Diversity in Teams: Collaboration and Performance in Experiments with Different Tasks*

Ornella Darova† and Anne Duchene‡

December 31, 2023

Abstract

We run two field experiments on team diversity in a large undergraduate economics class. We use a multidimensional measure of diversity based on gender, race, and migration status. Small groups with random compositions are generated and assigned team tasks. In the first experiment, tasks are creative, while they are analytical in the second one. We estimate the impact of diversity on teamwork quality and group performance. We find a significant U-shaped impact of diversity on teamwork quality in both experiments. However, the impact on performance depends on the type of assignment: it is positive for creative tasks, but negative for analytical ones. We interpret these results as the consequence of two conflicting forces: diversity is a source of creativity, but it can hamper communication and coordination between team members. When tasks are creative, the first (positive) force dominates; for analytical tasks, instead, communication challenges do. The U-shaped impact on teamwork quality suggests that *faultlines* – dividing lines that split a group into subgroups based on demographic characteristics – can cause inter-subgroup cohesion to break down, while very homogeneous or very heterogeneous groups collaborate better.

Keywords: Diversity, Knowledge Production, Creativity, Team Work, Education
JEL Codes: I21, J15, A22

*We thank helpful comments from Prof. Hanming Fang, Prof. Petra Todd, and from Prof. Francesco Agostinelli. This paper was presented at the Annual AEA Conference on Teaching and Research in Economic Education (CTREE), at the Annual Conference of the European Society for Population Economics (ESPE), at the Young Economists' Meeting (Masaryk University) and at a research seminar at the University of Pennsylvania. We thank the participants for their useful comments as well. Ornella Darova acknowledges the financial support of the Center for the Study of Ethnicity, Race and Immigration at the University of Pennsylvania. This study was registered in the AEA RCT Registry with ID AEARCTR-0009918 and digital object identifier (DOI) 10.1257/rct.9918-1.1.

†University of Pennsylvania
‡University of Pennsylvania

1 Introduction

In June 2023, the U.S. Supreme Court issued a ruling dismantling affirmative action in college admissions – a decision that might have significant implications beyond education and into the corporate workplace. The ruling comes at a time of unprecedented focus on diversity in education and in organizations, as minorities are increasingly represented in schools and the workforce, and cultural and institutional changes have increased gender diversity (Census Bureau, 2020). Simultaneously, learning and working environments have been shifting toward more and more teamwork and group problem-solving (Mathieu et al., 2014; Wuchty et al., 2007; Deming, 2017))[1]. As jobs in modern economies become increasingly complex and interdisciplinary, teams can outperform individuals by exploiting synergies between members (Garicano and Rossi-Hansberg, 2006; Lacerenza et al., 2018). In education, a large body of evidence shows the positive relationship between collaborative learning and student achievement, effort, persistence, and motivation (Springer et al., 1999; Johnson et al., 2007).

These changes raise an important question: do more diverse teams work better? Team diversity refers to variation among team members in distinct attributes, such as demographic endowments, socioeconomic characteristics, and educational background. On the one hand, diverse teams are more innovative and creative as they benefit from a wider pool of knowledge and experience, but on the other hand they face higher communication and coordination costs and can potentially lead to more conflicts (Lazear, 1999).

There is an extensive literature on the relationship between team diversity and performance, but it reveals mixed results[2]. Some papers highlight higher communication costs and lower trust levels among more heterogeneous individuals (Morgan and Várdy, 2009; Hamilton et al., 2012), while others show complementarities and benefits of information sharing (Mello and Ruckes, 2006; Prat, 2002; S. Horwitz and I. Horwitz, 2007). Kahane et al., 2013 find that hockey team performance positively correlates with the share of foreign team members, due to a larger pool of skills, but language and cultural differences override the gains from diversity if heterogeneity is particularly high.

This study aims to reconcile these mixed findings by producing causal assessments through two field experiments in a controlled environment: small random homework groups are generated in a large college undergraduate class.

There seems to be two key differences across previous studies that explain their inconsistent results. The first difference is the degree of creativity of the task performed by the team. Papers showing that diversity enhances team performance seem to focus on tasks that are highly creative or involve strategic and complex decision-making (Richard and Shelor, 2002; Jackson and Joshi, 2004; Wegge et al., 2008). In Freeman and Huang, 2015, nationally diverse

[1] According to Cross et al., 2021, collaborative work "has risen 50% or more over the past decade to consume 85% or more of people's work weeks".

[2] Alesina and La Ferrara, 2005 survey the literature on diversity and economic performance. For other detailed reviews, see Daan Van Knippenberg, 2004; Williams and O'Reilly, 1998, Simsarian Webber and Donahue, 2001, Jackson and Joshi, 2004, Guillaume et al., 2017.

research teams publish more often in high-impact journals, and in Ferrucci and Lissoni, 2019, migrant inventors increase team diversity and are associated with higher quality patents.[3] In an experimental setting similar to ours, Hoogendoorn and Praag, 2012 and Hoogendoorn, Oosterbeek, et al., 2013 find a positive impact of ethnic diversity in teams of undergraduate business students whose assignment is to start up a venture. By contrast, studies that find a negative impact of diversity focus on less creative, more routine tasks. In Leonard and D. Levine, 2006, age diversity has a negative effect on sales in retail stores. Lyons, 2017 finds that birthplace diversity hinders performance due to communication problems when tasks are highly specialized (see also Marx et al., 2021 and Hjort, 2014).

The second key difference is the portion of the diversity spectrum considered: the impact of diversity on team performance seems to be negative for relatively homogenous groups, and positive for highly diverse groups. The stream of literature on ethnic fractionalization initiated by Easterly and R. Levine, 1997 and Alesina, Devleeschauwer, et al., 2003 finds that low coordination and trust are associated with more heterogeneous societies, along with high levels of conflict (Alesina and La Ferrara, 2005). But this literature captures segregation and thus does not consider moderately to highly heterogeneous society. Similarly, Lyons, 2017 and Marx et al., 2021 find a negative impact of diversity on cohesion, but only consider groups that are either completely homogeneous or split in half across different nationalities. By contrast, Hoogendoorn and Praag, 2012 who find a positive impact of ethnic team diversity, consider moderate to high diversity by focusing on groups with at least 20% of non-Dutch members.

In this study, we attempt to reconcile these previously mixed findings by implementing two different experiments with comparable settings but different tasks, one more creative than the other. Moreover, we address the non-linear impact of diversity by building small groups with random compositions that represent the full spectrum of group diversity, from very homogeneous to very heterogeneous.

A key contribution of this paper is to build an index of teamwork quality – as an alternative to team performance – based on collaboration between members, balance of member contributions, and the absence of conflicts. We find that teamwork quality follows a U-shaped pattern, where very homogeneous and very heterogeneous groups show better teamwork, in both experiments. This result is consistent with the psychology and organizational behavior literature on *faultlines* (Lau and Murnighan, 1998; Carton and Cummings, 2013): faultlines are defined as hypothetical dividing lines that split a group into relatively homogeneous subgroups based on group members' alignment along multiple diversity dimensions (e.g., one subgroup with only white males and another with only Asian females). While such faultlines do not appear in very homogeneous and very heterogeneous groups, they might create coalitions or "splits" in intermediate groups, increasing the probability of conflict or lack of cooperation and ultimately hurting group cohesion.

