guarantee enough participants for the study. Economics students were chosen as the minority because the skill set of economists was disjoint from the skill set of nurses, suggesting a meaningful complementarity of skills in teams.[15] Of the 1104 registered students 503 signed up as available to participate in the experiment (387 Nursing and 116 Economics students). From this set, we randomly chose an identical number of female and male students from each degree and invited them to come on the day of the experiment. A total of 242 students participated in the study (62 economists and 180 nurses). The experiment took place in four sessions, within a single day, at the university campus. Each session was planned such that it would have at least one team of each type.

Payment After the main task, students completed a survey that asked them about their demographic and other information. Participants were incentivized to complete the survey by paying them 500 FCFA if the survey was completed. In addition to the above, each student received 500 FCFA (0.72$) for participating in the study, independently of performance.

The tests and survey were paper-based. The tests were corrected on the spot, so that payment was implemented shortly after finishing the survey. Room supervisors were instructed to place students far enough from each other in the individual and team-based test, so that it was not possible for them to cheat or hear what others were discussing. No incidents were reported.[16]

Randomization After consent and registration but before each session, participants were invited to wait in a room where all the participants for each session were randomly assigned to their teams at the same time. The mechanism for random assignment to teams was as follows. First, a unique ID with room and team assignment was created for each participant. This was printed on cards and placed in a bag. Since we wanted to stratify by gender of nurses and also randomize the gender of economists, we had to find a way to keep track of gender of participants but without making gender salient. To minimize gender salience, gender was never mentioned throughout the experiment. We asked participants (in no particular order and not by gender) to pick up their card from a bag one by one. Women and men chose their cards from different (but identically-looking) bags. Students were then guided towards their respective rooms as mentioned on their cards. A room supervisor then read out loud the instructions of the experiment. After finishing and handing-in the first test to their room supervisor, students were organized into teams as indicated by their ID card that they picked before the start of the experiment.

[15] In the recruitment announcement, it was stated that each participant would solve some simple tests for 2 hours and thirty minutes session. Students in both degrees were taking in-person exams in the weeks before the experiment. We took advantage of this opportunity to present the study to them and to register their interest and availability to participate in it.

[16] The rules were strict. A participant would be disqualified without any payment were she to break the rules, including talking to other participants (outside her team) or using her cellphone. In the team test the whole team would be disqualified if a single one of its members cheated.

C Appendix: Discussion of minorities in large vs small teams

While our focus is on the performance of nurses, our majoritarian group, the net benefit of diversity depends also on how minorities feel and perform. Abrupt policies that increase workplace diversity may have unintended consequences on the minority if there is limited integration and inclusion (Folke and Rickne, 2022). Unfortunately, the nature of team interaction in our experiment is too stylized and short to capture aspects of minority inclusion which are important to consider in reality. Nevertheless, the performance of our minority (economists) across team types gives us some insights into this issue. Figure C1 shows that economists' performance improvement is greatest in a large team which is gender homogeneous. However, being assigned to a large team where economists are a minority both in gender and skills does not improve their performance with respect to working in pairs.[17] This result complements the one for women: the combination of skill and gender diversity in large teams seem to create communication costs which are disproportionately higher than in other combinations.

