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

Darova, Ornella and Duchene, Anne

University of Pennsylvania, University of Pennsylvania

15 January 2024

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

Ornella Darova† and Anne Duchene‡

March 6, 2024

Abstract

We run two field experiments on team diversity in a large undergraduate economics class. Small groups with random compositions are generated and assigned team tasks. In the first experiment, tasks are creative and complex, while in the second one they are more standard. We use a multidimensional measure of diversity based on gender, race, and migration status. We estimate its impact on teamwork quality and group performance. We find a significant U-shaped effect on teamwork quality in both experiments. However, the impact on performance depends on the type of task: it is positive for creative tasks, but negative for standard 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 standard 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. These results allow us to build a comprehensive framework to better understand the impact of diversity on teamwork.

Keywords: Diversity, Knowledge Production, Creativity, Teamwork, 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-ASSA Conference, 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 and discussants 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).[1] Simultaneously, learning and working environments have been shifting toward more and more teamwork and group problem-solving (Wuchty et al., 2007; Mathieu et al., 2014; Deming, 2017).[2] 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 trends raise an important question: do more diverse teams work better? While there is an extensive literature studying this question, it has revealed mixed results so far.[3] As we describe in the next paragraphs, these inconsistent results are due to different lenses of analysis of both teamwork and diversity. Our objective is to build a comprehensive framework that factors in all these lenses, in order to identify when and why diversity can facilitate or hinder teamwork. We do so in the context of higher education, by estimating causal assessments in a controlled environment - a large college undergraduate class with randomized homework groups.

The first lens of analysis is the type of task performed by the team. Some studies highlight higher communication and coordination costs among more heterogeneous individuals (Morgan and Várdy, 2009; Hamilton et al., 2012), while others show complementarities and benefits of information sharing and a broader set of backgrounds (Prat, 2002; Mello and Ruckes, 2006; I. Horwitz and S. Horwitz, 2007; Kahane et al., 2013). We hypothesize that the outcome of those two opposing forces depends on the degree of task creativity. 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 research teams publish more often in high-impact journals; in Ferrucci and Lissoni, 2019, migrant inventors increase team diversity and are associated with higher quality patents; Vogel et al.,

[1] See also Eckel and Grossman, 2005.

[2] 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".

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

2014 find both higher gender and ethnic diversity of entrepreneurial teams to result in better funding.[4] In an experimental setting similar to ours, Hoogendoorn et al., 2012 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 standard tasks. Lyons, 2017 finds that birthplace diversity hinders performance due to communication problems when tasks are highly specialized (see also Leonard and D. Levine, 2006, Hjort, 2014 and Marx et al., 2021). We factor in different types of tasks by running two experiments, one with creative tasks and the other with standard tasks, and we compare results.

The second lens of analysis is the measure of teamwork. One branch of the literature analyzes team collaboration, trust, conflict and general group dynamics. This measure depends heavily on agents' preferences, particularly homophily, and it is typically studied in larger contexts than teams such as neighborhoods, cities or nations,[5] or in smaller one-on-one interactions, like games.[6] This literature generally finds segregation and demographic fractionalization to be associated with higher levels of conflict and lower trust. Another branch of the literature focuses on team performance per se, where diversity is considered as a technology in a production function combining team members' efforts. This measure is more likely to depend on the type of task performed, and therefore on the trade-off described in the previous paragraph between creativity gains and coordination costs. In this paper, we carefully distinguish between team dynamics and team performance, using tailored survey questions to team members for the former, and grades for the latter.

The third and final lens is the portion of the diversity spectrum being considered. Papers such as the ones mentioned above on fractionalization typically compare fully homogeneous groups to fundamentally segregated groups. Other papers, such as Hoogendoorn et al., 2012 who find a positive impact of ethnic team diversity, consider moderate to high diversity. In our experiments, we build small groups with random compositions of students that represent the full spectrum of group diversity, from very homogeneous to very heterogeneous. This more flexible functional form of diversity allows us to consider possible non-linear impacts of diversity on teamwork.

