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† Anne Duchene‡

September 9, 2024

Abstract

This study explores the impact of demographic diversity on teamwork in higher education. Using an experimental design, we find that diversity enhances team performance on creative tasks but hinders it on standard tasks. Additionally, diversity influences teamwork quality in a U-shaped pattern, suggesting that extremely homogeneous or heterogeneous groups collaborate more effectively, while moderately diverse groups face challenges. This paper contributes to understanding the distinct effects of diversity on team creativity and cohesion, emphasizing the role of demographic characteristics in shaping team dynamics.

JEL Codes: I21, J15, A22.

*We thank helpful comments from Hanming Fang, Petra Todd, David Gill, Francesco Agostinelli, Ashley Tharayil, Jamiella Brooks and Bruce Lenthall. 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. IRB Approval Number 850326 for this experiment was granted in 2021. Due to privacy concerns and conditions of our IRB approval, the data used in this study cannot be made publicly available. The data contains sensitive demographic information and detailed group assignments that could lead to the identification of individual participants. Researchers may contact the corresponding author for further inquiries about the data.

†University of Pennsylvania. Corresponding author: odarova@sas.upenn.edu.

‡University of Pennsylvania.

I. 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 demographic 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) and fostering creativity – a crucial skill that affects educational attainment and labor market outcomes (Gill and Prowse, 2021). In education, there is ample evidence showing the positive relationship between collaborative learning and student achievement, effort, persistence, and motivation (Springer et al., 1999; Johnson et al., 2007).

These trends have led to the growth of a large body of literature on the effects of demographic diversity on team performance.[3] While that literature has revealed mixed results so far, it conventionally presents diversity as a double-edged sword: more diverse teams benefit from more creativity and knowledge sharing, but they face higher communication and coordination costs. The divergence of results in the empirical literature has traditionally been interpreted through the lens of these two opposing forces and their relative strength.[4] But does diversity really foster team creativity and knowledge sharing? And does it really hamper team cohesion and increase the risk of conflict? This paper tests the consensus view of diversity's two opposing effects on teamwork in

[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] For detailed reviews, see Guillaume et al., 2017; I. Horwitz and S. Horwitz, 2007; Joshi and Roh, 2009; Williams and O'Reilly, 1998; Alesina and La Ferrara, 2005; Simsarian Webber and Donahue, 2001; Van Knippenberg et al., 2004.

[4] For creativity and information sharing, see (Prat, 2002; Mello and Ruckes, 2006; I. Horwitz and S. Horwitz, 2007). For communication and coordination costs, see Lazear, 1999; Morgan and Várdy, 2009; Hamilton et al., 2012.

the context of higher education. We exploit an experimental setting in a large undergraduate class, where students are randomly assigned to homework groups of three or four, with varying levels of diversity in terms of race, gender, and place of birth. This multi-dimensional measure of diversity departs from the previous literature, which typically examines a single demographic characteristic, and allows us to consider the full spectrum of group diversity, from very homogeneous to very heterogeneous groups.

To understand the effect of team diversity on creativity and knowledge sharing, we run two almost identical experiments that differ on the type of task performed. We find that more diverse groups perform better when the task is creative and complex, and worse when the task is standard. Consistently with the consensus view, this result suggests that diversity's positive impact on team performance hinges on gains from creativity. We then address the effect of team diversity on coordination and communication, by building an index of teamwork *quality*, based on collaboration between members, balance of member contributions, and the absence of conflicts.[5] We find that diversity has a U-shaped effect on teamwork quality. More precisely, very homogeneous and very heterogeneous groups work better together, while intermediate groups face more communication and coordination problems. While it refutes to some extent the consensus view, this result echoes the concept of *faultlines* developed in the psychology and organizational behavior literature.[6] 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 moderately diverse groups, increasing the probability of conflict or lack of cooperation and ultimately hurting group cohesion. Such pattern seems to be determined by preferences and, consistently, is not contingent on the type of task.

This paper proceeds as follows. Section II reviews the related literature. Section III details the

[5]Note that this interpretation of teamwork is related to the literature on social skills and team production. In particular, Deming, 2017 shows that social skills improve communication and collaboration, and thus decrease coordination costs.