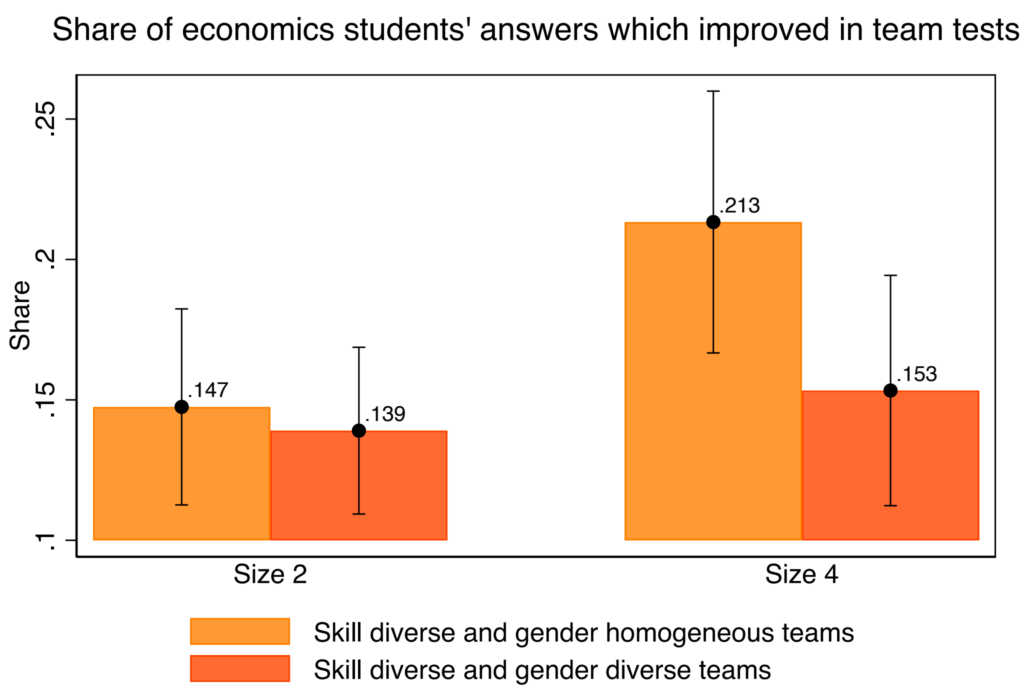

Figure C1: Share of minority students' answers that improved in teams of different sizes and composition.

[17]The number of observations becomes too small, but this gap is driven by female economists in male-dominated teams, who significantly decrease their improvement.

D Appendix: Theory extension and proofs

D.1 Extension: introducing gender diversity

This extension of the model considers the case in which the skill diverse team member is also of a different gender. Communication between members of a different gender has an additional cost $c_g(m)$, which we allow to vary with team size m. We make inter-gender communication costs a function of team size to capture the fact that in larger teams one gender may be discouraged from effectively participating in communication exchanges. Studies have shown that women's lead is affected when they are part of mixed-gender teams (Born et al., 2022), it might be the case that their confidence and initiative suffers in gender-diverse teams as these become larger.

We assume that men and women have the same ability distribution within types, thus there is no additional gain or loss from interacting with a person with a diverse skill set who is of the opposite gender. If there are n workers of the opposite gender in the team, then worker i's expected performance is: $E(z_i|m,k,n) = pZ_{1,m} + (1-p)Z_{2,k} - kc_o - nc_g(m)$.

If all type-2 workers are of the opposite gender with respect to type-1 workers, that is $k = n$, the expected performance of worker i in a skill and gender diverse team is:

$$E(z_i|m,k,k) = pZ_{1,m} + (1-p)Z_{2,k} - k(c_o + c_g(m))$$

The following proposition explains how worker's i expected performance changes in gender heterogeneous vs homogeneous teams of different sizes:

Proposition 3. *If $c_g(m)$ is independent of team size, the gain in expected performance for type-1 workers from introducing a same-gender team member should be the same in gender diverse and gender homogeneous teams. If $c_g(m)$ increases (decreases) with team size, then type-1 workers in gender diverse teams benefit less (more) from an increase in team size than those in homogeneous ones.*[18]

In the first case gender diversity does not influence the relationship between team size and skill diversity because it goes hand-in-hand with the costs of communicating with skill-diverse members.[19] . This is not the case if the inter-gender costs of communication vary with team size. The intuition is the following: if cost $c_g(m)$ decreases as team size increases, it becomes less costly to communicate with others. As a result, not only do type-1 workers enjoy a gain in performance due to the larger team size, but they also enjoy a reduction in communication cost and can spend more time discussing questions that matter rather than solving communication issues. Since gender homogeneous teams do not enjoy any cost reduction from this increase in team size, the gain in performance is larger in gender diverse teams than gender homogeneous ones. On the other hand, if cost $c_g(m)$ increases as team size increases, then expanding team size also entails additional costs. As a result, because it is more costly to communicate with others, performance in gender-diverse teams will increase less in team size than in gender-homogeneous ones.

