Experimental Economics (2023) 26:749–774
https://doi.org/10.1007/s10683-023-09800-2

ORIGINAL PAPER

When, and why, do teams benefit from self-selection?

Mira Fischer[1] · **Rainer Michael Rilke**[2] · **B. Burcin Yurtoglu**[3]

Received: 20 December 2021 / Revised: 2 March 2023 / Accepted: 3 March 2023 /
Published online: 28 March 2023
© The Author(s) 2023

Abstract

We investigate the effect of team formation and task characteristics on performance in high-stakes team tasks. In two field experiments, randomly assigned teams performed significantly better than self-selected teams in a task that allowed for an unequal work distribution. The effect was reversed if the task required the two team members to contribute more equally. Investigating mechanisms, we observe that teams become more similar in ability and report to cooperate better when team members can choose each other. We show how different levels of skill complementarity across tasks may explain our results: If team performance largely depends on the abilities of one team member, random team assignment may be preferred because it leads to a more equal distribution of skills across teams. However, if both team members' abilities play a significant role in team production, the advantage of random assignment is reduced, and the value of team cooperation increases.

Keywords Team performance · Self-selection · Field experiment

JEL Classification I21 · M54 · C93

This paper analyses two field experiments. The field experiments were pre-registered with the code AEARCTR-0002757 and AEARCTR-0003646 under the title "Peer selection and performance - A field experiment in higher education". We thank Steffen Loev, Marek Becker, and Andrija Denic for their extremely helpful assistance with the data. We also thank Bernard Black, Robert Dur, Ayse Karaevli, Simeon Schudy, Gari Walkowitz, participants of the Advances with Field Experiments Conference in Boston, and seminar participants at the Higher School of Economics in Moscow, Humboldt University of Berlin, University of Trier, University of Duisburg-Essen, University of Mannheim, Burgundy School of Business in Dijon, University of Amsterdam, and WHU - Otto Beisheim School of Management for their helpful comments and suggestions on earlier versions of this paper. Financial support by Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) is gratefully acknowledged. The replication material for the study is available at 10.17605/OSF.IO/6UYZK.

✉ Rainer Michael Rilke
rainer.rilke@whu.edu

Extended author information available on the last page of the article

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

1 Introduction

In today's highly complex economic environment, cooperation among individuals is crucial for organizational success. As businesses become increasingly global and cross-functional, the need for teamwork has been growing in all domains of work and life (Cross et al., 2016; O'Neill and Salas, 2018). Indeed, firms and organizations create value by providing mechanisms for people to work together and to take advantage of complementarities in their skills and interests (Lazear and Oyer, 2012), especially when they need to develop and introduce new products (Sivasubramaniam et al., 2012; Tabrizi, 2015). The nature and effectiveness of teamwork matter for productivity in diverse settings, ranging from entrepreneurial ventures (Reagans and Zuckerman, 2019) to the mutual fund industry (Patel and Sarkissian, 2017), and from medical practice (Geraghty and Paterson-Brown, 2018) to research projects seeking to achieve scientific breakthroughs (Wuchty et al., 2007).

Economists and management scholars have studied extensively the influence of various factors that affect team performance including the role of incentives (e.g., team bonuses or tournaments) and cooperation in teams. Prior research has found a positive impact of team bonuses and team piece rates (e.g., Englmaier et al., 2018; Erev et al., 1993; Friebel et al., 2017; Hamilton et al., 2003) and inconclusive evidence on the effects of team tournament incentives on productivity (e.g., Bandiera et al., 2013; Delfgaauw et al., 2022). While potentially crucial for team performance, only very few studies have investigated the effects of team assignment mechanisms. Chen and Gong (2018) found that university students who self-selected their teammates performed better on a presentation task than students who were randomly assigned to teams. Likewise, Boss et al. (2021) found that without task autonomy students who could freely choose with whom they worked performed better on a presentation task than randomly assigned student teams; with task autonomy no performance difference between self-selected and randomly assigned teams was found.

Different team tasks may require different combinations of inputs from team members and the team assignment mechanism may impact those. Thus the effects of self-selection and random assignment on team performance likely depend on the characteristics of the task. For example, when people are allowed to choose their teammates, they may match with people they like (e.g., Curranrini et al., 2009; Leider et al., 2009), which may increase productivity by improving the communication and cooperation in such teams as compared to randomly assigned teams.[1] However, descriptive evidence on teams suggests that a higher quality of collaboration within teams does not necessarily lead to higher productivity (Stewart and Barrick, 2000) and that the association between teamwork quality and team performance

[1] Laboratory experiments examining related questions in formal games found that high-ability individuals are motivated by expected personal gains rather than pro-social preferences when deciding to join a team (Cooper et al., 2021). However, cooperation in endogenously formed groups was found to be similar to the contribution levels in groups with exogenous matching (e.g., Chen, 2017; Gächter and Thöni, 2005; Guido et al., 2019). In a firm context, people appear to trade off both the pecuniary benefits of better cooperation and the non-pecuniary benefits of working in teams with friends against the pecuniary benefits of working with higher-ability team members (Bandiera et al., 2013; Hamilton et al., 2003).

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

depends on specific task characteristics. Hoegl and Gemuenden (2001) identify the innovativeness of the task as a key characteristic. They report that for tasks that have straightforward solutions and only involve moderate levels of technical novelty, uncertainty, and complexity, teamwork quality should have a lesser effect on team performance than for projects with greater degrees of freedom and higher complexity.

Furthermore, if people expect positive complementarities between their own and their teammate's abilities, self-selection would result in positive assortative matching on those abilities within the sample (Becker, 1973) and lead to productivity gains if actual complementarities are present.[2] However, self-selection may lead to lower productivity compared to random assignment if team members' skills are mutually substitutable, but assortative matching results in an inefficient concentration of high-ability workers in some teams while leading to the absence of high-ability workers in other teams.

As the effects of self-selection and random assignment on team performance likely hinge on the characteristics of the task, our study focuses on two different tasks when analyzing how these assignment mechanisms influence composition, collaborative effort, and performance of teams. We expect that when people are allowed to choose a teammate, they will tend to match with someone they like and who has similar abilities. Furthermore, we expect that when performing a task well requires high levels of skill complementarity and collaboration, self-selection leads to higher average team performance. On the contrary, when in fulfilling the team task the team members' abilities are substitutes, we hypothesize that self-selection is detrimental to average team performance, because it leads to an inefficient concentration of skills in some teams and collaboration is rendered relatively unimportant.

To test these hypotheses, we embedded two field experiments in a mandatory microeconomics course for first-year undergraduate students at a major German business school. The course consisted of two parallel study groups who were receiving the same course content from the same instructor. In the winter quarters of 2017/18 and 2018/19, two cohorts of students were randomly assigned to those study groups at the individual level. In the first week of the quarter, in one study group students were allowed to choose a teammate (treatment *Self*), while in the other study group students were randomly assigned to a team of two (treatment *Random*).

The teams had to work on two high-stakes tasks that were graded for correctness, conciseness, and coherence: either a written task that required the team to submit a written solution, or a video task that required the team members to submit a videotaped team solution in which each of the team member and their respective contribution to the solution was visible. In the written task format full substitutability of team members' contributions was thus possible as it was not verifiable which team

[2] Gary Becker's (1973) famous model showed that with one-dimensional characteristics, positive complementarities in those characteristics with respect to joint surplus lead to positive assortative matching. Many studies have substantiated this result empirically by showing that, for example, spouses assortatively match on educational attainment (Siow, 2015).

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

member contributed which part, while in the video format this was not the case as each team member took up half of the presentation time.

Using objective high-stakes skill measures from the university registry, we find that compared to teams that were randomly assigned, teams that were self-selected were more homogeneous in terms of their abilities. Additionally, they exhibited higher levels of perceived collaborative effort. Interestingly, our results show that self-selected teams performed significantly worse than randomly assigned teams on the written task but tended to do better on the video task.

