Institute for Economic Studies, Keio University

Keio-IES Discussion Paper Series

Peer Learning in Teams and Work Performance: Evidence from a Randomized Field Experiment

Kenju Kamei、John Ashworth

5 April, 2022
DP2022-005
https://ies.keio.ac.jp/en/publications/18310/

Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office@adst.keio.ac.jp
5 April, 2022

Peer Learning in Teams and Work Performance: Evidence from a Randomized Field Experiment

Kenju Kamei、John Ashworth

Keio-IES DP2022-005
5 April, 2022
JEL Classification: C93, J24, I23
Keywords: peer effects;dilemma;field experiment;teamwork;knowledge sharing

Abstract

A novel field experiment shows that learning activities in pairs with a greater spread in abilities lead to better individual work performance, relative to those in pairs with similar abilities. The positive effect of the former is not limited to their performance in peer learning material, but it also spills over to their performance in other areas. The underlying improvement comes from the stronger increased performance of those whose achievements were weak prior to peer learning. This implies that exogenously determining learning partners with different abilities helps improve productivity through knowledge sharing and potential peer effects.

Kenju Kamei
Faculty of Economics, Keio University
2-15-45 Mita, Minato-ku, Tokyo
kenju.kamei@keio.jp

John Ashworth
Department of Economics and Finance, Durham University
Durham University Business School, Mill Hill Lane, Durham DH1 3LB, United Kingdom

Peer Learning in Teams and Work Performance: Evidence from a Randomized Field Experiment

Kenju Kamei,[1,*] and John Ashworth[2]

[1] Faculty of Economics, Keio University

[2] Department of Economics and Finance, Durham University

April 2022

Abstract: A novel field experiment shows that learning activities in pairs with a greater spread in abilities lead to better individual work performance, relative to those in pairs with similar abilities. The positive effect of the former is not limited to their performance in peer learning material, but it also spills over to their performance in other areas. The underlying improvement comes from the stronger increased performance of those whose achievements were weak prior to peer learning. This implies that exogenously determining learning partners with different abilities helps improve productivity through knowledge sharing and potential peer effects.

JEL classification codes: C93, J24, M53, M54, I23

Keywords: peer effects, dilemma, knowledge sharing, field experiment, teamwork

Acknowledgement: The authors thank all the staff in the undergraduate office in Durham University for facilitating the peer review assessment activities in the 2019/20 academic year, and also Ashlee Bennett for organizing background data of the students. The authors thank Pedro Dal Bó and Katy Tabero for their helpful comments.

[*] Correspondence Author: Faculty of Economics, Keio University, 2-15-45, Mita, Minato-ku, Tokyo 108-8345, Japan. Email: kenju.kamei@gmail.com, kenju.kamei@keio.jp.

1. Introduction

Situations in which peers interact with each other to improve performance, while aiming to achieve individual goals, are ubiquitous whether in schools, daily lives, or the workplace. A key to achieving a Pareto optimal outcome is successful collaboration among peers. Benefits from peer interactions may be strong when they have heterogeneous abilities and skills if the less endowed interact with and learn from the highly endowed, because the former has larger room for improvement. A dilemma, nevertheless, exists since peer learning interactions may not be cost-free and benefits may be small for the highly endowed individuals.

Prior research suggests that people's individual performances, whether work-related or academic, are affected by other members in their peer groups due to peer pressure and learning through social interactions (e.g., Katz *et al.*, 2001; Sacerdote, 2001; Falk and Ichino, 2006; Mas and Moretti, 2009). For example, peer interactions and learning are powerful practices in firms and organizations (e.g., Ichniowski *et al.*, 1997; Hamilton *et al.*, 2003). However, how peer groups should be formed is not straightforward. First, peer interactions are known to lead to better overall work performance in a group with heterogeneous rather than homogeneous abilities and skills[1] under individual-based remunerations (i.e., each member is evaluated based on their own individual performance), even though the average abilities and skills are similar for the two types of groups (e.g., Carrell *et al.*, 2009; Lyle, 2009; Duflo *et al.*, 2011; Feld and Zölitz, 2017). The superiority of group heterogeneity is driven by high-ability (-achieving) members. They are able to and tend to form subgroups with like-minded high achievers, thereby leading to a further improvement of their performances. In contrast, low-ability (-achieving) members suffer, although, on average, overall performance is better thanks to the strong improvement of high types. It is therefore difficult to judge the efficacy of heterogeneity normatively. Second, groups with a greater spread in abilities are likewise known to have an advantage under group incentives, that is, when members are evaluated based on their group performance. The underlying mechanism is, however, different, and is the so-called mutual learning hypothesis: under group incentives, more able workers not only impose norms for the sake of other members, but they also teach less able workers (e.g., Ichniowski *et al.*, 1997; Hamilton *et al.*, 2003). Hence, the finding from the second branch of the literature implies that a way to form a socially-desirable peer group under individual-based remunerations may be to have a sufficiently small peer group with a greater spread in abilities, thereby preventing sub-groups from endogenously emerging and, at the same time ensuring a pairing between more able and less able

[1] The former (latter) group is also called a "heterogeneous" ("homogeneous") group, hereafter, in the paper.

workers. Such pairing may result in a Pareto improvement through effective mutual learning interactions. This paper experimentally manipulates interim achievement differences in pairs and then studies how the intervention affects individual work performance.

While it is unclear how the mutual learning hypothesis extends to the case of a small group under individual incentives, pairing a more able worker with a less able worker for peer learning has an advantage at least from a theoretical perspective. For example, experimental literature suggests that people have social image concerns and other-regarding preferences (e.g., Ariely *et al.*, 2009; Shang and Croson, 2009). Pairing with a high-ability worker may give the low-ability worker a particular incentive to work hard to avoid incurring disutility from social effects, such as shame and harming their social image (e.g., Bénabou and Tirole, 2006 and 2011; Bowles and Gintis, 2015). Such a pairing may also change reference points of the less able workers if they have poor work attitudes and are over-optimistic about their own performances (e.g., Abeler *et al.*, 2011; Svenson, 1981). High-ability workers may also try helping low-ability ones from inequality aversion if they are required to interact with each other (e.g., Fehr and Schmidt, 1999). A possible concern is, however, a perverse reaction by high-ability workers: those who are forcibly matched with low-ability ones contrary to their preferences may not contribute their best efforts in peer learning activities (e.g., Kamei and Markussen, 2020).

This paper provides clean evidence from a randomized field experiment in a classroom that pairing individuals with a large achievement difference, rather than with similar achievement levels, leads to better overall performance on average. All first-year undergraduate students in accounting and finance took a compulsory first-year introductory course in economics in Durham University during the 2019/20 academic year. This was a full-year course which began in the first term of the academic year. Students learned microeconomics in the first term, and macroeconomics in the second term. The students' performances were evaluated solely based on their own marks in a written examination which took place at the end of the academic year.

The students' learning activities were composed of attendance in lectures, engagement in bi-weekly mandatory seminars (whose size was 15 to 20 students), two formal written assignments (one for each term), and two peer review assessment activities (one for each team). In the peer review assessment activities, each student was paired with another in their respective seminar group. They independently attempted a problem set, after which they critically assessed their partner's solutions and then discussed the problem set in pairs by holding a meeting. Pair partners were changed between terms 1 and 2. An intervention was made for the peer review activities in term 2: while each student was randomly assigned their pair partner in half of the

seminar groups (treatment condition), pairs were formed so that their interim class performances were similar to each other in the other half of the seminar groups (control condition). Thus, pairs in the treatment condition had a greater variation in interim achievement levels between pair mates, relative to those in the control treatment. Partner assignment was random in all seminar groups for term 1.

The exam performance data show that students were higher-achieving on average in the treatment than in the control condition. This underscores the effectiveness of having a pair with a greater spread in prior performances for peer learning in organizations. A close look at the data reveals that the worse performers improved academic performance significantly more than their matched better performers in pairs belonging to the treatment condition. This suggests that consistent with the mutual learning theory, exogenously determining learning partners with different abilities helps improve productivity through knowledge sharing and positive peer effects. Further, interestingly, the students in the treatment groups showed better understanding of even the term 1 material in the examination, suggesting that their peer learning experiences in macroeconomics spilled over to their understanding of microeconomics. This may mean that worse performers improved study habits when revisiting the term 1 material for the exam. Such productivity spillovers within individuals further imply the importance of devising an effective pairing procedure.

The rest of the paper proceeds as follows: Section 2 describes the related literature, and then Section 3 discusses the background, the experimental design, and the procedure. Section 4 provides hypotheses based on behavioral models and 5 presents the results. Section 6 provides discussions and concludes.

2. Related Literature

This study is closely related to three branches of the literature in labor, organizational and personnel economics, experimental economics, and economic education: (a) peer interactions in heterogeneous teams, (b) peer pressure and productivity, and (c) mutual learning theory. While the literature in (a) seem to suggest that teams with homogeneous rather than heterogeneous abilities and skills lead to better performance, the literature in (b) and (c) suggests the opposite. However, prior research, on balance, suggests the superiority of heterogeneous teams.

2.1. Peer Interactions in Heterogeneous Setups

A key aspect of teamwork in teams is how students share knowledge and skills with their teammates. Such a peer interaction structure can be described by an asymmetric public goods game as students' knowledge and abilities differ according to their backgrounds and unobserved

characteristics, and knowledge sharing is a typical example of a social dilemma. A rich laboratory experiment literature suggests that collaboration would be less successful in a heterogeneous rather than a homogeneous setup, because the tension between highly and less endowed members is usually intense. For example, in a team where endowments are unequally distributed, members' endowment sizes are known to be negatively correlated with their levels of cooperativeness in the voluntary provision of a public good (e.g., Chan *et al.*, 1996; Kamei, 2018; Maurice *et al.*, 2013). For example, in Kamei (2018), while the average contribution of the highly endowed was less than 10% of their endowment, that of the least endowed was more than 50% of the endowment on average.

Further, a negative effect of productivity heterogeneity in teams among team members was also reported. For example, Fischbacher *et al.* (2014) experimentally showed that heterogeneity in returns, i.e., Marginal per capita return (MPCR), from contributing to a public good undermines cooperation, perhaps driven by members' pessimistic beliefs about others' contribution behaviors.[2]

In sum, this line of the literature suggests higher academic performances through more successful collaboration in homogeneous rather than heterogeneous teams. Having said that, the previous experiments listed above were conducted under *anonymous* conditions in a laboratory (where subjects' identities, such as faces and names, were not revealed to each other) and also subjects' individual contribution decisions were not verifiable since the experiments were designed based on public goods games with group sizes of at least three. Hence, the earlier findings may not apply to the setup of the present study since students' interactions within pairs in the peer review assessment activities are made with full transparency here. Literature in (b) in fact suggests the positive impact of observability – see Section 2.2. Most of the empirical research or randomized field experiments in literature (b) and (c) used non-anonymous setups, finding the superiority of team heterogeneity – see Sections 2.2 and 2.3.

2.2. Peer Pressure and Productivity

There is a substantial literature, both theoretically and empirically/experimentally, for the role of peer pressure on influencing behaviors. Much theoretical research suggests that having detailed information about members' behaviors per se, such as through peer review assessment

[2] The positive impact of homogeneity is not limited to the distribution of resources or productivity. Prior experimental research on sorting in public goods games found a higher level of cooperation when like-minded individuals (in terms of cooperative dispositions) are grouped together rather than otherwise, although whether cooperation norms prevail in a group depends on the group's cooperativeness – for example, see, Gunnthorsdottir *et al.* (2007) and Gächter and Thöni (2005). In Page *et al.* (2005), even the average contribution of groups where less cooperative subjects were grouped together were not that low under sorting.

activities in the present study, may trigger positive effects on work performance. For example, working in teams could trigger social image concerns among members, thereby encouraging hard work to be recognized by others (e.g., Bénabou and Tirole, 2006 and 2011). Because high observability makes it possible to compare self with others, teamwork could also trigger social effects, such as guilt, shame and pride, among members in teams (e.g., Bowles and Gintis, 2015), hence inducing them to study harder to achieve better academic performance.