In our experiment, female nurses in gender-diverse teams seem to face a larger increase in the inter-gender communication cost $c_g(m)$ as team size m increases than male nurses do in a symmetric team.

[18]One could also assume that the cost c_o of inter-type communication is a function of team size. A positive correlation with size would weaken the result in proposition 2, while a negative correlation would strengthen it. We leave this aside and focus on the empirically-relevant case of gender-related costs.

[19]The change in i's expected performance in a mixed-gender team is: $\Delta_k(m) = E(z_i|m+1,k,k) - E(z_i|m,k,k) = p[Z_{1,m+1} - Z_{1,m}]$. When the team is gender homogeneous ($n = 0$): $\Delta_0(m) = E(z_i|m+1,k,0) - E(z_i|m,k,0) = p[Z_{1,m+1} - Z_{1,m}]$.

D.2 Proofs

Proposition 1 The effect of group size on expected performance is concave.

Proof. Worker i's expected performance in a group of m workers is:

$$E(z_i|m) = pZ_{1,m} - (m-1)c$$

On the other hand, worker i's expected performance in a group of $m+1$ workers is:

$$E(z_i|m+1) = pZ_{1,m+1} - mc$$

Therefore, the gain in performance from having an additional worker in the group is:

$$\begin{aligned}\Delta(m) &= E(z_i|m+1) - E(z_i|m)\\ &= pZ_{1,m+1} - mc - (pZ_{1,m} - (m-1)c)\\ &= p(Z_{1,m+1} - Z_{1,m}) - c\end{aligned}$$

Since $Z_{1,m}$ is increasing and concave, it holds that $Z_{1,m+1} - Z_{1,m}$ decreases as m increases. Thus, $\Delta(m)$ decreases as m increases, which proves that $E(z_i|m)$ is concave. □

Proposition 2 For a type-1 worker, the gain in expected performance from being placed in a diverse team with k type-2 workers instead of in a same-sized homogeneous team is increasing in team size.

Proof. The difference between worker i's performance in the diverse and homogeneous team of size $m+k$ is: $\Delta_k(m) = p(Z_{1,m} - Z_{1,m+k}) + (1-p)Z_{2,k} - kc_o$ On the other hand, the difference between worker i's performance in the diverse and homogeneous team of size $l+k$ is:

$$\Delta_k(l) = p(Z_{1,l} - Z_{1,l+k}) + (1-p)Z_{2,k} - kc_o$$

We have:

$$\begin{aligned}\Delta_k(m) - \Delta_k(l) &= p(Z_{1,m} - Z_{1,m+k}) + (1-p)Z_{2,k} - kc_o - (p(Z_{1,l} - Z_{1,l+k}) + (1-p)Z_{2,k} - kc_o)\\ &= p\left[(Z_{1,m} - Z_{1,m+k}) - (Z_{1,l} - Z_{1,l+k})\right]\end{aligned}$$

Assuming that we have $m > l$, since $Z_{1,m}$ is concave, it holds that $|Z_{1,m} - Z_{1,m+k}| < |Z_{1,l} - Z_{1,l+k}|$. Since $Z_{1,m} - Z_{1,m+k} < 0$, it holds that $Z_{1,m} - Z_{1,m+k} > Z_{1,l} - Z_{1,l+k}$. Thus, $\Delta_k(m) - \Delta_k(l) > 0$, thereby proving that the relative contribution of k workers is increasing in m. □

Proposition 3 If $c_g(m)$ is independent of group size, the gain in expected performance for type-1 workers from introducing a same-gender group member should be the same in gender diverse and gender homogeneous groups. If $c_g(m)$ increases (decreases) with group size, then type-1 workers in gender diverse groups benefit less (more) from an increase in group size than homogeneous ones.