Our comprehensive framework of two experiments with different types of tasks allows us to analyze the effect of multi-dimensional diversity (according to race, gender, and migration status) on two measures of teamwork – performance and teamwork quality. We find that teamwork quality (an index based on collaboration between members, balance of member contributions, and the absence of conflicts) 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

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

[5] See Akerlof, 1997; Easterly and R. Levine, 1997; Alesina, Devleeschauwer, et al., 2003; Alesina and La Ferrara, 2005; Montalvo and Reynal-Querol, 2005. For a systematic review, see Dinesen et al., 2020.

[6] See Fershtman and Gneezy, 2001, Burns, 2006, Finseraas et al., 2019.

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. Such tendency is determined by preferences and should not be contingent on the type of task.

When we consider group performance, instead, we find that higher diversity translates into higher grades for creative tasks but lower ones for non-creative assignments, even after controlling for teamwork quality. We consider this finding to be suggestive of the fact that diversity can result in creativity gains when tasks are of creative nature, but communication costs prevail when assignments are not creative instead; such impact is not related to the group dynamics, but to diversity as a technology of production.

Section 2 details the experiment design and describes the setting, while section 3 presents the reduced-form empirical analyses and results. Section 4 concludes.

2 Experiment Design and Empirical Analysis

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. The course is one-semester long and typically enrolls approximately 600 students. 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. Students are from various geographic areas and the vast majority (around 90%) are in their first semester of college. Therefore, they generally do not know each other before the course begins. As suggested by Burns, 2012, this can reflect into higher salience of demographic features. Finally, this is not a female- or male-dominated class and different ethnicities/races are

largely represented.

At the beginning of the semester, the students are randomly assigned through an algorithm 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.

We run two almost identical experiments in two consecutive iterations of the same class. The only difference between the experiments is the type of task assigned to groups.

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 create their own exam question, 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 an old exam question, 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 degree and modality of collaboration (frequency, technology, location, division of labor etc.). 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)* personality traits (extroversion and openness[7], as the two most relevant Big Five personality traits in this setting),[8] gender, race/ethnicity, place of birth (POB), 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

[7]Respectively, these are responses to questions regarding how much they agreed in a scale from 0 to 10 with the sentences "I am able to make friends" and "I am open to suggestions of others".

[8]The Big Five are commonly used in psychology to characterize an individual's personality. They measure extroversion, agreeableness, openness to experience, conscientiousness, and neuroticism (opposite of emotional stability). For more details, see Borghans et al., 2008.

experience, including degree of collaboration, conflicts and workload distribution. This novel dataset allows the analysis of granular information about race and ethnicity: traditionally, the literature uses either the categories of the US Census, or the division in whites and non-whites, or URMs and not; this data collection involves, 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 demographic 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 quizzes, homework, participation and exams, but most importantly group scores - a key outcome for our analysis. We utilize the two first quizzes at the beginning of the semester as baseline measures of individual performance. 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 Appendix A.

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. 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 have specifications controlling for such individual characteristics; furthermore, we 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 (quizzes) grades as the grading was different across experiments, we are able to compare final course grades, which do not statistically differ. While we provide statistics for homophily in this table, we detail the definition and computation in section 3. In the same section, we further address the last rows of the table pertaining to the outcomes of the experiment.

	Experiment A					Experiment B					
Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max	Difference (Std.Err.)
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)
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	1	-3.898	1.409	*Standardized variable*
DD in Gender, Race, POB and Parents' POB	588	0	1	-3.301	1.814	629	0	1	-3.042	1.926	*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.

2.3 Diversity Measures

Different streams of literature have contributed to the implementation of diversity measurements. While earlier economic literature mostly concentrates on supply shocks of immigrants (Borjas, 2003) or the prevalence of minorities, more recent studies have tried to assess diversity per se. Different measures such as evenness and polarization (Fearon, 2003; Montalvo and Reynal-Querol, 2005), size dominance of groups and segregation (Hunt and Gauthier-Loiselle, 2010; Moser et al., 2014; Foged and Peri, 2016), and dispersion and richness (Brixy et al., 2020), have been taken into consideration, mostly being uni-dimensional. Let us briefly consider the uni-dimensional diversity with respect to race, gender and migration status according to a typical implementation 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. This formula is not ideal for our purposes as the 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 build an index that is multidimensional and is designed to more directly measure diversity in terms of dissimilarity between members of a small group. Considering gender and race/ethnicity together, along with place of birth, 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, which typically concentrates only on one dimension.