[3] See also Hamilton et al., 2012, Ozgen et al., 2012 and Ozgen et al., 2013.

We further analyze the impact of diversity on group performance and find that a higher diversity translates into higher grades for creative tasks but lower ones for non-creative assignments. This result highlights how mixed findings regarding diversity impact in the literature may stem from different types of tasks. Diversity can result in creativity gains when tasks are of creative nature, but communication costs prevail when assignments are not creative instead.

A significant contribution of this paper is on the measure of diversity itself. While earlier economic literature mostly concentrates on supply shocks of immigrants (Borjas, 2003) or the prevalence of minorities, more recent studies have tried to identify diversity per se. Different measures such as evenness and polarization (Fearon, 2003; Montalvo and Reynal-Querol, 2005), size dominance of groups and segregation (Foged and Peri, 2016; Hunt and Gauthier-Loiselle, 2010; Moser et al., 2014), and dispersion and richness (Brixy et al., 2020), have been taken into consideration, mostly being uni-dimensional. In this study, we build a new *multidimensional* measure of diversity based on pairwise distances among individuals with respect to race, gender, and place of birth of the students and their parents. This methodology is explained in detail in the following section.

Section 2 details the experiment details and describes the setting, then presents the reduced-form empirical analyses and a discussion of the group work strategies. Section 3 presents the conclusions and a discussion of the study.

2 Empirical Analysis & Experiments Results

2.1 Experiments

We study the effect of group composition on collaboration and performance using data from an introductory undergraduate microeconomics course taught at a private university in the fall semester. The course typically enrolls approximately 600 students in the fall semester of each year. Every week, students attend two lectures taught by the main instructor in a large auditorium, and a smaller recitation with fewer than 25 students taught by a teaching assistant (TA).

This is an ideal setting for analyzing the effect of diversity for several reasons. Given the size of the groups, students are induced to have some degree of interaction, and this experiment allows us to observe these dynamics closely. Moreover, the class is an introductory undergraduate course that teaches the fundamentals of microeconomics; students take the course for various reasons, from fulfilling a general education requirement to majoring in economics. They typically choose a wide range of majors after this class. Most students are in their first year, and introductory microeconomics is the first economics course that they take in college. Therefore, students generally do not know each other before the course begins. Finally, this is not a female- or male-dominated class and different ethnicities/races are largely represented.

About three weeks into the semester, the students are randomly assigned to

groups of three or four within their recitation section. Every other week, groups send a written presentation to their TA, and then present orally in recitation. Two separate experiments performed on these group projects differ in the type of skills required.

In the first experiment (**Experiment A**), groups alternate between two types of open-ended exercises. The first type is directly targeted to exam preparation: groups write their own exam question (multiple choice), and send it with its solution to their TA, before presenting it and explaining the solution in class. The second type connects concepts to the real world: groups are given a prompt on a current event or a policy debate, which they must research and write a short paragraph about, send it to their TA, before presenting and discussing it in class.

In the second experiment (**Experiment B**), there is only one type of exercise. Groups are given a close-ended old exam question (multiple-choice), and they must send their TA a detailed solution, before a group member presents the explanation in class. While this task requires problem-solving skills, it does not involve the same creative thinking as in Experiment A.

Each group project is graded by the group's TA on the basis of completion, effort and correctness. All group members get the same grade, unless one name was left out of the submission (in which case that student gets a 0). All group project scores account for 10 percent of a student's course grade. These groups are self-directed and members are not assigned specific roles, so they can autonomously choose the strategy they prefer to work together. Other aspects regarding the class, such as the instructor, the demographic composition of teaching assistants, the material and the structure remained basically unchanged.

2.2 Data

The paper employs a novel data source. A survey was administered to 547 students out of 588, corresponding to a response rate of 93% for the first experiment, and 604 out of 629 (96% response rate) for the second experiment. The survey contains i) baseline information on personality traits (openness and extroversion, as the two most relevant Big Five traits in this setting), students' preferences on classroom practices and learning activities, gender, race/ethnicity, place of birth, parents' place of birth and daily financial stress, FGLI (First Generation Low Income) status, previous background in Economics, ii) outcomes of interest for our analysis regarding group work experience, including degree of collaboration, conflicts and workload distribution. Differently from most of the previous literature, this novel dataset allows the analysis of granular information about the race/ethnicity: traditionally, what has been used are either the categories of the US Census, or the division in whites and non-whites, or in URMs and not URMs; this data collection involved, instead, detailed information including additional categories such as East Asian, South Asian, Middle Eastern, North African, etc, and the possibility to select more than one race. We allow for the selection of a range of gender identities as well, but observe very few cases outside of the "male" or "female" categorization. Questions regarding demo-

graphic aspects of students were asked at the end of the survey, as advocated by Gilovich et al., 2013, to avoid the possibility of stereotype threat, a relevant concern in this context.

The survey is merged to rich administrative data containing individual grades throughout the semester, including several quizzes, homework assignments, midterm and final exams, participation scores and other grades for reading assignments and discussion boards, but most importantly group scores - a key outcome for our analysis. Most of the components determining grades are automatised on a virtual platform, leaving very little room for instructor or TA possible discrimination or bias. In addition, administrative data contain which recitation and presentation group each student is (randomly) assigned to, the gender and race/ethnicity of their TA, and an identifier for their TA.

We show summary statistics in Table 1. Students are split across 167 random groups in experiment A, and 163 random groups in experiment B. A detailed list of the key variables' construction is provided in the Appendix.

The two panels in the summary statistics show as well how the two groups are highly comparable along the illustrated key dimensions through a simple t-test. Notice that although there was no randomisation to allocate students to one experiment versus the other, we still find a limited number of observable differences in the baseline variables. It is worth noting that we will have specifications controlling for such individual characteristics; furthermore, we will show how having students that declare to be on average more "open to suggestions from others" in the non-creative experiment B is going to possibly moderate our coefficients of interest instead of inflating them, offering a conservative assessment. Moreover, while we cannot compare directly baseline grades as the grading was different across experiments, we are able to compare the final grades, which do not statistically differ. While we provide statistics for homophily in this table, we will provide a detailed description of the definition and its operationalization in the results section. In the same section, we will further address the last rows of the table pertaining to the outcomes of the experiment.

In our response rate analysis, we advocate for a conservative approach by refraining from relying solely on the crude rate: instead, we propose incorporating responses categorized as "I don't know" for our main outcomes (such as the collaboration within teams) within the missing data designation. Given this categorization, for experiment A we have a 88.4% response rate, while for experiment B we have 88.9%.

We display in Table 2 means comparisons along with a t-test on their difference for key variables that we have for all participants and do not find any evidence of differential attrition except for baseline grade being significant at the 10% level in Experiment B. We furthermore regress the dummy for survey respondents on the two diversity measures we build - which we will discuss in the following subsections. We do not find concerning coefficients for neither of the experiments in Table 3.