[6]Lau and Murnighan, 1998; Carton and Cummings, 2013

experimental design and describes the setting. Section IV presents and discusses the reduced-form empirical analyses and results, and Section V concludes.

II. Contribution to the Literature

This paper contributes to a large literature on diversity in teams, at the intersection of economics, psychology and organizational behavior. Empirical studies analyzing the impact of demographic diversity on team performance have produced mixed findings. These inconsistencies are typically attributed to the relative strength of diversity's two opposing effects on performance, as both a booster to creativity and an obstacle to team cohesion. But to our knowledge, these two effects have not been disentangled empirically. Our paper contributes to the existing literature in three ways.

First, we address the impact of diversity on creativity. Papers that find a positive impact of diversity on team performance seem to focus on tasks that are highly creative or involve strategic and complex decision-making. For example, Freeman and Huang, 2015 show that nationally diverse research teams publish more often in high-impact journals.[7] In Vogel et al., 2014, more gender and ethnic diversity in entrepreneurial teams result in better funding. 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.[8] By contrast, studies that find a negative impact of diversity on performance focus on less creative, more standard tasks. For example, Lyons, 2017 finds that birthplace diversity hinders performance when tasks are highly specialized.[9] In this paper, we factor in different types of tasks by running two experiments on two different cohorts of the same class. The experiments' design is identical, except for the type of task: in one experiment, teams perform creative and complex tasks, while in the other experiment,

[7] Similarly, Ferrucci and Lissoni, 2019 find that R&D teams with more migrant inventors are associated with higher quality patents.

[8] See also Richard and Shelor, 2002; Jackson and Joshi, 2004; Wegge et al., 2008; Hamilton et al., 2012; Ozgen et al., 2012; Ozgen et al., 2013. Note the exception of Dutcher and Rodet, 2022, who find that demographic diversity does not have a measurable effect on team creativity, when accounting for diversity of experience.

[9] See also Leonard and Levine, 2006, Hjort, 2014 and Marx et al., 2021.

tasks are standard. This distinction allows us to test whether the positive impact of diversity on team performance really hinges on creativity.

Second, we address the impact of diversity on team communication, coordination and cohesion. This question is studied by two branches of the literature. The first branch is the literature on social trust and conflict, mostly focused on large communities like neighborhoods, cities or nations.[10] Those papers generally find segregation and demographic fractionalization to be associated with higher levels of conflict and lower trust, mainly because of agents' homophily – they trust those who are different from themselves less than those who are more similar (Dinesen et al., 2020). The second branch is the psychology and organizational behavior literature on *faultlines*.[11] Faultlines are hypothetical dividing lines that can potentially split a group into more homogenous subgroups, based on one or more demographic attributes. While faultlines seem to disappear altogether at minimum and maximum levels of diversity, they are present in moderately heterogeneous teams where subgroups can be the source of team conflict. In this paper, we analyze the impact of diversity on team cohesion by building an index of teamwork *quality*, based on collaboration between members, balance of member contributions, and the absence of conflicts. We find that the impact of diversity on teamwork quality is not necessarily negative, and follows a pattern that is more consistent with the second branch of the literature on faultlines. One reason why our results differ from the first branch of the literature may be a non-linear impact of diversity - an aspect which we carefully address in the paper.

Finally, we address a shortcoming of the literature about the the range of diversity considered and the dimensions included in such measure. Most papers focus on a single isolated demographic attribute (e.g., gender, age, ethnicity). But as argued by Jackson and Joshi, 2004, each team member's identity is likely defined by the confluence of multiple attributes. In this paper, we offer an alternative measure of diversity, as the degree of dissimilarity between team members with respect to three demographic characteristics taken together: gender, race/ethnicity and place of birth.

[10]For a review of this literature, see Dinesen et al., 2020. See also Fershtman and Gneezy, 2001, Burns, 2006, Finseraas et al., 2019, that study trust in one-on-one experimental games.

[11]Lau and Murnighan, 1998; Carton and Cummings, 2013; Chiu and Staples, 2013

Using that index, we encompass a broad range of diversity, from very homogeneous to very heterogeneous groups – unlike most studies that focus on a specific portion of the diversity spectrum. Indeed, some papers only consider the right-hand side of the spectrum, like Hoogendoorn et al., 2012 (in which the least diverse teams have at least 20% of non-native members). Conversely, other papers focus on lower end of the diversity spectrum, like the literature on social trust mentioned above (which typically compares fully homogeneous groups to segregated groups), or Lyons, 2017 and Marx et al., 2021, which only consider groups that are either completely homogeneous or that are split in half across different nationalities.

