Group Formation and Performance: Field Experimental Evidence

Roy Chen and Jie Gong*

March 2016

Abstract

Given the prevalence of teamwork in organizations, it is important to understand how group formation impacts team productivity. To address this question, we experimentally examine the effect of group formation on performance by randomly assigning students in an undergraduate business course to one of three treatments: (1) groups are assigned randomly; (2) groups are assigned to maximize skill complementarity; or (3) groups are determined by the students. Our results show that, when given the choice, students form groups based on their social networks. We further find that these groups perform better than randomly-assigned groups and about as well as those with maximized skill complementarity. We also find some evidence that this performance difference is due to the higher effort expended by students in self-selected groups.

*Chen: Department of Economics, Faculty of Arts and Social Sciences, National University of Singapore, AS2 #06-02, 1 Arts Link, Singapore 117570, ecscwr@nus.edu.sg. Gong: Department of Strategy and Policy, NUS Business School, National University of Singapore, 15 Kent Ridge Drive, Singapore, 119245, gong@nus.edu.sg. We would like to thank Yan Chen, Nancy Kotzian, Jin Li, Tracy Liu, Ivan Png, and seminar participants at the National University of Singapore, and the 2015 Economic Science Association World Meetings, Sydney for helpful discussions and comments, and Xuyuan Liu and Bo Shen for research assistance. The financial support from the National University of Singapore is gratefully acknowledged.
1 Introduction

Teamwork is pervasive in today’s workplaces. Indeed, Lazear and Shaw (2007) report that over 80 percent of firms use teams as part of their organizational approach, and that over 70 percent of firms use self-directed teams. Inherent in this team focus is the belief that teams can produce results that exceed the total capabilities of individual members. Given the importance placed on teams as a means to meet firm goals, firms must think carefully about how such teams are formed in order to accomplish the desired tasks.

One key question in designing work teams is how group membership is determined. On the one hand, the firm better understands the technology of team production. On the other hand, allowing employees to form their own teams lowers the firm costs involved in collecting and processing information about team members. However, despite the cost savings from decentralizing the decision, it is unclear whether doing so improves team productivity. It is possible that self-chosen teams may see a decrease in team productivity, as workers may have team composition incentives that are not perfectly aligned with those of the firm.

However, it is also plausible that self-chosen teams would increase productivity. For example, if workers choose team members with whom they are socially connected, they may be better able to mitigate any free riding behavior. Evidence from the social identity literature in economics suggests that allowing employees to form their own groups may enhance their sense of attachment to their groups (Chen, 2016), and people are less likely to shirk their group responsibilities when they feel more attached to a group (see, for example, Akerlof and Kranton, 2000; Eckel and Grossman, 2005; Chen and Li, 2009). Socially-connected members may also know more information about each other, such as work style or personality, that is generally hidden to the firm but that affects team communication and coordination.

In our study, we use a classroom experiment to examine the impact of decentralized group formation on overall group productivity. Specifically, we vary
the team formation process for 685 undergraduates enrolled in a business school course and compare their group performance. Our treatment methodology closely resembles the group formation process found in the workplace and thus has applicability for organizations. At the beginning of the course, the students are separated into smaller sections of approximately 24 students each. We then apply one control and two treatment conditions. In the control condition, the four-person groups are randomly assigned. In the first treatment condition, subjects can form four-person groups freely with any individuals in the same section. In the second treatment condition, we determine the groups based on an algorithm designed to match students with complementary skills. These latter teams reflect the workplace approach to team design, where managers use information collected through the hiring or performance appraisal process to take advantage of worker skill complementarity.

We first elicit from all subjects their demographics and personal skill rankings in four skill areas that we identify as important for success in completing the group project. We then randomize sections into one of six treatments in a $3 \times 2$ design. In the group formation dimension, we vary whether the four-student groups are formed by random assignment ("random treatment"), an algorithm that prioritizes skill complementarity ("algorithm treatment"), or student choice ("self-chosen treatment"). In the information dimension, we vary whether we tell the subjects about the skill rankings of the other students in their section, as such information may affect how students choose their groups. Once the groups are formed, we ask the students to report any prior connections that exist between group members.

The results of our experiment show that the group formation process has a significant effect on group performance. People who choose their own groups perform significantly better than those who are randomly assigned, and about as well as those who are assigned a group by the algorithm. These results provide evidence that decentralized group formation does not harm group productivity. Furthermore, taking into account the cost of eliciting worker information and ap-
plying the algorithm in the group formation process, these results suggest that, in the end, decentralized group formation is more efficient for a firm.

In terms of how students choose to form their groups, we find that students base their decisions on social connections, at the expense of skill complementarity, even when skill information is available. The fact that self-chosen groups perform as well as algorithm groups provides evidence that social connections can compensate for a lack of skill complementarity. When we examine why self-selected teams perform better than those that are randomly assigned, we find some evidence that the higher performance is attributed to a higher level of effort exerted by students in self-chosen groups. Along the information dimension, we find no effect of skill information on either group formation selection or group performance.