Proof. $c_g(m)$ **is independent from group size**
This case means that c_g is a constant. The change in expected performance from having an

additional type-1 worker of same gender as worker i in a gender diverse group is:

$$
\begin{aligned}
\Delta_k(m) &= E(z_i|m+1,k,k) - E(z_i|m,k,k) \\
&= pZ_{1,m+1} + (1-p)Z_{2,k} - k(c_o + c_g) - (pZ_{1,m} + (1-p)Z_{2,k} - k(c_o + c_g)) \\
&= p(Z_{1,m+1} - Z_{1,m})
\end{aligned}
$$

On the other hand, we can obtain the change in expected performance from the same addition in a gender homogeneous group;

$$
\begin{aligned}
\Delta_0(m) &= E(z_i|m+1,k,0) - E(z_i|m,k,0) \\
&= p(Z_{1,m+1} - Z_{1,m})
\end{aligned}
$$

The change in expected performance from having an additional type-2 worker of same gender as worker i in a gender diverse group is:

$$
\begin{aligned}
\Delta_k(k) &= E(z_i|m,k+1,k) - E(z_i|m,k,k) \\
&= pZ_{1,m} + (1-p)Z_{2,k+1} - (k+1)c_o - kc_g - (pZ_{1,m} + (1-p)Z_{2,k} - k(c_o + c_g)) \\
&= (1-p)(Z_{2,k+1} - Z_{2,k}) - (k+1-k)c_o \\
&= (1-p)(Z_{2,k+1} - Z_{2,k}) - c_o
\end{aligned}
$$

Again, we can obtain the change in expected performance from the same addition in a gender homogeneous group; $\Delta_0(k) = (1-p)(Z_{2,k+1} - Z_{2,k}) - c_o$ Therefore, it holds $\Delta_k(m) = \Delta_0(m)$ and $\Delta_k(k) = \Delta_0(k)$, that is that the change in expected performance from having an additional same-gender worker should be the same in both gender diverse and homogeneous groups.

$c_g(m)$ increases with group size ($c'_g > 0$)

The change in expected performance from having an additional type-1 worker of same gender as worker i in a gender diverse group is:

$$
\begin{aligned}
\Delta_k(m) &= E(z_i|m+1,k,k) - E(z_i|m,k,k) \\
&= pZ_{1,m+1} + (1-p)Z_{2,k} - k(c_o + c_g(m+k+1)) - (pZ_{1,m} + (1-p)Z_{2,k} - k(c_o + c_g(m+k))) \\
&= p(Z_{1,m+1} - Z_{1,m}) - k(c_g(m+k+1) - c_g(m+k))
\end{aligned}
$$

From there, we can obtain the change in expected performance following the same addition in a gender homogeneous group;

$$\Delta_0(m) = p(Z_{1,m+1} - Z_{1,m})$$

As a result, the difference in change in gender diverse and homogeneous groups is:

$$
\begin{aligned}
\Delta_m(k) &= \Delta_k(m) - \Delta_0(m) \\
&= p(Z_{1,m+1} - Z_{1,m}) - k(c_g(m+k+1) - c_g(m+k)) - p(Z_{1,m+1} - Z_{1,m}) \\
&= -k(c_g(m+k+1) - c_g(m+k))
\end{aligned}
$$

The change in expected performance from having an additional type-2 worker of same gender as worker i in a gender diverse group is:

$$\begin{aligned}\Delta_k(k) &= E(z_i|m,k+1,k) - E(z_i|m,k,k)\\ &= pZ_{1,m} + (1-p)Z_{2,k+1} - (k+1)c_o - kc_g(m+k+1) - (pZ_{1,m} + (1-p)Z_{2,k} - k(c_o + c_g(m+k)))\\ &= (1-p)(Z_{2,k+1} - Z_{2,k}) - c_o - k(c_g(m+k+1) - c_g(m+k))\end{aligned}$$