Given the small sample sizes, we perform power calculations adjusted for the strong cluster intraclass correlation. For a power of 80% and a significance

at the 1% level to detect an impact of a point on the teamwork quality we need a minimum number of clusters amounting to 143 with an average of 4 members for cluster which amounts to a total of 572 observations. Given that experiments involve about 160-170 clusters with about 600 observations with full information, we believe we have ability to discern an impact of this magnitude.

2.3 Diversity Measures

Different streams of literature have contributed to the operationalization of measuring diversity, including information theory, biology and economics. Statistical representations of diversity may reflect different aspects: notably, richness, but also evenness/dominance: in other words, one can capture the variety of races and/or genders, and/or the distributional aspects of it. Let us briefly consider the uni-dimensional diversity with respect to race, gender and migration status according to a typical operationalization that is found in the literature (Østergaard et al., 2011; Parrotta et al., 2014): the Shannon diversity index (Shannon, 1948), originally introduced to measure entropy, which is formulated in the following manner:

$$H = -\sum_{i=1}^{C} p_i ln_2(p_i) \tag{1}$$

where C is the number of distinct categories and p_i is the proportion of individuals belonging to category i for the reference population. However, this formula is not ideal for our purposes: notice that this index is maximized when the groups have even subgroups. According to this measure, in our dataset we find that, for instance, a group that has two white individuals and two Hispanic individuals will correspond to the same quantity of entropy as a group that has one South Asian, one white, one Middle Eastern/North African and one that is East Asian. We believe, therefore, that this measure is not a satisfactory measure of diversity in such a context.

We build an index that is multidimensional and is designed to more directly measure diversity in terms of dissimilarity between members of a group. Considering gender and race/ethnicity together, along with migration status, is key for this study that investigates how homogeneous versus heterogeneous individuals work together for common goals, and this represents an innovation with respect to most of the existing literature on the topic of diversity, typically concentrated only on the race/ethnicity dimension.

More specifically, the dissimilarity index we propose is a multidimensional measure of diversity that takes into consideration race, gender, place of birth of the students and of their parents. We take pairwise distances across all individuals by group and compute the average distance for each of them. As we have only categorical variables, the dissimilarity index will be the following:

$$DD = \frac{1}{\binom{n}{2}} \sum_{i>j}^{N} \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(x_{ik} \neq x_{jk}) \tag{2}$$

where N is the number of individuals in the group, K is the number of characteristics being included in the index, and x_{ik} is the realization of characteristic k for individual i. One could easily extend this measure of dissimilarity to ordinal or continuous variables through the use of pairwise distances between individuals through Gower dissimilarity indexes (Gower, 1971), but for the sake of the characteristics we are interested in, this formula is sufficient. This is our main measure of diversity throughout the analysis. With this measure, if we consider again the previous example, it is clear that the group with four different ethnicities or races instead of two would have a higher index of diversity, as desired, as the average of pairwise differences would be higher. We show the distribution of the two dissimilarity measures we take into consideration for our samples in Figures 1 and 2.

2.4 Randomization Balance

We test the randomization balance by regressing our diversity measures on baseline characteristics, including demographics, socio-economic status and baseline grade (composed of the sum of the two first quizzes students took at the very beginning of the class, before being assigned to groups. We furthermore regress diversity on two personality traits that students self-declared in the survey and that are relevant for group working: one is a score from 0 to 10 for how much students agree with the statement "I am able to make friends", for their extroversion, while the other is an analogue score for the statement "I am open to the suggestions of others". We find that none of the covariates predicts the treatment; the only exception is FGLI status, which is positively associated with the first measure of diversity, DD in Gender and Race, at the 10% level for Experiment A. Results are shown in Table 4.

2.5 Experiment Results

The specification we employ for our analysis is the following:

$$Y_{ij} = \alpha DD_j + \beta DD_j^2 + \gamma X_i + \delta X_j + \epsilon_{ij}$$

where Y_{ij} is the outcome of student i assigned to group j, DD_j is the dissimilarity measure, X_i is a rich battery of individual controls (gender identity, URM identity, dummy for the place of birth being the US versus abroad for both respondents and their parents, baseline performance, socio-economic status, personality traits, homophily and whether they studied economics before) and X_j is a vector of group controls (team aggregates for the individual controls - gender composition, URM prevalence, average baseline performance, standard deviation of baseline performance, fraction of students born outside of the US or with parents born outside, average personality traits and homophily). We explore two measures of dissimilarity: the first one is based only on gender and race, while the second one one also includes place of birth and parents' place

of birth. The errors are clustered at the group level j. For group outcomes, we employ a similar specification, but at the group level; therefore, there is obviously no need to cluster errors.

2.5.1 Impact of diversity on teamwork quality

We start by investigating the impact of diversity on the teamwork quality that students reported. In particular, we elicit students' opinion on the overall degree of collaboration they experienced within groups as a measure from 0 to 10, whether the workload distribution was balanced in the group, and whether they had conflicts within groups.

Our primary outcome measure is constructed as a Principal Component Analysis (PCA) index, amalgamating three standardized survey-reported dimensions of teamwork quality: the degree of collaboration, the workload distribution balance, and the lack of conflicts within groups. We find that both measures of demographic diversity manifest a distinctive U-shaped impact on teamwork quality, and this pattern is consistent for both experiments. This indicates that groups at the extremities of homogeneity or heterogeneity tend to declare more serene teamwork, irrespective of the nature of the task undertaken. We show regression results in Table 5.

We further complement our analysis with graphs showing the U-shaped impact that we described. As we find analogous results among experiments A and B, we pooled the samples together. In Figure 3 we show the distribution of our outcomes of interest, the teamwork quality, for three terciles of the two different measures of diversity, controlling for the usual battery of individual and group-level covariates. We can observe how in both cases there is a higher density of low levels of teamwork quality for the second tercile group, which includes individuals that were assigned to a group with intermediate diversity. In the other pair of figures, 4, we show the distribution of teamwork quality over the spectrum of both diversity measures along with the quadratic prediction and the 95% confidence interval shaded area. For further insight into this pattern, we provide a detailed breakdown of the impact on each individual component of the index in the Appendix. Notice that aggregating different components of the same measure we focus our analysis in represents also a strategy to deal with the multiple hypothesis issue.

We investigate heterogeneous impacts on females and underrepresented minority (URM) students. We do not find any of these categories to be differently impacted. However, we should take these results with caution as we have limited power to detect effects for subgroups of our sample.

To interpret these results, we invoke the well-established "group *faultlines*" theory (Lau and Murnighan, 1998; Carton and Cummings, 2013; Chiu and Staples, 2013), which posits the presence of hypothetical dividing lines within groups, predicated upon salient demographic attributes. Faultines can potentially split group's members on the basis of one or more features. For instance, if there is a group of four members, two of which are, say, East Asian, and two of which are white Caucasians, there will be a clear faultline with respect

to the race/ethnicity. This theory predicts that faultlines become stronger and more stable as more attributes align themselves in the same way, particularly if attributes are highly correlated - which is consistent with our results typically finding coefficients of larger magnitudes when considering measures of dissimilarity that involve more factors.

In the context of our exogenous groups, the emergence of endogenous subgroup formations along these faultlines seems to contribute to suboptimal group dynamics, manifesting as worsened teamwork quality. In the words of Lau and Murnighan, group fragmentation resulting from clear faultlines has the potential to worsen group cohesiveness and interaction, forming internal split coalitions with homophilous relationships that can hinder the pursuit of common goals.