III. Experiment Design and Empirical Analysis

A. Experiments

We study the effect of group composition on collaboration and performance using data from an introductory undergraduate microeconomics course taught at a large 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).

At the beginning of the semester, students are randomly assigned through a computer-based algorithm to groups of three or four within their recitation section. Every other week, each group sends a written project to its TA, and then presents it orally in recitation.

This is an ideal setting to analyze the effects 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.

We run two almost identical experiments with two different cohorts in the same class. The only difference between the experiments' design is the type of task assigned to groups[12].

In the first experiment (Experiment A), groups alternate between two types of open-ended exercises. The first type entails creating an exam question based on the chapter covered that week.[13] The other type consists in researching and writing a paragraph about a given prompt on a current event or a policy debate.[14]

In the second experiment (Experiment B), there is only one type of exercise. Groups are given an existing exam question, for which they must explain the solution. 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.

[12]A pre-analysis plan was prepared prior to the commencement of this study and attached to the AEA RCT Registry Record. The only substantial deviation from the plan was the conduct of the experiment in two iterations with differing task types—one creative and the other standard. Although the assignment to either the creative or standard task experiment was not randomized, this approach allowed us to contrast the effects of task type on team performance under varying levels of diversity.

[13]The question is multiple choice with at least four possible answers, and the group must also explain its solution. Evaluation of the group's work takes into account the level of difficulty and originality of the question, and correctness of its solution.

[14]Here is an example of prompt for the chapter on firms' costs and competition: "Last Spring in New York city, Starbucks announced that while it would keep some stores open, it would close some stores permanently, and it would shutter about 20 others a year later when their leases ended. What was the relationship between cost and revenue for each of these three categories of Starbucks stores? Why do you think Starbucks is moving to a new model of to a smaller, pickup model?"

B. Data

The paper employs a novel data source on two cohorts of students, each corresponding to a given experiment. A survey was administered to students in each cohort: in the first one (experiment A), 547 out of 588 students responded (a response rate of 93%), while in the second one (experiment B) 604 out of 629 students responded (96% response rate). The survey contains *i)* personality traits (extroversion and openness[15], as two relevant Big Five personality traits in this setting),[16] 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 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 between whites and non-whites, or between URMs and others.[17] 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, in-class participation and exams, but most importantly scores on group work - 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

[15]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".

[16]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.

[17]URM stands for Underrepresented Minority, an individual who identifies as African American/Black; American Indian/Alaskan Native; Hispanic; Native Hawaiian/other Pacific Islander; or more than one race when at least one of the preceding URM categories above.

determining grades are automatised on a virtual platform, leaving negligible 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 the first cohort (experiment A), and 163 random groups in the second (experiment B). A detailed list of the key variables' construction is provided in Appendix A.

The type of group work is different in the two experiments: groups perform creative tasks in the first cohort and standard tasks in the second cohort, without the benefit of randomization. Despite the non-random assignment to one experiment or the other, the summary statistics across both panels reveal a high degree of comparability between the cohorts along key dimensions, as confirmed by simple t-tests. It is worth noting that we estimate 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 slightly different across experiments, we are able to compare final course grades, which do not statistically differ. By analyzing two separate cohorts, our study design effectively eliminates potential order-of-exposure effects, ensuring that any observed differences in outcomes are attributable to the experimental conditions rather than the sequence of exposure. In section IV, 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)
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)

Significance levels: * p<0.1 ** p<0.05 *** p<0.01.

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.

C. 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. Most of those measures are uni-dimensional, which is problematic in our study, given the size of groups (3-4 individuals) and the possible formation of subgroups within those groups.

To show why, 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.

To address that issue, 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 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 repre-

sents 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.

Note that we focus on dissimilarity according to those three demographic attributes because conspicuous characteristics are likely to determine how team members identify each other early on in their interactions and how they form social networks. As noted by Tsui et al., 1992, one's similarity on visible and relatively immutable traits influences feelings of identification.

