Self-selection of peers and performance*

Lukas Kiessling Jonas Radbruch Sebastian Schaube

March 30, 2020

Abstract

In many natural environments, carefully chosen peers influence individual behavior. Using a framed field experiment at secondary schools, we examine how self-selected peers affect performance in contrast to randomly assigned ones. We find that self-selection improves performance by approximately 16-19% of a standard deviation relative to randomly assigned peers. Our results document peer effects in multiple characteristics and show that self-selection changes these characteristics. However, a decomposition reveals that variations in the peer composition contribute only little to the estimated average treatment effects. Rather, we find that self-selection has a direct effect on performance.

Keywords: Framed field experiment, Peer effects, Self-selection, Peer assignment rules
JEL-Codes: C93, D01, I20, J24, L23

*Lukas Kiessling: Max Planck Institute for Research on Collective Goods, lkiessling@coll.mpg.de; Jonas Radbruch: University of Bonn and IZA, j.radbruch@uni-bonn.de; Sebastian Schaube: University of Bonn, sebastian.schaube@uni-bonn.de. We thank Viola Ackfeld, Philipp Albert, Thomas Dohmen, Lorenz Goette, Ingo Isphording, Sebastian Kube, Pia Pinger, Ulf Zölitz and audiences at Bonn, MBEPS 2017, VfS 2017, ESA Europe 2017, COPE 2018, ESA World 2018, IZA World Labor Conference, IZA Brown Bag, Rady Spring School in Behavioral Economics 2017, Bonn-Mannheim Ph.D. Workshop, 20th IZA Summer School in Labor Economics, 12th Nordic Conference on Behavioral and Experimental Economics, Max Planck Institute for Research on Collective Goods, EEA 2018, Bergen, and ESWM 2018 for helpful feedback and comments. We also thank the schools and students that participated in the experiments. This research was undertaken while all authors were at the University of Bonn. We did not obtain an IRB approval for this project because at the time of the experiment there did not exist an IRB at the University of Bonn's Department of Economics. However, we would like to stress that the schools' headmasters approved the study, written parental consent was required for students to take part in the study and participation was voluntary. Moreover, the experiment is in line with the requirements of the BonnEconLab. Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through CRC TR 224 (projects A01 and A02) is gratefully acknowledged. This study is registered in the AEA RCT Registry and the unique identifying number is: "AEARCTR-0005562".

"The first thing I would do every morning was look at the box scores to see what Magic did. I didn't care about anything else."
– Larry Bird

1 Introduction

Basketball hall of famer Larry Bird motivated himself to train harder not by focusing on any player but rather by looking at his rival Magic Johnson's performance during the previous night's game. Similarly, seeing a specific classmate study long and continuously might also help to concentrate on one's own work. In various dimensions of life – ranging from students in educational settings (Sacerdote, 2001) over cashiers in supermarkets (Mas and Moretti, 2009) and fruit pickers on strawberry fields (Bandiera, Barankay, and Rasul, 2009, 2010) to fighter pilots during World War II (Ager, Bursztyn, and Voth, 2016) – people affect each other through their presence, performance and choices. Yet, these social influences often stem from specific persons – roommates, frequently interacting coworkers, friends, or former colleagues – that individuals select themselves. This is in stark contrast with settings in which peers are randomly or exogenously assigned. But what actually changes once we allow peers to be self-selected? In general, these settings differ in two aspects: first, self-selection changes with whom one interacts; and, second, having the opportunity to self-select peers fundamentally changes the mode of peer assignment from exogenous (or random) assignment to self-selection. Both of these channels potentially alter an individual's motivation and behavior.

In this paper, we study how different peer assignment rules – self-selection versus random assignment – affect individual performance. In doing so, we examine a key feature of many peer effect studies, namely the absence of self-selection. In a first step, we document differences in performance between treatments which allow for self-selection or random assignment of peers. Subsequently, we analyze the underlying mechanisms. For this purpose, we decompose performance improvements into their two possible sources: an indirect effect stemming from changes in the peer composition and a direct effect from being able to self-select rather than being assigned to a specific peer.

In order to study the effects of self-selection, we conducted a framed field experiment (Harrison and List, 2004) with over 600 students (aged 12 to 16) in physical education classes of German secondary schools. Students took part in two running tasks (suicide runs) – first alone, then with a peer – and filled out a survey in between that elicited preferences for peers, personal characteristics, and the social network within each class. Our treatments exogenously varied the peer assignment in the second run

using three different peer assignment rules. We implemented a random matching of pairs (RANDOM) as well as two matching rules that used elicited preferences to implement two notions of self-selection: first, the classroom environment enabled students to state preferences for known peers (*name-based preferences*); and second, using a running task yielded direct measures of performance and thus could be used to select peers based on their relative performance in the first run (*performance-based preferences*). Using these two sets of preferences, we implemented two treatments with self-selection of peers by matching students based on either their name-based preferences (NAME) or preferences over relative performance (PERFORMANCE).