Laboratory experiments support the idea that high visibility encourages socially-desirable behaviors under certain conditions: e.g., voluntary contributions to a public good (e.g., Samek and Sheremeta, 2014), direct punishment of norm violators (e.g., Kamei and Putterman, 2015), third party enforcement of social norms (e.g., Kamei, 2018), and charitable-giving (e.g., Ariely *et al.*, 2009; Soetevent, 2005). For example, in a real-effort experiment where no material incentives were associated with effort provision, Ariely *et al.* (2009) demonstrated that worker subjects exert stronger efforts for charity when they have to tell others about their own donation amounts than when their identities and acts remain anonymous. In the context of this study, students can realize their partners' achievement levels in the peer review assessment activities, which would give the less able student in a pair an incentive to work hard *in* and *after* the peer review activities so that she can avoid incurring disutility due to such information effects from being seen her weaker performances by her pair mate or other students outside the peer review activities. This kind of behavioral effect would be larger, the larger ability difference a pair has. Hence, it can be predicted that the learning effects of peer review activities would be on average larger in pairs with heterogeneous than with homogeneous abilities.

The information effect in pairs is not limited to the social effects. For example, information about peers' strong efforts and/or performance may change the reference points of the less able students (e.g., Abeler *et al.*, 2011). For example, Shang and Croson (2009) found that people donate larger amounts when they are informed of others' active charitable-giving behaviors, even if own donation amounts are kept private (also see Croson and Shang [2008]). This line of the literature again suggests that students' learning outcomes in the present paper would on average be higher in pairs with a greater spread in abilities and skills, since the less able students can become aware of their academic positions through the peer review assessment activities, thereby updating their expected study behaviors. This prediction may be reinforced by the fact that workers are on average known to have overconfident and biased beliefs about own ability (e.g., Langer, 1975; Svenson, 1981; Larkin *et al.*, 2012).

During the last two decades, economists have devoted considerable efforts into identifying peer pressure and interactions in realistic setups using randomized field experiments. Prior

research can be broken into either (i) peer effects with (possible) social interactions – simply "peer effects," hereafter (e.g., Katz *et al.*, 2001; Sacerdote, 2001; Carrell *et al.*, 2009; De Grip and Sauermann, 2012)[3] or (ii) peer effects without direct interactions – "pure peer effects," hereafter (e.g., Falk and Ichino, 2006; Guryan *et al.*, 2009; Brune *et al.*, forthcoming).

The research area of (i), the closest to the present experiment, suggests positive peer effects. For example, based on random assignment of university students to dorms and roommates, Sacerdote (2001) found that assigned peers have a strong impact on own academic achievement measured by grade point average – see also Carrell *et al.* (2009) who showed similar academic peer effects for random assignment of students to large peer groups called squadrons. In the context of residential neighborhoods, Katz *et al.* (2001) likewise found positive peer effects, based on a research design with random assignment of housing vouchers to poor families. In the context of the workplace, De Grip and Sauermann (2012) studied peer effects among call agents (who work individually) in call centers. They found that work-related training not only improves the worker productivity, but it also leads to an increase in the productivity of coworkers that did not participate in the training program. These positive peer effects can be thought of as having two components: (a) the mere effects from high visibility and the presence of peers, and (b) local social interaction effects.

While these studies successfully provide insights into the role of peers, suggesting that peer effects lead to similar behavioral outcomes among peer mates, the recent literature further advances the mechanisms behind peer effects by exploring how it is affected by team heterogeneity. Its main finding is on the superiority of heterogeneous teams in terms of abilities, and peer effects differ largely by the ability and achievement level. The research discusses that high-ability (-achieving) individuals gain more than low-ability (-achieving) ones in a heterogeneous team (e.g., Lyle, 2009; Duflo *et al.*, 2011; Carrell *et al.*, 2013; Feld and Zölitz, 2017; Booij *et al.*, 2017). This is because the former tend to benefit more from like-minded high types in social interactions within a heterogeneous team, as they *selectively* form sub-teams with like-minded high types, while low-ability (-achieving) individuals are hurt in peer interactions. Such sorting is not possible in homogeneous teams. The proportion of females in a peer group is also known to enhance positive peer effects (e.g., Black *et al.*, 2013; Lavy and Schlosser, 2011; Lu and Anderson, 2015).

Albeit quite convincing, one issue for some randomized field experiments with a large scale, such as Sacerdote (2001), is its internal validity of the research in that there are no controls

[3] Roughly speaking, this can also be called social effects in the term used by Charles Manski (e.g., Manski, 1993).

for the ways in which individuals interact with each other locally. This means that one is unable to underpin what kinds of social interactions exactly triggered the positive peer effects in these studies. The present paper uses a microenvironment that clearly specifies who interacts with whom, and in what way. As discussed in Section 3, each student, divided into either treatment or control groups, has the same learning activities with their assigned partner using a pre-determined pair activity. The intervention may create positive peer interactions locally afterwards to achieve their individual goals.

Another common feature of the prior research is that peer groups are large in most papers: for example, the peer group size is a squadron consisting of 30 students in Carrell *et al.* (2009), 35 cadets per company in Lyle (2007, 2009), and 10-15 students per classroom section in Feld and Zölitz (2017). While a large peer group size is ubiquitous in our real lives, a smaller peer group is also equally common (e.g., pair work in an academic environment such as a classroom, police officers patrolling in pairs). The present paper uses the minimum peer group size – a two-person pair. The use of the smallest size has a methodological advantage as a better control: its setup makes it possible to specify who interacts with whom, while classifying a better or worse performer in each pair, and excluding possible formation of sub-teams. This setup may lead to a result different from the earlier established finding on the heterogeneous treatment effects. For example, the worse performer may benefit more through peer learning since her paired better performer has no choice but to interact with him in the activity. This paper supplements a large body of the prior research with randomized field experiments by providing clean evidence on the role of team heterogeneity in peer interactions when the peer group size is sufficiently small.

It is not possible to identify how large the pure peer effects would be according to literature (i). This question was investigated by using clever setups with high internal validity by various sets of authors, for example, in the context of part-time jobs by Falk and Ichino (2006), pluckers working in an agricultural firm by Brune *et al.* (forthcoming), and professional sports tournaments by Guryan *et al.* (2009). Most studies, such as Falk and Ichino (2006) and Brune *et al.* (forthcoming), found positive pure peer effects, although some studies (e.g., Guryan *et al.*, 2009) did not find such effects. While the sizes of pure peer effects vary largely across the labor markets and different contexts, a meta-analysis in fact showed that pure peer effects would be on average positive (Herbst and Mas, 2015).

2.3. Mutual Learning Theory

A rich body of the empirical literature in labor and personnel economics discussed that social interactions (e.g., communication) among peers and training help improve individual

productivity. For example, Ichniowski *et al.* (1997) documented that innovative human resource management practices used in steel finishing lines improved productivity. The practices include enhanced communication practices among workers and skills training, combined with many other aspects such as high involvement in teams and employment security. Using data from a garment plant, Hamilton *et al.* (2003) found that, while given a choice workers sorted into teamwork, rather than individual work, teams with a greater spread in abilities were more productive under *team*-based remuneration. They discussed that this phenomenon can be explained by an intra-team bargaining model (i.e., high-ability workers impose strong norms) and the mutual learning theory (i.e., high-ability workers teach less-ability ones).

While the prior research was successful in showing the importance of incentives, peer learning and team heterogeneity, it is not possible to measure the effects of peer learning in isolation, since the practices contain at least several dimensions at the same time. It is also not clear how more able workers teach less able workers if each worker is compensated based only on their *individual* performance.

The impact of productivity spillovers was also empirically identified in specific sectors, suggesting that it may be positive. For example, Arcidiacono *et al.* (2017) showed that productivity spillovers among professional players play an important role in the team outcomes in the National Basketball Association. Mas and Moretti (2009), using high-frequency scanner data, showed that introducing a high-skilled cashier to a shift would improve other cashiers' productivity in supermarket chain stores. See also Azoulay *et al.* (2010) for research productivity spillovers among academics and Jackson and Bruegmann (2009) for the case of teaching effectiveness in elementary schools. Nevertheless, it is unclear exactly what kinds of social interactions or learning (if any) triggered positive effects on those involved in the prior studies.

3. Background, Experimental Design and Procedure

The Introduction to Economics module (ECON1101) is a core compulsory module that all the first-year undergraduate students in finance and accounting (a total of 250 to 350 students dependent on the year) must take at Durham University. A "module" is a term used in the United Kingdom to refer to a course. Teaching in a module is organized and implemented by a teaching team. As the very first economics module, students learn the basic principles of economics, and the module serves as a foundation for upper-year core modules in micro- and macroeconomics. Even though it is the first introductory course, because Durham University is highly ranked in the United Kingdom with high quality entrants, students learn some technical aspects, such as mathematical calculations for Cournot competition, the Solow growth model and the IS-LM

model. ECON1101 consists of ten weeks to study microeconomics (term 1) and nine weeks to study macroeconomics (term 2) – see Figure 1. The students' performances are evaluated solely based on one written examination (summative assessment, hereafter) which takes place at the end of the academic year (at the end of May). The maximum (minimum) mark given is 100 (0). Each student will be given an academic grade based on their *own* mark. The grade is: first class (≥ 70), second class (below 70 but ≥ 50), third class (below 50 but ≥ 40), or fail (below 40). Both the raw mark and grade will be written in the student's official transcript.

The summative assessment consists of three parts: Parts A, B, and C. Part A consists of two short-answer questions, each of which accounts for 10% of the assessment mark. Part B (C) consists of two questions from microeconomics (macroeconomics), and each student must select one of the two questions. Parts B and C each account for 40% of the assessment mark. Most questions are essay-type (the summative assessment can be found in Supplementary materials Section A.7). The examination is held online and students need to complete the problem set in a 48 hours window when the problem set is distributed.[4] There is a rigorous word limit: the maximum word count is 3,750 words (markers will stop reading once the maximum word count is reached).[5] Students are instructed not to copy and paste from textbooks or lecture notes (copies and pastes are not given marks). A plagiarism check is also performed for each script by the undergraduate office and also by the Turnitin software. Kamei was the module leader for ECON1101 for the 2019/20 academic year and Ashworth was the department head when the randomized field experiment was planned before the academic year started (he was the department head from 2016 to 2019). The experiment was designed and implemented using the students of this module for the 2019/20 academic year.

Students have four key learning activities. The first one is weekly two-hour lectures. Kamei delivers lectures in term 2, while another faculty member does so in term 1.[6] The lectures are held in a large lecture hall, and all students take the same sessions. The lectures are designed to introduce the key economic concepts and methods, and to present the technical analysis in action. They are always accompanied by presentation slides (and sometimes also mathematical

[4] The summative assessment is released on May 25, 2020 at 9 am, with the deadline being May 27 at 9 am.
[5] Marking is operated through a rigorous double-marking procedure. The university appoints first markers and second markers in this module. When each first marker finishes marking their assigned set of scripts, second markers independently determine marks on randomly selected scripts (see the University's Learning and Teaching handbook).
[6] Lecturing of macroeconomics is the only teaching duty of Kamei. Kamei is responsible for the management of the module (e.g., coordinating with the other lecturer in term 1, monitoring the teaching work of seminar tutors during terms 1 and 2). The allocation of teaching and the make-up of teaching teams and duties within the teams are determined by the department with the members of staff informed of the team and responsibilities for the model prior to the start of the 2019/20 academic year.

handouts) whose electronic files are distributed to every student through a Blackboard (DUO) prior to the lectures. The lecture time is not designed as interactive (rather one-way delivering of key concepts), although students have some opportunities to raise issues and confirm their understanding of the analysis. Other than lectures in terms 1 and 2, there are two additional lectures (one for microeconomics, and the other for macroeconomics) in term 3, i.e., revision periods, as reviews before the summative assessment.

Second, students have one-hour mandatory seminars in every other week (a total of four seminars for microeconomics in term 1 and another four for macroeconomics in term 2). The students in accounting and finance are allocated to one of the 16 seminar groups by the undergraduate office.[7] Each seminar group has around 15 to 20 students. There are three tutors in the module. One tutor is responsible for seven seminar groups in term 1, while another tutor is responsible for the other nine groups in term 1.[8] The third tutor is responsible for all 16 seminar groups and uses the same instruction across the 16 groups in term 2. This would minimize the possible effects of common shocks (e.g., Lyle, 2007). Students' learning could, nevertheless, be affected by their assigned seminar group since they have social interactions there. Hence, seminar group clustering is included for all data analyses such as regression analyses in Section 5.