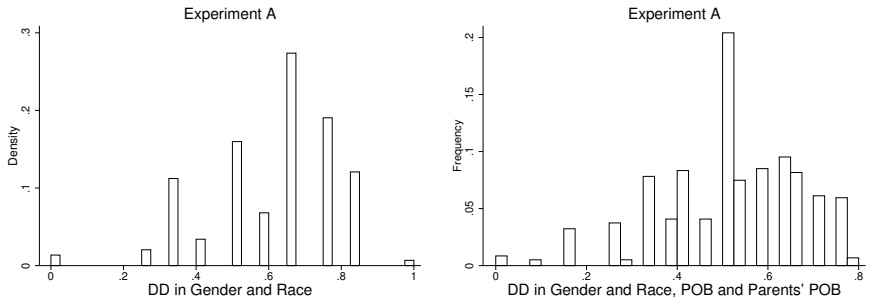


Figure 1: Experiment A. Distribution of the two different dissimilarity (DD) measures for the groups in our dataset according to 1) race/ethnicity and gender; 2) race/ethnicity, gender, and migration status.

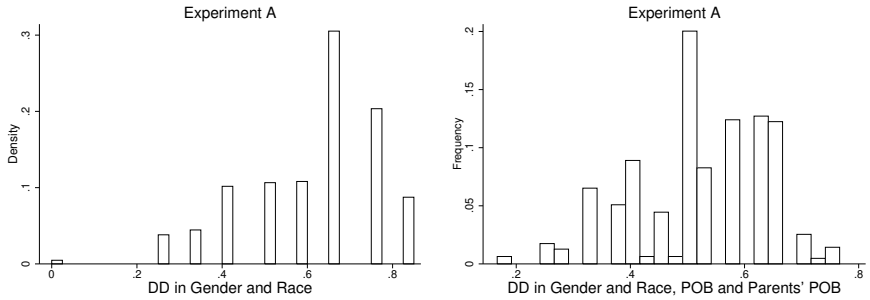


Figure 2: Experiment B. Distribution of the two different dissimilarity (DD) measures for the groups in our dataset according to 1) race/ethnicity and gender; 2) race/ethnicity, gender, and migration status.

D. 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 the first cohort (experiment A) and 88.9% for the second (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
Significance levels: * 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

Significance levels: * 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

* Significance levels: 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 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.

IV. 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 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.

A. 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 cohesion, irrespective of the nature of the task undertaken. We show regression results in Table 5. In Figure 3, we employ cubic B-spline regression with three knots to model the relationship between teamwork quality and diversity, controlling for the usual battery of individual and group-level covariates. This method allows for a more flexible fit compared to a simple quadratic form, capturing potential higher degrees non-linearities. The shaded areas represent the 95% confidence intervals, calculated using the standard errors of the predicted values. The results confirm the U-shaped relationship between teamwork quality and diversity for both diversity measures, as evidenced by the spline fits, as we don't observe additional inflec-

tion points or significant deviations. 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.

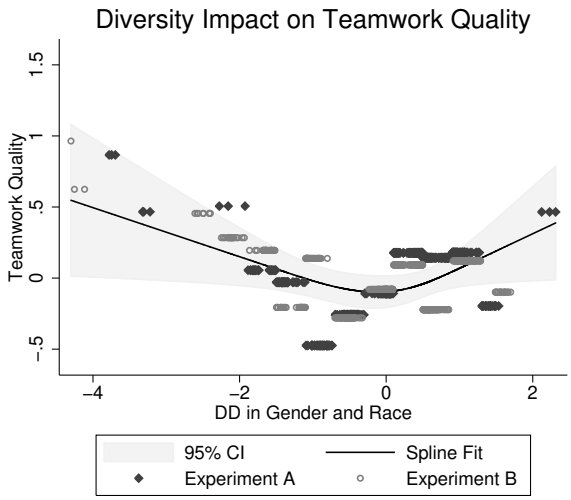
Experiment A				
	(1)	(2)	(3)	(4)
DD in Gender and Race	0.0105	0.0924		
	(0.0605)	(0.0741)		
Quadratic DD in Gender and Race		0.0854**		
		(0.0345)		
DD in Gender, Race, PoB and Parents' PoB			-0.0185	0.0338
			(0.0684)	(0.0782)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.0643*
				(0.0346)
F-statistic for Quadratic Term		6.13**		3.45*
Prob > F		0.014		0.065
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	429	429	429	429
Experiment B				
	(1)	(2)	(3)	(4)
DD in Gender and Race	-0.0295	0.0578		
	(0.0459)	(0.0609)		
Quadratic DD in Gender and Race		0.0783**		
		(0.0303)		
DD in Gender, Race, PoB and Parents' PoB			-0.0729	-0.00415
			(0.0510)	(0.0604)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.109***
				(0.0405)
F-statistic for Quadratic Term		6.68**		7.21***
Prob > F		0.011		0.008
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	493	493	493	493