These findings can be explained by a simple formal model that demonstrates that the likely benefits of self-selection come into play only if the contributions of both team members are needed to complete the task. In other words, if the skills needed to perform a task are, in principle, fully substitutable (if one team member can produce an answer, the skills of the other team member are useless), this task is, on average, performed better by randomly assigned teams because random assignment raises average maximum team ability as it prevents assortative matching. However, if the skills needed to perform a task are sufficiently complementary (the higher the skills of one team member the more productive the skills of the other team member), self-selected teams may outperform randomly assigned teams because cooperation of assortative matching of good team members creates synergies that may make up for inferior performance of teams with only less skilled team members. Additionally, if self-selected teams are on average higher motivated because they can work with someone they like, this is an additional channel that may contribute to self-selected teams superior performance vis-à-vis randomly assigned teams in certain tasks.

This study contributes to the small body of experimental work on the effects of different team assignment mechanisms on team performance in real-world settings. Chen and Gong (2018) used a field experiment in a large undergraduate class at the National University of Singapore to show that self-selection led to a process of team formation that was based on the members' social connections rather than on their skills and resulted in higher effort towards the task. University students who self-selected their teammates performed better on an oral presentation task than students who were randomly assigned to teams as well as students who were assigned to teams using an algorithm designed to maximize skill complementarity. Boss et al. (2021) used a field experiment in a compulsory undergraduate introductory Business and Entrepreneurship class at a large public university in Germany and found that teams with autonomy over choosing either ideas or team members outperformed teams who could choose neither. Performance was measured in a pitch deck presentation task, submitted in the form of slides. The effect of choosing which idea to present was significantly stronger than the effect of choosing teams. However, teams who could choose both ideas and team members did not perform better than teams who could choose neither. Büyükboyaci and Robbett (2019) investigated a related question in a laboratory experiment. They found that when specialization was not possible, self-selection had no effect on performance; and that the option to specialize had a positive effect on performance, which was significantly magnified when agents had a say in who joined their team.

Our study makes two major contributions to the existing work. First, while existing studies use one type of task, which was an oral presentation in the study of Chen

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

and Gong (2018) but a written presentation in the study of Boss et al. (2021), and find that self-selected teams outperform randomly selected teams on an assigned task (but not if the task is also self-selected, Boss et al., 2021), our study varies the tasks and reaches a more general conclusion. When we use a video-taped presentation task, we find that self-selected teams outperform randomly assigned teams. However, when the task format is written and individual contributions are not verifiable, we observe that the result does not hold as randomly assigned teams now tend to outperform self-selected teams. Second, we offer a theoretical explanation for the observed results. Self-selection (compared to random selection), if it affects skill composition and motivation of teams, can have positive or negative effects on a team's performance depending on the task's production function, specifically its skill complementarity and collaborative effort intensity.

Our paper also speaks to a broader literature on team diversity and its effects on team performance. Task-related team heterogeneity (Van Knippenberg and Mell, 2016) can enhance team effectiveness through mutual learning in that more able members transfer their skills to less able ones and by mitigating moral hazard problems through peer pressure (Hamilton et al., 2003). In contrast, aspects of demographic diversity such as functional background, gender, and race do not correlate systematically with team performance (Apesteguia et al., 2012; Hoogendoorn et al., 2013). The motivation and ability composition channels that we study in our paper enrich the potential channels through which such task-related aspects of diversity may influence performance.

Our findings also add to previous studies on how self-selection of teams affects outcomes other than team performance. Some work shows that self-selection enhances cooperation in public-goods dilemmas (Gross et al., 2016; Page et al., 2005). Self-selection may also allow dishonest people who are willing to violate rules to increase personal profit by abandoning those who are not willing to cooperate (Reuben and Stephenson, 2013). Instead, they can search for a like-minded dishonest partner who is willing to violate rules for mutual profit (Charroin et al., 2022; Gross et al., 2018). On a different note, in a framed field experiment in schools, Kiessling et al. (2021) study the question of how self-selection versus random assignment influences individual performance in a running task and find that students run faster when they are allowed to self-select a peer to run against.

The paper proceeds as follows: Sect. 2 presents a slightly formalized exposition of how random team assignment versus self-selection may affect team performance on different tasks; Sect. 3 describes the field experiment; Sect. 4 presents the results; and Sect. 5 concludes.

2 Team performance on different tasks: relative importance of team members' abilities and collaborative effort

Though our field setting did not allow us to impose a specific production function for the team tasks, and we do not intend to test a theoretical model of team performance, we use a short, slightly formalized exposition that captures the key features of our experiment to facilitate the development of our hypotheses. To illustrate

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

how the composition of the team members' abilities and the intensity of their collaboration may affect the team's performance depending on the type of task they are engaged in, we assume a hypothetical setting that involves two team tasks that vary in their production function: Two individuals, denoted as *i* and *j,* form a team. Each teammate has a uni-dimensional cognitive ability level a_i and a_j, and the team can invest collaborative effort (q).

We assume that a team's output, which determines their score, s, is given by:

$$max(a_i, a_j)^{\alpha} \cdot min(a_i, a_j)^{\beta} \cdot q^{\gamma}. \quad (1)$$

α, β, and γ represent the elasticities of the score with respect to the ability of the more able teammate, the less able teammate, and the collaborative effort, respectively. In other words, these parameters measure the responsiveness of the team's output to a change in the levels of the team members' abilities and collaborative effort. At the level of the multiplicative terms this is a standard Cobb-Douglas function with three inputs and allows us to capture the intuition that the division of labor and the level of collaboration tasks require may vary and that the productivity of one input may depend on the level of another input, i.e. that inputs are complements. At the level of the sub-terms, $max(a_i, a_j)$ and $min(a_i, a_j)$,the function assumes substitutability of the team mates' cognitive abilities a_i and a_j. Combining complementary and the substitutive features, this function helps us to differentiate between two extreme types of tasks.[3]

First, if the structure of a task requires that both team members implement a solution together, even if one team member's ability is more important in finding the solution ($\alpha > \beta$), the abilities of both team members, as well as the quality of their collaboration, matter for team performance; thus, $\alpha > 0$, $\beta > 0$ and $\gamma > 0$. Therefore, the team's score on this kind of task – i.e., a task in which the team members' abilities and collaborative efforts are complements – is determined by: $s_C = max(a_i, a_j)^{\alpha} \cdot min(a_i, a_j)^{\beta} \cdot q^{\gamma}$.

Second, if the task is best done by one person alone, the ability of the most able team member may be of paramount importance for team performance; thus, in such cases, the ability of the other team member and team collaboration may not matter. Under these assumptions, $\alpha = 1$, $\beta = 0$ and $\gamma = 0$. The team's score on this kind of task—i.e., a task in which the team members' abilities are substitutes—is thus given by: $s_{NC} = max(a_i, a_j)$.

If the score of one individual depends positively on the productivity of their teammate, there is an incentive for subjects to match with a high-ability teammate. If the matching is two-sided—i.e., if all individuals can actively search for a teammate—subjects will assortatively match by ability. This tendency results in high-ability individuals forming teams with other high-ability individuals, and low-ability individuals forming teams with other low-ability individuals. If the productivity of

[3] Additive production functions of the type $\alpha \cdot a_i + \beta \cdot a_j + \gamma \cdot q$, sometimes used in abstract lab experimental team production tasks, assume full substitutability of inputs and do not allow us to capture tasks that require positive input levels of all team members.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

one individual additionally depends on the team's collaborative efforts, there is an incentive to choose teammates who are likely to put in considerable effort. In line with this theoretical result, the empirical literature has suggested that when subjects are allowed to choose their teammates, they tend to choose teammates who have similar abilities, and with whom they are acquainted (Ai et al., 2016; Chen and Gong, 2018; Leider et al., 2009). Based on this reasoning, we would expect to find that the maximum ability is, on average, lower in self-selected teams than in randomly assigned teams, because high-ability individuals tend to cluster in some of the teams. At the same time, we would expect the levels of collaborative effort to be higher in self-selected teams, as the team members may enjoy working together more, and may thus work together more productively than team members in randomly assigned teams.