The result is the convex impact of diversity that we observe, suggesting that diversity per se does not inherently precipitate conflict and cohesion deficits; rather, it is the emergence of fragmentation and polarization along these faultlines that gives rise to these adverse outcomes.

Homophily For the faultlines theory to be applicable to this context, it must be the case that homophily, a preference to create social networks with similar individuals in a biased manner beyond the effect of relative population sizes (Coleman, 1958), is a phenomenon that is found to be present among the students that are part of the sample we consider. We employ the definitions in Currarini et al., 2009 to quantify this phenomenon. We ask students to indicate the races and genders of closest friends in the University. We compare the fractions of same types friends to the fractions of those types in the whole undergraduates' population. If the former is larger than the second, we categorize the individual as featuring homophily. We repeat the same process using instead the fractions of those types in our sample. As some types are under- or over-represented in the class with respect to the broader university population, these comparisons do not necessarily correspond; in particular, females are under-represented in the class. Moreover, the race/ethnicity types are more granular in our survey data. We find a very strong evidence of homophily across all types in Tables 7, 8, 9 and 10.

2.5.2 Impact of diversity on group performance

Transitioning our focus to the impact of diversity on group performance, we adhere to a similar analytical framework, albeit at the group level. Group performance is herein quantified through the assessment of grades for group projects. Results are shown in Table 6.

In this case, our empirical exploration yields divergent results contingent upon the experimental context. In Experiment A, characterized by tasks demanding higher levels of creativity, both measures of diversity exhibit a positive influence on group scores. Conversely, in Experiment B, featuring more analytical tasks resembling conventional examination exercises, diversity exerts a negative impact on group performance.

Notably, we supplement our analysis with specifications that control for teamwork quality. While we discern a robust association between teamwork quality and group performance, accounting for this variable only partially influences the observed coefficients. Ergo, while teamwork quality bears a positive association with superior performance, it is not the sole conduit through which diversity impinges on group grades. Notice, furthermore, that in the case of experiment A, including this control more significantly affects the quadratic terms, possibly accounting for the groups working well together at the right extreme of the distribution of diversity; in the case of experiment B, instead, while the quadratic terms are never significant, the linear terms are slightly magnified: in other words, linear estimates not accounting for teamwork quality are slightly up-ward biased as some groups on the highest end of the distribution are compensating by working well together.

While the lack of additional data to test channels directly impedes further empirical testing, we interpret the results through the lens of the existing literature. When the group performance principally hinges on creativity, the positive impact stemming from a diversified array of backgrounds supersedes the concomitant communication costs. In contrast, when tasks lean towards mechanistic and resemble standard examination exercises lying on adherence to predefined rules and methodologies, the creativity dividends emanating from demographic diversity are eclipsed by the attendant communication and coordination hurdles. In other words, when there is only one correct response to an assignment, diversity is not going to help - if anything, it can represent an obstacle.

Related to this interpretation, going back to Table 1, we can appreciate the overall average outcomes for the experiment. Notice that the average degree of collaboration declared by the experiments' participants was systematically higher in Experiment A, with a creative nature. At the same time, the workload was distributed more equally on average in this experiment: as one would expect, in this case more students felt like every component of the team gave a contribution, which is consistent with the idea of creativity gains coming from different points of view and more members putting complementary efforts towards the production of the assignment. As a final point, notice that the composition of students in the second experiment declared to be generally more open to suggestions, when compared to students of experiment A. Given this aspect, we may be underestimating the negative impact of diversity on teamwork when it comes to analytical tasks.

2.5.3 Discussion

This comprehensive analysis underscores the intricate interplay between diversity, teamwork quality, and group performance, shedding light on the multifaceted dynamics that govern the collective efficacy of diverse teams. The results prompt a broader reflection on the role and adaptability of evaluation metrics in the modern educational landscape. While standardized assessments have their merits, particularly in objectively gauging specific competencies, they

may inadvertently downplay the significance of diversity, which often thrives on the fusion of contrasting viewpoints and approaches. The challenges posed by diverse teams in mechanical evaluation contexts invite discussions on how to refine assessment methodologies to ensure that they are sensitive to the unique dynamics engendered by diverse collaborations.

As the academic and corporate world increasingly embrace diversity, especially in a modern economy based more and more on knowledge production and complex tasks, it becomes incumbent upon institutions to evolve their evaluation frameworks to better recognize and harness the potential inherent in a variety of backgrounds and perspectives.

3 Conclusions

We exploit two field experiment in an undergraduate class of introductory Economics with random small groups to study the impact of diversity on team working and group performance, comparing different types of knowledge production: creative tasks (experiment A) and non-creative ones (experiment B). We find a U-shaped impact of group diversity on teamwork quality for both experiments, suggesting very homogeneous groups and very heterogeneous groups work better as teams. We then analyze the impact of group diversity on group performance, and we find that it is positive for creative tasks, and negative for analytical ones.

Our results allow us to address three sources of inconsistencies in the literature on the impact of diversity on teamwork: (i) the focus on only part of the diversity spectrum, which overlooks possible non-linearities of the impact; (ii) the lack of a clear distinction between different types of tasks; (iii) the entanglement between teamwork quality and performance.

The U-shaped impact of diversity on teamwork quality that we find for both experiments can potentially explain the mixed results of prior studies on team diversity. The literature on fractionalization, which finds a low degree of collaboration among more diverse groups, compares fully homogeneous communities to societies where there are clear fractions (i.e., moderately heterogeneous). In other words, it captures the negative slope part of our U-shaped curve, but ignores fully heterogeneous communities, and therefore the positive gradient part of the curve. Most of the literature on teamwork that we review also captures only part of the diversity spectrum: the impact of diversity on performance is generally negative in studies that compare homogeneous groups to groups split in half, and positive in studies that compare mixed groups to completely heterogeneous groups.

We furthermore contribute to the literature on team diversity by highlighting the critical role played by the type of task being performed (creative or not): we show the task nature determines which of the two forces associated with diversity – creativity gains vs communication challenges – prevails.

Our findings corroborate the theory that endogenous sub-groups formation and entrenchment can be detrimental to teamwork: this novel experimental

evidence on the impact of diversity with respect to important demographic characteristics can inform optimal team composition. When forming teams with common tasks, possible faultlines should be taken into account and avoided to maintain group cohesion. Furthermore, different types of tasks prescribe different optimal degrees of diversity. Finally, as diverse teams work better on creative tasks, teachers might consider including these learning tools as inclusive practices in the classroom.

References

Alesina, Alberto, Arnaud Devleeschauwer, et al. (2003). "Fractionalization". In: *Journal of Economic Growth* 8.2, pp. 155–94.

Alesina, Alberto and Eliana La Ferrara (Sept. 2005). "Ethnic Diversity and Economic Performance". In: *Journal of Economic Literature* 43.3, pp. 762–800.

Borjas, George J. (2003). "The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market". In: *The Quarterly Journal of Economics* 118.4, pp. 1335–1374.

Brixy, Udo, Stephan Brunow, and Anna D'Ambrosio (2020). "The unlikely encounter: Is ethnic diversity in start-ups associated with innovation?" In: *Research Policy* 49.4.