Our results have applicability to workplace team formation for several reasons. First, our subjects are representative of real Singaporean workers because they are business students at the National University of Singapore who will soon be joining the workforce. Second, our group formation treatments reflect possible firm approaches to team assignment. In our algorithm treatment, we obtain the same type of skill information as that obtained by firms during the interview process. In the self-chosen treatment, we mimic a decentralized group formation process that firms can undertake. Finally, the task concerns real-world phenomena and requires the delivery of a presentation, and therefore mirrors projects undertaken by teams in real organizations such as consulting firms.

Our experiment contributes to the extensive literature on decentralization in large organizations. Within this field, Mookherjee (2006) provides a comprehensive survey of studies that examine decentralization from a mechanism design standpoint. These studies emphasize that decentralization provides benefits through improved communication, but may create costs for a firm arising from the principal-agent problem. However, there are fewer empirical examinations of the impact of decentralization on productivity. Ichniowski, Shaw and Prentushi (1997) and Ichniowski and Shaw (1999, 2003) document that U.S. businesses have
increased the use of innovative human resource management practices, delegating production decisions to worker teams, and find that this has a significant impact on productivity. In another line of studies, Bresnahan, Brynjolfsson and Hitt (2002) and Hubbard (2000) find that lowers the cost of communication and information processing affect firms’ decentralization decision. This study contributes to the literature by providing field experimental evidence on the effect of decentralizing the group formation decision on worker productivity.

Our study also makes several contributions to the teamwork literature. First, the superior performance of the groups in our algorithm treatment provides support for economic models of skill complementarity. In one study on worker skill complementarity, Hamilton, Nickerson and Owan (2003) find that heterogeneous teams in a garment factory in California are more productive than teams with members of homogeneous ability. In another study of undergraduates in the Netherlands, Hoogendoorn, Parker and van Praag (2014) find an initial increase in productivity for teams with greater ability dispersion, with an ultimate productivity decline as dispersion becomes too wide. Our study extends their results by examining skill complementarity in multiple dimensions. That is, in our algorithm treatment, we identify four different skills and place people into groups such that each skill is represented by at least one group member. In this way, our study better mirrors the economic theory approach to team configuration highlighted by Lazear and Shaw (2007).

In addition, our results lend support to previous findings that self-selected groups perform better. For example, Chen (2016) finds that coordination in a minimum-effort game in a laboratory setting is improved when subjects are allowed to choose their own groups. Similarly, Blasco et al. (2013) find that coders allowed to choose their own groups in an online field experiment perform better on a coding task.

Interestingly, our finding that individuals choose groups based on social connections is different from the results of Hamilton, Nickerson and Owan (2003) who find that individuals choose groups based on ability. This difference in find-
ings could be explained by the incentive levels in the different studies.\textsuperscript{1} Indeed, Bandiera, Barankay and Rasul (2013), in a field experiment with U.K. fruit pickers, find that switching the incentives from a piece-rate model to a rank or tournament model leads pickers to form groups with others of similar ability rather than groups of their friends. This suggests that, with stronger incentives, individuals may be more likely to choose their groups based on skill complementarity rather than social connections.

2 Experimental Design

In this section, we discuss the setting and design of our experiment, as well as our data. Our experiment addresses the question of how different group formation protocols impact group performance. Our subject pool consists of 685 students in a large undergraduate class at the National University of Singapore (NUS). The class meets in both a large lecture format, delivered weekly for 13 weeks, and smaller discussion sections, which meet weekly starting in the third week of lecture. Each student attends one of 29 discussion sections for the duration of the class.\textsuperscript{2} The discussion sections, led by teaching assistants, primarily review the material presented in the lectures.

As part of their course requirements, students must present a group project in their discussion section worth 25 percent of their final grades. Given the importance of university grades in Singapore for future career prospects,\textsuperscript{3} students have

\textsuperscript{1}In Hamilton, Nickerson and Owan (2003), factory workers were originally paid a piece rate, but then chose their own groups and were paid a group piece rate. In our context, students are rewarded based on group performance, a potentially weaker incentive.

\textsuperscript{2}A university-wide balloting system assigns students to sections based on their submitted preference rankings. In our randomization procedure, we account for the possibility that some sections may be more popular than others.

\textsuperscript{3}Pan, Soon and Wong (2015) find that a 1 unit increase in cumulative average points at the National University of Singapore (an increase in grade average from B- to B+ or B+ to A) is associated with a 12.3 percent increase in monthly salary.
a strong incentive to perform as well as possible on the project. With four students per group and a total of six groups per section, three groups present in Week 8 of the discussion section and three groups present in Week 10. The presentations are given in two different weeks due to time constraints. As such, the groups that present in Week 8 are given a different topic than the groups that present in Week 10, with the week in which a group presents randomly determined. The groups’ performance on this project constitutes our main outcome variable. While the groups are graded by their sections’ respective teaching assistants (TAs), we ensure consistent performance measurement for our purposes by having two research assistants (RAs) assess the performance of every group in every section.