We are combining the above strands of reasoning to formulate the following hypotheses with respect to the effects of self-selection versus random assignment: We expect to observe that the performance of randomly assigned teams is, on average, better and more homogeneous than the performance of self-selected teams if they are solving a task in which the team members' abilities are substitutes, and in which collaboration is unimportant. The written task was designed to be such a task and we thus expect randomly assigned teams to perform better in it than self-selected teams. Furthermore, we expect to find that the benefit of randomly assigned teams over self-selected teams is smaller when they are performing a task in which the team members' abilities are complements, and collaboration matters. If α is sufficiently small (i.e., if the ability of the lower-ability team member, as well as the collaborative effort of both team members are sufficiently important for the team to perform well), then self-selected teams may even outperform randomly assigned teams in the latter task. The video task was designed to require high levels of inputs of both team members and strong collaboration and we thus expect self-selected teams to perform better in it than randomly assigned teams.

3 Study

3.1 Context and background

The field experiment was conducted with students of the BSc program at a well-known German business school between October 2017 and April 2019. The business school offers university education in business administration, with degrees at the BSc, MSc, MBA, and PhD levels, as well as executive education programs. The school has around 2000 students. At the BSc level, the school offers the International Business Administration program. In academic year 2017/2018, a total of 672 students were enrolled in the program, 26% of whom were female.

Studying the impact of team formation mechanisms on team performance requires an environment in which participants can choose teammates, in which the selection mechanism can be exogenously varied, and in which team performance can be objectively measured. The environment of the business school class we studied fulfilled all of these criteria, while allowing us to maintain a high degree of control.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Furthermore, to observe self-selection not only on demographic characteristics, but also on ability, we needed a sample of participants who were already acquainted with each other. This was the case for our student subjects, given that at the point in time when they were attending the class, they had already completed courses together, and had ample opportunities to get to know each other through extracurricular activities (e.g., through student societies and sports teams; and through involvement in music, drama, political campaigning, or community work) that took place at the business school.

3.2 Experimental timeline and treatments

The field experiments took place in the Microeconomics I course, with two cohorts of first-year students in the BSc program in International Business Administration participating. In each cohort, students were randomly assigned to one of two separate classes, both taught by the same instructor (one in the morning and one in the afternoon of the same day). During the first week, students learned that to fulfill the course requirements, they had to complete two tasks in teams of two and pass an exam at the end of the quarter. The instructor did not announce any task-specific details about the team tasks in the first week. Students were informed that these tasks were take-home assignments that had to be completed during study hours. Further, they knew that they would have to complete both tasks with the same team member and that re-matching was not permitted.

For each cohort, in one class—i.e., the *Self* treatment—the instructor told students on the first day to form a team with a fellow student of their choice. Students had to write down their team's composition and submit it to the instructor before the second meeting. In the other class—i.e., the *Random* treatment—students were randomly assigned to a team of two, and they were informed of their team's composition by email before the second meeting.

The first team task was assigned to the students in mid-November, and had to be completed by early December. The second team task was assigned to the students in early December, and had to be completed by the end of January. The final exam took place in March. During the course, students received no feedback on their performance on the team tasks. After the final exam, the feedback consisted only of the students' overall course grades. Upon request, students could also receive detailed information about both their team's performance on the different tasks, and their individual performance on the exam. Figure 1 displays the timeline of the experiment.

In the winter quarter of 2017/18 (Experiment I, $n = 190$, 31% female) the students completed two written team tasks. In the winter quarter of 2018/19 (Experiment II, $n = 192$, 29% female), the first task was a written task, and the second task was a video task. Across the two experiments, the first task was identical, and the students were supposed to submit their solutions in written form. By contrast, the second task differed across the two experiments, although it had a very similar content. In Experiment I, students were supposed to submit their solutions in written form; whereas in Experiment II, students were required to videotape their solution.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

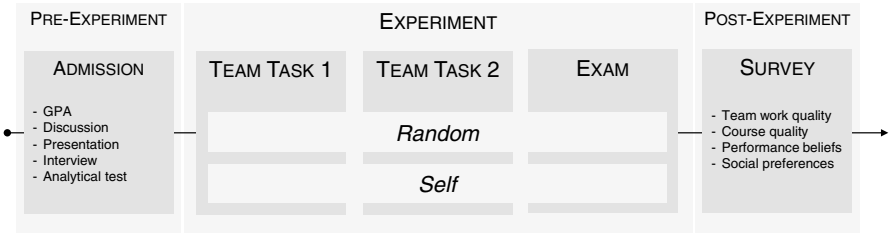


Fig. 1 Sequence of events and data sources. Figure displays the variables and the sequence of events in the experiments. The sequence of events is the same for both *Experiment I* and *Experiment II*

This design allowed us to identify interaction effects of the team formation mechanism with the task characteristics, as well as heterogeneous trends in collaboration across the treatments.

Our study design has two major advantages. First, having two written tasks in the first experiment allows us to check for differential time trends in teams' performance in the two treatments. Second, repeating the written task of the first experiment in the second experiment allows us to compare the performance of the two experimental samples on this task and to investigate the difference-in-differences in team performance in the two treatments across both experiments. This study design allows us to show that it is primarily the task difference that tends to reverse the performance between randomly assigned and self-selected teams rather than time trends or differences between samples.

3.3 Task formats

Given that we expected the effect of the team assignment mechanism on team performance to hinge on the degree to which the abilities of both team members and their levels of collaboration mattered for productivity, we aimed to design two types of task formats that required the same levels of cognitive ability but that differed in the extent to which they required inputs and collaboration from both team members. We chose to use microeconomics exercise sets that required very similar cognitive skills to complete but for which the solutions were submitted in different formats, text or video. Students had to submit their solutions to both tasks in teams of two and received a common team score for each. The students' submissions for both types of tasks were evaluated based on whether they gave correct, concise, and coherent answers to the microeconomics problems. However, the instructions for the video format contained the additional requirement that both team members and their respective contributions to the solution should be visible. Because of this stipulation, there was a high degree of complementarity between team members' skills in the video task by construction, even if the equal split in time does not imply exactly equal contributions. Genuine cooperation was needed to implement a consistent and coherent solution. In the written format, the students were required to submit a written solution and it was not verifiable which team member completed which part or whether they divided the work at all. In principle, they only had to reach an

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

agreement about which team member was better at providing a solution, in which case the input of the other team member was not needed. Thus, the written task allowed for full substitutability of team members' contributions.

The *written task* consisted of problems for which students had to submit written solutions. These problems called for the application of the theoretical knowledge that the students had acquired during lectures, such as analyzing demand patterns, calculating market outcomes, or designing pricing strategies. Providing a solution involved explaining the theoretical background, applying a correct approach to the solution, and performing a series of calculations that possibly included one or two graphs. In addition, the instructions for the written tasks specified that the students had to present their written answers clearly. The answers could be either typed or handwritten, but they had to be legible.

The *video task* consisted of questions for which students had to submit their solutions in a five-minute video. The questions required a level of microeconomics skills very similar to that required in the written task, and the solutions also consisted of explaining the theoretical background, applying a correct approach to the solution, and performing a series of calculations. The teams were allowed to use whiteboards, graphs, illustrations, and slides to make their videos more effective. In addition, the instructions specified that the video should be comprehensible; i.e., that the presenters' speech should be understandable. The instructions further stated that the teams could use their smartphones to produce the video, and that the technical quality of the video itself would not be graded. Finally, and crucially, the instructions stated that both team members, along with their individual contributions, had to be visible in the video. The lecturer explained that videos in which only one team member could be seen giving the presentation were not acceptable. All students' submissions met this criterion, and were thus evaluated for their correctness, conciseness, and coherence.[4]

3.4 Data

Data for the study were gathered from three sources (see Fig. 1). The pre-experiment data contained the students' high school performance (GPA) and their performance on the business school's admission tests. Both the GPA and the results of the admission tests were independent measures of each student's academic ability prior to the experiment, as they were not affected by their peers at the business school. Moreover, our endline data includes information on each student's performance on the two team tasks and on the final course exam at the end of the quarter. The data also included information on each student's perceptions of the cooperative behavior within their team, their relationship with their team member, their evaluation of the teacher's performance in the course, and an incentivized measure of pro-sociality collected through a post-experiment survey that was conducted after the final exam and before the students received feedback about their performance.