A problem set is distributed to students one week prior to a given seminar session, and students are expected to attempt those questions before the class. Seminars are designed to be interactive. While a seminar tutor explains the answer, students are also invited to discuss their answers in their seminar group. This means that the unit of independent observation is seminars. Students' attempts before the seminar are not checked (nor are these activities subject to students' final marks). However, attendance at seminars is mandatory and registers are taken in all sessions as an important academic commitment.[9]

Third, students have one formative assessment – simply "formative" hereafter – in each term (two pieces in total). Students must answer a problem set and submit their answers officially to the university. The term 1 formative asks questions on the producer theory, while the term 2 assessment asks those on the short-run macroeconomics – see the problem set in Supplementary materials Section A.3 and A.5. The seminar tutors mark the scripts online, and

[7] The undergraduate office (using a University algorithm) is fully responsible for the allocation. Anyone in the teaching team, including the module leader (Kamei), is not at all involved in the allocation process.
[8] As students' performances might be affected by the difference in the seminar tutor in term 1, term 1 tutor assignment is controlled by having a dummy variable when estimating treatment effects in data analysis (Section 5).
[9] Students were also informed: "Failure to attend without a prior arrangement will be noted and any student who misses a seminar without having made a prior arrangement should attend their tutor's next consultation hour. Persistent absences will result in our instigating formal monitoring processes."

provide individual feedback to each student. While the marks are not be counted towards their final marks, they are recorded in the students' information sheet in the university. The formative aims to help students understand the material and its applications in a structured way, as well as help them prepare for the summative assessment. An intervention was made in term 2 utilizing the students' performances in the term 1 formative (further details below).

The fourth activity is a peer review assessment (PRA), for which a randomized control trial is implemented (the above three learning activities are identical to all students). This activity starts earlier than the formative assessment in each term. In the PRA activity, students are first distributed a problem set whose format is the same as the formative (e.g., on October 28, 2019 for term 1), attempt the problems independently, and then submit their scripts officially to the university (e.g., the submission deadline is November 8, 2019 for term 1). The term 1 peer review assessment asks questions on the consumer theory, while the term 2 assessment asks those on long-run macroeconomics. Notice that the topic of the PRA is different from that of the formative. The problem set can be found in Supplementary materials Section A.2 and A.4. Once the submission deadline passes, students are informed of their pair partner (e.g., on November 11, 2019 for term 1). The seminar tutors collect students' scripts of the respective seminar groups from the undergraduate office and give students their partners' scripts during the following seminars. Each pair works together to discuss the problem set and their scripts. The procedure uses a proforma prepared by the authors – see Supplementary materials Section A.6 for the proforma in term 2 as an example. Specifically, each student critically assesses their partner's script by completing Part A of the proforma for each question, has a meeting in their pair, gives their partner the proforma so that the partner can write afterwards what they learned from the activity, and also jointly decide on an agreed mark for the script. Students must submit the proforma officially to the undergraduate office. The students go through the whole process *without* seeing answer sheets of the PRA problem set: a solution to the problem set is distributed to students in each term *after* the deadline of proforma submission passes. Each pair is encouraged to study and find answers by themselves if both the students in the pair did not solve the problem set. The timeline of the assessment, along with other module activities, can be found in Figure 1 (also in Supplementary materials Section A.1).

In term 1, in the PRA activities, all students are randomly broken into pairs in their respective seminar groups. When the number of students in a given seminar is odd, there is one team consisting of three students. In term 2, by contrast, seminar groups are randomly assigned either the random matching condition ("treatment condition") or the sorting condition ("control condition") so that the number of seminar groups in each condition is 8 out of 16. On the one

hand, students in a group with the treatment condition are randomly broken into pairs as in term 1. The pairing is completely random, implemented through computer random number generation.[10] On the other hand, students in the control condition are sorted in descending order according to their term 1 formative assessment marks, and pairs (teams) are formed so that their marks are adjacent to each other.[11] Any student is *not* aware of this paring process.

In order to study how the PRA activities affect their focus on preparing for the summative assessment, one question comes from the consumer theory (long-run macroeconomics) and the other comes from the producer theory (short-run macroeconomics) in Part B (Part C) in the summative assessment. In other words, the topics of the four questions appearing in Parts B and C are covered by the two formatives and two PRA in the module. The students were not given any information regarding which topics would be tested in the summative assessment.

To maximize the external and internal validity of the project, it is essential for students to engage in the learning activities without knowing the presence of on-going experiments or the matching differences by seminar group. The institutional review board (IRB) at Durham University, however, asked the authors to seek consent from the students and explain a possible research activity. As a compromise, the research team simply includes a generic consent statement in the proforma of the term 1 peer review assessment without writing any substance of the research as follows:

> Consent:
> Your assessment marks may be used for the purpose of further research and to enhance the learning experience for the programme. I consent this possibility.
> Your Signature: _____

The performances of only those who gave us consent are used for the study (this procedure has been approved by the IRB).[12] Hence, students in the module are not aware of any experimental aspects. Students are simply explained the learning objective of the PRA activities using the materials in Supplementary materials Section A.1.

To further enhance the internal validity of the experiment, all module teaching team members perform lecturing, tutoring and consultation following the module outline and the

[10] There is an exception in which the same four pairs as in term 1 were formed by chance and hence the pairing was further randomly changed to ensure that all students had different partners for terms 1 and 2.
[11] The same eight pairs as in term 1 was formed by sorting. Their pairing was adjusted by the authors, swapping the partners among similar interim marks so that all students had different partners.
[12] Another condition in obtaining the IRB approval is for us not to make any student's proforma publicly available.

requirement set by Kamei, without being informed of the on-going experiment. Moreover, all administrative staff members in the undergraduate office are likewise uninformed of the presence of the experiment. Note that Kamei does not serve as a tutor of any student in the module. Seminar tutors deal with all aspects of the peer review assessment activities as well as direct interactions with students in seminars. Kamei's minimum interactions with the students help minimize a possible unconscious bias that might have occurred had he tutored any of the students.

4. Hypothesis

Either a behavioral model that incorporates social effects, such as shame and pride (Section 4.1), or a behavioral model with interdependent motives (Section 4.2), predicts that (i) students perform better on the summative assessment in pairs with a greater spread in interim performances, and (ii) the worse performers in the treatment condition improve performances through working with their matched better performers, in the setup of this study. Both the motives would be plausible ones. This section mathematically illustrates the mechanism of each motive in isolation to derive the hypothesis of the paper based on behavioral effects.

4.1. Social Effects among Peer Mates

Assume first the following payoff functional forms to describe the situation prior to engaging in the peer review assessment activities for high type (*h*) and low type (*l*):

$$G_h(y_h) = g(y_h) - \gamma_h {y_h}^2. \tag{1}$$

$$G_l(y_l) = g(y_l) - \gamma_l {y_l}^2. \tag{2}$$

Here, $g(y_i)$ is the student's interim performance gauged by the formative in the module, $y_{i\in\{h,l\}}$ is the effort put so far by type *i* to learn class materials, and $\gamma_i {y_i}^2$ is the cost associated with the effort provision, and the cost function is assumed to be quadratic. For simplicity, further assume that:

$$g(y_i) = \alpha_i + \beta_i y_i. \tag{3}$$

Parameter values are set so that $\alpha_h > \alpha_l > 0$, $\beta_h > \beta_l > 0$, and $0 < \gamma_h < \gamma_l$ (i.e., the high type has higher returns from study effort and lower unit effort costs than the low type).

Each type's optimal interim effort provision can then be derived by using the first-order conditions for (1) and (2):

$${y_h}^* = \frac{\beta_h}{2\gamma_h}. \tag{4}$$

$${y_l}^* = \frac{\beta_l}{2\gamma_l}. \tag{5}$$

Note that $\frac{\beta_h}{2\gamma_h} > \frac{\beta_l}{2\gamma_l}$ and the second-order conditions are satisfied. Hence, it can be predicted that the high type exerts stronger effort and has a better interim achievement level than the low type.

As the PRA activities help improve each other's performance through mutual learning, the payoff function after the activities can be re-written for each type as below:

$$\pi_h(y_h, e_h|e_l) = G_h(y_h) + a_l e_l + b_h e_h - c_h {e_h}^2. \quad (6)$$

$$\pi_l(y_l, e_l|e_h) = G_l(y_l) + a_h e_h + b_l e_l - c_l {e_l}^2. \quad (7)$$

In these equations, $e_{i\in\{h,l\}}$ is the effort level provided by *i* to teach his/her partner through critical assessment of the partner's script and discussions. The cost function is assumed to be quadratic for peer learning as well, i.e., $c_i {e_i}^2$, such that $c_l > c_h$ (the unit effort costs in the PRA activities are higher for the low than for the high type). The term $b_i e_i$ refers to student *i*'s own benefit from engaging in the PRA activities. By contrast, $a_l e_l$ ($a_h e_h$) is the benefit that the high (low) type receives from her (his) matched low (high) type in the pair, and $a_h > a_l$.

Equations (6) and (7) suggest that the PRA activities do not change each type's optimal effort choice decision regarding y_i, because $\frac{\partial \pi_i(y_i, e_i|e_j)}{\partial y_i} = \frac{\partial G_i(y_i)}{\partial y_i}$ for $i = h, l$. However, the optimal choice would change if it is additionally assumed that each type receives a positive (negative) utility when their own class performance is better (worse) than their pair partner's. For an illustrative purpose, consider the following utility function:

$$\theta_h(y_h, e_h|e_l) = \pi_h(y_h, e_h|e_l) + \mu_h[\{g(y_h) + a_l e_l + b_h e_h\} - \{g(y_l) + a_h e_h + b_l e_l\}]^2. \quad (8)$$

$$\theta_l(y_l, e_l|e_h) = \pi_l(y_l, e_l|e_h) - \mu_l[\{g(y_h) + a_l e_l + b_h e_h\} - \{g(y_l) + a_h e_h + b_l e_l\}]^2. \quad (9)$$

The equation in the squared bracket ($= \{g(y_h) + a_l e_l + b_h e_h\} - \{g(y_l) + a_h e_h + b_l e_l\}$) is the intra-pair difference in the achievement level after the PRA activities. The assumption here is that the high (low) type has feelings of pride (shame), which leads to a positive (negative) utility. μ indicates each type's utility weight on the social effects. μ_h and μ_l are both positive, such that $\mu_h < \mu_l$. This condition means that the impact of shame is stronger than that of pride. The presence of such social effects does influence their self-study efforts. This can be seen by using the first-order conditions for utilities (8) and (9) as follows:

$$\frac{\partial \theta_h}{\partial y_h} = \beta_h - 2\gamma_h y_h + 2\mu_h \beta_h [\{g(y_h) + a_l e_l + b_h e_h\} - \{g(y_l) + a_h e_h + b_l e_l\}] = 0. \quad (10)$$

$$\frac{\partial \theta_l}{\partial y_l} = \beta_l - 2\gamma_l y_l + 2\mu_l \beta_l [\{g(y_h) + a_l e_l + b_h e_h\} - \{g(y_l) + a_h e_h + b_l e_l\}] = 0. \quad (11)$$

Notice that

$$\frac{\partial \theta_h}{\partial y_h}\Big|_{y_h{}^*=\frac{\beta_h}{2\gamma_h}, y_l{}^*=\frac{\beta_l}{2\gamma_l}} = 2\mu_h\beta_h\left[\alpha_h - \alpha_l + \frac{\beta_h{}^2}{2\gamma_h} - \frac{\beta_l{}^2}{2\gamma_l} + a_l e_l + b_h e_h - a_h e_h - b_l e_l\right].$$

$$\frac{\partial \theta_l}{\partial y_l}\Big|_{y_h{}^*=\frac{\beta_h}{2\gamma_h}, y_l{}^*=\frac{\beta_l}{2\gamma_l}} = 2\mu_l\beta_l\left[\alpha_h - \alpha_l + \frac{\beta_h{}^2}{2\gamma_h} - \frac{\beta_l{}^2}{2\gamma_l} + a_l e_l + b_h e_h - a_h e_h - b_l e_l\right].$$

Here, the squared bracket (the difference in achievement level between the high and low types after the PRA activities) is considered as still positive because $\alpha_h > \alpha_l$, $\frac{\beta_h{}^2}{2\gamma_h} > \frac{\beta_l{}^2}{2\gamma_l}$ and $0 < \mu_h < \mu_l$. This suggests that the optimal self-study effort exerted by type *i* (denoted as $y_i{}^{**}$) is greater than $y_i{}^* = \frac{\beta_i}{2\gamma_i}$. This means that both the high and low types work harder driven by the social effects. Conditions (10) and (11) can also be simplified to:

$$\frac{y_h - \frac{\beta_h}{2\gamma_h}}{y_l - \frac{\beta_l}{2\gamma_l}} = \frac{\frac{\mu_h\beta_h}{\gamma_h}}{\frac{\mu_l\beta_l}{\gamma_l}} = \frac{\mu_h}{\mu_l}\frac{\gamma_l\beta_h}{\gamma_h\beta_l}.$$

By the assumptions on β and γ, $\frac{\gamma_l\beta_h}{\gamma_h\beta_l} > 1$. Thus, the above condition implies that the low type shows a stronger improvement than the high type if the effect of shame is large enough that $\mu_l > \frac{\gamma_l\beta_h}{\gamma_h\beta_l}\mu_h$.