Significance levels: * p<0.1 ** p<0.05 *** p<0.01.

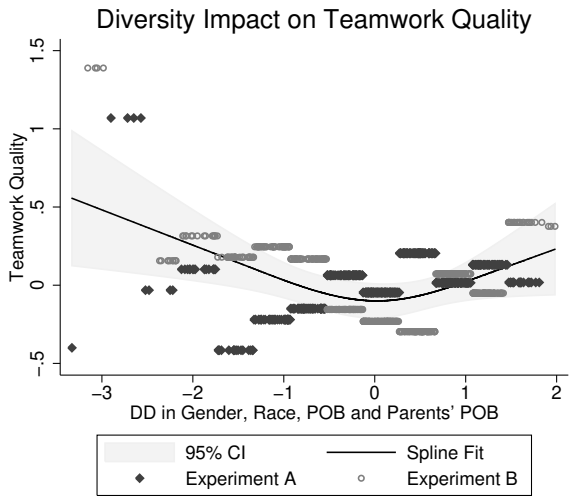
Table 5: Impact of diversity on teamwork quality. This is a PCA variable aggregating three self-reported measures through surveys: the degree of collaboration within teams, the absence of conflicts, and the equal distribution of the workload.

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.

These results echo the well-established "group *faultlines*" theory (Lau and Murnighan, 1998;



(a) Scatterplot of teamwork quality and DD in Gender and Race.



(b) Scatterplot of teamwork quality and DD in Gender, Race, POB, and Parents' POB.

Figure 3: Scatterplots of teamwork quality and two diversity measures controlled for individual and group regressors.

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.

1. Homophily

For the faultlines theory to be applicable to this context, it must be the case that group members display out-group-aversion (Alesina and La Ferrara, 2002; Olsson et al., 2005), or homophily – "the tendency of individuals to associate with similar others" (Lawrence and Shah, 2020).

To test for the presence of this phenomenon, we ask students to indicate the races and genders of their closest friends in the University. We then define homophily by employing the same definitions as Currarini et al., 2009: 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 a comparison with the fractions of those types in our sample in particular. 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. These results confirm the well-established result from previous research that demographic homophily (e.g. based on race/ethnicity, gender, age) is an important determinant of social networks (Chetty et al., 2022; Currarini et al., 2009).

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.

B. 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 depending on the experimental context. In experiment A, characterized by more creative and complex 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. This suggests that teamwork quality is not the sole channel through which diversity affects group grades. In particular, there must be a direct impact of diversity itself as an input in the production of the final output, and this impact must differ by task type.

For creative tasks, the positive effect may come from creativity gains: a more diverse array of backgrounds and knowledge corresponds to a higher performance. The estimates and the F-statistic of the quadratic term for diversity measures suggest that adding such term to the model