More specifically, within each group, we take pairwise distances across all pairs of individuals; then, we characterize groups by the average of the pairwise distances. As we have only categorical variables, the dissimilarity index is therefore given by:

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

where n is the number of individuals in the group, i and j are distinct members of the same group, K is the number of characteristics being included in the diversity 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, differently from entropy. We provide an illustrative example in Appendix A. Furthermore, this measure allows a finer degree of granularity, allowing us to distinguish between homogeneous, moderately homogeneous, and fully heterogeneous groups. We show the distribution of the two dissimilarity measures we take into consideration for our samples in Figures 1 and 2. We then standardize this measure for the analysis to facilitate the interpretation of coefficients.

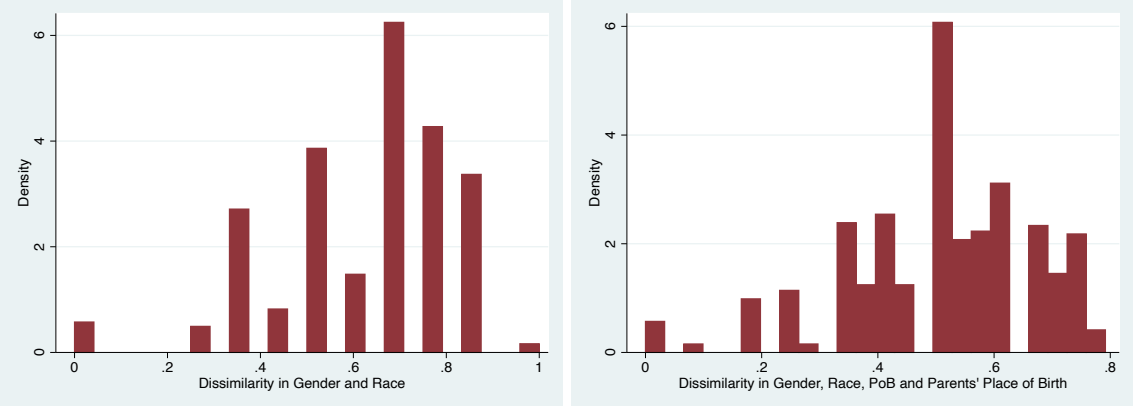


Figure 1: Experiment A. Distribution of the two different dissimilarity (DD) measures for the groups in our dataset according to 1) race/ethnicity and gender; 2) race/ethnicity, gender and migration status.

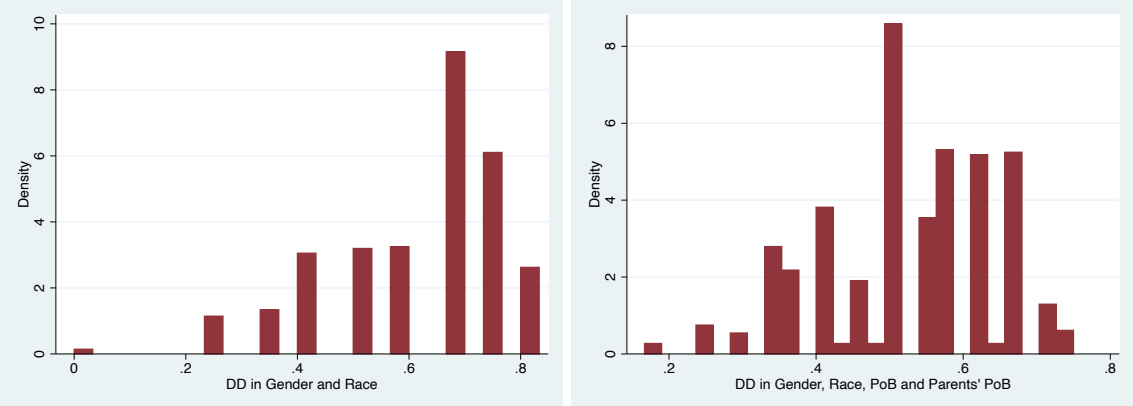


Figure 2: Experiment B. Distribution of the two different dissimilarity (DD) measures for the groups in our dataset according to 1) race/ethnicity and gender; 2) race/ethnicity, gender and migration status.

2.4 Attrition and Randomization Balance

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. This categorization, slightly pushes down response rates to 88.4% for experiment A and 88.9% for experiment B.