These analytical implications can be summarized as Proposition 1 below. Proposition 1 suggests that peer review activities with a greater spread in interim performances lead to an stronger effect than those with similar performances, as the social effects operate strongly in the former than in the latter.

Proposition 1: *Suppose that the high (low) type receives a positive (negative) utility due to feelings of pride (shame) through the PRA activities. Then, the PRA activities encourage both the high and low types to study harder to improve own understanding of class materials, accordingly resulting in performance improvements. The positive effect on the low type is stronger than on the high type if the low type is concerned about shame large enough that* $\mu_l > \frac{\gamma_l\beta_h}{\gamma_h\beta_l}\mu_h$.

4.2. Interdependent Motives

The mutual learning hypothesis states that high-ability workers teach less-ability ones, thereby improving the performances of especially the latter. To see this, for tractability, let us simplify the payoff functions (6) and (7) as follows:

$$\pi_h(e_h|e_l) = k_h + a_l e_l + b_h e_h - c_h e_h{}^2. \quad (12)$$

$$\pi_l(e_l|e_h) = k_l + a_h e_h + b_l e_l - c_l {e_l}^2. \quad (13)$$

Here, k_h and k_l are the students' interim performances before engaging in the PRA activities, such that $k_h > k_l > 0$. Each type's optimal effort provision in the PRA activities can be derived by using the first-order conditions for (12) and (13):

$${e_h}^* = \frac{b_h}{2c_h}. \quad (14)$$

$${e_l}^* = \frac{b_l}{2c_l}. \quad (15)$$

Notice that ${e_h}^*$ and ${e_l}^*$ are each dependent only on their own payoff parameters. This means that without considering interdependent concerns, each type's effort provision in peer learning would not be affected by their partner's ability or payoff parameters.

In order to show how the prediction may change by the inclusion of interdependent motives, as an illustration, assume now that each type's utility function can be expressed as follows:

$$u_h(e_h|e_l) = \pi_h - x_h \cdot (\pi_h - \pi_l)^2. \quad (16)$$

$$u_l(e_l|e_h) = \pi_l - x_l \cdot (\pi_l - \pi_h)^2. \quad (17)$$

This means that a student prefers to have *similar* academic performances between pair mates. x_i is the utility weight on the interdependent preferences. Each type's optimal effort provision can again be derived by differentiating utilities (16) and (17) with respect to e_i:

$$\frac{\partial u_h}{\partial e_h} = [1 - 2x_h \cdot (\pi_h - \pi_l)](b_h - 2c_h e_h) + 2a_h x_h(\pi_h - \pi_l) = 0. \quad (18)$$

$$\frac{\partial u_l}{\partial e_l} = [1 - 2x_l \cdot (\pi_l - \pi_h)](b_l - 2c_l e_l) + 2a_l x_l(\pi_l - \pi_h) = 0. \quad (19)$$

For simplicity, further assume that x_h and x_l are small enough that the interdependent preferences do not reverse the intra-pair relative payoff standing, i.e., $\pi_h > \pi_l$, in equilibrium. Condition (18) suggests that the optimal effort level of high type *h* (${e_h}^{**}$) is greater than ${e_h}^*$ if x_h is sufficiently large. By contrast, Condition (19) suggests that the optimal effort level of low type *l* (${e_l}^{**}$) is always lower than ${e_l}^*$. Hence, the students' interdependent motives help shrink the intra-pair performance difference through the PRA activities under certain conditions.

<u>The size of intra-pair interim performance difference (i.e., $k_h - k_l$) and peer learning outcome:</u>

With this framework, it can be shown that a larger intra-pair difference in the interim performance induces the high type *h* to put more efforts in improving her matched low type. To show this, for simplicity, assume the following:

Assumption 1: $b_h = b_l = 0$.

Assumption 2: $c_h = c_l = c$, but $a_h > a_l$.

Assumption 1 means that effort provision by the high type in the PRA activities is purely costly without resulting in her own private benefits. Assumption 2 is merely a normalization: while the unit effort cost is the same for the high and low types, the impact of the PRA activities (a_h, a_l in Equations (12) and (13)) differs by the type.

Proposition 2: *Consider Assumptions 1 and 2. Then, the larger interim performance difference a pair has (i.e., the higher k_h, relative to k_l, the high type has), the greater effort the high type exerts in improving the performance of her matched low type.*

Proof: Assumption 1 implies that $e_h{}^* = e_l{}^* = 0$ from Equations (14) and (15), and therefore $\pi_h(e_h{}^*|e_l{}^*) = k_h > k_l = \pi_l(e_l{}^*|e_h{}^*)$.

Under these two assumptions, Condition (18) reduces to the following:

$$f(e_h|k_h, k_l, a_h, c) \coloneqq -ce_h + [2ce_h x_h + a_h x_h](k_h - ce_h{}^2 - k_l - a_h e_h) = 0 \qquad (20)$$

Appling the Implicit Function Theorem to (20), it can be found that the larger k_h the high type *h* has, the greater effort level *h* exerts in improving the performance of the low type:

$$\frac{\partial e_h}{\partial k_h} = -\frac{\partial f/\partial k_h}{\partial f \partial e_h} = -\frac{2ce_h x_h + a_h x_h}{-c + 2cx_h(k_h - ce_h{}^2 - k_l - a_h e_h) - (2ce_h x_h + a_h x_h)(2ce_h + a_h)}$$

$$= -\frac{2ce_h x_h + a_h x_h}{-c + \frac{2c^2 e_h}{2ce_h + a_h} - (2ce_h x_h + a_h x_h)(2ce_h + a_h)}$$

(Note: the above equality is obtained by using Condition (20))

$$= \frac{2ce_h x_h + a_h x_h}{\frac{ca_h}{2ce_h + a_h} + (2ce_h x_h + a_h x_h)(2ce_h + a_h)}$$

> 0. □

5. Results

284 students enrolled in the module at the beginning of the academic year. Almost all the students remained in the module when final module registrations emerge after the student review period.[13] The total number of students was 279 at the beginning of term 2. 92.8% of them gave consent for their data to be used for possible research. As a result, the subject pool includes 129 and 130 students in the treatment and control conditions, respectively. Table 1 summarizes students' performances in term 1 formative assessments. It indicates that almost all the students

[13] This is as expected since ECON1101 is a compulsory module.

submitted the formative assessments. This is not a surprise since the submission of the formative assessment was compulsory. The average mark of the formative assessment was 70.38 in the treatment condition, somewhat lower than that in the control condition (72.05). However, the difference is not statistically significant according to a Somers' D test with seminar group clustering (two-sided $p = 0.721$).[14] This means that the random allocation of matching conditions was successful. In the subject pool, 123 and 127 students in the treatment and control conditions, respectively, completed summative assessments by the deadline.

Table 1 also includes information on the proportions of female students and those of British students. It shows that the proportion of female students is somewhat smaller in the treatment than in the control condition. The proportion of British students is somewhat larger in the former than in the latter. Prior research suggests that having a higher proportion of female peers helps improves performance (e.g., Black *et al.*, 2013; Lavy and Schlosser, 2011; Lu and Anderson, 2015). If this is applicable for the student pool of this study, students may tend to achieve better academic performances in the control than in the treatment condition in the end.[15] However, the differences in these proportions are not significantly different. Thus, these demographics are sufficiently balanced between the two conditions.

As explained in Section 3, students' attendance in bi-weekly seminar activities were set mandatory in the module. Some students missed the seminars, nevertheless (Figure 2). The average attendance rates show similar trends for the treatment and control conditions. The rates were high at the beginning of each term, and then gradually declined from seminar to seminar. The attendance rates were around 70% (somewhat over 60%) in the fourth seminar in term 1 (term 2). A Somers' D test with seminar group clustering found that the difference in the average seminar attendance rate was not significant between the treatment and control conditions at two-sided $p = 0.470$. This means that students' exposure to the seminar activities were also balanced.

The learning outcome in this module can be measured based on students' performances in the summative assessment at the end of the academic year. The data show that despite the slightly weaker interim (term 1 formative) performance in the treatment than in the control condition (Table 1), students in the treatment condition achieved stronger performance in the summative assessment. As shown in panel A of Figure 3, the difference in the average mark between the two matching conditions was around three points and is significant at two-sided $p = 0.031$ according to a Somers' D test with seminar group clustering. In particular, while students

[14] Testing based on Somers' D is identical to the Mann-Whitney test if clustering is not included.

[15] As will be explained in this section, students performed significantly better in the treatment than in the control condition despite the somewhat lower percentage of female students in the former.

in the bottom 10% showed similar achievements in the two conditions, the rest (90% of students) showed higher achievements in the treatment than in the control condition.

In term 2, a small fraction of students did not submit peer review assessments.[16] Even if students did not submit the PRAs, they were still encouraged to meet with their partners and discuss the problem set within the pairs as an academic commitment, meaning that there might have still been some effect. However, having an effective discussion could be difficult without having their partners' scripts. Hence, it would be useful to study a possible treatment effect while limiting data to pairs in which both pair mates submitted the PRAs. Panel B of Figure 3 reports the cumulative distributions for the restricted dataset. It shows an almost similar pattern to panel A and the performance difference is significant at two-sided $p = 0.034$ according to a Somers' D test with seminar group clustering. This suggests that the size of the treatment effect was not affected by the omission of those who were not able to complete the PRA activities.

Care needs to be exercised when formally studying the treatment effect of intervention, especially because not all students completed the summative assessments (i.e., submitted by the deadline). The data indicate that among the 259 students who gave consent, nine students (3.47% of the subject pool) did not complete the assessment. No submission for "good cause" is treated as different from zero marks in the exam with the university.[17] As their marks were unobserved, a Heckman two-stage selection model was used to control for a possible impact of the selection bias although the effect of selection bias seems to be very small. Considering that students' effort levels put in the module may explain their decisions to complete the exam, their attendance rates in seminar activities and submission records of formative assessments are included as independent variables in the first-stage selection equation.[18] As shown in Table 2, the model was estimated when using all data (columns I.i and I.ii) or using students who submitted the term 2 PRAs (columns II.i and II.ii). The t2 (term 2) random matching dummy is included in all the specifications to estimate the effect of treatment intervention.

The estimation first reveals that, as expected, students' efforts exerted in the module are good predictors for their completion of the summative assessment: the number of formative

[16] 11 students (4.23% of the students in the subject pool) did not submit the assessments in term 2.

[17] For example, if a student had a valid reason, such as illness, for not attending, there would be a blank in their transcript; they can take a resit exam as their first attempt. The authors do not have accessible data regarding reasons for not attending the exam or resit marks.

[18] A given student's seminar attendance rate was calculated based on the eight seminar activities in the module. It should be acknowledged that in one seminar group, the seminar tutor failed to take attendance in the eighth seminar; thus, the attendance rates of students in that group were calculated based on the records in the other seven seminars. Results reported in this paper do not change qualitatively even if students' attendance rates are calculated based on the seven seminar activities for all seminar groups.

assessments not submitted has a significantly negative coefficient in the first-stage selection equation. In addition, as shown in columns I.i and I.ii, the seminar attendance rate variable has a significantly positive coefficient when all eligible subjects are considered.

Second, and most important, regardless of whether the students' term 1 formative marks are controlled for, the t2 random matching dummy consistently has a significantly positive coefficient in the second stage regression. This suggests that working in pairs with different abilities leads to a higher performance than pairing students whose achievement levels were similar to each other, in support of the view from the peer-effect hypothesis and the mutual learning theory.[19]

Result 1: *Students in pairs with a great spread in interim achievement levels performed better on the summative assessment compared with those in pairs whose interim performances were similar to each other.*

What drove the positive impact of the PRA activities in term 2? The only difference between the two conditions is the use of random matching or sorting in the pairing process. Students in the control (sorting) condition were divided into pairs so that their term 1 formative assessment marks were similar to each other. By contrast, pairing was randomly formed for students in the treatment (random matching) condition. With this difference in the matching protocol, the average intra-pair absolute difference in the formative mark was more than four times in the treatment than in the control condition: it was 25.16 (5.81) marks with clustered standard errors of 3.28 (1.07) marks in the former (latter). Panel A of Supplementary materials Figure B.2 reports the histogram of absolute individual performance differences by the matching condition. It clearly indicates that the differences spread widely to the right in the treatment condition, while they are concentrated around 0 in the control condition. Hence, these patterns confirm that, as intended, pairs in the treatment condition had a greater variation in interim achievement levels between pair mates, compared with those in the control treatment.