Experiment A				
	(1)	(2)	(3)	(4)
DD in Gender and Race	0.0138	0.0239**		
	(0.00923)	(0.0117)		
Quadratic DD in Gender and Race		0.0113*		
		(0.00614)		
DD in Gender, Race, POB			0.0133	0.0237*
			(0.00954)	(0.0125)
Quadratic DD in Gender, Race, POB				0.0136**
				(0.00624)
F-statistic for Quadratic Term		3.37*		4.78**
Prob > F		0.068		0.030
Group Controls	Y	Y	Y	Y
Observations	167	167	167	167
Mean Teamwork Quality	0.0348**	0.0327**	0.0359**	0.0334**
	(0.0153)	(0.0148)	(0.0156)	(0.0151)
DD in Gender and Race	0.0133	0.0201*		
	(0.00858)	(0.0106)		
Quadratic DD in Gender and Race		0.00750		
		(0.00489)		
DD in Gender, Race, POB			0.0156	0.0234*
			(0.00961)	(0.0122)
Quadratic DD in Gender, Race, POB				0.0104*
				(0.00549)
F-statistic for Quadratic Term		2.35		3.58*
Prob > F		0.127		0.060
Group Controls	Y	Y	Y	Y
Observations	167	167	167	167
Experiment B				
	(1)	(2)	(3)	(4)
DD in Gender and Race	-0.0725**	-0.100*		
	(0.0318)	(0.0580)		
Quadratic DD in Gender and Race		-0.0259		
		(0.0292)		
DD in Gender, Race, POB			-0.0807**	-0.101**
			(0.0352)	(0.0446)
Quadratic DD in Gender, Race, POB				-0.0406
				(0.0350)
F-statistic for Quadratic Term		0.79		1.35
Prob > F		0.376		0.247
Group Controls	Y	Y	Y	Y
Observations	163	163	163	163
Mean Teamwork Quality	0.111**	0.121**	0.106*	0.127**
	(0.0547)	(0.0563)	(0.0537)	(0.0545)
DD in Gender and Race	-0.0710**	-0.111*		
	(0.0317)	(0.0586)		
Quadratic DD in Gender and Race		-0.0371		
		(0.0301)		
DD in Gender, Race, POB			-0.0743**	-0.103**
			(0.0334)	(0.0429)
Quadratic DD in Gender, Race, POB				-0.0591*
				(0.0356)
F-statistic for Quadratic Term		1.51		2.75
Prob > F		0.220		0.100
Group Controls	Y	Y	Y	Y
Observations	163	163	163	163

Significance levels: * p<0.1 ** p<0.05 *** p<0.01.

Table 10: Impact of diversity on the group score for group assignments.

contributes to explaining the outcome in the creative tasks experiment, but not in the non-creative one. In particular, there seems to be some suggestive evidence of convexity for the creative tasks: in other words, diversity's marginal impact is increasing, which is consistent with creativity gains coming from complementarity of team members' contributions. On the other hand, the quadratic term does not seem to contribute to explain the group score in Experiment B.

Related to this interpretation, we can go back to Table 1 to 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 creative tasks. 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 member of the team contributed, which is consistent with the idea of creativity gains coming from different points of view and more members making complementary efforts towards the production of the assignment. As a final point, notice that students in the second experiment declared to be generally more open to suggestions, compared to students in experiment A. Therefore, we may be underestimating the negative impact of diversity on teamwork when it comes to standard tasks.

However, there is still a puzzle to be solved: if diverse groups coordinate well as teams, why can't they perform better when given standard tasks (as in Experiment B)? While we lack additional data to test channels empirically, we can interpret the results through the lens of the existing literature to identify two possible explanations.

One possibility is that the teamwork quality measure we build does not capture further communication or crowding costs that are associated with both standard tasks and high diversity. 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 hurdles prevail. In other words, when there is only one correct response to an assignment, diversity not only does not help - if anything, it can represent an obstacle, even in absence of group conflicts or of lack of team coordination.

Another possible explanation is related to the social identity theory, in particular group-contingent social preferences, according to which sharing a common social identity increases group effort, as shown by R. Chen and Y. Chen, 2011. In our framework, this theory implies that as a team becomes more diverse, there is less group identity and therefore a lower effort: group identity is the most salient and leads to maximum effort in fully homogeneous teams, while effort is minimal in very diverse teams with no social group identity.

Overall, the impact of diversity on effort would then depend on three effects: the positive or constant effect of creativity gains (depending on the type of task), the hump-shaped effect of coordination costs, and the negative effect of group identity on effort.[18] For creative and complex tasks, performance increases with diversity if creativity gains are relatively strong compared to the loss of group identity. On the other hand, standard task yield no or little creativity gains, so performance decreases with diversity due to the loss of group identity. We show the three different forces associated with diversity and teamwork in Figure 4.[19]