[4] The exercise sets can be found in the Online Appendix.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

3.4.1 Pre-experiment ability measures

Our pre-experiment ability measures came from the business school's student registry; specifically, from its admissions data.[5] The business school's program, which is known to be highly competitive, uses a selective admissions procedure. In the first step of the admissions process, applicants to the BSc program provide basic demographic information and their high school grade point average (GPA).[6] The admissions office ranks applicants by their GPA, and invites the top 10% to an admissions day, where the applicants take a written test designed to measure their analytical reasoning (quantitative) skills. They also take an oral test that has a presentation, a group discussion, and an interview component, and is intended to measure the applicants' communication, social, and problem-solving skills, and to assess whether they are a good fit for the program. The components of the oral test are each rated by two independent evaluators, whose ratings are then averaged.

For our analysis, we will use the students' *GPA* scores, the quantitative part of the admission test (henceforth called the *Analytical Test*), and an aggregate measure of the oral part of the admission test (henceforth called the *Admission Test*).[7]

3.4.2 Team outcomes and grading

Each student's team's performance on both tasks and their individual exam determined their final grade. Each team received a common grade for their performance per team task. For each task teams could achieve between 0 and 15 points. For the individual exam—written at the end of the course—a student could achieve between 0 and 70 points, thus a maximum of 100 points for the entire course.

A teaching assistant who had previous experience with the course but who was unaware that an experiment was taking place and thus blind to the treatments, graded the students' performance on both tasks. To ensure comparability between the video and the written task the correctness of the students' solutions was graded

[5] The business school had no institutional review board at the beginning of the project; thus we could not obtain a formal IRB approval. In the students' contract with the school, they consented to the anonymous processing of their data. The agreement stipulates that the university can use the administrative data for statistical and scientific purposes. Moreover, the variation was implemented with the permission of the business school's Assistant Dean of Studies and is within the normal range of changes the private business school regularly implements to improve its teaching.

[6] The German GPA (Abiturnote) ranges from 4.0 (sufficient) to 1.0 (excellent) grade and is the most important criterion for university admission in Germany (e.g., Fischer and Kampkötter, 2017). Our sample had an average GPA of 1.79 (SD = 0.504). For our analysis, we inverted the GPA so that higher values indicated better high school performance.

[7] For each student, we averaged the scores over all components and standardized them.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

on a scale between 0 and 15 for both tasks.[8] Given the highly competitive setting, there was no incentive for students to share work across groups.[9]

3.4.3 Post-experiment survey

On the day following the final exam, we invited the students to take part in an online post-experiment survey.[10] This survey elicited the students' perceptions of the quality of the collaboration in their team, of their relationship with their team member, and of the teaching. To incentivize participation, we used a raffle in which one survey participant was picked randomly to receive a 200 EUR reward. For an incentivized measure of the students' pro-sociality, we asked the students what fraction of this amount they would like to donate to UNICEF if they won.

4 Results

We begin our analysis by establishing the internal validity of our experimental approach. We show that the student sample did not differ between the treatments on any observable variables elicited before the experiments. Next, we test our modeling assumptions by investigating how the two assignment mechanisms affected team composition and cooperation. We then proceed by focusing on our primary outcome measure, and test how the two assignment mechanisms affected team performance. Finally, we investigate the relative importance of the team members' abilities for team performance.

4.1 Randomization checks

Table 1 provides an overview of the properties of our sample in the treatments and the experiments. We show separate summary statistics for Experiment I and Experiment II, and pooled statistics for both experiments. The table shows that the randomization was successful in producing highly similar groups based on observable characteristics, such as high school performance (GPA) and performance on the admission tests. The only characteristic that differed significantly between treatments in Experiment II was the percentage of female students ($p = .038$, χ^2-test, one-sided).[11] We therefore provide results from two regression specifications, both with and without controlling for gender (and other observables).

[8] As we were concerned that the ratings of the video task might suffer from low reliability due to the video format, we subjected them to a validation exercise. In this exercise, two additional independent raters rated the videos based on the same instructions as those used by the teaching assistant who made the original assessments. The additional ratings had correlations of 0.72 and 0.71 with the original rating, and a correlation of 0.81 between each other. Thus, the reliability of the presentation ratings can be considered satisfactory. All results were found to be robust to using these additional ratings.

[9] In addition to having a highly competitive admission process, the school also has a competitive grading system, where the average grade in each course has to be within a narrow corridor, which disincentivizes information sharing and cooperation across groups as it worsens own grades at given performance levels. If a lecturer notices that answers of several groups are suspiciously similar, they have to investigate the case for cheating and plagiarism, which, if confirmed, are punished with failed grades and, if repeated, exclusion from the program.

[10] The survey was accessible until just before the exam grades were published; which usually takes up to 6 weeks.

[11] Unless otherwise stated, all p-values are based on two-sided tests.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Table 1 Randomization checks

Variable	Experiment I			Experiment II			Experiment I + II		
	Self	Random	*p* value	Self	Random	*p* value	Self	Random	*p* value
GPA	0.056	−0.054	0.195	0.074	−0.084	0.523	0.065	−0.069	0.134
Analytical test	0.016	−0.016	0.889	−0.056	0.064	0.429	−0.021	0.023	0.668
Admission test	−0.037	0.037	0.480	0.037	−0.042	0.592	0.001	−0.001	0.864
% female	0.287	0.323	0.593	0.356	0.220	0.038	0.323	0.273	0.282

Descriptive statistics of pre-experiment data, admission test scores. GPA is inverted and z-standardized, with a higher GPA indicating better school performance. The Analytical test, and the Admission test are z-standardized. The *p* values are from a two-sided Mann-Whitney U (MWU) test comparing differences in mean ranks between the two treatments. The *p* values for the comparison of % female are from a one-sided χ^2-test. Unstandardized values (Table A.1) and a correlation matrix (Table A.3) appear in the Online Appendix

4.2 Team formation

To test our modeling assumptions, we first investigate how the two assignment mechanisms affected team ability composition and cooperation.

4.2.1 Ability composition

In this subsection, we investigate how allowing team members to self-select affected the team composition. We begin by looking at how students (in the *Self* treatment) formed teams. To do so, we used pre-experiment registry data on each student's ability (measured as their performance on the various tasks in the admission test and their GPA), and gender. For each team and measure *m*, we calculate the absolute difference between both teammates

$$m_{ij} = |x_i - x_j|$$

where *i* and *j* indicate the teammates.

Thus, lower absolute differences indicate that the teammates were more similar, and higher values indicate that they were less similar. If the students in the *Self* treatment were matched on certain measures, we would observe a higher degree of similarity, i.e., a lower average absolute difference, than in the *Random* treatment. Moreover, as a reference point, we calculated the average absolute difference after simulating the matching of each student with all potential teammates from the respective treatment. This simulation provided us with information about what a hypothetical within-sample random team composition might look like.

The results appear in Table 2. The first column shows the absolute differences for all measures in *Self*, while the second column shows the absolute differences for the simulated *Random* "treatment" and the third column for the actual *Random* treatment. A comparison of the values suggests that the students sorted themselves into teams with students of similar levels of ability and of the same gender. More specifically, we observed that the self-selected teams were more similar in terms of their

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Table 2 Self selection and composition of teams

Variable	Observed		Simulation
	Self	Random	Random
GPA	0.977	1.092	1.118***
% Female	0.204	0.348	0.409**
Analytical test	1.012	1.150	1.173***
Admission test	1.235	1.070	1.096

The table displays the average absolute differences between teammates on the pre-experiment observables. Simulation *Random* denotes the average absolute difference for the respective variable from a simulation in which we pairwise matched all students within a treatment within an experiment. Significance indicators: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

GPAs, their scores on the written admission test, and their gender. These differences were significant (GPA: $p = .0044$; Analytical test: $p = .0155$; Female: $p = .0075$, MWU test). Interestingly, we did not find significant differences between the *Self* or *Random* treatments for the oral parts of the admission test.[12]

4.2.2 Perceived cooperation

A second mechanism that we hypothesized to be affected by the treatment variation and to mediate team performance was the quality of cooperation. In our post-experiment survey, we asked students to evaluate their collaboration experience in their team during the course (see Table 3 for an overview of all of the questions).