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 do not find concerning coefficients for neither of the experiments in Table 3.

	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	0.369	(0.483)	0.443	(0.500)	-0.074	(0.061)
Female	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	-0.00924	
	(0.0140)	
DD in Gender, Race, POB and Parents' POB		0.000369
		(0.0131)
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.

We test the randomization balance by regressing our diversity measures on baseline characteristics, including demographics, socio-economic status, personality traits and baseline grade. 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.

	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 (0-10)	-0.0180 (0.0791)	-0.0132 (0.0852)	520
Open to suggestions of others (0-10)	-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	-0.00495 (0.0183)	0.00126 (0.0189)	575
Female	0.00541 (0.0183)	0.00694 (0.0188)	575
Born abroad	0.00304 (0.0166)	-0.00273 (0.0171)	575
Parents born abroad	0.00259 (0.0170)	0.000956 (0.0176)	554
Able to make friends (0-10)	0.0265 (0.0716)	0.0280 (0.0736)	554
Open to suggestions of others (0-10)	0.0155 (0.0571)	0.0241 (0.0588)	545
FGLI (First Generation Low Income)	0.0216 (0.0169)	-0.000505 (0.0174)	545
Financial aspects daily source of stress	0.0281 (0.0218)	0.0141 (0.0223)	575
Baseline grade	-0.00496 (0.0409)	-0.00911 (0.0421)	575

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

Table 4: Randomization Balance.

Given the small sample sizes, we perform power calculations adjusted for the strong cluster intra-class 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.

3 Experiment Results

The specification we employ for our analysis is the following:

$$Y_{ig} = \alpha + \beta DD_g + \gamma DD_g^2 + \delta X_i + \eta X_g + \epsilon_{ig}$$

where Y_{ig} is the outcome of student i assigned to group g, DD_g is the dissimilarity measure of group g, 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 grade, socio-economic status, personality traits, homophily and whether they studied economics before) and X_g is a vector of group controls (team aggregates for the individual controls - gender composition, URM prevalence, average baseline grades, standard deviation of baseline grades, 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/ethnicity, while the second one one also includes place of birth and parents' place of birth. The errors are clustered at the group level g. For group outcomes, we employ a similar specification, but at the group level. For binary outcomes, we use a similar logistical regression.

3.1 Impact of diversity on teamwork quality

We start by investigating the impact of diversity on teamwork quality. This outcome is constructed as a Principal Component Analysis (PCA) index, amalgamating three standardized survey-reported dimensions of teamwork quality: the degree of collaboration (on a scale from 0 to 10), 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 report more serene teamwork, irrespective of the nature of the task undertaken. We show regression results in Table 5. In Figure 3, we show the distribution of teamwork quality, controlling for the usual battery of individual and group-level covariates, 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 considering this aggregated measure also

represents a strategy to deal with the multiple hypothesis issue.

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	0.0578	
	(0.0609)	
Quadratic DD in Gender and Race	0.0783**	
	(0.0303)	
DD in Gender, Race, POB and Parents' POB		-0.00415
		(0.0604)
Quadratic DD in Gender, Race, POB and Parents' POB		0.109***
		(0.0405)
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 teamwork 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.

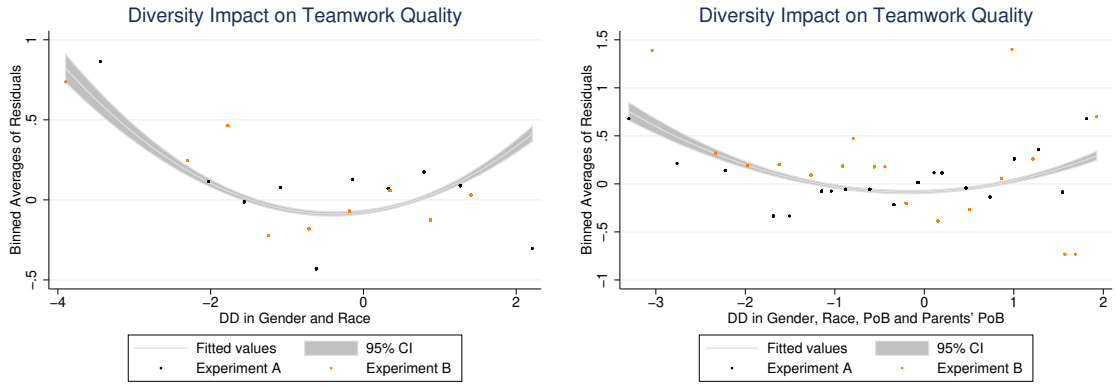


Figure 3: Scatterplots of teamwork quality and two diversity measures respectively (DD in Gender and Race and DD in Gender, Race, POB and Parents' POB), controlled for the usual battery of individual and group regressors.