In order to study how the treatment effect differs by the intra-pair relative term 1 performance standing in the treatment condition, the data in the treatment condition was split

[19] The gender composition in pairs may affect the size of peer effects (e.g., Black *et al.*, 2013; Lavy and Schlosser, 2011; Lu and Anderson, 2015). However, in principle, individual characteristics do not need to be controlled in this study since matching conditions were randomly assigned to groups and the proportion of female students was not significantly different between the treatment and control conditions (Table 1). Nevertheless, an additional regression was conducted as a robustness check while controlling for available demographic variables (own gender, pair partner's gender, and the interaction between the two gender variables, students' nationality). The estimation found qualitatively similar results, i.e., positive effects of having a greater spread in abilities in a pair – see Supplementary materials Table B.1 for the detail.

into two sets: the "better performers" and the "worse performers." The better (worse) performer is defined as a student whose term 1 formative mark is better (worse) than his/her matched partner'.[20] The average term 1 formative marks of the better and worse performers were 80.35 and 57.81, respectively, in the treatment condition (the average mark in the control condition was in the middle between the two, and it was 72.05 as already discussed in Table 1) – see also panel B of Supplementary materials Figure B.2. Intriguingly, the performance data in the summative assessment by the relative standing reveal that the better and worse performers had similar achievements in the end, scoring 71.31 and 71.94 points, respectively, each of which was better than the average summative performance under sorting (Table 1).[21] This seems to suggest that working in a team with a greater spread in abilities strongly supports the learning of the worse performer.

A regression was performed by including two indicator variables – the better performer and the worse performer dummies – as independent variables, to formally study the role of relative performance standing in the treatment condition. Having the two dummies make it possible to identify how the aggregate positive effect of term 2 random matching (Result 1) differs by student's interim achievement standing. The reference group is the students in the control condition.

Table 3 reports the estimation results. The regression only uses the data from pairs in which both the pair mates submitted term 1 formative assessments, since otherwise it is not possible to judge which student was the better or worse performer in the interim stage. A Heckman selection model was again used in the analysis in order to deal with unobserved summative marks of some students. As shown in columns I.i and II.i, when only the two performance dummies are used as independent variables in the second stage equation, the dummies obtain weakly significantly or significantly positive coefficients. This implies that not only the worse but also the better students might have similarly benefited from the PRA activities in the treatment condition. However, this interpretation is misleading. As shown in columns I.ii and II.ii, once students' term 1 formative assessment marks are controlled for, only the worse performer dummy obtains a significantly positive coefficient at the 5% level or better. It follows that the significant coefficients for the better performer dummy in columns I.i and II.i

[20] There was no student whose term 1 formative mark was exactly the same as his/her partner's.

[21] The absolute size of marks are not perfectly comparable between the formative and summative assessments because the former has only one problem set while the latter has multiple problem sets.

are driven by the differences in the term 1 formative marks, and the worse performers in pairs are the ones that mainly benefited from the PRA activities.[22]

It should be worth noting that the coefficient estimates for the better performer dummy are not negative, but close to zero. This means that the better performers were not hurt by being matched with the worse performers.

The empirical results from columns I.ii and II.ii of Table 3 can be rationalized by a frequently-used behavioral model that assumes that the utilities of students may be influenced by some social effects, such as shame, guilt, and pride, or some interdependent preferences, such as inequality aversion (see Section 4 for an illustrative theoretical analysis). This is because under such assumptions, the worse performer in a pair has more non-material incentives than the paired better performer to improve performance due to feelings of shame under certain conditions. It can also be predicted that consistent with the mutual learning hypothesis, the more able have interdependent motives to exert efforts in improving her matched student's academic performance in the PRA activities. Which motives matter more, social effects such as shame, or the mutual learning motives, would depend according to the students' characteristics and preferences, and the contexts.

Result 2: *The worse performers in the treatment condition improved performance through working with the better performers in the peer review assessment activities.*

While the average treatment effect was quite strong on the worse performers, one may wonder more precisely how their improvements depend on the sizes of intra-pair interim achievement differences. For example, if the abilities are too different among pair mates, improvements may be weak as they may not be able to effectively communicate with each other due to the lack of basic technical skills or fundamentally different work attitudes. Alternatively, the weaker the interim performance they have compared with the better performers, the worse performers may benefit more, considering that they have more room for improvement and also even worse performers have certain competency being admitted to the university based on the admission criteria.[23] This question can be examined by looking at the relationship between (a)

[22] With the same reason as written in footnote 198, an additional regression was conducted while controlling for demographic information as a robustness check, finding almost the same as in Table 3 (see Supplementary materials Table B.2).

[23] The students' pre-university data, such as A-level grades, were unavailable due to its confidentiality, making it impossible to study possible effects of pre-university achievement levels. Having said this, these background data were in any case not used in any aspect of the module. The unavailability of the data would also not affect the findings of the paper since seminar group allocations of students were random (footnote 7) and the background data were not required for pairing in the PRA activities.

student *i*'s performance relative to his/her pair partner *j*' in the term 1 formative assessment (i.e., $x_i = f_i - f_j$) and (b) *i*'s performance improvement gauged by the summative assessment relative to *i*'s interim mark (i.e., $y_i = s_i - f_i$).[24]

Figure 4 reports the relationship by matching condition. Three interesting patterns emerge. First, the more behind a student was in the interim stage, the larger improvement she achieved in the end through the term 2 PRA activities in the treatment condition (see the region where $x_i < 0$ in panel A). This resonates with the idea that peer learning is an effective way to improve the performance of poorer workers. This tendency is also seen for some small number of pairs in the control condition where intra-pair interim performance difference happened to be large despite the sorting process (see the region where $x_i < 0$ in panel B). Second, the peer-learning benefit of better performers did not depend on the size of the interim performance differences they had compared with their matched weak performers. In the region where $x_i > 0$ of panel A, the slope of the fitted curve becomes flatter as *x* becomes large. The small elasticity of the better performers' marks may be driven by a ceiling effect (the maximum mark is bounded above at 100). This also importantly means that forming pairs with larger ability differences would not hurt the better performers. Third, and as anticipated, observations were more concentrated around $x_i = 0$, and the variation in *y* was much smaller in the control than in the treatment condition. Specifically, the standard errors (with seminar group clustering) were 2.414 and 1.568 marks in the treatment and control treatments, respectively. This suggests that the peer-learning effects are more homogeneous in the control than in the treatment condition.

The impact of the PRA activities can also be seen in the students' choices in the summative assessment. Students selected one of the two questions in Part B, and likewise in Part C. Among the two questions in each Part, one question came from the topics in the PRA, while the other came from the topics in the formative. In Parts B and C of the summative assessment, strikingly, 98.0% and 91.6% of the students, respectively, selected questions whose topics were covered by the PRA (Figure 5.A). Hence, the PRA activities are more effective in deepening the learning and/or enhancing their study motivations than the formatives that are simply marked by tutors.

A closer look at students' performances by the Part reveals two further interesting patterns (Figures 5.B and 5.C). First, the students in pairs with a great spread in abilities

[24] As noted in footnote 21, the absolute size of the summative mark (s_i) is not fully comparable to that of the term 1 formative mark (f_i), whose aspect makes the interpretation of the size of y_i difficult. However, y_i {$= s_i - f_i$} is still a nuanced measure of academic improvement for across-subject comparisons. For example, given the value of *f*, the higher academic improvement a student has, the greater *s* and therefore the greater *y* a student has.

(treatment condition) performed markedly better in the long-run macro question in Part C, the topic of term 2 PRA, relative to those in pairs with similar interim performances (control condition). It should be acknowledged that students' marks for the short-run macro question were not that different between the two conditions as seen in panel C, whose result might have been driven by its small sample size (panel A).

Second, and equally important, the positive effect of term 2 peer learning spills over to their learning of term 1 materials. Those in the treatment condition performed much more strongly in Part B than those in the control condition, whether they selected the consumer theory question (the topic in term 1 PRA) or the producer theory question (the topic in term 1 formative). Notice that each student had two PRA partners – one for term 1 activities, and the other for term 2 activities. There were no differences in the partner assignment procedure for the term 1 activities between the treatment and control conditions.

Lastly, students performed well in the short-answer questions in Part A of the summative assessment regardless of the matching condition: their average mark in Part A does not significantly differ by the matching condition (14.89 for the treatment condition; 14.61 for the control condition). This means that students in both matching conditions likely put in sufficient effort, being able to grasp the basic concepts and applying the concepts to economic questions.

In order to formally study treatment effects on students' performances in Part B, and also in Part C, a Heckman selection model was estimated (Table 4). The dependent variable is either their performances in Part B (column I) or Part C (column II), and the independent variables are the same as in Tables 2 and 3. The model was estimated without splitting the data further to those who selected Question 3 or 4 (Question 5 or 6) because the attempt here is to study the overall impact on their performance on the term 1 (term 2) materials. The estimation clearly shows that students performed better in the treatment than in the control condition for both Parts B and C – see the t2 random matching dummy variable. Consistent with Result 2, the positive effect under term 2 random matching was driven by an improved performance exhibited by the worse performer in the pairs – see columns I.iv and II.iv.

As a further robustness check, nevertheless, the same model was also estimated when using students' marks in the PRA materials, namely Question 3 (Question 5), as the dependent variable. It finds almost the same results as Table 4 – see Supplementary materials Table B.3.[25] Hence, it can be concluded that the spill-over effects of term 2 peer learning on term 1 materials

[25] Additional regressions were further performed to study the impact of term 2 random matching on the students' scores in the formative materials, i.e., Question 4 in Part B (Question 6 in Part C). However, the selection model was not able to be estimated due to a small number of selected data.

are significant. The spill-over effects can be thought of as being driven indirectly by the worse performers' improved learning in the PRA discussions, and/or by their study habits reformed through the PRA experiences.

Result 3: *The students in the treatment condition performed better than those in the control condition not only for Part C but also for Part B of the summative assessment. This suggests that the positive treatment effects of pair heterogeneity through the term 2 PRA activities spillover to their learning of the term 1 materials.*

5. Conclusion

Using a randomized field experiment in a classroom, this paper found that exogenously pairing two individuals with different interim achievement levels leads to better overall performance through peer learning, compared with pairing together those whose achievements are similar to each other, even though average achievement levels at the interim stage were similar for the two kinds of peer groups. The positive impact of pair heterogeneity was driven by a strong performance improvement of the less able, rather than of the more able, in a pair. The more able was also not hurt by being matched with the less able. This result is consistent with the bargaining and mutual learning hypotheses: more able workers can not only impose productive norms for the sake of their teams, but they also teach less able workers (e.g., Ichniowski *et al.*, 1997; Hamilton *et al.*, 2003). An important aspect here is that the prediction from these theories holds even if each student is evaluated based on their *individual* performance. A detailed look at the data further revealed that the positive effect of peer learning activities in term 2 was not limited to their understanding of term 2 materials (macroeconomics), but it also spilled over to their understanding of term 1 materials (microeconomics). This underlines great importance in devising effective pairing when implementing human resource management practices such as communication, peer learning, and skill training in organizations.

The role of team heterogeneity is an active research agenda in the recent literature. Based on randomized field experiments, it suggests that while peer effects are stronger in teams with a greater spread in abilities, the positive effects tend to be limited to high-ability individuals in a large team since they endogenously choose to interact with like-minded high types (e.g., Carrell *et al.*, 2009; Carrell *et al.*, 2013; Lyle, 2009; Duflo *et al.*, 2011; Feld and Zölitz, 2017). The low-ability individuals are therefore hurt in such teams. This matching situation was modeled in the sorting condition of the present study. The random matching condition of this paper showed stronger work performance than the sorting condition, and the worst performer in the former *benefited* more than the better performer. With the smallest peer group size, each student had no choice but to interact with their assigned peer in a mandatory peer learning activity. In addition,

each pair was given a pre-determined task for learning activities. This setup effectively prevented the less able from being excluded from the more able students in the module.