We asked students to agree or disagree (on a 7-point Likert scale) with several statements aimed at capturing various aspects of team collaboration and organization. More specifically, we asked questions about the perceived quality of the cooperation and the pleasure of working together.[13]

Table 3 displays the results from the post-experiment survey, pooled for Experiments I and II, and for each experiment separately. When asked about their experience during the task, students in *Self* reported that they communicated more ("We communicated a lot"; $p < .0001$, MWU test) and that they cooperated better ("We helped each other a lot"; $p = .0188$) than the students in *Random*. Moreover, they indicated that the teammates' contributions were more equally distributed ("Both team members contributed equally"; $p = .021$), and that both teammates exerted effort ("Both team members exerted effort"; $p = .002$). These comparisons clearly

[12] The Admission test score also did not correlate with the performance on the different team tasks (see Table A.3). In principle, it is possible that we were lucky in the team composition in *Random* For this reason, we show the results of the simulation in the second column of the table. Comparing *Self* with the results of our simulation yielded similar results.

[13] We also ask a battery of questions about the perceived teaching quality, which might have influenced performance. However, we found no significant differences between the treatments and experiments, which suggests that the lecturer's teaching was of the same quality in both classes and experiments.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Table 3 Overview of survey items and survey results

Survey item	Experiment I			Experiment II			Experiment I + II		
	Random		Self	Random		Self	Random		Self
A. Perceived quality of cooperation (1=Not agree, 7=Completely agree)									
We communicated a lot	5.16	$<^{*}$	6.08	5.58	$<^{**}$	6.14	5.35	$<^{***}$	6.11
We helped each other a lot	5.41	$<^{**}$	5.95	5.82	$<$	6.07	5.60	$<^{**}$	6.01
Both team members exerted effort	5.46	$<^{**}$	6.07	5.93	$<^{*}$	6.34	5.67	$<^{***}$	6.20
Both team members contributed equally	5.12	$<^{*}$	5.59	5.44	$<$	5.87	5.26	$<^{**}$	5.73
Our individual skills complemented very well	4.99	$<^{**}$	5.56	5.26	$<$	5.52	5.11	$<^{**}$	5.54
Our team was very stressed	2.87	$<^{*}$	3.30	2.53	$<$	2.69	2.71	$<$	3.00
Our team was very motivated	5.53	$<$	5.78	5.70	$<$	5.85	5.61	$<$	5.81
The mood in our team was good	5.79	$<^{*}$	6.10	6.19	$<$	6.27	5.98	$<$	6.18
The coordination of our team was very good	5.03	$<^{*}$	5.38	5.56	$<$	5.59	5.27	$<$	5.49
I was dominant in leading the team	4.49	$>$	4.22	4.30	$>$	4.21	4.40	$>$	4.22
One person was dominant in leading the team	4.10	$>$	3.84	4.14	$>$	3.85	4.12	$>$	3.84
B. Attitude towards the other (1=Not agree, 7=Completely agree)									
My team member is a friend	3.82	$<^{***}$	6.25	4.33	$<^{***}$	6.06	4.06	$<^{***}$	6.15
I knew the team member very well before the course	2.60	$<^{***}$	6.19	2.93	$<^{***}$	5.66	2.75	$<^{***}$	5.93
C. Perceived teaching quality (1=Not agree, 7=Completely agree)									
I learned a lot from the professor to complete the exercises	5.57	$<$	5.66	5.77	$>$	5.34	5.66	$>$	5.50
The professor asked questions to test our understanding	5.72	$<$	5.74	5.88	$>$	5.55	5.79	$>$	5.65
Professor was too fast in explaining the contents	2.41	$<$	2.62	2.58	$<$	2.86	2.49	$<$	2.74
The lecturer spend too much time on simple things	3.28	$<$	3.30	3.16	$<$	3.38	3.22	$<$	3.34
The professor gave too complicated answers	2.24	$<$	2.23	2.18	$<$	2.62	2.21	$<$	2.42
D. Social preferences									
Donation to UNICEF (in €)	100.12	$>$	97.59	96.21	$<$	97.46	98.34	$>$	97.53
Observations	68		73	57		71	125		144

Table reports descriptive statistics of student responses in the post-experimental survey. p-values stem from a two-sided MWU test for a comparison of averages between *Self* and *Random*. Significance indicators: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

show that the teams in *Random* used a different approach to solving the problem sets than the teams in *Self*, who are possibly assigning more work to the more able teammate but who are also cooperating less.[14]

Furthermore, we found that the students' moods ("The mood in our team was good"; $p = 0.189$), levels of stress ("Our team was very stressed."; $p = 0.134$) and motivation levels ("Our team was very motivated"; $p = 0.151$) for the teams in *Self* did not differ from those for the teams in *Random*.

Although the students in *Self* were more likely to report being friends ("My team member was a friend"; $p < 0.0001$) or having been acquainted with their teammate before the course ("I knew the team member very well before the course"; $p < 0.0001$), it was not the overall pleasure of working together, but rather the higher level of cooperation, that was different between the teams in the two treatments.[15]

4.3 Team performance

Our primary outcome measure, team performance, is the score that the teams received for their work on two separate tasks during the quarter. We summarize our results in Fig. 2, which plots the standardized average team score for each task by treatment, and also shows individual exam performance. The left panel shows the outcomes for Experiment I, while the right panel shows the outcomes for Experiment II.

For Experiment I, in which the solutions to the first and the second team tasks had to be submitted in written form, the figure indicates that, on average, the teams in *Random* performed better than the teams in *Self*. A non-parametric comparison of average team scores yielded a significantly lower score for the teams in *Self* than for the teams in *Random* ($p = 0.007$,MWU test). The results of a non-parametric test for the equality of variances between the treatments underlined this pattern, and showed that the variance of team performance was significantly larger in the *Self* treatment ($p = 0.002$, Levene's test).[16] We also observed no change in performance over time. A comparison of the average performance

[14] As Table 3 shows, these differences between the treatments were mostly driven by the reports of students from Experiment I. As the video clip in Experiment II required each teammate to cooperate and to appear in the video to present the results, the students might have tried to fulfill this expectation. Therefore, a desirability bias might explain why we did not find as strong a difference in self-reported cooperation in Experiment II as we did in Experiment I. Our finding that the average ratings of cooperation also tended to be higher in *Self* in Experiment II than in Experiment I points in the same direction.

[15] 70% (Experiment I: 74%, Experiment II: 67%) of the students responded to our request to participate in the survey. We tested and found no significant difference in the fraction of participating students between the *Random* and the *Self* treatment (Experiment I: $p = 0.282$, Experiment II: $p = 0.261$, χ^2 test). Furthermore, participation in the survey was balanced in terms of GPA ($p = 0.146$, $p = 0.466$, MWU test), the analytical test scores ($p = 0.091$, $p = 0.334$, MWU test), and gender ($p = 0.730$, $p = 0.822$, χ^2 test).

[16] A separate analysis of the first and the second team task yielded similar significant differences in averages (1st team task: $p = 0.011$, 2^{nd} team task: $p = 0.068$, MWU test), and (marginally) significant differences in variances (1st team task: $p = 0.104$, 2nd team task: $p = 0.001$, Levene's test). A detailed pairwise comparison appears in Table A.2 in the Online Appendix.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

between the first and the second team tasks revealed no significant differences (*Self*: $p = 0.885$, *Random*: $p = 0.9291$, MWU test).

First, for Experiment II, the figure indicates that the teams in *Self* performed worse on the written task than those in *Random*, while the effect appears to have reversed when the teams were working on the video task. Indeed, in the first team task of Experiment II, we replicated the observed pattern of Experiment I. The teams in *Self* performed marginally significantly worse than those in *Random* when the task was written ($p = 0.064$, MWU test), but this time the variances were not significantly different ($p = 0.193$, Levene's test). Figure 2 appears to show that the average team performance was higher in *Self* than in *Random* in the second team task (the video task). However, the results of non-parametric tests comparing the mean and the variance of average team performance between *Self* and *Random* did not reject the null hypothesis that the performance in both treatments was equal ($p = 0.156$, MWU test; $p = 0.381$, Levene's test). This time, however, we observed a large change in performance between the two tasks. The performance of the self-selected teams was marginally significantly better on the video task than on the written task ($p = 0.0790$, Wilcoxon Signed Rank test; hereafter, WSR test), while the performance of the randomly assigned teams was marginally significantly worse on the video task than on the written task ($p = 0.0556$).