From there, we can obtain the change in expected performance following the same addition in a gender homogeneous group;

$$\Delta_0(k) = (1-p)(Z_{2,k+1} - Z_{2,k}) - c_o$$

As a result, the difference in change in gender diverse and homogeneous groups is:

$$\begin{aligned}\Delta_k(k) &= \Delta_k(k) - \Delta_0(k)\\ &= (1-p)(Z_{2,k+1} - Z_{2,k}) - c_o - k(c_g(m+k+1) - c_g(m+k))\\ &- ((1-p)(Z_{2,k+1} - Z_{2,k}) - c_o)\\ &= -k(c_g(m+k+1) - c_g(m+k))\end{aligned}$$

Thus, we see that if $c_g(.)$ increases with group size, we have $\Delta_m(k) < 0$ and $\Delta_k(k) < 0$, that is the increase in expected performance from an extra same-gender colleague is greater in gender homogeneous groups than gender diverse ones. The opposite holds if $c_g(.)$ decreases with group size. $\square$

Appendix: For Online Publication
Instructions to the participants

Instructions for the supervisors of the experiment "Group structure and performance".

Wait for the participants of the session to arrive in the room: make sure that the correct room is indicated on the students' placards (green for economics and pink for nursing). Make sure that all participants are seated as far apart as possible. Complete the record sheet. Read the instructions to the participants.

Good morning! In this room we are going to perform 3 types of tests:

- the first will be individual and will have 25 questions
- the second will be done in groups and will have 25 questions
- each question has only one correct option.
- if a question is answered with more than one option, or with the answer chosen in a different way from that indicated in the document, the answer will be considered wrong.
- the third consists of a survey to be completed with personal information.

The results of these tests will be used in an independent study and will not be handed over to the University, so it has no consequence on your academic results. All the information you provide will be completely confidential.

You will have a fixed payment per person of 500 FCFA, regardless of the test result.

Each of you will have to answer different questions: some are from the subject of your course of study, others from the subject of your classmates' course, and some general knowledge questions.

The first test will be taken individually, and you will be paid 80 francs per correct answer, up to a maximum of 2000 FCFA. The time is 30 minutes. I will let you know when there are 10 minutes left until the time is up. When I announce that time is up, everyone should stop writing, even if you haven't finished answering all the questions on the test, then I will collect the sheets. You may not use phones or talk to colleagues; if anyone is caught breaking the rules, they will be expelled, and will forfeit their right to financial compensation. If anyone has any questions (but not about the content of the questions), they must call me and ask me in a low voice, and I will decide whether the question is pertinent and extends to everyone or not.

Make sure all participants have their phones switched off. Deliver the individual tests, different for nursing and economics students. **Before starting the tests, all participants have to write on pages 1 and 3 of their tests their university registration number and study ID number (which they find on their test papers).** When everyone has finished the record, the supervisor should start counting the 30 minutes for solving the individual test (mark the time on the personal record sheet). Caution! DO NOT SAY THAT IT IS BETTER TO ANSWER A QUESTION RANDOMLY IF THEY DO NOT KNOW THE ANSWER. Warn when there are 10 minutes left before the end of the test. Collect individual tests. Record the time of completion on the sheet. Hand in tests to Jair (or substitute) who will bring in group tests. Organise the room into groups; the members of the groups have to sit at the corresponding table. DO NOT GIVE INDICATIONS ABOUT THE DESIRABILITY OF MORE OR LESS COLLABORATION BETWEEN MEMBERS OF THE SAME GROUP IN THE GROUP TEST. Read the instructions to the participants.

Now the same test will be repeated in groups. Each person has to complete **his or her own personal sheet with the** same questions, but in this test, they can communicate ONLY with their group mate(s). At the end of the group test, only one test of one of the group members will be randomly chosen and marked. All members of the group will be paid according to the number of correct answers in this test (80 francs per correct question, up to a maximum of 2000 FCFA). The time is 30 minutes. I will always tell you when there are 10 minutes left to go. When I announce that time is up, you must stop writing, even if you have not finished answering all the questions, and I will collect the sheets. You may not use phones or talk to colleagues from other groups or speak with a loud voice. Classmates from other groups in the same room should not be able to hear what the other groups are discussing. If one of you is found breaking the rules, all members of the group to which you belong will be expelled, without financial compensation. If anyone has a doubt (but it cannot be about the content of the questions), they must call me and ask me in a low voice, and I will decide whether the doubt is pertinent and extended or not to everyone then I will communicate it to the whole room.