Carton, Andrew M. and Jonathon N. Cummings (2013). "The impact of subgroup type and subgroup configurational properties on work team performance." In: *The Journal of applied psychology* 98.5, pp. 732–58.

Chiu, Yi-Te and D. Sandy Staples (2013). "Reducing Faultlines in Geographically Dispersed Teams: Self-Disclosure and Task Elaboration". In: *Small Group Research* 44.5, pp. 498–531.

Coleman, James S. (1958). "Relational Analysis: The Study of Social Organizations with Survey Methods". In: *Human Organization* 17.4, pp. 28–36.

Cross, Rob et al. (2021). "Collaboration overload is sinking productivity". In: *Harvard Business Review*.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin (2009). "An Economic Model of Friendship: Homophily, Minorities, and Segregation". In: *Econometrica* 77.4, pp. 1003–1045.

Daan Van Knippenberg Caarsten K.W. De Dreu, Astrid C. Homan (2004). "Work group diversity and group performance: An integrative model and research agenda". In: *Journal of Applied Psychology* 89, pp. 1008–1022.

Deming, David J. (2017). "The growing importance of social skills in the labor market". In: *Quarterly Journal of Economics* 132.4, pp. 1593–1640.

Easterly, William and Ross Levine (1997). "Africa's Growth Tragedy: Policies and Ethnic Divisions". In: *The Quarterly Journal of Economics* 112.4, pp. 1203–1250.

Fearon, James D. (2003). "Ethnic and Cultural Diversity by Country". In: *Journal of Economic Growth* 8.2, pp. 195–222.

Ferrucci, Edoardo and Francesco Lissoni (2019). "Foreign inventors in Europe and the United States: Diversity and Patent Quality". In: *Research Policy* 48.9.

Foged, Mette and Giovanni Peri (Apr. 2016). "Immigrants' Effect on Native Workers: New Analysis on Longitudinal Data". In: *American Economic Journal: Applied Economics* 8.2, pp. 1–34.

Freeman, Richard and Wei Huang (2015). "Collaborating with people like me: Ethnic coauthorship within the United States". In: *Journal of Labor Economics* 33.1, pp. 289–318.

Garicano, Luis and Esteban Rossi-Hansberg (2006). "Organization and inequality in a knowledge economy". In: *Quarterly Journal of Economics* 121.4, pp. 1383–1435.

Gilovich, Thomas et al. (2013). *Social Psychology*. 3rd. New York: W. W. Norton & Company.

Gower, John C. (1971). "A General Coefficient of Similarity and Some of Its Properties". In: *Biometrics* 27.4, pp. 857–871.

Guillaume, Yves R. F. et al. (2017). "Harnessing demographic differences in organizations: What moderates the effects of workplace diversity?" In: *Journal of Organizational Behavior* 38, pp. 276–303.

Hamilton, Barton H., Jack A. Nickerson, and Hideo Owan (2012). "Diversity and Productivity in Production Teams". In: *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*. Ed. by Alex Bryson. Vol. 13. Emerald Group Publishing Limited, pp. 99–138.

Hjort, Jonas (Dec. 2014). "Ethnic Divisions and Production in Firms". In: *The Quarterly Journal of Economics* 129, pp. 1899–1946.

Hoogendoorn, Sander, Hessel Oosterbeek, and Mirjam Van Praag (2013). "The impact of gender diversity on the performance of business teams: Evidence from a field experiment". In: *Management Science* 59.7, pp. 1514–1528.

Hoogendoorn, Sander and Mirjam van Praag (July 2012). *Ethnic Diversity and Team Performance: A Field Experiment*. Tinbergen Institute Discussion Papers. Tinbergen Institute.

Horwitz, Sujin and Irwin Horwitz (2007). "The Effects of Team Diversity on Team Outcomes: A Meta-Analytic Review of Team Demography". In: *Journal of Management* 33.6, pp. 987–1015.

Hunt, Jennifer and Marjolaine Gauthier-Loiselle (Apr. 2010). "How Much Does Immigration Boost Innovation?" In: *American Economic Journal: Macroeconomics* 2.2, pp. 31–56.

Jackson, Susan E. and Aparna Joshi (2004). "Diversity in social context: a multi-attribute, multilevel analysis of team diversity and sales performance". In: *Journal of Organizational Behavior* 25, pp. 675–702.

Johnson, David, Roger Johnson, and Karl Smith (2007). "The state of cooperative learning in postsecondary and professional settings". In: *Educational Psychology Review,* 19.1, pp. 15–29.

Kahane, Leo, Neil Longley, and Robert Simmons (Mar. 2013). "The Effects of Coworker Heterogeneity on Firm-Level Output: Assessing the Impacts of Cultural and Language Diversity in the National Hockey League". In: *The Review of Economics and Statistics* 95.1, pp. 302–314.

Lacerenza, Christina et al. (2018). "Team development interventions: Evidence-based approaches for improving teamwork". In: *American Psychologist* 73.4, pp. 517–531.

Lau, Dora C. and J. Keith Murnighan (1998). "Demographic Diversity and Faultlines: The Compositional Dynamics of Organizational Groups". In: *The Academy of Management Review* 23.2, pp. 325–340.

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

Leonard, Jonathan and David Levine (2006). "The Effect of Diversity on Turnover: A Large Case Study". In: *ILR Review* 59.4, pp. 547–572.

Lyons, Elizabeth (2017). "Team Production in International Labor Markets: Experimental Evidence from the Field". In: *American Economic Journal: Applied Economics* 9.3, pp. 70–104.

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

Mathieu, John E. et al. (2014). "A Review and Integration of Team Composition Models: Moving Toward a Dynamic and Temporal Framework". In: *Journal of Management* 40.1, pp. 130–160.

Mello, Antonio S. and Martin E. Ruckes (2006). "Team Composition". In: *The Journal of Business* 79.3, pp. 1019–1039.

Montalvo, José G. and Marta Reynal-Querol (June 2005). "Ethnic Polarization, Potential Conflict, and Civil Wars". In: *American Economic Review* 95.3, pp. 796–816.

Morgan, John and Felix Várdy (Mar. 2009). "Diversity in the Workplace". In: *American Economic Review* 99.1, pp. 472–85.

Moser, Petra, Alessandra Voena, and Fabian Waldinger (Oct. 2014). "German Jewish Émigrés and US Invention". In: *American Economic Review* 104.10, pp. 3222–55.

Østergaard, Christian R., Bram Timmermans, and Kari Kristinsson (2011). "Does a different view create something new? The effect of employee diversity on innovation". In: *Research Policy* 40.3, pp. 500–509.

Ozgen, Ceren, Peter Nijkamp, and Jacques Poot (2012). "Immigration and innovation in European regions". In: *Migration Impact Assessment*. Edward Elgar Publishing. Chap. 8, pp. 261–298.

— (2013). "The impact of cultural diversity on firm innovation: evidence from Dutch micro-data". In: *IZA Journal of Migration* 2.18.

Parrotta, Pierpaolo, Dario Pozzoli, and Mariola Pytlikova (2014). "Labor diversity and firm productivity". In: *European Economic Review* 66.C, pp. 144–179.