More evidence for this change in behavior across types was provided by a difference-in-difference analysis. Calculating the difference between the first and the second team task for both experiments and comparing them between treatments yielded a significant difference for Experiment II ($p = 0.0086$, MWU test), but no significant difference for Experiment I ($p = 0.9483$).

Furthermore, the figure also shows that the exam performance was unaffected by the treatment. Neither the average student performance nor the variance of the student performance on the final exam differed significantly across treatments (Experiment I: $p = .455$, MWU test; $p = .995$, Levene's test; Experiment II: $p = .984$, MWU test, $p = .603$, Levene's test). This finding indicates that the team assignment mechanism did not have a spillover effect on exam performance. It can also be seen as evidence that the effectiveness of teaching did not differ between the two treatment groups, and, therefore, that the lecturer's behavior was unlikely to have influenced the different levels of team performance.[17]

Second, we ran regressions controlling for pre-experiment observables to verify these observations. Table 4 shows the results of OLS regressions with standard errors clustered at the team level, where the the dependent variable is the team performance (z-standardized) for the both team tasks, separately. In models (1)–(3), we predicted the team performance on the first team task. Model (1) included only

[17] To check whether students perceived the quality of the teaching differently between the two treatments, we included four items in our post-experimental survey. We did not observe a significant difference between the two experimental conditions for any of these questions. This finding supports our assumption that the teacher had no influence on the study results. In Panel C of Table 3, we display the respective items and results.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

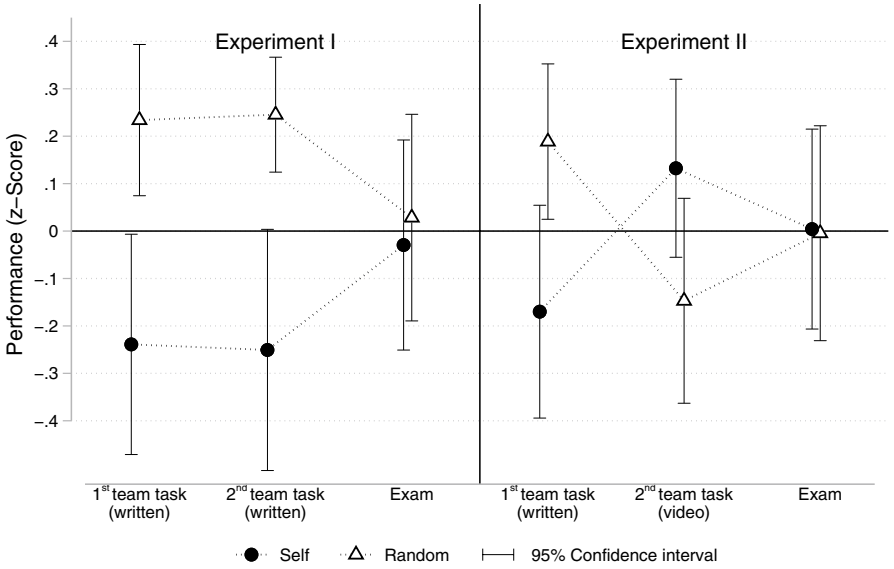


Fig. 2 Team assignment, performance, and task characteristics. Figure shows the average team performance (z-standardized) for the tasks in our experiments. The left panel shows the results from Experiment I, while the right panel shows the results from Experiment II

a dummy variable for the *Self* treatment ("1 if *Self*"). The self-selected teams performed, on average,.415 ($p = .004$; CI[− 0.694; − 0.136]) standard deviations worse on the first task than the randomly assigned teams. Model (2) included a dummy variable for the experiment ("1 if Experiment II") and an interaction term of the *Self* treatment and the experiment ("1 if *Self* x Experiment II") to control for potential interactions. While both of these control variables remained insignificant, the coefficient on the treatment dummy *Self* remained significant and almost unchanged at -.473 ($p = .02$; CI[− .870; − 0.077]), which indicates that for the first task, the treatment effect was not significantly different across the experiments. In model (3), we included additional controls, and found that the treatment effect was not affected by their inclusion ($\beta = -0.478$, $p = 0.018$; CI[− 0.875; − 0.082]). Interestingly, we found that the students' GPAs, but not their admission test scores, predicted the team performance.

Next, we studied the second team task. The regression results appear in models (4)–(6). In model (4), we pooled observations from both experiments (ignoring the type of task), and included only a treatment dummy. Consistent with the results of the non-parametric analysis, we found no significant effect of self-selection, which suggests that a meaningful investigation of the effects of the team assignment process on performance should take into account the task characteristics. After we controlled for the experiment and interacted with the treatment, we found that the teams in the *Self* treatment in Experiment I performed.496 ($p = 0.014$.; CI[− 0.892; − 0.100]) standard deviations worse on the second task than the teams in the *Random* treatment (model 5). We thus found very similar treatment effects for the first and the second tasks in

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Experiment I, which suggests that there was no heterogeneous learning across treatments, and that the ordering of the tasks did not matter. In addition, after adding up the first and the third coefficients in model (5), we found that in Experiment II, the teams in *Self* tended to perform.279 standard deviations better than the teams in *Random on the second task (video task)*. In line with the non-parametric analysis, a joint F-test showed that this difference was not significant ($p = 0.1774$). Adding additional control variables did not significantly change the coefficients. Interestingly, we again found that GPA positively predicted the performance on the second team task.

4.3.1 Heterogeneity analysis

The results from our survey suggest that self-selected teams cooperated more than randomly assigned teams. While this additional cooperation might just have been inefficient for teams with high-ability students, this could have created an additional negative effect for teams with low-ability students. Therefore, we explore the question do students of varying abilities derive differing benefits from the self-selection process? To explore this question we split the sample at the median high school GPA. We find (in Table 5) that the *Self* treatment had significant negative effects on team performance on the first written task (model 1) for both low-ability (β = -.491,p=.040; CI [-0.979; -0.026]) and high ability students (β = -.445, p=0.001; CI [-0.669; -0.169]). These effects were not significantly different from each other. Furthermore, the *Self* treatment had a significantly negative effect on the team performance of high-ability (β = -0.371, p =.067; CI [-0.749; 0.05]) and low-ability (β = -0.782, $p = 0.010$; CI [−1.377; -0.0198]) students on the second written task (model 2, Experiment I only). These effects were ot significantly different from each other. The *Self* treatment had no differential effect on the team performance of high-ability ($\beta = 0.297$, $p = .174$; CI[− 0.141;.73]) and low ability ($\beta = -0.006$, $p = 0.984$; CI[− 0.564;.573]). Overall, the results of this heterogeneity analysis suggest that allowing for self-selection into groups harmed the performance of both low- and high-ability students on the written task.

4.3.2 Testing the relative importance of abilities

Up to now, we have made the following observations: First, we showed that the skill composition and cooperation of the teams differed between *Random* and *Self*; i.e., that in the latter treatment, students tended to choose a partner with similar skills and reported higher levels of cooperation. Second, we showed that the teams in *Random* performed better on the written task, for which less skill complementarity was needed; and that the team performance on the video task, for which more skill complementarity was needed, did not differ between *Random* and *Self*.