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. For instance, if there is a group of four members, two of which are East Asian, and two of which are white Caucasians, there will be a clear faultline with respect to the race/ethnicity. In the words of Lau and Murnighan, 1998, group fragmentation resulting from clear faultlines has the potential to hinder group cohesiveness and interaction, forming internal split coalitions with homophilous relationships that can worsen teamwork quality. 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.

3.1.1 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 slightly 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 6, 7, 8 and 9.

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 6: 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 7: 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 8: Homophily by gender - Experiment A

Gender	Homophilic (University)
Male	81.5%
Female	80.1%

Gender	Homophilic (Class)
Male	68.5%
Female	89.9%

Table 9: Homophily by gender - Experiment B

3.2 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 10.

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

We supplement our analysis with specifications that control for teamwork quality. While we discern a robust positive association between teamwork quality and group performance, accounting for this variable only partially influences the observed coefficients. In both experiments these estimates reveal that the groups in the middle of the distribution of diversity were downward biased: in the controlled specifications those groups' performance in relationship to diversity is pushed up. This is consistent with previous results showing those groups feature relatively lower collaboration, partially impacting performance.

However, coefficients corresponding to linear impacts are only minimally affected. While teamwork quality bears a positive association with superior performance, it is not the sole conduit through which diversity impinges on group grades. Therefore, there must be a direct impact of diversity itself as

an input in the production of the final output. 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 individuals supersedes the concomitant communication costs. In contrast, when tasks lean towards mechanistic and adhere to predefined rules and methodologies, attendant communication and coordination hurdles prevail. 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.

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	-0.100*	
	(0.0580)	
Quadratic DD in Gender and Race	-0.0259	
	(0.0292)	
DD in Gender, Race, POB and Parents' POB		-0.101**
		(0.0446)
Quadratic DD in Gender, Race, POB and Parents' POB		-0.0406
		(0.0350)
Group Controls	Y	Y
Observations	163	163
DD in Gender and Race	-0.111*	
	(0.0586)	
Quadratic DD in Gender and Race	-0.0371	
	(0.0301)	
DD in Gender, Race, POB and Parents' POB		-0.103**
		(0.0429)
Quadratic DD in Gender, Race, POB and Parents' POB		-0.0591*
		(0.0356)
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 10: 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.

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 standard tasks.

4 Discussion and Conclusion

This comprehensive analysis underscores the intricate interplay between diversity, teamwork quality, and group performance. The impact of demographic diversity on teamwork quality is orthogonal to the type of task, which is consistent with the hypothesis that it is driven by inner preferences or primitives: it depends on group dynamics, not on the nature of the final output. When it comes to the impact on group performance instead, the results depend on the type of output itself. When controlling for teamwork quality, the estimates for linear coefficients are only marginally affected. This suggests that they are driven prominently by a direct impact of demographic diversity on production, instead of group dynamics.

It is important to note that the duration of teamwork in our experiments is relatively short – one semester – which may exacerbate the distinction between teamwork quality and performance. While this is typically the case in higher education, it might be longer in organizations when coworkers collaborate on long-term projects, like patents and innovation for example. In that case, the quality of teamwork and team performance might become more similar.

The analyses of the effect of two measures of diversity - excluding or including place of birth for respondents and their parents - do not provide noticeably different results. This suggests visible demographic features play the main role. This may be specific to the context of the experiments where participants are undergraduate students and relatively proficient in a common language.

The results offer valuable insights for educational and corporate institutions about how teams should be designed and assessed. If teachers and managers want to maximize team performance and collaboration, they need to consider the type of task involved. While standard assessments have their advantages, particularly in objectively gauging specific competencies, they may downplay the significance of creative knowledge production, which often thrives on the complementarity of efforts,

and as we find, on diversity. This suggests that assessing students and employees on teamwork rather than individual performance might create a more inclusive learning and working environment – in particular in an economy where knowledge production and tasks are becoming increasingly complex.