weidmann

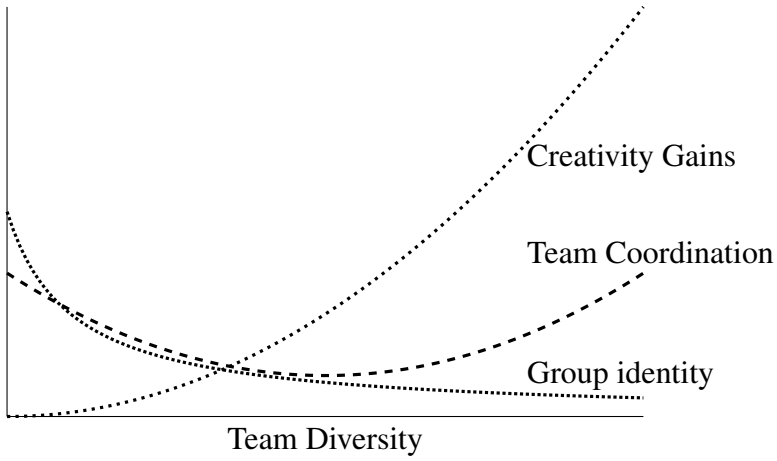


Figure 4: Illustration of the different forces associated with diversity and teamwork.

[18]The direct link between coordination costs and team effort is suggested by Deming, 2017 who shows that social skills decrease coordination costs, and Weidmann and Deming, 2021 who show that social skills have a positive impact on team effort.

[19]Note that in this reasoning, we are assuming that team incentives (i.e. group grades) have the same impact on team effort, regardless of the type of task. However, as Englmaier et al., 2018 suggest, the effectiveness of bonus incentives might be different when teams perform non-routine, complex tasks.

V. Conclusion

This comprehensive analysis underscores the intricate interplay between diversity, teamwork quality, and group performance. The impact of demographic diversity on teamwork quality is independent of 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. However, when it comes to the impact of diversity on group performance, 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 team cohesion may play a larger role in longer term collaborations. Nevertheless, it's pertinent to recognize that many group work scenarios in higher education settings also unfold within analogous one-semester courses. Similarly, projects of comparable length are commonplace in various work-related contexts.

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 possibly corporate, institutions about how teams should be designed and assessed. If leaders aim 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 spillovers, and as we find, on diversity. This suggests that assessing students on teamwork rather than only on individual performance might create a more inclusive learning environment – and appropriate in particular in

an economy where knowledge production and tasks are becoming increasingly complex.

References

Alesina, A. and E. La Ferrara (2002). "Who trusts others?" In: *Journal of Economic Literature* 85.2, pp. 207–234.

— (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.

Chen, R. and Y. Chen (2011). "The Potential of Social Identity for Equilibrium Selection". In: *American Economic Review* 101.6, pp. 2562–2589.

Chetty, R. et al. (2022). "Socialcapital II: determinants of economic connectedness". In: *Nature* 608.7921, pp. 122–134.

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.

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.

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.

Dutcher, G. and C. S. Rodet (2022). “Which two heads are better than one? Uncovering the positive effects of diversity in creative teams”. In: *Journal of Economics & Management Strategy* 31.4, pp. 884–897.

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.

Englmaier, F. et al. (2018). “The Effect of Incentives in Non-Routine Analytical Team Tasks”. In: *CESifo Working Paper* 6903.

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.

Gill, D. and V. Prowse (2021). “The Creativity Premium”. In: *IZA Discussion Paper* 14421.

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.

Joshi, A. and H. Roh (2009). “The role of context in work team diversity research: a meta-analytic review”. In: *Academy of Management Journal* 52.3, pp. 599–627.

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.

Lawrence, B. S. and N. P. Shah (2020). “Homophily: Measures and meaning”. In: *Academy of Management Annals* 14.2, pp. 513–597.

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

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.

Olsson, A. et al. (2005). “he Role of Social Groups in the Persistence of Learned Fear”. In: *Science* 309.29, pp. 785–787.

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.

— (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.

Tsui, A. S., T. D. Egan, and C. A. O’Reilly (1992). “Being different: Relational demography and organizational attachment”. In: *Administrative Science Quarterly* 37.4, pp. 549–579.

Van Knippenberg, D., C. K. De Dreu, and A. C. Homan (2004). “Work group diversity and group performance: An integrative model and research agenda”. In: *Journal of Applied Psychology* 89, pp. 1008–1022.

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.