To better understand the role of individual skills in the team tasks, we now focus on the relationship between skills and team performance in *Random.* We concentrate our analysis on this treatment, since we can be sure that in *Random*, the composition of the team members' abilities is exogenous and is not confounded by other factors of the team member selection process, unlike in *Self*. We operationalized each student's ability with their exam score, as this measure captures abilities immediately

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Table 4 Regressions for team performance

Independent variables	Dependent variable:					
	Performance on 1st team task			Performance on 2nd team task		
	(Exp. I and II: written)			(Exp. I: written, II: video)		
	(1)	(2)	(3)	(4)	(5)	(6)
1 if *Self*	−.415***	−.473**	−0.478**	−0.107	−0.496**	−0.521***
	(0.141)	(0.201)	(0.201)	(0.145)	(0.201)	(0.199)
1 if Experiment II		−0.045	0.031		−0.392**	−0.320*
		(0.163)	(0.149)		(0.177)	(0.163)
Self x Experiment II		0.114	0.029		0.775***	0.728**
		(0.283)	(0.276)		(0.288)	(0.286)
Controls						
GPA			0.117**			0.123**
			(0.057)			(0.061)
1 if female			−0.024			−0.250
			(0.119)			(.154)
Admission Test			0.006			−.080
			(.052)			(.053)
Constant	0.212***	0.234**	0.240**	0.054	0.245***	0.332***
	(0.081)	(0.114)	(0.121)	(0.089)	(0.086)	(0.107)
Observations	382	382	377	382	382	377
R-squared	0.043	0.044	0.067	0.003	0.041	0.067

Note: Columns (1)–(3) show OLS regressions of z-standardized team performance on the first task. In both experiments, the students had to submit a written solution to the task. Columns (4)–(6) show OLS regressions of z-standardized team performance on the second task. In Experiment I, the students had to submit a written solution to the task; while in Experiment II, the students had to submit a video clip. The control variables are GPA, admission test scores, and gender. GPA and Admission test have been z-standardized. Standard errors clustered on teams are in parentheses. As a robustness check, we ran the same regressions with team level average control variables (see Table A.4 in the Online Appendix) yielding similar results. Significance indicators: $^{***}\ p < 0.01$, $^{**}\ p < 0.05$, $^{*}\ p < 0.1$

relevant to the team tasks and is not affected by the treatment assignment.[18] To measure the students perceived collaborative effort we ran a principal-component factor analysis (PCA). Including all items from the post-experimental survey a PCA retained one factor (Eigenvalue 5.78) that we call "collaborative effort". Since we predict performance on the team-level we take the average perceived collaborative effort for each team as our independent variable. Results are displayed in 6. To estimate the Cobb-Douglas production function (1) in section 2 we linearize the function by log-transforming the performance on the team tasks as well as the ability and cooperation measures.

[18] The students' GPAs or scores on the analytical test contained more noise, but yielded qualitatively similar results.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Table 5 Heterogeneity analysis

Independent variables	Dependent variable: Performance on...		
	... 1st team task	... 2nd team task	
	(Exp. I + II: written)	(Exp. I: written)	(Exp. II: video)
	(1)	(2)	(3)
1 if *Self*	−0.491**	−0.782***	−0.006
	(0.238)	(0.296)	(0.293)
1 if GPA > Median	0.156	0.037	0.037
	(0.163)	(0.283)	(0.265)
Self x (GPA > Median)	0.046	0.411	0.303
	(0.237)	(0.268)	(0.313)
Constant	0.160	0.348**	−0.048
	(.142)	(0.152)	(0.207)
Self + *Self* x (GPA > Median)	−0.445***	−0.371*	0.297
	(0.133)	(0.200)	(0.217)
Observations	377	189	188
R-squared	0.070	0.130	0.043

This table displays the result of a OLS regression analysis (robust standard errors clustered on the team level in parentheses). All specifications include GPA, Admission test, and female as control variables. All scores have been z-standardized. Significance indicators: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

For the written task (models 1–3), we observe that—if anything—the maximum ability tends to positively influence the team outcome. For the same models, the coefficients for the minimum ability are close to zero and tend to be negative. For the video task (model 5), we observe that both coefficients tend to be positive.[19] We observe in models (2), (4) and (6), that the coefficients for collaborative effort are very small.

While the signs of the coefficients on $\max(a_i;a_j)$ and $\min(a_i;a_j)$ are consistent with our line of reasoning, both variables lack statistical significance, and the predictive power of the model is low. Given the small sample size, we cautiously interpret these results as being mildly suggestive of differences in the relationship between the composition of the team members' abilities and the team performance across the two tasks.

[19] In Figure A.2, we display the linear relationships between the team performance and individual abilities. The black line shows the relationship for the team member with the highest ability ($\max(a_i, a_j)$), and the gray line shows the relationship for the team member with the lowest ability ($\min(a_i, a_j)$). In line with our reasoning above, we observed a positive relationship between the team performance and the maximum ability for the task that required low levels of skill complementarity. It appears that the minimum ability had no impact on the teams' outcomes. For the video task, in which higher levels of skill complementarity were required, both the maximum and the minimum ability had a positive impact on the team performance.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

5 Conclusion

This paper has provided evidence from field experiments that studied how team formation processes influenced team performance. We used data on students' individual characteristics and behavior at a business school to examine the effects on team performance of varying both the team formation process and the task characteristics. The results of our randomized field experiments add a new dimension to the debate on the effects of the team formation process on team performance. Previous experiments did not use objective ability measures to capture team formation patterns, and they did not offer an explanation for the observed effects of the team formation process on the team members' performance on different tasks. By contrast, we used data on student ability generated prior to the experiments to study how the team formation process affected the teams' abilities and social composition, which, in turn, affected the teams' cooperation and performance on team tasks with different skill complementarities.

We found that the team formation mechanism significantly affected team composition, cooperation, and performance.[20] When subjects were allowed to choose their teammate, they matched with someone of similar abilities and reported higher levels of cooperation. This selection pattern proved to be performance-enhancing when the underlying task required a high degree of skill complementarity an cooperation. In contrast, the random assignment of teammates tended to improve team performance when the task required little or no skill complementarity and cooperation. After the students completed the team tasks, we measured the individual performance of the subjects, and found no differences between the team formation mechanisms, which indicates that the effect observed at the team level did not translate into individual performance differences.

Spillover and interference are common problems in field experiments. In principle, students in our experiment could talk to each other about the tasks and potentially share their answers with friends across teams and not just collaborate with their own team mate. The institutional setting within which our experiments are embedded, however, is one that discourages cooperation across teams because the business school's grading method enhances competitive behavior. Study regulations require the average grade to be within a narrow corridor to limit grade inflation, which disincentivizes cooperation across groups as it worsens own grades at given own performance levels. Furthermore, examination regulations consider direct information sharing and other forms of explicit cooperation as cheating and/or plagiarism which leads to a failing grade and possibly to more severe consequences, including a failing grade in the overall course module and disciplinary action. No cheating and/or plagiarism was detected during the experiments. Nevertheless, we cannot fully exclude the possibility that a subset of students might have shared information.

[20] In this study, we focused on the contrast between self-selection and random assignment. An alternative approach would be to assign subjects based on algorithms that maximize team performance (e.g., Wei et al., 2023). For tasks with low collaboration intensity, this could be an algorithm that maximizes the differences in the team members' abilities.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Table 6 Team performance, individual abilities, and collaborative effort in *Random*

Independent variable	Dependent variable: Log Performance on...					
	...1st team task		...2nd team task			
	(1)	(2)	(3)	(4)	(5)	(6)
$\max(a_i;a_j)$	0.164	0.106	0.116	0.029	0.359	0.350
	(0.187)	(0.189)	(0.161)	(0.188)	(0.294)	(0.300)
$\min(a_i;a_j)$	−0.072	−0.043	−0.025	0.069	.108	.061
	(0.070)	(0.065)	(0.054)	(0.066)	(0.137)	(0.149)
Collaborative effort		0.028		0.091		−0.024
		(0.025)		(0.074)		(0.039)
Constant	2.100***	2.112***	2.110***	1.898***	1.205*	1.385**
	(0.431)	(0.467)	(0.364)	(0.337)	(0.655)	(0.670)
Observations	92	82	48	43	44	39
Experiment	I+II	I+II	I	I	II	II
Task format	written	written	written	written	video	video
R^2	0.020	0.039	0.009	0.129	0.084	0.083

Note: The table displays regression coefficients (Standard errors are in parentheses) of OLS regressions. Models in columns (1) and (2) predict the log transformed team performance on the first task across both experiments in *Random*. Both models include a dummy variable for the experiments, which was insignificant ($\beta = [-.0000188;0.006]$) and is not displayed. Columns (3) to (6) show OLS regressions for the second task. Significance indicators: $^{***}\ p < 0.01$, $^{**}\ p < 0.05$, $^{*}\ p < 0.1$

Our study offers insights for managers and team leaders; i.e., for individuals who decide how teams are put together in firms and other organizations. It suggests that if managers want to maximize team performance, they first need to consider the type of task involved before deciding whether employees should be able to self-select their team mates. Given that randomly assigned teams can produce superior outcomes for tasks that are characterized by low collaboration intensity, our findings also reveal a weakness in the trends towards more "agile work practices" (e.g., Mamoli and Mole, 2015), which give employees the freedom to choose their working groups regardless of the circumstances.