Table 1: Experimental Design

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Number of sections</th>
<th>Number of four-student groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Assignment</td>
<td>No Information</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Skill Information</td>
<td>5</td>
</tr>
<tr>
<td>Algorithm Assignment</td>
<td>No Information</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Skill Information</td>
<td>5</td>
</tr>
<tr>
<td>Student Chosen</td>
<td>No Information</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Skill Information</td>
<td>5</td>
</tr>
</tbody>
</table>

We employ a $2 \times 3$ factorial design, with all subjects in the same section also in the same treatment. That is, we randomly assign different sections into one of six treatments. A summary of our experimental design is displayed in Table 1.

---

4Since not all sections have exactly 24 students, some groups may consist of sizes other than four students. We exclude all groups of size other than four from our analysis (45 students total).

5In our analysis, we control for topic fixed effects so that we can effectively compare the performance of groups presenting on the same topic.

6We stratify the randomization based on section popularity and tutor.
Along one dimension, we vary the method by which the groups are formed for the group project according to one of three possibilities. In the “Random” treatment, students are placed into four-person groups randomly. In the “Self-chosen” treatment, students are asked to choose their own groups of four people from the students in their section. In the “Algorithm” treatment, we assign students to four-person groups based on their responses to a demographic survey, administered to all students during the first lecture.

One of the questions in this survey asks students to rank their own proficiency in four different skills that we identify as important for the group project, also highlighted in the rubric given to the students for the group project: presentation, research, quantitative analysis, and economic theory. Since our “Algorithm” treatment is based on the principle that it is better to have people with complementary skills working together on a group project, we design our algorithm so that an ideal group would consist of four students with distinct highest-ranked skills.

For our algorithm, we treat group formation as a linear assignment problem, which is a classic problem in combinatorial optimization, as described by Munkres (1957):

The personnel-assignment problem is the problem of choosing an optimal assignment of $n$ men to $n$ jobs, assuming that numerical ratings are given for each man’s performance on each job. An optimal assignment is one which makes the sum of the men’s ratings for their assigned jobs a maximum.

We use the solution to this linear assignment problem, called the Munkres or Hungarian algorithm, to determine our groups in the “Algorithm” treatment. Specifically, for a section of 24 students, we assign each student to one of four roles (i.e., each section has 6 presentation roles, 6 research roles, etc.). To create each group, we then randomly choose one student from each role. Because the Munkres algorithm maximizes the students’ ratings for their assigned roles at the section level, this method of group formation is Pareto optimal.\footnote{This method is Pareto optimal in the sense that no student can change groups to increase a}
In addition to varying how groups are formed, we vary whether we give students information regarding the self-reported skills of the other students in their section. For students in the “Random” and “Algorithm” treatments, this information has no bearing on students’ group assignments, but could reveal some information about the other members of their groups. For students in the “Self-chosen” treatment, this information might affect their group selection choices.

Our experimental design allows us to test for several effects of group formation. First, we can verify whether skill diversity affects group performance by comparing the “Random” and “Algorithm” treatments. Subjects in these two treatments experience identical experimental procedures (since these subjects are all given a group assignment with no further explanation), so any differences in performance can be attributed to the group assignment process. Secondly, by comparing the “Random” and “Self-chosen” treatments, we can observe whether students who choose their own groups perform better than those who are randomly assigned to groups. We can also observe whether students in the “Self-chosen” treatment choose their groups based on social connections or on member ability. Lastly, by comparing the performance in the “Information” and “No information” treatments, particularly for the subjects in “Self-chosen” groups, we can observe whether subjects use information about group members to determine their groups and, by extension, to improve their performance.

Table 2 summarizes our experiment timeline as well as our data. In the first lecture of the semester, all students in the course complete a demographic survey, which includes a question regarding how they would rank their skill in presentation, research, quantitative analysis, or economic theory (a copy of the survey is included in Appendix A.1).

The student sections begin in Week 3 of the lecture. At the first section meeting, an experimenter assigns the students to four-person groups (or asks the stu-

---

9
Table 2: Experiment Timeline and Data Collected

<table>
<thead>
<tr>
<th>Week</th>
<th>Task(s)</th>
<th>Data collected</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>In-class survey</td>
<td>Skill rankings, Demographics</td>
</tr>
<tr>
<td>3</td>
<td>Group assignment</td>
<td>Network information</td>
</tr>
<tr>
<td></td>
<td>Topic 1 announcement</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Topic 2 announcement</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Topic 1 presentations</td>
<td>Topic 1 performance</td>
</tr>
<tr>
<td>10</td>
<td>Topic 2 presentations</td>
<td>Topic 2 performance</td>
</tr>
<tr>
<td>16</td>
<td>Final exam</td>
<td>Effort</td>
</tr>
</tbody>
</table>

Students to form their own groups) and collects information about the pre-existing social networks within each group (a copy of the survey is included in Appendix A.2). To ensure consistency in the delivery of this information, the same experimenter attends all section meetings that week. To minimize possible experimenter demand effects, the experimenter reads from a script when speaking to the students. After all groups are formed, students fill out a short “network survey” indicating if they know any of their groupmates outside of class. Half of the groups in each section are then randomly assigned to work on Topic 1, while the other half are assigned to work on Topic 2. Topic 1 is given to the students at the end of Week 3 and Topic 2 is given to the students at the end of Week 5.