References

Akerlof, G. A. (1997). "Social Distance and Social Decisions". In: *Econometrica* 65.5, pp. 1005–1027.

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

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

Borghans, L. et al. (2008). "The Economics and Psychology of Personality Traits". In: *Journal of Human Resources* 43.4, pp. 972–1059.

Borjas, G. 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, U., S. Brunow, and A. D'Ambrosio (2020). "The unlikely encounter: Is ethnic diversity in start-ups associated with innovation?" In: *Research Policy* 49.4.

Burns, J. (2006). "Racial stereotypes, stigma and trust in post-apartheid South Africa". In: *Economic Modelling* 23.5, pp. 805–821.

— (2012). "Race, diversity and pro-social behavior in a segmented society". In: *Journal of Economic Behavior Organization* 81.2, pp. 366–378.

Carton, A. M. and J. 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, Y.-T. and D. S. Staples (2013). "Reducing Faultlines in Geographically Dispersed Teams: Self-Disclosure and Task Elaboration". In: *Small Group Research* 44.5, pp. 498–531.

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

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

Currarini, S., M. O. Jackson, and P. 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, A. C. H. (2004). "Work group diversity and group performance: An integrative model and research agenda". In: *Journal of Applied Psychology* 89, pp. 1008–1022.

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

Dinesen, P. T., M. Schaeffer, and K. M. Sønderskov (2020). "Ethnic Diversity and Social Trust: A Narrative and Meta-Analytical Review". In: *Annual Review of Political Science* 23.1, pp. 441–465.

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

Eckel, C. C. and P. J. Grossman (2005). "Managing diversity by creating team identity". In: *Journal of Economic Behavior Organization* 58.3, pp. 371–392.

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

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

Fershtman, C. and U. Gneezy (2001). "Discrimination in a Segmented Society: An Experimental Approach". In: *The Quarterly Journal of Economics* 116.1, pp. 351–377.

Finseraas, H. et al. (2019). "Trust, ethnic diversity, and personal contact: A field experiment". In: *Journal of Public Economics* 173, pp. 72–84.

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

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

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

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

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

Guillaume, Y. 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, B. H., J. A. Nickerson, and H. Owan (2012). "Diversity and Productivity in Production Teams". In: *Advances in the Economic Analysis of Participatory and Labor-Managed Firms*. Vol. 13. Emerald Group Publishing Limited, pp. 99–138.

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

Hoogendoorn, S., H. Oosterbeek, and M. Van Praag (2012). "Ethnic Diversity and Team Performance: A Randomized Field Experiment". In: *Academy of Management Proceedings* 2012.1.

Horwitz, I. and S. 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, J. and M. Gauthier-Loiselle (2010). "How Much Does Immigration Boost Innovation?" In: *American Economic Journal: Macroeconomics* 2.2, pp. 31–56.

Jackson, S. E. and A. 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, D., R. Johnson, and K. Smith (2007). "The state of cooperative learning in postsecondary and professional settings". In: *Educational Psychology Review,* 19.1, pp. 15–29.

Kahane, L., N. Longley, and R. Simmons (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, C. et al. (2018). "Team development interventions: Evidence-based approaches for improving teamwork". In: *American Psychologist* 73.4, pp. 517–531.

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

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

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

Marx, B., V. Pons, and T. Suri (2021). "Diversity and team performance in a Kenyan organization". In: *Journal of Public Economics* 197.

Mathieu, J. 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, A. S. and M. E. Ruckes (2006). "Team Composition". In: *The Journal of Business* 79.3, pp. 1019–1039.

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

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

Moser, P., A. Voena, and F. Waldinger (2014). "German Jewish Émigrés and US Invention". In: *American Economic Review* 104.10, pp. 3222–55.

Østergaard, C. R., B. Timmermans, and K. 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, C., P. Nijkamp, and J. Poot (2012). "Immigration and innovation in European regions". In: *Migration Impact Assessment*. Edward Elgar Publishing. Chap. 8, pp. 261–298.

Ozgen, C., P. Nijkamp, and J. Poot (2013). "The impact of cultural diversity on firm innovation: evidence from Dutch micro-data". In: *IZA Journal of Migration* 2.18.