Weidmann, B. and D. J. Deming (2021). “Team players: How social skills improve team performance”. In: *Econometrica* 89.6, pp. 2637–2657.

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.

A. Construction of Key Variables

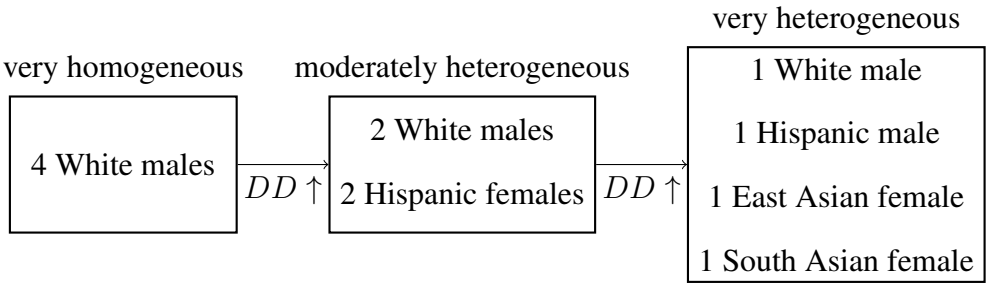
- *URM*: binary variable equal to 1 if Black and/or Hispanic/Latinx selected among the options in the survey question "What is the race/ethnicity that you identify with?", and 0 otherwise. We complement this information by administrative records if the information is not provided by the student through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

- *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.

B. Sub-Components of Teamwork Quality

Experiment A				
	(1)	(2)	(3)	(4)
DD in Gender and Race	-0.0161	0.0742		
	(0.0630)	(0.0658)		
Quadratic DD in Gender and Race		0.102**		
		(0.0420)		
DD in Gender, Race, PoB and Parents' PoB			-0.0714	-0.0259
			(0.0674)	(0.0743)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.0612
				(0.0387)
F-statistic for Quadratic Term		5.86**		2.50
Prob > F		0.0165		0.115
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	493	493	493	493
Experiment B				
	(1)	(2)	(3)	(4)
DD in Gender and Race	-0.0105	0.181		
	(0.143)	(0.174)		
Quadratic DD in Gender and Race		0.172*		
		(0.0939)		
DD in Gender, Race, PoB and Parents' PoB			0.00137	0.165
			(0.144)	(0.167)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.258**
				(0.127)
F-statistic for Quadratic Term		3.35*		4.11**
Prob > F		0.0692		0.0443
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	493	493	493	493

Significance levels: * 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				
	(1)	(2)	(3)	(4)
DD in Gender and Race	0.00427	0.171		
	(0.141)	(0.194)		
Quadratic DD in Gender and Race		0.191**		
		(0.0943)		
DD in Gender, Race, PoB and Parents' PoB			-0.0590	0.0517
			(0.159)	(0.192)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.157*
				(0.0924)
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	429	429	429	429
Experiment B				
	(1)	(2)	(3)	(4)
DD in Gender and Race	-0.128	0.0000927		
	(0.0992)	(0.133)		
Quadratic DD in Gender and Race		0.133		
		(0.113)		
DD in Gender, Race, PoB and Parents' PoB			-0.219**	-0.0942
			(0.110)	(0.129)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.246**
				(0.0991)
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	493	493	493	493

Significance levels: * 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				
	(1)	(2)	(3)	(4)
DD in Gender and Race	0.103	0.157		
	(0.126)	(0.181)		
Quadratic DD in Gender and Race		0.0598		
		(0.0837)		
DD in Gender, Race, PoB and Parents' PoB			0.0607	0.0939
			(0.131)	(0.155)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.0449
				(0.0814)
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	493	493	493	493
Experiment B				
	(1)	(2)	(3)	(4)
DD in Gender and Race	0.0478	0.225		
	(0.115)	(0.155)		
Quadratic DD in Gender and Race		0.177**		
		(0.0895)		
DD in Gender, Race, PoB and Parents' PoB			-0.141	-0.123
			(0.122)	(0.140)
Quadratic DD in Gender, Race, PoB and Parents' PoB				0.0378
				(0.102)
Individual Controls	Y	Y	Y	Y
Group Controls	Y	Y	Y	Y
Observations	536	536	536	536

Significance levels: * 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.