During Weeks 8 and 10, two research assistants attend each section to view the group presentations. These assistants are not told the details of the experiment. They score each presentation on 10 different questions in 4 categories, corresponding to the 4 skills.\(^8\) Finally, to assess student effort, we ask each student to report the number of hours each group member spent working on the assignment.

\(^8\)As mentioned earlier, the students’ actual grades on the group project were determined by their TAs’ scores of their projects, not the RAs’ scores. The TAs and RAs used different rubrics, with only the TAs’ rubrics revealed to the students. The RAs’ rubric is included in Appendix A.3.
In order to incentivize students to tell the truth, we set up a game in the style of a Keynesian beauty contest. In this contest, students are given points based on how well their reported hours match the average reported hours of the other members of their group. In other words, students are incentivized to report what they believe others will report. To prevent collusion, we ask this question to students during the final exam for the class. Although this game has many Nash equilibria (where everyone’s reports match exactly), the instructor, using a similar example in a lecture, emphasized that the case where everyone tells the truth is a focal point. We also tell students on the final exam that “it is a Nash equilibrium for everyone to report the actual number of hours each person worked on the project.” The question posed is included in Appendix A.4.

3 Results

In this section, we first present our main results for the effects of group formation on group composition and performance. We then examine a possible channel through which group formation might affect performance by analyzing how group effort varies across the different treatments.

Our analyses are based on our data for the four-person groups, which include 640 students across 160 groups. In addition, we cluster standard errors at the section level to control for any possible dependency of decisions and performance across groups within a section. All non-parametric tests are two-sided Mann-Whitney $U$ tests, and the unit of observation is the group.

---

If a section does not have exactly 24 students due to either original enrollment or subsequent dropping of the course by a student, then that section will have some non-four-person groups. It is possible that different treatments have different dropout rates, but we test the likelihood of having fewer than four students in a group and find no differences across treatments. Due to unforeseen circumstances, we do not have the grades for one section. We exclude this section from the regression sample.
3.1 Group Composition

In the “Self-chosen” treatment, we examine whether students sort into groups based on network connections, skill complementarity, or demographic characteristics. Figure 1 presents the level of network connections and skill complementarity within a group by treatment. A student’s number of connections is the number of people she reports she knows outside the class. In order to calculate skill complementarity, we apply the Hungarian algorithm to find each member’s optimal role and sum the rank the student assigns to that role. Therefore, a smaller number indicates better skill complementarity.

Figure 1 shows that students tend to sort themselves into groups based on social connections rather than skill complementarity. Specifically, we see that self-chosen groups have more connected members than either randomly-assigned or algorithm-determined groups (Figure 1a). This comparison (network in self-chosen groups > network in control) is significant ($p < 0.01$). However, the comparison between the randomly-assigned and algorithm-determined groups is not significant ($p = 0.46$). On average, a student in the self-chosen treatment knows one more person in her group than a student in either the random or algorithm treatment. Figure 1b shows that the algorithm-determined groups have greater skill complementarity than either the self-chosen or randomly-assigned groups ($p < 0.01$ for both comparisons). However, there is no significant difference in skill complementarity between the randomly-assigned and self-chosen groups ($p = 0.45$).

We next run a series of OLS regressions across groups, accounting for any TA fixed effects in all analyses. The estimated treatment effects on group composition are presented in Table 3. Our dependent variables include the following group characteristics: (1) the average number of connections a member has within her group; (2) the sum of skill ranks; (3) the average student age; (4) the proportion of men in the group; (5) the proportion of business majors in the group; (6) the proportion of Chinese students in the group, and (7) the proportion of students born in Singapore in the group. Our independent variables include the self-chosen and
algorithm treatment dummies, with randomly-assigned groups as the reference.

The results of our regressions confirm our observations from Figure 1. First, we find that self-chosen groups have more network connections (Table 3, column 1) but about the same level of skill complementarity (column 2) as randomly-assigned groups. As for demographic characteristics, our results show that students do not sort into groups based on age, gender, discipline of study, ethnicity, or birthplace. We further find that our algorithm-determined groups exhibit greater skill complementarity—a result of our experimental design—but similar demographic patterns to our randomly-assigned groups. This finding reinforces our observation that skills are not directly correlated with any of our observed demographic characteristics.