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

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

Richard, O. C. and R. 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, C. E. (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423.

Simsarian Webber, S. and L. 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, L., M. E. Stanne, and S. 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.

Vogel, R. et al. (2014). "Funding decisions and entrepreneurial team diversity: A field study". In: *Journal of Economic Behavior Organization* 107. Empirical Behavioral Finance, pp. 595–613.

Wegge, J. 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, K. Y. and C. 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, S., B. F. Jones, and B. Uzzi (2007). "The Increasing Dominance of Teams in Production of Knowledge". In: *Science* 316.5827, pp. 1036–1039.

Appendix A Key Variables' Construction

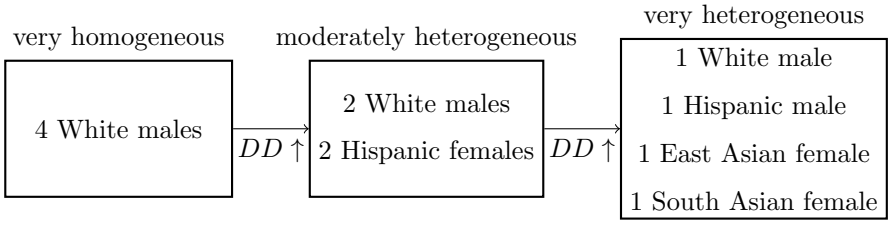
- *URM*: binary variable equal to 1 if Black and/or Hispanic/Latinx selected among the options in the survey question "What is the race/ethnicity that you identify with?", and 0 otherwise. We complement this information by administrative records if the information is not provided by the student through survey.

- *Female*: binary variable equal to 1 if "Female" is selected in the survey question "What is the gender that you identify with?", and 0 otherwise. 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?". Students could choose whether they were born in the United States or abroad.

- *Parents born abroad*: variable constructed through the survey question "Where were your parent(s)/guardian(s) born?". Students could choose whether at least one parent/guardian was born abroad.

- *DD in Gender and Race*, *DD in Gender, Race, Place of Birth and Parents' Place of Birth*: explained in detail in the subsection regarding diversity measures. Built with the package "cluster" in R. We provide an illustrative example below.



- *Able to make friends*: we asked the survey 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 one of the Big Five personality traits, 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 one of the Big Five personality traits, openness.

- *FGLI (First Generation Low Income)*: binary variable asked through survey "Do you identify yourself as a FGLI (First Generation Low Income) student?", equal to 1 if the respondent says yes, 0 otherwise.

- *Financial aspects daily source of stress*: binary variable asked through survey "Are financial aspects a source of concern or stress for you in your daily life?", equal to 1 if the respondent says yes, 0 otherwise.

- *Baseline grade*: sum of the grades from the first two quizzes, completed by students individually at the beginning of the semester.

- *Race/ethnicity-based homophily* and *Gender-based homophily*: explained in detail in the subsection regarding homophily in section 3.

- *Female TA* and *URM TA*: administrative records. We build the URM category consistently with the student-related definition.

- *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 binary variable "No conflict" which we aggregate in the PCA index for the teamwork quality.

- *Equally distributed workload*: binary variable asked through survey "Do you think the workload was typically distributed equally among the group members?", equal to 1 if the respondent says yes, 0 otherwise.

Appendix B Sub-Components of Teamwork Quality

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	0.181	
	(0.174)	
Quadratic DD in Gender and Race	0.172*	
	(0.0939)	
DD in Gender, Race, POB and Parents' POB		0.165
		(0.167)
Quadratic DD in Gender, Race, POB and Parents' POB		0.258**
		(0.127)
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.

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	0.0000927	
	(0.133)	
Quadratic DD in Gender and Race	0.133	
	(0.113)	
DD in Gender, Race, POB and Parents' POB		-0.0942
		(0.129)
Quadratic DD in Gender, Race, POB and Parents' POB		0.246**
		(0.0991)
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.

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	0.225	
	(0.155)	
Quadratic DD in Gender and Race	0.177**	
	(0.0895)	
DD in Gender, Race, POB and Parents' POB		-0.123
		(0.140)
Quadratic DD in Gender, Race, POB and Parents' POB		0.0378
		(0.102)
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.