Prat, Andrea (2002). "Should a team be homogeneous?" In: *European Economic Review* 46.7, pp. 1187–1207.

Richard, Orlando C. and Roger M. Shelor (2002). "Linking top management team age heterogeneity to firm performance: juxtaposing two mid-range theories". In: *The International Journal of Human Resource Management* 13.6, pp. 958–974.

Shannon, Claude E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423.

Simsarian Webber, Sheila and Lisa M. Donahue (2001). "Impact of highly and less job-related diversity on work group cohesion and performance: A meta-analysis". In: *Journal of Management* 27, pp. 141–162.

Springer, Leonard, Mary Elizabeth Stanne, and Samuel S. Donovan (1999). "Effects of Small-Group Learning on Undergraduates in Science, Mathematics,

Engineering, and Technology: A Meta-Analysis". In: *Review of Educational Research* 69.1, pp. 21–51.

Wegge, Jurgen et al. (2008). "Age and gender diversity as determinants of performance and health in a public organization: The role of task complexity and group size". In: *Journal of Applied Psychology* 93.6, pp. 1301–1313.

Williams, Katherine Y. and Charles A. O'Reilly (1998). "Demography and diversity in organizations: A review of 40 years of research". In: *Research in Organizational Behavior* 20, pp. 77–140.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi (2007). "The Increasing Dominance of Teams in Production of Knowledge". In: *Science* 316.5827, pp. 1036–1039.

	Experiment A					Experiment B					
Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max	Difference (St.Error)
Baseline Variables											
URM	588	.345	.476	0	1	629	.377	.485	0	1	-0.032 (0.028)
Female	585	.429	.495	0	1	629	.461	.499	0	1	-0.032 (0.029)
Born abroad	540	.239	.427	0	1	597	.268	.443	0	1	-0.029 (0.026)
At least one parent born abroad	542	.638	.481	0	1	597	.687	.464	0	1	-0.048* (0.028)
I am able to make friends (0-10)	546	6.824	2.09	0	10	604	6.879	1.983	0	10	-0.055 (0.120)
Open to suggestions of others (0-10)	546	7.44	1.625	2	10	604	7.627	1.569	0	10	-0.188** (0.094)
Race/ethnicity-based homophily	526	.835	.372	0	1	584	.849	.358	0	1	-0.015 (0.022)
Gender-based homophily	527	.729	.445	0	1	591	.785	.411	0	1	-0.056** (0.026)
Economics classes before college	545	.413	.493	0	1	604	.416	.493	0	1	-0.003 (0.029)
FGLI (First Generation Low Income)	518	.172	.378	0	1	576	.181	.385	0	1	-0.009 (0.023)
Financial aspects daily source of stress	526	.39	.488	0	1	568	.405	.491	0	1	-0.015 (0.030)
Baseline grade	587	4.066	.698	0	5	629	1.986	.132	0	2	*Different grading*
Classroom Features											
Female TA	588	.315	.465	0	1	629	.316	.465	0	1	-0.002 (0.027)
URM TA	588	.252	.434	0	1	629	.251	.434	0	1	0.001 (0.025)
Diversity Measures											
DD in Gender and Race	588	0	1	-3.444	2.21	629	0	.157	-.612	.221	*Standardized variable*
DD in Gender, Race, PoB and Parents' PoB	588	0	1	-3.301	1.814	629	0	.117	-.357	.226	*Standardized variable*
Experiment Outcomes											
Degree of team collaboration (0-10)	545	6.829	2.26	0	10	575	5.89	2.577	0	10	0.939*** (0.145)
Conflicts in the group	545	.182	.386	0	1	575	.141	.348	0	1	0.041* (0.022)
Equally distributed workload	466	.749	.434	0	1	575	.631	.483	0	1	0.118*** (0.029)
Final grade	588	86.293	9.981	40.7	100	629	86.787	9.442	48.83	100	-0.495 (0.557)

Table 1: Summary statistics and balance between experiments A and B. We display here key baseline variables, classroom features, our two measures of diversity and key experiment outcomes.

	Non Attrited		Attrited		Difference	
	Mean	St. Deviation	Mean	St. Deviation	Difference	St. Error
	Experiment A					
URM	0.340	(0.474)	0.382	(0.490)	-0.042	(0.061)
Female	0.440	(0.497)	0.338	(0.477)	0.102	(0.065)
Baseline grade	0.024	(0.981)	-0.184	(1.130)	0.207	(0.130)
Female TA	0.321	(0.467)	0.265	(0.444)	0.056	(0.060)
URM TA	0.248	(0.432)	0.279	(0.452)	-0.031	(0.056)
Group Score	0.023	(0.946)	-0.178	(1.341)	0.202	(0.129)
Observations	520		68		588	
	Experiment B					
URM respondent	0.369	(0.483)	0.443	(0.500)	-0.074	(0.061)
Female respondent	0.467	(0.499)	0.414	(0.496)	0.053	(0.063)
Baseline grade	0.027	(0.784)	-0.217	(2.020)	0.244*	(0.127)
Female TA	0.322	(0.468)	0.271	(0.448)	0.051	(0.059)
URM TA	0.250	(0.434)	0.257	(0.440)	-0.007	(0.055)
Group Score	0.023	(0.938)	-0.183	(1.395)	0.205	(0.127)
Observations	559		70		629	

Sample Means with Std. Dev. in brackets and Difference in Means with Std. Err. in brackets
* p<0.1 ** p<0.05 *** p<0.01

Table 2: Statistical differences between non attrited and attrited students' baseline characteristics. We use variables that we have for all students - basic demographics, grades, and classroom features.

	Attrited	Attrited
Experiment A		
DD in Gender and Race	-0.0143	
	(0.0143)	
DD in Gender, Race, PoB and Parents' PoB		0.0257
		(0.0159)
Group Controls	Y	Y
Observations	584	584
Experiment B		
DD in Gender and Race	-0.0588	
	(0.0891)	
DD in Gender, Race, PoB and Parents' PoB		0.00314
		(0.112)
Group Controls	Y	Y
Observations	629	629

* p<0.1 ** p<0.05 *** p<0.01

Table 3: Impact of the two diversity measures we employ on survey attrition.