Interestingly, our findings suggest that subjects do not seem to use information on other students’ skill rankings to improve the skill complementarity of their self-chosen groups. Specifically, for our “Self-chosen” treatments, we find no significant differences in either the level of skill complementarity or the number of connections for our “Information” and “No information” groups ($p > 0.5$ for both network and skill comparisons). Similarly, we find no difference in group performance for our “Information” and “No information” groups. Consequently, we merge these two groups in our subsequent analyses.
Table 3: OLS: Group Composition by Experimental Treatment

<table>
<thead>
<tr>
<th></th>
<th>(1) Network connections</th>
<th>(2) Skill ranks</th>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-chosen</td>
<td>1.009***</td>
<td>0.213</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>(0.153)</td>
<td>(0.155)</td>
<td>(0.143)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>0.044</td>
<td>-0.829***</td>
<td>-0.059</td>
</tr>
<tr>
<td></td>
<td>(0.139)</td>
<td>(0.243)</td>
<td>(0.131)</td>
</tr>
<tr>
<td>Constant</td>
<td>0.481***</td>
<td>6.158***</td>
<td>19.966***</td>
</tr>
<tr>
<td></td>
<td>(0.138)</td>
<td>(0.242)</td>
<td>(0.104)</td>
</tr>
</tbody>
</table>

Observations 160 160 160 160 160 160 160

$R^2$ 0.268 0.158 0.049 0.022 0.044 0.023 0.048

Notes: 1. We control for TA fixed effects.
2. Robust standard errors clustered by section are in parentheses.
3. Significant at * 10 percent; **5 percent; *** 1 percent.

3.2 Treatment Effects on Group Performance

We next examine whether group performance differs across treatment. To measure group performance, we have our research assistants score each group on a 100 point scale. Each group’s total score is assessed across four categories: research, theory, statistics, and presentation, with each category accounting for 25 percent of the group’s total score. We present the rubric used by the RAs in Appendix A.3.

Figure 2 presents the total and category scores for the groups in our experiment. From Figure 2a, we see that our self-chosen groups perform better than our randomly-assigned groups and approximately as well as our algorithm-determined groups. We also find that both our self-chosen and algorithm-determined groups perform better than our randomly-assigned groups ($p = 0.052$ and $p = 0.045$).

\footnote{We find a Pearson’s correlation of $r = 0.794$ between our two RAs, which indicates good correlation between their scoring of the groups.}
The self-chosen and algorithm-determined groups’ scores are not significantly different ($p = 0.94$). The results in Figure 2b show that this superior performance is observed across all four categories, with the largest difference observed in the statistics category ($p < 0.01$ comparing either the self-chosen or algorithm-determined group to the randomly-assigned group).

Table 4 reports the OLS estimates of our treatment effects on group performance. Column 1 presents the results of treatment on the total performance score while columns 2 to 5 break down the score by category. All regressions control for TA, topic (topic 1 or 2), section meeting day (Monday to Friday) and time (morning or afternoon) fixed effects, and presentation order within a section. Across our three groups, we find that both the self-chosen and algorithm-determined groups outperform the randomly-assigned groups, with the average treatment effect about 3.285 points higher for the total score, significant at the 5 percent level. Evaluating at the mean score of random groups (42.636), we find that this effect is equivalent to about a 7.7 percent increase in performance. The results in columns 2 to 5 show positive treatment coefficients for all four categories, with the effect on “statistics” performance estimated very precisely. These results suggest that all four categories contribute to the superior performance of our treatment groups, and that “statistics” performance is the main contributor to overall group perfor-
Table 4: OLS: Treatment Effects on Group Performance

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Research</td>
<td>Theory</td>
<td>Statistics</td>
<td>Presentation</td>
</tr>
<tr>
<td>Self-chosen</td>
<td>3.285**</td>
<td>0.392</td>
<td>0.293</td>
<td>1.832***</td>
<td>0.769*</td>
</tr>
<tr>
<td></td>
<td>(1.484)</td>
<td>(0.347)</td>
<td>(0.621)</td>
<td>(0.595)</td>
<td>(0.386)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>3.250*</td>
<td>0.682</td>
<td>0.328</td>
<td>1.760**</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td>(1.788)</td>
<td>(0.420)</td>
<td>(0.704)</td>
<td>(0.690)</td>
<td>(0.474)</td>
</tr>
<tr>
<td>Constant</td>
<td>51.224***</td>
<td>9.694***</td>
<td>15.47***</td>
<td>10.48***</td>
<td>15.58***</td>
</tr>
<tr>
<td></td>
<td>(3.302)</td>
<td>(1.042)</td>
<td>(0.964)</td>
<td>(1.319)</td>
<td>(0.813)</td>
</tr>
</tbody>
</table>

Observations: 160

$R^2$: 0.226 0.196 0.212 0.187 0.192

Notes: 1. We control for TA, presentation order, topic and time fixed effects.
2. Robust standard errors clustered by section are in parentheses.
3. Significant at * 10 percent; **5 percent; *** 1 percent.
mance.