This finding has a policy implication regarding effective learning practices. As already discussed, sub-groups tend to endogenously emerge among like-minded peers when the peer group size is large and their skills and abilities are heterogeneous. While the present study suggests a simple solution, namely, forcedly pairing more able with less able individuals, the so-called "tracking" has been proposed to help enhance productivity in the literature to date. For example, using a clear field experiment in primary schools, Duflo *et al.* (2011) demonstrated that when students are divided into sections based on prior achievement levels, even those assigned to low-achievement peer groups can improve academic performance. They argue that teachers can better tailor their instruction levels and methods if students are sorted based on their academic skills. The finding of the present paper, nevertheless, suggests that the effect of such tracking may not be strong under certain conditions if peer learning is an important element to achieve a given goal.[26] The field experiment in this study suggests that the more able would effectively teach the less able if they are forced to be paired through a fair process, which results in a performance improvement of the latter. This kind of pairing can be nested and implemented in multiple activities in organizations. Such positive effects of heterogeneity cannot be obtained if low-achieving individuals are simply grouped together.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. Reference Points and Effort Provision. *American Economic Review*, 101(2), 470-92.

Arcidiacono, Peter, Josh Kinsler, and Joseph Price, 2017. Productivity Spillovers in Team Production: Evidence from Professional Basketball. *Journal of Labor Economics*, 35(1), 191-225.

Ariely, Dan, Anat Bracha, and Stephan Meier, 2009. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, 99, 544-55.

Azoulay, Pierre, Joshua Graff Zivin, and Jialan Wang, 2010. Superstar Extinction. *Quarterly Journal of Economics*, 125(2), 549-589.

Bénabou, Roland, and Jean Tirole, 2006. Incentives and Prosocial Behavior. *American Economic Review*, 96(5), 1652-1678.

[26] It may also be difficult for teachers to adjust their teaching methods in low-achievement peer groups under certain conditions because low-ability students tend to report higher levels of satisfaction with their teachers' pedagogical practices (Lavy *et al.*, 2012).

Bénabou, Roland, and Jean Tirole, 2011. Identity, Morals, and Taboos: Beliefs as Assets. *Quarterly Journal of Economics*, 126(2), 805-855.

Black, Sandra, Paul Devereux, and Kjell Salvanes, 2013. Under Pressure? The Effect of Peers on Outcomes of Young Adults. *Journal of Labor Economics*, 31(1), 119-153.

Booij, Adam, Edwin Leuven, Hessel Oosterbeek, 2017. Ability Peer Effects in University: Evidence from a Randomized Experiment. *Review of Economic Studies*, 84(2), 547-578.

Bowles, S., Gintis, H., 2005. "Prosocial emotions," in L. Blume, S. Durlauf (Eds.), *The Economy as a Complex Evolving System III: Essays in Honor of Kenneth Arrow*, Oxford University Press, Oxford: 337-367.

Brune, Lasse, Eric Chyn, and Jason Kerwin, forthcoming. Peers and Motivation at Work Evidence from a Firm Experiment in Malawi. *Journal of Human Resources*.

Carrell, Scott, Richard Fullerton, James West, 2009. Does Your Cohort Matter? Measuring Peer Effects in College Achievement. *Journal of Labor Economics*, 27(3), 439-464.

Carrell, Scott, Bruce Sacerdote, James West, 2013. From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation. *Econometrica*, 81(3), 855-882.

Chan, Kenneth, Stuart Mestelman, Rob Moir, and Andrew Muller, 1996. The Voluntary Provision of Public Goods under Varying Income Distributions. *Canadian Journal of Economics*, 29(1), 54-69.

Croson, Rachel, and Jen Shang, 2008. The impact of downward social information on contribution decisions, *Experimental Economics*, 11, 221-233.

De Grip, Andries, and Jan Sauermann, 2012. The Effects of Training on Own and Co-worker Productivity: Evidence from a Field Experiment. *The Economic Journal*, 122(560), 376-399.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5), 1739-1774.

Falk, Armin, and Andrea Ichino, 2006. Clean Evidence on Peer Effects. *Journal of Labor Economics*, 24(1), 39-57.

Fehr, Ernst, and Klaus Schmidt, 1999. A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics*, 114(3), 817-868.

Fischbacher, Urs, Simeon Schudy, Sabrina Teyssier, 2014. Heterogeneous reactions to heterogeneity in returns from public goods, *Social Choice and Welfare*, 43, 195-217.

Gächter, Simon, and Christian Thöni, 2005. Social Learning and Voluntary Cooperation among Like-minded People. *Journal of the European Economic Association*, 3, 303-314.

Gunnthorsdottir, Anna, Daniel Houser, and Kevin McCabe, 2007. Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization*, 62(2), 304-315.

Guryan, Jonathan, Kory Kroft, and Matthew Notowidigdo, 2009. Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments. *American Economic Journal: Applied Economics*, 1(4), 34-68.

Hamilton, Barton, Jack Nickerson, and Hideo Owan, 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.

Herbst, Daniel, and Alexandre Mas, 2015. Peer effects on worker output in the laboratory generalize to the field. *Science*, 350(6260), 545-549.

Ichniowski, Casey, Kathryn Shaw, and Giovanna Prennushi, 1997. The Effects of Human Resource Management Practices on Productivity: A Study of Steel Finishing Lines. *American Economic Review*, 87(3), 291-313.

Jackson, Kirabo, and Elias Bruegmann, 2009. Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics*, 1(4), 85-108.

Kamei, Kenju, 2018. The Role of Visibility on Third Party Punishment Actions for the Enforcement of Social Norms. *Economics Letters*, 171, 193-197.

Kamei, Kenju, and Louis Putterman, 2015. In Broad Daylight: Fuller Information and Higher-Order Punishment Opportunities Can Promote Cooperation. *Journal of Economic Behavior & Organization*, 120, 145-159.

Kamei, Kenju, 2018. Promoting Competition or Helping the Less Endowed? Distributional Preferences and Collective Institutional Choices under Intragroup Inequality. *Journal of Conflict Resolution*, 62(3), 626-655.

Kamei, Kenju, and Thomas Markussen, 2020. Free Riding and Workplace Democracy – Heterogeneous Task Preferences and Sorting. Durham University Business School Working Paper No. 1, 2020.

Katz, Lawrence, Jeffrey Kling, and Jeffrey Liebman, 2001. Moving to opportunities in Boston: Early results of a randomized mobility experiment. *Quarterly Journal of Economics*, 116(2), 607-54.

Langer, Ellen, 1975. The Illusion of Control. *Journal of Personality and Social Psychology*, 32(2), 311-328.

Larkin, Ian, Lamar Pierce, and Francesca Gino, 2012. The psychological costs of pay-for-performance: Implications for the strategic compensation of employees. *Strategic Management Journal*, 33(10), 1194-1214.

Lavy, Victor, Daniele Paserman, and Analia Schlosser, 2012. Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom. *The Economic Journal*, 122(559), 208-237.

Lavy, Victor, and Analia Schlosser, 2011. Mechanisms and Impacts of Gender Peer Effects at School. *American Economic Journal: Applied Economics*, 3(2), 1-33.

Leuven, Edwin, Hessel Oosterbeek, Joep Sonnemans, and Bas van der Klaauw, 2011. Incentives versus Sorting in Tournaments: Evidence from a Field Experiment. *Journal of Labor Economics*, 29(3), 637-658.

Lu, Fangwen, and Michael Anderson, 2015. Peer Effects in Microenvironments: The Benefits of Homogeneous Classroom Groups. *Journal of Labor Economics*, 33(1), 91-122.

Lyle, David, 2007. Estimating and Interpreting Peer and Role Model Effects from Randomly Assigned Social Groups at West Point. *Review of Economics and Statistics*, 89(2), 289-299.

Lyle, David, 2009. The Effects of Peer Group Heterogeneity on the Production of Human Capital at West Point. *American Economic Journal: Applied Economics*, 1(4), 69-84.

Manski, Charles, 1993. Identification of Endogenous Social Effects: The reflection problem. *Review of Economic Studies*, 60(3), 531-542.

Mas, Alexandre, and Enrico Moretti, 2009. Peers at Work. *American Economic Review*, 99(1), 112-145.

Maurice, Jonathan, Agathe Rouaix, Marc Willinger, 2013. Income Redistribution and Public Good Provision: An Experiment. *International Economic Review*, 54(3), 957-975.

Page, Talbot, Louis Putterman, and Bulent Unel, 2005. Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency. *The Economic Journal*, 115(506), 1032-53.

Sacerdote, Bruce, 2001. Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics*, 116 (2), 681-704.

Samek, Anya, and Roman Sheremeta, 2014. Recognizing contributors: an experiment on public goods. *Experimental Economics*, 7, 673-690.

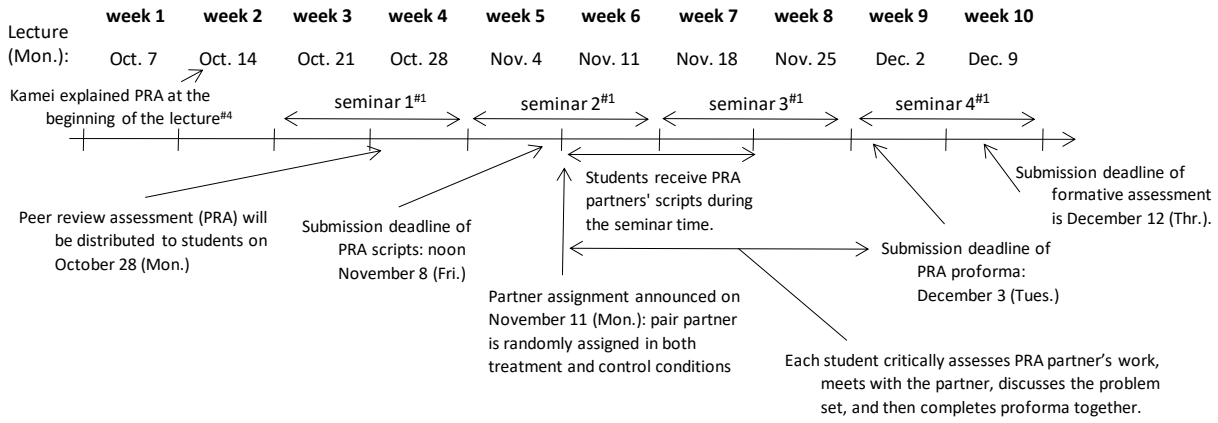
Shang, Jen, and Rachel Croson 2009. A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods. *The Economic Journal*, 119(540), 1422-1439.

Soetevent, Adriaan R., 2005,Anonymity in giving in a natural context—a field experiment in 30 churches, *Journal of Public Economics*, 89, 2301-2323.

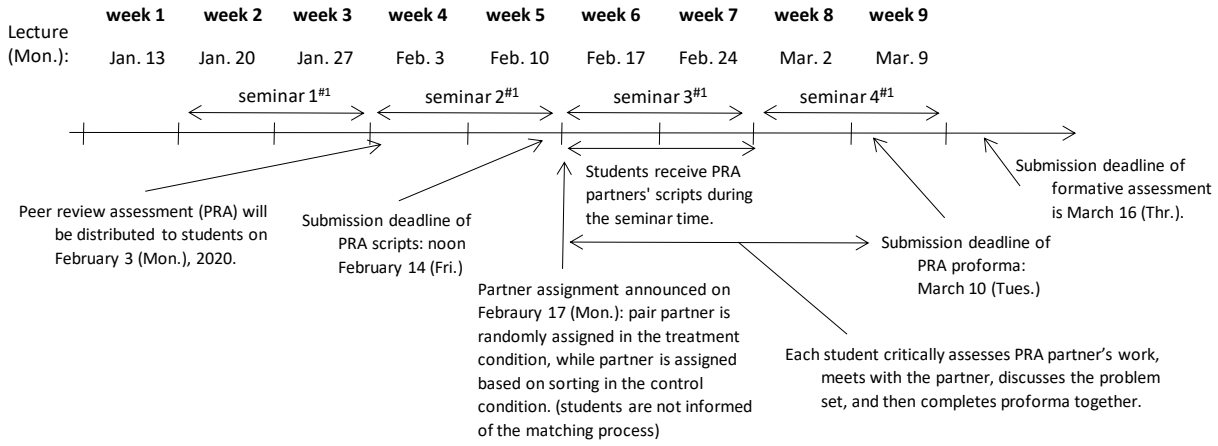
Svenson, Ola, 1981. Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, 47, 143-48.