	DD in Gender and Race	DD in Gender, Race PoB and Parents' PoB	Observations
Experiment A			
URM	0.0106 (0.0191)	-0.00222 (0.0206)	520
Female	0.00958 (0.0202)	0.0127 (0.0218)	520
Born abroad	0.0000254 (0.0166)	0.00233 (0.0179)	520
Parents born abroad	-0.00519 (0.0184)	0.00174 (0.0198)	522
Able to make friends	-0.0180 (0.0791)	-0.0132 (0.0852)	520
Open to suggestions of others	-0.0124 (0.0634)	-0.0142 (0.0683)	520
FGLI (First Generation Low Income)	0.0328* (0.0182)	0.0305 (0.0199)	498
Financial aspects daily source of stress	0.0206 (0.0238)	0.0131 (0.0257)	507
Baseline grade	-0.00585 (0.0399)	-0.00351 (0.0430)	520
Experiment B			
URM	-0.0315 (0.117)	0.0107 (0.161)	575
Female	0.0345 (0.116)	0.0591 (0.160)	575
Born abroad	0.0194 (0.106)	-0.0233 (0.146)	577
Parents born abroad	0.0165 (0.109)	0.00814 (0.150)	578
Able to make friends	0.169 (0.456)	0.238 (0.627)	575
Open to suggestions of others	0.0984 (0.364)	0.205 (0.500)	575
FGLI (First Generation Low Income)	0.137 (0.107)	-0.00430 (0.148)	554
Financial aspects daily source of stress	0.179 (0.139)	0.120 (0.190)	545
Baseline grade	-0.0316 (0.261)	-0.0776 (0.358)	575

* p<0.1 ** p<0.05 *** p<0.01

Table 4: Randomization Balance

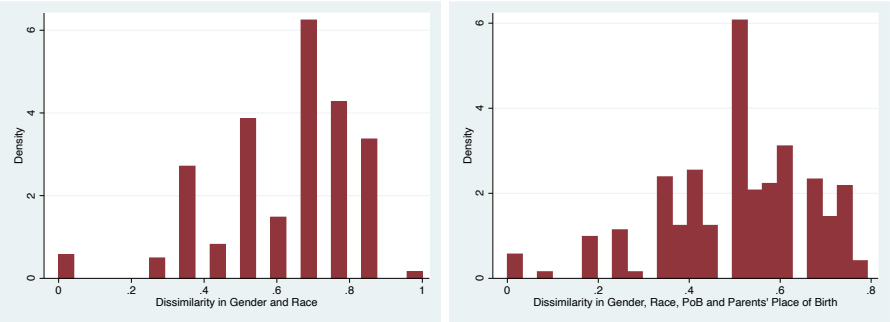

Figure 1: Experiment A. Distribution of the two different dissimilarity (DD) measures for the groups in our dataset according to 1) race/ethnicity and gender; 2) race/ethnicity, gender and migration status.

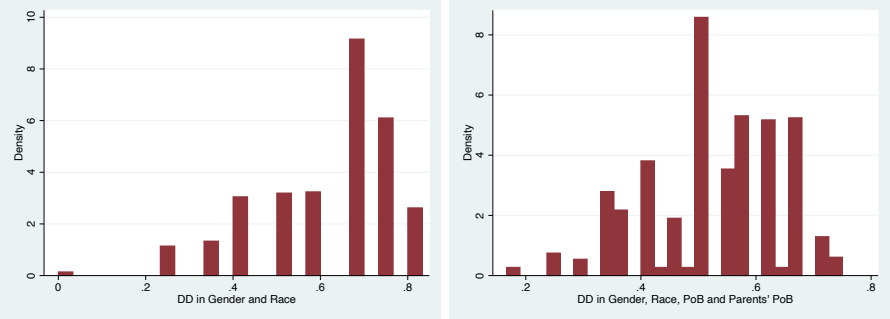

Figure 2: Experiment B. Distribution of the two different dissimilarity (DD) measures for the groups in our dataset according to 1) race/ethnicity and gender; 2) race/ethnicity, gender and migration status.

	(1)	(2)
Experiment A		
DD in Gender and Race	0.0924	
	(0.0741)	
Quadratic DD in Gender and Race	0.0854**	
	(0.0345)	
DD in Gender, Race, PoB and Parents' PoB		0.0338
		(0.0782)
Quadratic DD in Gender, Race, PoB and Parents' PoB		0.0643*
		(0.0346)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	429	429
Experiment B		
DD in Gender and Race	0.368	
	(0.388)	
Quadratic DD in Gender and Race	3.176**	
	(1.229)	
DD in Gender, Race, PoB and Parents' PoB		-0.0353
		(0.514)
Quadratic DD in Gender, Race, PoB and Parents' PoB		7.892***
		(2.938)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	493	493

* p<0.1 ** p<0.05 *** p<0.01

Table 5: Impact of diversity on team work quality. This is a PCA variable aggregating three self-reported measures through surveys: the degree of collaboration within teams, the absence of conflicts and the equal distribution of the workload.

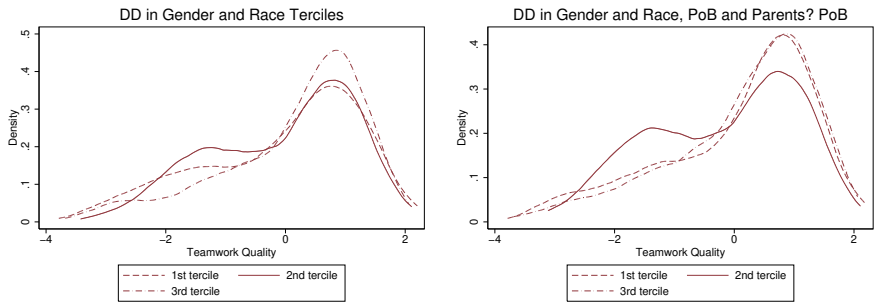

Figure 3: Pooled Experiments A and B. Distribution of teamwork, controlled for the usual battery of individual and group regressors, for the three terciles of diversity (measured respectively by DD in Gender and Race and DD in Gender, Race, PoB and Parents' PoB). The full line corresponds to the groups that have an intermediate intensity of diversity.

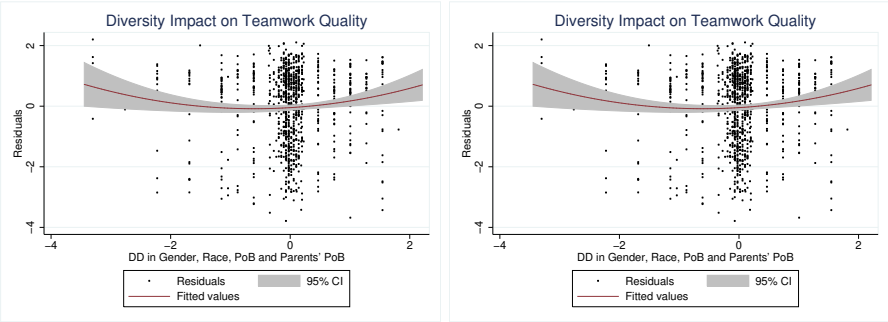

Figure 4: Pooled Experiments A and B. Scatters of teamwork quality and two diversity measures respectively (DD in Gender and Race and DD in Gender, Race, PoB and Parents' PoB), controlled for the usual battery of individual and group regressors.