Deliver the tests in groups, different for nursing and economics students. **Before starting the tests, all participants have to write on pages 1 and 3 of their tests their university registration number and study ID number (which they find on their test papers).** When everyone has finished their registration, the 30 minutes to complete the individual test begins (mark the time on the personal registration sheet). DO NOT SAY THAT IT IS BETTER TO ANSWER A QUESTION RANDOMLY IF THEY DO NOT KNOW THE ANSWER. Warn when there are 10 minutes to go. Collect tests as a group. Record the time of completion on the sheet. Hand in tests to Jair (or substitute) who will bring in the surveys. Read the instructions.

They will now complete a survey with personal information.
You will have a payment of 500 francs if the survey is completed in all its parts, in addition to the previous payments. You have 25 minutes to deliver the survey. I will let you know when 5 minutes are left. When you have finished completing the survey, you may hand it in and go to the auditorium, where you will be called for your payments.

Deliver the surveys, the same for all participants. Before starting, all participants should write on pages 1 and 3 of their tests their university registration number and study identification number (which they find on their test papers). When everyone has finished registering, start the 25 minutes to complete the survey (mark the time on the personal registration sheet). Help participants if they have any questions. Warn when there are 5 minutes to go. When the participants have finished completing the surveys, they can leave. Record the finishing time of the last participant on the sheet. When everyone has finished, take the surveys and the personal sheet to Lucia. Then the supervisors should go to the auditorium for the end of work evaluation meeting.

Examples of questions from the experiment

We included 120 different questions in the experiment: 40 general-knowledge questions were designed by professors at the partner university and the Ministry of Finance of Guinea-Bissau. Nursing questions and Economics questions were designed by professors from the respective departments. All questions were carefully checked for accuracy and then randomized into the different sessions and tests. As every question was only included in one session, no student saw or heard about the questions she or he faced in the experiment before taking the first test.

Examples for each group of questions include:

1. General-knowledge:

   - How many countries are there in Africa?
     (a) 72
     (b) 4
     (c) 46
     (d) 54
     (e) none of the previous answers
   - What is a water molecule composed of?
     (a) one hydrogen atom and one oxygen atom
     (b) two hydrogen atoms and one oxygen atom
     (c) one hydrogen atom and two oxygen atoms
     (d) two hydrogen atoms and two oxygen atoms
     (e) none of the above
   - Which of the following animals is impossible to find in Guinea-Bissau?
     (a) tortoise
     (b) chimpanzee
     (c) shark
     (d) giraffe
     (e) manatee

2. Economics:

   - The amount of currency in circulation in the economy is determined by...
     (a) the central bank
     (b) commercial banks
     (c) the General Direction for Taxes
     (d) financial markets
     (e) fractional reserve systems
   - The interest rate without discounting inflation is the...
     (a) real interest rate
     (b) nominal interest rate

(c) re-discount rate
(d) the tax pressure rate
(e) the discount rate

- The stock of equipment and structures used to produce goods and services is called...
  (a) technical knowledge
  (b) physical capital
  (c) innovation
  (d) current assets
  (e) human capital

3. Nursing:

- The following are factors that can produce pressure ulcers...
  (a) poor circulation and diabetes
  (b) urinary incontinence
  (c) generalized edema and paralysis
  (d) the three alternatives are correct
  (e) the three alternatives are incorrect

- The following is a basic measure for the control of contagious diseases...
  (a) vaccinate contacts
  (b) develop lectures
  (c) treat sources of infection
  (d) carry out immunological tests
  (e) carry out scientific research

- We consider "universal receiver" someone with the following blood type...
  (a) A-
  (b) B+
  (c) AB+
  (d) O-
  (e) the four alternatives are correct