Our results regarding statistics performance may reflect student perceptions of the difficulty of the statistical element of the project. In our student skill survey, we find that statistics is most often ranked as the skill in which students feel least proficient (57.8 percent of students), and least often ranked as the skill in which they feel most proficient (9.5 percent of students). These relatively low self-rankings of the statistics skill lead to a dearth of group members who can specialize in the statistics role. For our algorithm treatment, we are able to spread out the group members who list statistics as one of their top skills into different groups. In fact, 73 percent of our algorithm-determined groups have at least one member who ranks statistics as her first or second best skill, compared to only 54 percent of our randomly-assigned (self-chosen) groups.

Lastly, we conduct a Wald test comparing the coefficients of our “Self-chosen” and “Algorithm” treatments. The results of this analysis indicate that we cannot reject the null hypothesis that the two coefficients are the same ($p > 0.5$ in columns 1 to 5). This reinforces our conclusion that our self-chosen groups perform as well as our algorithm-determined groups.

### 3.3 Effort

Finally, we examine whether differences in group performance can be explained by differences in effort across the treatment groups. Using the reported hours worked from the effort survey, we calculate group effort by adding the hours reported by each student for each group member, i.e., her perceived man-hours worked by the group. Students have an incentive to report the actual hours worked,

---

11We compare these self-assessments to those for the other three skills: 1) presentation (28.1 percent rank it first, 20.2 percent rank it last), 2) research (26.7 percent rank it first, 13.4 percent rank it last), and 3) economic theory (35.6 percent rank it first, 8.6 percent rank it last).

12This lack of group members who consider themselves to be proficient in a specific skill only substantially affects the statistics skill. For the other three skills, at least 90 percent of groups in each treatment have a member who ranks the respective skill as either first or second.
but they may not perfectly observe each other's effort if they spend some time working alone on the project. However, any measurement error due to not observing hours would tend to bias our estimates downward, i.e., work against finding any differences in effort across groups.

Table 5: OLS: Treatment Effects on Group Members' Effort

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef. of variance</td>
<td>Minimum hours</td>
<td>Maximum hours</td>
<td></td>
</tr>
<tr>
<td>Self-chosen</td>
<td>0.124*</td>
<td>-0.015</td>
<td>0.185</td>
<td>0.247</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
<td>(0.017)</td>
<td>(1.027)</td>
<td>(1.022)</td>
</tr>
<tr>
<td>Algorithm</td>
<td>0.087</td>
<td>-0.009</td>
<td>0.301</td>
<td>0.553</td>
</tr>
<tr>
<td></td>
<td>(0.079)</td>
<td>(0.015)</td>
<td>(1.104)</td>
<td>(1.224)</td>
</tr>
<tr>
<td>Constant</td>
<td>4.002***</td>
<td>0.116</td>
<td>5.987</td>
<td>6.312</td>
</tr>
<tr>
<td></td>
<td>(0.777)</td>
<td>(0.108)</td>
<td>(6.654)</td>
<td>(7.115)</td>
</tr>
<tr>
<td>Observations</td>
<td>615</td>
<td>615</td>
<td>615</td>
<td>615</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.063</td>
<td>0.056</td>
<td>0.044</td>
<td>0.053</td>
</tr>
</tbody>
</table>

Notes: 1. We control for student demographics.
2. We control for tutor, presentation order, topic and time fixed effects.
3. Robust standard errors clustered by section are in parentheses.
4. Significant at * 10 percent; ** 5 percent; *** 1 percent.

Table 5, column (1) presents the estimates of a set of OLS regressions. The dependent variable is the logarithm of total man-hours.\(^{13}\) Note that we include the same group characteristics as in Table 4 as well as individual demographic characteristics such as age, gender, discipline of study, ethnicity, and birthplace. We find that self-chosen groups invest 12.4 percent more hours on the project than do randomly-assigned groups. This effect is significant at the 10 percent level.\(^{14}\)

\(^{13}\)We use the log transformation because the reported hours are lognormally distributed. Meanwhile, it automatically drops observations with zero hours. These are unlikely to be the actual effort exerted, given that every group, at minimum, prepares presentation slides.

\(^{14}\)If we exclude one outlier who reports 100 hours per member, the estimate increases to 13.8
Evaluating at the mean of the control (37.83 man-hours), the difference is about 4.69 hours per group. We find no significant difference in total effort between our algorithm-determined and randomly-assigned groups.

To gain further insight into effort as a mechanism for the effect of group formation on performance, we examine the dispersion of effort among group members. We measure dispersion as the coefficient of variation (the ratio of the standard deviation to the mean), and minimum and maximum reported hours across the four members of a group. As shown in Table 5, we find no significant difference between our treatment and control groups in the distribution of effort within a group.