Moreover, our results provide insights into the trade-off between diversity and employee autonomy. When managers want to create a more inclusive work environment by forming more diverse teams, random team assignment might prove more beneficial. Our field experiment showed that students are more likely to match with teammates of the same gender when they are allowed to self-select. This finding suggests that self-selection might create not just inequalities in abilities across teams, but also less gender-diverse teams.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10683-023-09800-2.

Funding Open Access funding enabled and organized by Projekt DEAL.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ai, W., Chen, R., Chen, Y., Mei, Q., & Phillips, W. (2016). Recommending teams promotes prosocial lending in online microfinance. *Proceedings of the National Academy of Sciences, 113*(52), 14944–14948.

Apesteguia, J., Azmat, G., & Iriberri, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science, 58*(1), 78–93.

Bandiera, O., Barankay, I., & Rasul, I. (2013). Team incentives: Evidence from a firm level experiment. *Journal of the European Economic Association, 11*(5), 1079–1114.

Becker, G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy, 81*(4), 813–846.

Boss, V., Dahlander, L., Ihl, C., & Jayaraman, R. (2021). *Organizing entrepreneurial teams: A field experiment on autonomy over choosing teams and ideas*. Organization Science.

Büyükboyaci, M., & Robbett, A. (2019). Team formation with complementary skills. *Journal of Economics & Management Strategy, 28*(4), 713–733.

Charroin, L., Fortin, B., & Villeval, M. C. (2022). Peer effects, self-selection and dishonesty. *Journal of Economic Behavior & Organization, 200*, 618–637.

Chen, R. (2017). Coordination with endogenous groups. *Journal of Economic Behavior & Organization, 141*(5), 177–187.

Chen, R., & Gong, J. (2018). Can self selection create high-performing teams? *Journal of Economic Behavior & Organization, 148*, 20–33.

Cooper, D. J., Saral, K., & Villeval, M. C. (2021). Why join a team? *Management Science, 67*(11), 6980–6997.

Cross, R., Rebele, R., & Grant, A. (2016). *Collaborative overload*. Harvard Business Review.

Curranrini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica, 77*(4), 1003–1045.

Delfgaauw, J., Dur, R., Onemu, O., & Sol, J. (2022). Team incentives, social cohesion, and performance: A natural field experiment. *Management Science, 68*(1), 230–256.

Delfgaauw, J., Dur, R., & Souverijn, M. (2020). Team incentives, task assignment, and performance: A field experiment. *The Leadership Quarterly, 31*(3), 101241.

Englmaier, F., Grimm, S., Schindler, D. & Schudy, S. (2018). *The effect of incentives in non-routine analytical team tasks-evidence from a field experiment*. CESifo Working Paper Series (6903).

Erev, I., Bornstein, G., & Galili, R. (1993). Constructive intergroup competition as a solution to the free rider problem: A field experiment. *Journal of Experimental Social Psychology, 29*(6), 463–478.

Fischer, M., & Kampkötter, P. (2017). Effects of German universities' excellence initiative on ability sorting of students and perceptions of educational quality. *Journal of Institutional and Theoretical Economics, 173*(4), 662.

Friebel, G., Heinz, M., Krüger, M., & Zubanov, N. (2017). Team incentives and performance: Evidence from a retail chain. *American Economic Review, 107*(8), 2168–2203.

Gächter, S., & Thöni, C. (2005). Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association, 3*(2), 303–314.

Geraghty, A., & Paterson-Brown, S. (2018). Leadership and working in teams. *Surgery (Oxford), 36*(9), 503–508.

Gross, J., Leib, M., Offerman, T., & Shalvi, S. (2018). Ethical free riding: When honest people find dishonest partners. *Psychological Science, 29*(12), 1956–1968.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Gross, J., Méder, Z. Z., Okamoto-Barth, S., & Riedl, A. (2016). Building the Leviathan–Voluntary centralisation of punishment power sustains cooperation in humans. *Scientific Reports, 6*(1), 1–9.
Guido, A., Robbett, A., & Romaniuc, R. (2019). Group formation and cooperation in social dilemmas: A survey and meta-analytic evidence. *Journal of Economic Behavior & Organization, 159*, 192–209.
Hamilton, B. H., Nickerson, J. A., & Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy, 111*(3), 465–497.
Hoegl, M., & Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science, 12*(4), 435–449.
Hoogendoorn, S., Oosterbeek, H., & Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science, 59*(7), 1514–1528.
Kiessling, L., Radbruch, J., & Schaube, S. (2021). Self-selection of peers and performance. *Management Science, 68*(11), 8184–8201.
Lazear, E. P., & Oyer, P. (2012). Chapter 12: Personnel economics. In R. Gibbons & J. Roberts (Eds.), *The handbook of organizational economics* (pp. 479–519). Princeton University Press.
Leider, S., Möbius, M. M., Rosenblat, T., & Do, Q.-A. (2009). Directed altruism and enforced reciprocity in social networks. *Quarterly Journal of Economics, 124*(4), 1815–1851.
Mamoli, S., & Mole, D. (2015). *Creating great teams: How self-selection lets people excel*. Pragmatic Bookshelf.
O'Neill, T. A., & Salas, E. (2018). Creating high performance teamwork in organizations. *Human Resource Management Review, 28*(4), 325–331.
Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency. *The Economic Journal, 115*(506), 1032–1053.
Patel, S., & Sarkissian, S. (2017). To group or not to group? Evidence from mutual fund databases. *Journal of Financial and Quantitative Analysis, 52*(5), 1989–2021.
Reagans, R., & Zuckerman, E. W. (2019). Networks, diversity, and productivity: The social capital of corporate R &D teams. *Organization Science, 12*(4), 502–517.
Reuben, E., & Stephenson, M. (2013). Nobody likes a rat: On the willingness to report lies and the consequences thereof. *Journal of Economic Behavior & Organization, 93*, 384–391.
Siow, A. (2015). Testing Becker's theory of positive assortative matching. *Journal of Labor Economics, 33*(2), 409–441.
Sivasubramaniam, N., Liebowitz, S. J., & Lackman, C. L. (2012). Determinants of new product development team performance: A meta-analytic review. *Journal of Product Innovation Management, 29*(5), 803–820.
Stewart, G. L., & Barrick, M. R. (2000). Team structure and performance: Assessing the mediating role of intrateam process and the moderating role of task type. *Academy of Management Journal, 43*(2), 135–148.
Tabrizi, B. (2015). *75 % of cross-functional teams are dysfunctional*. Harvard Business Review.
Van Knippenberg, D., & Mell, J. N. (2016). Past, present, and potential future of team diversity research: From compositional diversity to emergent diversity. *Organizational Behavior and Human Decision Processes, 136*, 135–145.
Wei, A., Chen, Y., Mei, Q., Ye, J., & Zhang, L. (2023). Putting teams into the gig economy: A field experiment at a ride-sharing platform. *Management Science*. https://doi.org/10.1287/mnsc.2022.4624
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036–1039.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press

Authors and Affiliations

Mira Fischer[1] · Rainer Michael Rilke[2] · B. Burcin Yurtoglu[3]

Mira Fischer
mira.fischer@wzb.eu

B. Burcin Yurtoglu
burcin.yurtoglu@whu.edu

[1] WZB Berlin Social Science Center, Reichpietschufer 50, 10115 Berlin, Germany

[2] Economics Group, WHU - Otto Beisheim School of Management, Burgplatz 2, 56176 Vallendar, Germany

[3] Finance Group, WHU - Otto Beisheim School of Management, Burgplatz 2, 56176 Vallendar, Germany

https://doi.org/10.1007/s10683-023-09800-2 Published online by Cambridge University Press