	(1)	(2)
Experiment A		
DD in Gender and Race	0.0239**	
	(0.0117)	
Quadratic DD in Gender and Race	0.0113*	
	(0.00614)	
DD in Gender, Race, PoB and Parents' PoB		0.0237*
		(0.0125)
Quadratic DD in Gender, Race, PoB and Parents' PoB		0.0136**
		(0.00624)
Group Controls	Y	Y
Observations	167	167
DD in Gender and Race	0.0201*	
	(0.0106)	
Quadratic DD in Gender and Race	0.00750	
	(0.00489)	
DD in Gender, Race, PoB and Parents' PoB		0.0234*
		(0.0122)
Quadratic DD in Gender, Race, PoB and Parents' PoB		0.0104*
		(0.00549)
Mean Teamwork Quality	0.0327**	0.0334**
	(0.0148)	(0.0151)
Group Controls	Y	Y
Observations	167	167
Experiment B		
DD in Gender and Race	-0.639*	
	(0.369)	
Quadratic DD in Gender and Race	-1.051	
	(1.182)	
DD in Gender, Race, PoB and Parents' PoB		-0.863**
		(0.380)
Quadratic DD in Gender, Race, PoB and Parents' PoB		-2.947
		(2.535)
Group Controls	Y	Y
Observations	163	163
DD in Gender and Race	-0.705*	
	(0.373)	
Quadratic DD in Gender and Race	-1.504	
	(1.222)	
DD in Gender, Race, PoB and Parents' PoB		-0.878**
		(0.365)
Quadratic DD in Gender, Race, PoB and Parents' PoB		-4.283*
		(2.584)
Mean Teamwork Quality	0.121**	0.127**
	(0.0563)	(0.0545)
Group Controls	Y	Y
Observations	163	163

* p<0.1 ** p<0.05 *** p<0.01

Table 6: Impact of diversity on the group score for group assignments. We show our canonical specification and a further specification that controls for the teamwork quality PCA index.

Race/ Ethnicity	Homophilic (University)
White	78.8%
Black	81.1%
Asian	88.6%
Hispanic	79.5%

Race/ Ethnicity	Homophilic (Class)
White	78.7%
Black	81.1%
East Asian	88.4%
South Asian	91.9%
Hispanic	81.8%
Middle E./North A.	65.2%

Table 7: Homophily by race/ethnicity - Experiment A

Race/ Ethnicity	Homophilic (University)
White	86.8%
Black	85.3%
Asian	85.2%
Hispanic	74.2%

Race/ Ethnicity	Homophilic (Class)
White	86.8%
Black	85.3%
East Asian	88.8%
South Asian	85.9%
Hispanic	74.2%
Middle E./North A.	68%

Table 8: Homophily by race/ethnicity - Experiment B

Gender	Homophilic (University)
Male	81.0%
Female	75.0%

Gender	Homophilic (Class)
Male	62.4%
Female	86.2%

Table 9: Homophily by gender - Experiment A

Gender	Homophilic (University)
Male	81.5%
Female	80.1%

Gender	Homophilic (Class)
Male	68.5%
Female	89.9%

Table 10: Homophily by gender - Experiment B

Appendix A Key Variables' Construction

- *URM*: Black and/or Hispanic/Latinx selected among the options in the question "What is the race/ethnicity that you identify with?". We complement this information by administrative records if the information is not provided by the student through survey.
- *Female*: dummy constructed for "Female" being selected in the question "What is the gender that you identify with?". We complement this information by administrative records if the information is not provided by the student through survey.
- *Born abroad*: variable constructed through the survey question "Where were you born"?
- *Parents born abroad*: variable constructed through the survey question "Where were your parent(s)/guardian(s) born?"
- *Able to make friends*: we asked the respondent to pick a value from 0 to 10 representing how much the sentence "I am able to make friends" describes them. This is meant to capture the personality trait regarding extroversion.
- *Open to suggestions of others*: we asked the respondent to pick a value from 0 to 10 representing how much the sentence "I am open to the suggestions of" describes them. This is meant to capture the personality trait regarding openness, as contextualized in teamworking.
- *FGLI (First Generation Low Income)*: asked through survey "Do you identify yourself as a FGLI (First Generation Low Income) student?".
- *Financial aspects daily source of stress*: asked through survey "Are financial aspects a source of concern or stress for you in your daily life?"
- *Baseline grade*: sum of the grades from the first two quizzes, completed by students individually.
- *Race/ethnicity-based homophily* and *Gender-based homophily*: explained in detail in the subsection regarding the impacts of diversity on teamwork quality in the results section.
- *Female TA* and *URM TA*: administrative records. We build the URM category consistently with the student-related definition. Notice that concretely the experiment involved Hispanic/Latinx TAs.
- *DD in Gender and Race*, *DD in Gender, Rac, PoB and Parents' Pob*: explained in detail in the subsection regarding diversity measures. Operationalized through the package "cluster" in R.

- *Degree of team collaboration*: asked through survey "How would you grade the degree of collaboration in your group? - From 0 (no collaboration) to 10 (full collaboration)".
- *Conflicts in the group*: asked through survey "Were there any tensions or conflicts within your group?". We then employ the absence of conflicts to build the variable "No conflict" which we employ in the PCA index for the teamwork quality.
- *Equally distributed workload*: asked through survey "Do you think the workload was typically distributed equally among the group members?".

Appendix B Sub-Components of Teamwork Quality

	(1)	(2)
Experiment A		
DD in Gender and Race	0.0742	
	(0.0658)	
Quadratic DD in Gender and Race	0.102**	
	(0.0420)	
DD in Gender, Race, PoB and Parents' PoB		-0.0259
		(0.0743)
Quadratic DD in Gender, Race, PoB and Parents' PoB		0.0612
		(0.0387)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	493	493
Experiment B		
DD in Gender and Race	1.152	
	(1.106)	
Quadratic DD in Gender and Race	6.962*	
	(3.806)	
DD in Gender, Race, PoB and Parents' PoB		1.402
		(1.422)
Quadratic DD in Gender, Race, PoB and Parents' PoB		18.73**
		(9.241)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	493	493

* p<0.1 ** p<0.05 *** p<0.01

Table 11: Impact of diversity on the degree of collaboration within groups, as self-reported through surveys.

	(1)	(2)
Experiment A		
DD in Gender and Race	0.171	
	(0.194)	
Quadratic DD in Gender and Race	0.191**	
	(0.0943)	
DD in Gender, Race, PoB and Parents' PoB		0.0517
		(0.192)
Quadratic DD in Gender, Race, PoB and Parents' PoB		0.157*
		(0.0924)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	429	429
Experiment B		
DD in Gender and Race	0.000590	
	(0.848)	
Quadratic DD in Gender and Race	5.375	
	(4.566)	
DD in Gender, Race, PoB and Parents' PoB		-0.803
		(1.101)
Quadratic DD in Gender, Race, PoB and Parents' PoB		17.81**
		(7.184)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	493	493

* p<0.1 ** p<0.05 *** p<0.01

Table 12: Impact of diversity on equal workload distribution within teams, as self-reported through surveys.

	(1)	(2)
Experiment A		
DD in Gender and Race	0.157	
	(0.181)	
Quadratic DD in Gender and Race	0.0598	
	(0.0837)	
DD in Gender, Race, PoB and Parents' PoB		0.0939
		(0.155)
Quadratic DD in Gender, Race, PoB and Parents' PoB		0.0449
		(0.0814)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	493	493
Experiment B		
DD in Gender and Race	1.433	
	(0.988)	
Quadratic DD in Gender and Race	7.183**	
	(3.628)	
DD in Gender, Race, PoB and Parents' PoB		-1.046
		(1.192)
Quadratic DD in Gender, Race, PoB and Parents' PoB		2.743
		(7.365)
Individual Controls	Y	Y
Group Controls	Y	Y
Observations	536	536

* p<0.1 ** p<0.05 *** p<0.01

Table 13: Impact of diversity on presence of conflict within groups, as self-reported through surveys.