Figure 1: *Timeline of Peer Review Activities*

A. Term 1 (Microeconomics)[#2]

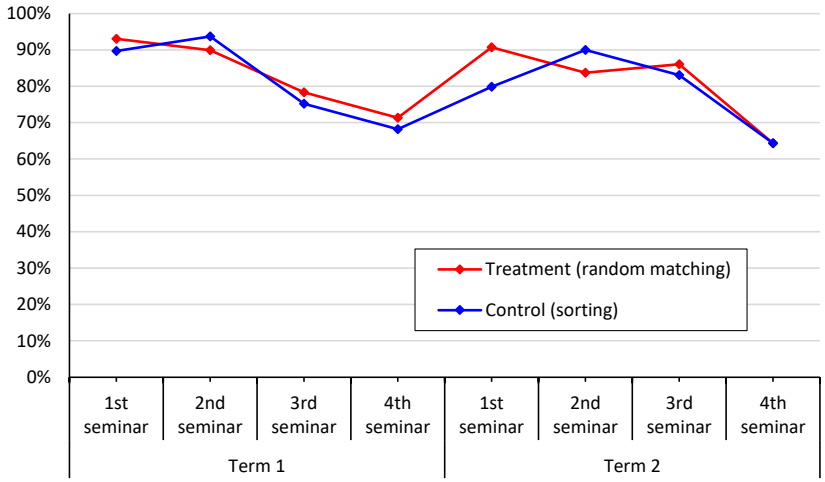


B. Term 2 (Macroeconomics)[#3]



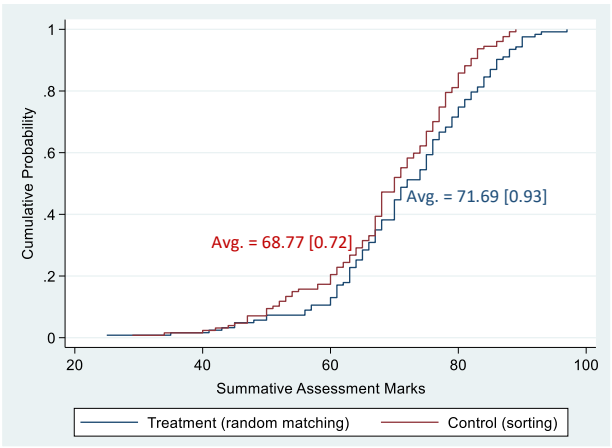
Notes: [#1] Each seminar is scheduled during a two weeks window. The date of biweekly seminars (Monday, Tuesday or Thursday) differs by the seminar group, set by the undergraduate office. [#2] Students' conditions are identical for the treatment and control seminar groups in term 1. [#3] There is only one difference between the treatment and control groups in term 2. In a control seminar group, students are sorted in descending order according to term 1 formative marks; two students with the closest marks are paired. By contrast, in a treatment group, students are randomly broken into pairs for the peer review activities irrespective of their term 1 formative mark. [#4] As discussed, another faculty member was responsible for the lecturing in term 1. PRA was explained by Kamei using the materials in Supplementary materials A.1 at the onset of the second lecture in term 1.

Figure 2: *Trends of Average Seminar Attendance Rates*

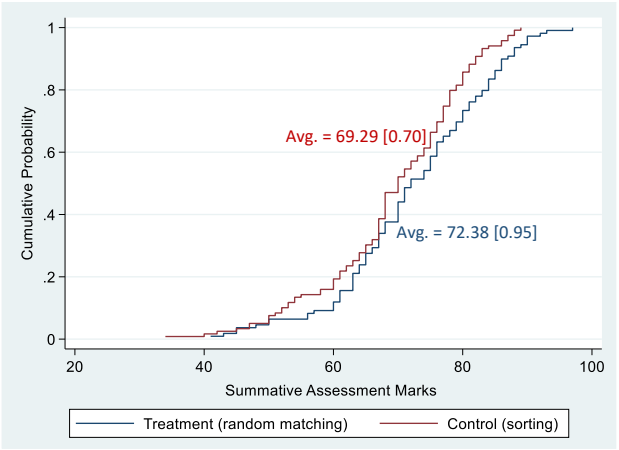


Note: The attendance rates were calculated based on eligible students (those who gave consent).

Figure 3: *Distribution of Students' Summative Assessment Marks*



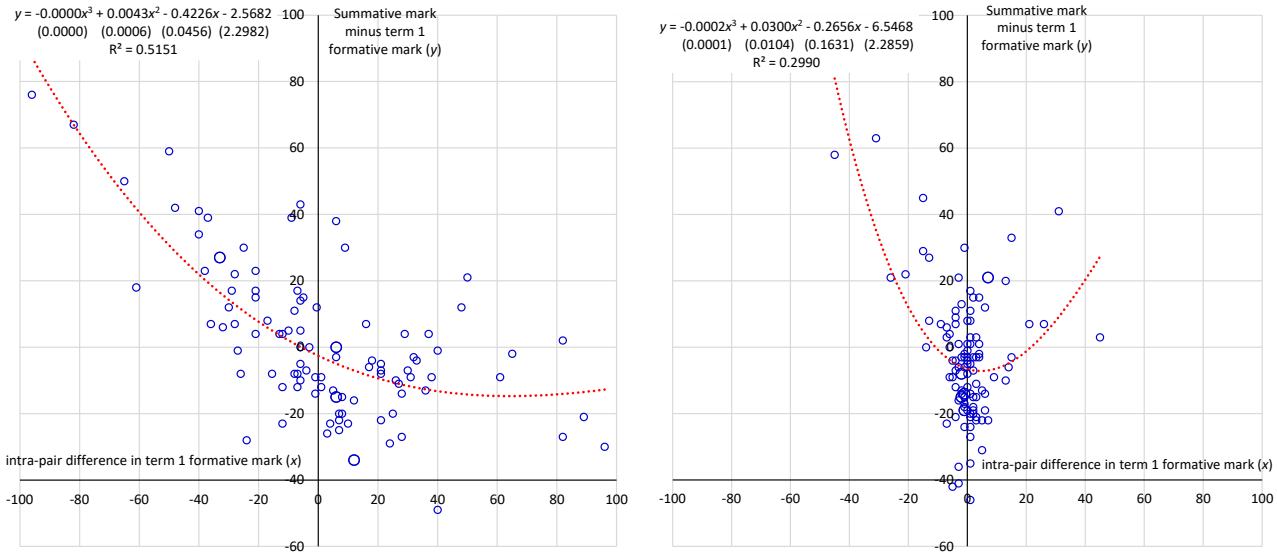
(A) All eligible students in the subject pool



(B) Students who submitted term 2 PRA and whose partner also submitted the assessment

Notes: The numbers in squared brackets are standard errors clustered by seminar group ID. The number of eligible observations in panel A is not the same as the size of the subject pool since a small number of students did not complete the summative assessments (Table 1).

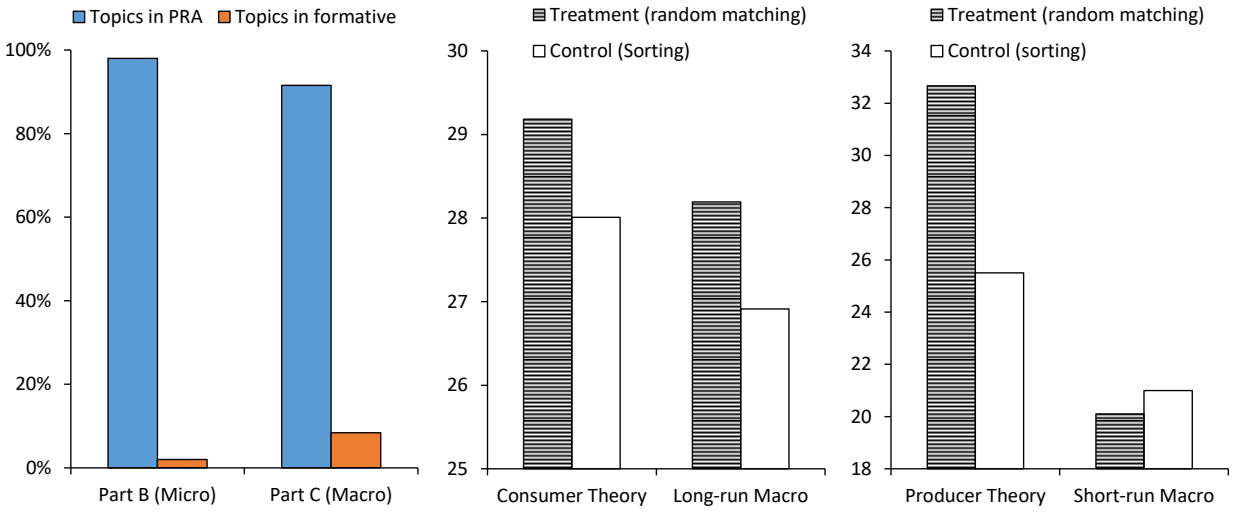
Figure 4: *Interim Achievement Differences and Improvement of Performances*



(A) Treatment condition (random matching) (B) Control condition (sorting)

Notes: The size of each point indicates its frequency. Almost all points have the frequency of one (i.e., one student). The numbers in parentheses in the polynomial equations in the figures are robust standard errors clustered by seminar group ID.

Figure 5: *Students' Decisions to Select Questions and Performances in the Summative Assessment*



(A) Distribution of students' selection (B) Avg. marks for topics in PRAs (C) Avg. marks for topics in formatives

Notes: The maximum mark in each part is 40. The topics in the PRA are the consumer theory (term 1) and long-run macroeconomics (term 2). The topics in the formatives are the producer theory (term 1) and short-run macroeconomics (term 2). As in other analyses, those who gave consent and completed the summative assessment were used to calculate the distributions of problem selection and average marks.

Table 1: *Summary of Conditions*

Treatment	A. Treatment condition (Random matching)	B. Control condition (Sorting)	C. Total	Two-sided *p* for H_0: A = B[#5]
a. Number of seminar groups	8	8	16	---
b. Total number of students[#2]	139[#1]	139	278	---
c. Students who gave consent (subject pool)				
c.i. Number of the students	129 (92.8%)	130 (93.5%)	259 (93.2%)	---
c.ii. Number of female students out of c.i	59	72	131	0.1365
c.iii. Number of British students out of c.i	44	33	77	0.1363
c.iv. Number of those who submitted term 1 formatives out of c.i	122 (87.8%)	121 (87.1%)	243 (87.5%)	---
c.v. Avg. term 1 formative mark (out of 100) [#3,#4]	70.38	72.05	71.21	0.721
c.vi. Number of those who submitted summative assessments out of c.i	123 (95.4%)	127 (97.7%)	250 (96.5%)	---
c.vii. Number of those who submitted summative assessments out of c.iv	116 (83.5%)	118 (84.9%)	234 (84.2%)	---
c.viii. Avg. term 1 formative mark for c.vii[#3]	70.84	71.71	71.28	0.867

Notes: [#1] The number of the students was 142 at the beginning of term 2 when treatment allocations were made. Three students in the treatment condition were, however, were not assigned any pair partners because one student requested exemption from the activity due to disability, another was found to have been suspended from the university (this student was not able to attend any academic activities in term 2), and the other withdrew from the module at the beginning of term 2. [#2] When the number of students in a given seminar group was odd, one interaction unit was a three-student team where PRA scripts were swapped among them. [#3] The cumulative distributions of formative assessment marks can be found in Supplementary material Section B.1. [#4] The average term 2 formative marks were similar for the two conditions (60.4 in the treatment condition; and 59.6 in the control condition). It should be worth noting that students were distributed, and started to work on, the term 2 formative assessment on February 24, 2020, which was before the peer review assessment activities were completed. Hence, it is not surprising to see no effects of the treatment interventions on term 2 formative performances. [#5] Fisher's exact tests for rows c.ii and c.iii, and Somers' D with seminar group clustering for rows c.v and c.viii.