4 Conclusion

In this study, we use a classroom experiment to manipulate how groups are formed in order to examine the effect of group formation on group composition, effort, and performance. Our results show that students tend to select groups based on social connections rather than member skills. We further find that these self-chosen groups perform better than do randomly-assigned groups, and about as well as algorithm-determined groups with optimal skill complementarity. We also find evidence that the superior performance of self-chosen groups is due to a higher amount of effort exerted within these groups. These findings are consistent with the prediction that social connections can mitigate free riding behavior in teamwork.

Regarding the decentralization of group formation in the workplace, our findings suggest that allowing employees to form their own workgroups can lead to a similar level of productivity as manager-determined groups. Delegating group formation may take advantage of hidden information, such as social connections, which are valued in group work. Furthermore, taking into account the costs associated with collecting worker information and designing formation rules, self-
chosen groups provide cost savings for a firm.

In theory, a centralized mechanism should be able to mimic the outcome of any decentralized system, a direct implication of the Revelation Principle. However, the validity of this theory relies on the absence of any communication or information processing costs. In contrast, in actual organizational contexts, groups are formed based on rich information about participants; this information is collected at a cost. The choice between centralization and decentralization therefore reflects a trade-off between the communication costs involved in obtaining information when forming groups and the potential incentive problems if workers act in their own self-interest when forming their own groups. When decision-making is delegated to the workers, our findings suggest that they sort by friendship, which in turn may overcome the incentive problems in teams, making decentralization the superior organizational arrangement.

Finally, it should be noted that the effectiveness of self-chosen teams may depend on other policies adopted by the firm. For example, Bandiera, Barankay and Rasul (2013) find that self-chosen teams may select differently depending on the strength of their incentives. Together with our results, these findings suggest that a firm that adopts a decentralized group formation policy must also carefully choose the incentive scheme that it employs. This is an interesting empirical question that should be examined in the future.

References


A Appendices

A.1 Skills Survey

BSP 1005 Survey

Please answer the following questions.

1. Name: ..........................................................

2. Matriculation number: ......................................

3. Home Faculty: ................................................

4. What is your age? ______

5. What is your gender?  □ Male  □ Female

6. Which of the following best describes your racial or ethnic background?
   □ Chinese  □ Malay  □ Indian  □ Other

7. Were you born in Singapore?  □ Yes  □ No

8. What was the most recent school you attended before joining NUS?
   (Examples: Victoria JC, Nanyang Polytechnic, etc.)

   ______________________________

9. Have you taken any economics classes before? If so, at what level did you take your most advanced class?
   (Examples: JC, Polytechnic, etc.)

   ______________________________

10. At what level did you take your most advanced mathematics class?
    (Examples: JC, Polytechnic, etc.)

    ______________________________

    (See next page →)
11. A major grading component of BSP 1005 is a group project. The following four skills are essential for completing the project. **Please rank these skills by how competent you feel in each.** This information may be shared with your potential group members.

(1) Presentation skills: ability to effectively communicate and present information

(2) Research skills: ability to find relevant background information and data

(3) Quantitative skills: ability to use statistical tools (such as Excel) to analyze data

(4) Economic theory skills: ability to apply what you have learned in the module to the project

Example: If you feel you are best at economic theory, next best at quantitative methods, next best at presenting, and worst at research, then you would say that your own ranking is:

\[ 4 > 3 > 1 > 2 \]

Your own ranking:

\[ ___ > ___ > ___ > ___ \]
A.2 Group Network

Group 1

Do any members of your group know each other outside of this module?
### A.3 Presentation Rubric

<table>
<thead>
<tr>
<th>Category</th>
<th>Item</th>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>Background</td>
<td>policy, institutional, and cultural details</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence</td>
<td>how relevant and authoritative are their arguments?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data Collected</td>
<td>amount, scope and usefulness of data collected</td>
<td></td>
</tr>
<tr>
<td>Theory</td>
<td>Framework</td>
<td>unifying framework?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Economic Analysis</td>
<td>how reasonable is their analysis?</td>
<td></td>
</tr>
<tr>
<td>Statistics</td>
<td>Methods</td>
<td>how detailed and appropriate is their econometrics?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistical Analysis</td>
<td>how good is their answer and explanation?</td>
<td></td>
</tr>
<tr>
<td>Presentation</td>
<td>Flow</td>
<td>how well did their presentation flow?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Professionalism</td>
<td>were they and their slides professional?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Engagement</td>
<td>was it interesting?</td>
<td></td>
</tr>
<tr>
<td>Factual</td>
<td>How many people spoke during the presentation?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questions</td>
<td>How did they divide the presentation? By topic / section / time</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>How much time did they take?</td>
<td>&lt;15 / =15 / &gt;15</td>
<td></td>
</tr>
</tbody>
</table>
A.4 Effort Elicitation

Recall the flat tire example we discussed in lecture. This question is a similar task regarding your group project.

a. Please write down the number of hours each person in your group spent on the group project using the table below. As long as you fill out this table, you will receive 2 marks.

b. You can receive up to 4 more marks based on how well your reported hours match the average hours your group members report. The better you match the average, the more marks you receive.

Just as in the flat tire example, it is a Nash equilibrium for everyone to report the actual number of hours each person worked on the project. Your responses to this question will not affect anyone’s grade on the group project. (For the exact method we will use to calculate your marks, see the next page).

<table>
<thead>
<tr>
<th>Name of Group Member</th>
<th>Number of Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>_____ hours</td>
</tr>
<tr>
<td>You</td>
<td></td>
</tr>
<tr>
<td></td>
<td>_____ hours</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>_____ hours</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>_____ hours</td>
</tr>
</tbody>
</table>

27
The number of marks you receive is determined by the following procedure:

1. For each member of your group, we will compare two numbers: a) the number reported by you for that member and b) the average number reported by the others in your group for that member.
2. We calculate the absolute difference between these two numbers.
3. We now have this difference for each member of your group, and we take the maximum of these differences.
4. The number of marks you receive is based on this maximum difference:
   - You receive 4 marks if this difference is at most 1 hour.
   - You receive 3 marks if this difference is between 1 hour and 2 hours.
   - You receive 2 marks if this difference is between 2 hours and 4 hours.
   - You receive 1 mark if this difference is between 4 hours and 6 hours.
   - You receive 0 marks if this difference is more than 6 hours.

Example:

<table>
<thead>
<tr>
<th>Group Member</th>
<th>Your Report</th>
<th>Average of Others’ Reports</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>You</td>
<td>y hours</td>
<td>a hours</td>
<td></td>
</tr>
<tr>
<td>Member 2</td>
<td>w hours</td>
<td>b hours</td>
<td></td>
</tr>
<tr>
<td>Member 3</td>
<td>x hours</td>
<td>c hours</td>
<td></td>
</tr>
<tr>
<td>Member 4</td>
<td>y hours</td>
<td>d hours</td>
<td></td>
</tr>
</tbody>
</table>

Call the maximum of these differences \( D = \max(|y-a|, |w-b|, |x-c|, |y-d|) \). You will receive marks \( P \) based on the value of \( D \):

\[
P = \begin{cases} 
4, & \text{if } D \leq 1 \\
3, & \text{if } 1 < D \leq 2 \\
2, & \text{if } 2 < D \leq 4 \\
1, & \text{if } 4 < D \leq 6 \\
0, & \text{if } D > 6 
\end{cases}
\]
A.5 Pareto Efficiency of the Algorithm Treatment

Here we prove that our method of group formation for the algorithm treatment is Pareto optimal, in the sense that no students can change groups to increase a group’s total ratings without decreasing another group’s total ratings.

Proof. Suppose that there exists a change in the formation of groups that makes one group better off without making any other groups worse off. Let us start with a simple case in which the change involves two students: student $i$ from group $m$ and student $j$ from group $n$. Without loss of generality, assume that after swapping the two students, group $m$ has strictly higher skill ratings and group $n$ has weakly higher ratings. The other four groups in the section remain unchanged.

After swapping the two students, group $m$ and group $n$ each experience a change related to the roles within each group. If every student performs the same role in the new group as initially assigned, the skill ratings for each member remain the same and so do the total ratings for each group.

If the two students change their roles upon entering their new groups, this leads to strictly higher ratings for group $m$ and weakly higher ratings for group $n$. This will increase the students’ ratings for their roles at the section level, contradicting the optimal assignment solved by the Munkres algorithm, which already maximizes the sum of all students’ ratings at the section level.

For changes involving more than two students, the proof is similar: if the roles within groups remain the same, there is no improvement in groups’ total ratings. If roles change, then there is a new assignment of students to roles such that, at the section level, total ratings are improved. This result again contradicts the optimal assignment solved by the Munkres algorithm.

A.6 Implementation of the Algorithm Treatment

For each section, we have 24 students, whom we must assign to 24 jobs. We treat these jobs as 6 copies of 4 skill-based roles (i.e., there are 6 presentation roles, 6 research roles, 6 quantitative roles, and 6 economic theory roles). The output is
a role assignment for all 24 students, with 6 students assigned to each role. This calculation is performed by the Munkres/Hungarian algorithm, where the input is the cost for each student-job combination and the algorithm must minimize overall costs.

The cost of each student-job combination is determined such that the roles corresponding to the student’s best skill have the lowest cost, the roles corresponding to the student’s second best skill have the next lowest cost, and so on. In particular, we give the rank of 1 a cost of 0, the rank of 2 a cost of 1, the rank of 3 a cost of 25, and the rank of 4 a cost of 601. This ensures that, when given the option of placing 1 student in her third-ranked role (with all others in their best role), or placing all students in their second-ranked role, the algorithm will choose the latter option.

Once the algorithm gives us each student’s job assignment, we form groups by randomly placing one student from each role into a group, so that each group has 1 student assigned to each different role. This process should, in principle, create groups that are balanced with a diverse set of skills.