Table 2: *Treatment Effects of Term 2 Random Matching*

(A) Second Stage Regression (Treatment effect)

Dependent variable: Summative assessment mark of student *i*

Data:	(I) All data		(II) Both *i* and *i*'s partner submitted term 2 PRA	
Independent variable:	(i)	(ii)	(i)	(ii)
(a) t2 random matching dummy {=1(0) for the treatment (control) condition}	3.10*** (1.19)	3.30*** (1.03)	2.43** (1.16)	2.50** (1.02)
(b) A dummy that equals 1 if *i* did not submit term 1 formative assessment	---	4.26 (5.80)	---	6.20 (6.55)
(c) Interaction term: (1 – variable (b)) × term 1 formative assessment mark	---	0.16*** (.040)	---	0.16*** (.05)
A three-student team dummy {= 1(0) if a student was assigned to a three(two)-student team}	-6.13* (3.45)	-0.89 (3.47)	-4.50 (4.00)	0.30 (4.03)
Constant	70.15*** (1.62)	59.54*** (3.00)	70.39*** (1.09)	59.51*** (3.35)
# observations	259	259	236	236
# selected	250	250	228	229
Log pseudolikelihood	-1005.01	-995.13	-906.96	-902.24

(B) Selection equation that explains whether student *i* submits the summative assessment (i.e., the submission is observed)

Data:	(I) All data		(II) Both *i* and *i*'s partner submitted term 2 PRA	
Independent variable:	(i)	(ii)	(i)	(ii)
(a) t2 random matching dummy {=1(0) for the treatment (control) condition}	0.04 (.09)	0.17** (0.08)	-0.21 (.33)	-0.15 (0.29)
(b) A dummy that equals 1 if *i* did not submit term 1 formative assessment	---	5.09 (n.a.)	---	4.87 (3.15)
(c) Interaction term: (1 – variable (b)) × term 1 formative assessment mark	---	0.002 (0.003)	---	0.005 (0.01)
(d) The number of formative assessments not submitted {= 0, 1, 2}	-0.61*** (0.03)	-0.36*** (0.02)	-0.60** (0.29)	-0.86* (0.47)
(e) Seminar attendance rate ∈ [0, 1]	0.23*** (0.01)	0.50*** (0.02)	0.18 (0.68)	0.28 (0.71)
A three-student team dummy {= 1(0) if a student was assigned to a three(two)-student team}	0.09 (0.27)	4.15 (n.a.)	4.70*** (.86)	4.14** (1.86)
Constant	1.80*** (0.14)	1.16*** (0.26)	2.28*** (0.62)	1.93*** (0.65)

Notes: Estimations of the Heckman two-stage selection model with robust standard errors clustered by seminar group ID. The numbers in parentheses are standard errors. In addition to the independent variables listed in the table, a term 1 tutor dummy was controlled in both stages of regressions since there were two seminar tutors in term 1. A three-student team dummy was also added as a control since there was one such team in a session whose number of students was odd. In columns II.i and II.ii, only observations in which a student submitted term 2 peer review assessment and his/her partner also submitted it were used as data. Equations I.i, I.ii, II.i, and II.ii of panel B are the selection equations of columns I.i, I.ii, II.i, and II.ii, respectively, of panel A. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

Table 3: *Mechanism behind the Positive Impact of Term 2 Random Matching*

(A) Second Stage Regression

Dependent variable: Summative assessment mark of student *i*

Data:	(I) All data		(II) Both *i* and *i*'s partner submitted term 2 PRA	
Independent variable:	(i)	(ii)	(i)	(ii)
(a) Better performer dummy: $\mathbf{1}\{x_i > x_j$, random matching$\}$[#1]	1.67*	0.08	2.48**	0.27
	(.89)	(0.97)	(1.05)	(1.22)
(b) Worse performer dummy: $\mathbf{1}\{x_i < x_j$, random matching $\}$[#2]	2.78*	5.34***	2.56	3.64**
	(1.54)	(1.65)	(1.57)	(1.72)
(c) Term 1 formative assessment mark (x_i)	---	0.20***	---	0.19***
		(0.05)		(0.06)
A three-student team dummy {= 1(0) if a student was assigned to a three(two)-student team}	-6.67	0.42	-5.94	0.89
	(4.70)	(3.76)	(5.19)	(4.54)
Constant	71.06***	57.53***	71.36***	57.83***
	(1.18)	(3.58)	(1.15)	(3.90)
# observations	243	243	226	226
# selected	234	234	218	218
Log pseudolikelihood	-935.95	-926.36	-858.79	-859.79

(B) Selection equation that explains whether student *i* submits the summative assessment (i.e., the submission is observed)

Data:	(I) All data		(II) Both *i* and *i*'s partner submitted term 2 PRA	
Independent variable:	(i)	(ii)	(i)	(ii)
(a) Better performer dummy: $\mathbf{1}\{x_i > x_j$, random matching$\}$[#1]	-0.38***	-0.24*	-0.44***	-0.42
	(0.08)	(0.13)	(0.10)	(0.53)
(b) Worse performer dummy: $\mathbf{1}\{x_i < x_j$, random matching$\}$[#2]	0.12	-0.08	0.12	0.14
	(0.12)	(0.17)	(0.13)	(0.40)
(c) Term 1 formative assessment mark (x_i)	---	-0.006	---	0.010
		(0.006)		(0.014)
(d) The number of formative assessments not submitted {= 0, 1, 2}	-0.57***	-0.51***	-.69***	-1.02**
	(0.03)	(0.14)	(0.04)	(0.50)
(e) Seminar attendance rate $\in$ [0, 1]	0.24***	0.41***	0.81***	0.31
	(0.01)	(0.09)	(0.05)	(0.65)
A three-student team dummy {= 1(0) if a student was assigned to a three(two)-student team}	0.13	n.a.[#3]	3.81	3.25
	(0.37)		(n.a.)	(3.39)
Constant	1.71***	1.95***	1.42***	1.69***
	(0.14)	(0.65)	(0.15)	(0.95)

Notes: Estimations of the Heckman two-stage selection model with robust standard errors clustered by seminar group ID. The numbers in parentheses are standard errors. In addition to the independent variables listed in the table, a term 1 tutor dummy was controlled in both stages of regressions since there were two seminar tutors in term 1. Only observations in which a student submitted term 1 formative and his/her partner also submitted it were used as data. A three-student team dummy was also added as a control since there was one such team in a session whose number of students was odd. [#1] $\mathbf{1}\{x_i > x_j$, random matching$\}$ is an indicator variable which equals 1 if $x_i > x_j$ and *i* is in the treatment condition; 0 otherwise. Here, x_i (x_j) is *i*'s (*i*'s partner *j*'s) term 1 formative assessment mark [#2] $\mathbf{1}\{x_i < x_j$, random matching$\}$ is an indicator variable which equals 1 if $x_i < x_j$ and *i* is in the treatment condition; 0 otherwise. The reference group is observations in the control condition. Equations I.i, I.ii, II.i, and II.ii of panel B are the selection equations of columns I.i, I.ii, II.i, and II.ii, respectively, of panel A. [#3] The three-student team dummy was not included in the selection equation since otherwise the model was not converged. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.

Table 4: *The Impact of Term 2 Random Matching by Part in the Summative Assessment*

(A) Second Stage Regression

Dependent variable:	(I) Mark of student *i* in Part B (Micro)				(II) Mark of student *i* in Part C (Macro)			
Independent variable:	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
(a) t2 random matching dummy {=1(0) for the treatment (control) condition}	1.49** (0.75)	1.62** (0.65)	---	---	1.35** (0.55)	1.01* (0.57)	---	---
(b) A dummy that equals 1 if *i* did not submit term 1 formative assessment	---	3.77 (2.84)	---	---	---	-.20 (3.55)	---	---
(c) Interaction term: (1 – variable (b)) × term 1 formative assessment mark	---	0.08*** (0.02)	---	---	---	0.05** (0.02)	---	---
(d) Better performer dummy: **1{**$x_i > x_j$, random matching}[#1]	---	---	1.14** (0.49)	0.32 (0.47)	---	---	0.01 (0.65)	-0.21 (0.66)
(e) Worse performer dummy: **1{**$x_i < x_j$, random matching}[#2]	---	---	1.20 (1.07)	2.27** (1.02)	---	---	0.95 (0.82)	2.19*** (0.85)
(f) Term 1 formative assessment mark (x_i)	---	---	---	0.09*** (0.03)	---	---	---	0.06*** (0.02)
A three-student team dummy {= 1(0) if a student was assigned to a three(two)-student team}	-0.35 (2.12)	1.98 (1.92)	-0.96 (2.54)	1.71 (2.15)	-4.03*** (1.10)	-1.58 (1.65)	-3.37 (2.19)	-2.24 (2.21)
Constant	28.72*** (0.73)	23.14*** (1.64)	28.94*** (0.48)	22.59*** (1.98)	26.98*** (0.53)	23.68*** (1.52)	27.53*** (0.62)	23.79*** (1.74)
# observations	259	259	243	243	259	259	243	243
# selected	250	250	234	234	250	250	234	234
Log pseudolikelihood	-870.62	-863.83	-815.16	-807.72	-827.49	-835.17	-777.63	-761.32

(B) Selection equation that explains whether *i* submits the summative assessment (i.e., the submission is observed)

Dependent variable:	(I) Mark of student *i* in Part B (Micro)				(II) Mark of student *i* in Part C (Macro)			
Independent variable:	(i)	(ii)	(iii)	(iv)	(i)	(ii)	(iii)	(iv)
(a) t2 random matching dummy {=1(0) for the treatment (control) condition}	-0.20* (0.11)	0.001 (0.09)	---	---	0.26*** (0.08)	0.32 (0.27)	---	---
(b) A dummy that equals 1 if *i* did not submit term 1 formative assessment	---	0.26 (0.39)	---	---	---	6.32*** (0.88)	---	---
(c) Interaction term: (1 – variable (b)) × term 1 formative assessment mark	---	0.001 (0.003)	---	---	---	0.004 (0.01)	---	---
(d) Better performer dummy: **1{**$x_i > x_j$, random matching}[#1]	---	---	-0.17*** (0.06)	-0.24 (1.15)	---	---	-0.38 (0.35)	-0.03 (0.11)
(e) Worse performer dummy: **1{**$x_i < x_j$, random matching}[#2]	---	---	-0.05 (0.14)	-0.23 (0.48)	---	---	-0.37 (0.41)	-0.46 (0.14)
(f) Term 1 formative assessment mark (x_i)	---	---	---	0.00 (0.02)	---	---	---	0.003 (0.004)
(g) The number of formative assessments not submitted {= 0, 1, 2}	-0.27*** (0.01)	-0.28*** (0.01)	-0.40*** (0.02)	-0.47 (0.96)	-0.16*** (0.01)	-0.55* (0.29)	-0.54** (0.27)	-0.58*** (0.04)
(h) Seminar attendance rate $\in$ [0, 1]	0.00 (0.00)	-0.26*** (0.010)	-0.13*** (0.01)	-0.08 (0.09)	0.00 (0.00)	0.05 (0.65)	0.09 (0.59)	0.61*** (0.04)
A three-student team dummy {= 1(0) if a student was assigned to a three(two)-student team}	-0.086 (0.28)	-0.036 (0.26)	2.70 (n.a.)	3.85 (n.a.)	4.45 (n.a.)	6.04*** (0.35)	4.78*** (0.47)	6.52 (n.a.)
Constant	1.76*** (0.12)	1.76*** (0.22)	1.85*** (0.10)	1.99*** (0.29)	1.26*** (0.10)	2.00*** (0.52)	2.25*** (0.52)	.83*** (0.29)

Notes: Estimations of the Heckman two-stage selection model with robust standard errors clustered by seminar group ID. The numbers in parentheses are standard errors. All data are used. In addition to the independent variables listed in the table, a term 1 tutor dummy was controlled in both stages of regressions since there were two seminar tutors in term 1. For columns I.iii, I.iv, II.iii, and II.iv, only observations in which a student submitted term 1 formative and his/her partner also submitted it were used as data. [#1] **1**{$x_i > x_j$, random matching} is an indicator variable which equals 1 if $x_i > x_j$ and *i* is in the treatment condition; 0 otherwise. Here, x_i (x_j) is *i*'s (*i*'s partner *j*'s) term 1 formative assessment mark [#2] **1**{$x_i < x_j$, random matching} is an indicator

variable which equals 1 if $x_i < x_j$ and *i* is in the treatment condition; 0 otherwise. Equations I.i, I.ii, I.iii, I.iv, II.i, II.ii, II.iii and II.iv of panel B are the selection equations of columns I.i, I.ii, I.iii, I.iv, II.i, II.ii, II.iii and II.iv, respectively, of panel A. Results change little when the demographic information is added as controls. *, **, and *** indicate significance at the .10 level, at the .05 level and at the .01 level, respectively.