We find that self-selection of peers leads to an average performance improvement of 16–19 percent of a standard deviation relative to randomly assigned peers. While students in RANDOM also improve their performance from the first to the second run, the improvements with self-selected peers almost double. Self-selection changes the peer composition, e.g., students predominantly interact with friends in NAME, but tend to choose others with a similar past performance in PERFORMANCE. Based on this finding, we decompose the overall treatment effect into an indirect effect that is due to the peer's altered characteristics and a direct effect of being able to self-select a peer. Although we observe substantial peer effects in multiple dimensions (e.g., in relative performance in the first run), a peer's characteristics do not explain treatment differences resulting in an indirect effect close to zero. Instead, our estimates provide evidence for a direct effect of peer self-selection on performance. Borrowing from self-determination theory (Deci and Ryan, 1985, 2000), we interpret this direct effect as a positive effect of having autonomy: being able to self-select peers has a psychological effect that enhances intrinsic motivation and improves subsequent performance. Finally, we simulate other exogenous peer assignment rules that seek to maximize or minimize the productivity differences between students to increase aggregate performance. We document that these alternative policies yield performance improvements close to those observed with randomly assigned peers and lower than those with peer self-selection. These findings thus support our interpretation that self-selection of peers may have an intrinsic value beyond changes in the peer composition.

Our results have three main contributions to the literature on peer effects, social interactions, and autonomy. First, we show that self-selection changes with whom people interact and thereby affects the overall composition of the reference or peer group. Second, we present evidence that self-selection of peers can directly affect behavioral outcomes and productivity. This highlights a novel channel through which the selection of peers can affect behavior. We thereby provide the first clean evidence on autonomy in a field setting. Third, we document that peer effects may be present

in multiple dimensions and discuss how this limits the effectiveness of exogenous peer assignment rules.

We document a strong causal difference in performance between widely-used randomly assigned peer groups and self-selected peers.[1] This focus on random peer assignment is understandable given that researchers aim to identify a clean causal effect of being exposed to peers. However, similar to what has been found in previous studies exploring the selection of students into peer groups (e.g., Cicala, Fryer, and Spenkuch, 2018; Tincani, 2017), our results indicate that the relevant *and* self-selected peer within a group does not correspond to a random peer. Rather, this systematic selection helps to understand why the impact of certain peer groups differs compared to others: friends and non-friends may have differential effects (Chan and Lam, 2015; Lavy and Sand, 2019) and only persons with specific characteristics may affect performance (Aral and Nicolaides, 2017).[2] In light of our results, such differential peer effects can be due to self-selection of relevant peers. Related to our paper, Chen and Gong (2018) study self-selection of team members and document, consistent with our findings, that teams form endogenously along the social network outperform randomly assigned ones. Yet, we move beyond their work in at least three dimensions. First, we focus on a setup with a single peer and individual incentives. Thus, we restrict the possible sources of peer effects to that single peer. Second, we lever a rich dataset of individual characteristics and provide evidence that several attributes of randomly assigned peers matter. Third, by eliciting preferences for peers, we observe a normally unobserved dimension – the fit of a peer. Taken together, these features allow us to move beyond their results by documenting that peer self-selection in itself may constitute a channel through which peers can influence behavior.

Moreover, our findings help to reconcile mixed evidence on the effectiveness of interventions changing class or work-group compositions to exploit peer effects (e.g., Booij, Leuven, and Oosterbeek, 2017; Carrell, Sacerdote, and West, 2013; Duflo, Dupas, and Kremer, 2011; Garlick, 2018). In our setup, the combination of two effects – the change in the peer composition and the multidimensionality of peer effects – has only a small impact on aggregate performance. More specifically, we move beyond peer effects in a single dimension and allow several characteristics such as produc-

[1] The literature on peer effects builds on (conditional) random assignment to identify peer effects and circumvent statistical issues outlined in Manski (1993). See also Sacerdote (2011) and Herbst and Mas (2015) for literature reviews on peer effects in education and a comparison of peer effects from field and lab settings, respectively.

[2] In a companion paper, Kiessling, Radbruch, and Schaube (2020), we study the peer selection process in more depth and relate the selection of peers to individual-level determinants.

tivity, friendship ties, and personality measures to exert peer effects.[3] Our results show that there are sizable peer effects apart from productivity and we highlight consequences of such multidimensional peer effects that hold independently of our specific setting. In particular, if policy-makers assign groups based on peer effects in a single dimension only, they neglect the fact that all assignment rules simultaneously change other peer characteristics, potentially giving rise to peer effects apart from the targeted dimension. These effects can counterbalance each other and lead to a net effect that is in our case close to zero and in general ambiguous. Hence, studies analyzing peer interactions and peer assignment policies need to take into account not only a potential direct effect of self-selection, but also the multidimensionality of peer effects.

Our findings also contribute to the literature studying the effects of autonomy and decision rights on behavioral outcomes. Specifically, we provide the first field evidence that self-selection (of peers) can have a direct effect that increases performance beyond its instrumental value of changing peer characteristics. Therefore, we complement laboratory studies by Bartling, Fehr, and Herz (2014) and Owens, Grossman, and Fackler (2014), which demonstrate that people are willing to pay for autonomy, i.e., the opportunity to actively select relevant aspects of their decision environment (Deci and Ryan, 1985). Similarly, autonomy in the workplace is associated with higher wages and employee happiness (Bartling, Fehr, and Schmidt, 2013) and leads to increased labor supply (Chevalier et al., 2019), while removing autonomy has been found to have negative consequences on employee effort (Falk and Kosfeld, 2006).[4] Our results highlight an additional channel through which autonomy might provide value to employers or policy-makers: the freedom to choose one's own peers or teammates can boost performance similar to other non-monetary incentives such as recognitions and awards (Bradler et al., 2016; Kosfeld and Neckermann, 2011), framing of rewards (Levitt et al., 2016) or personal goals (Corgnet, Gómez-Miñambres, and Hernán-González, 2015; Koch and Nafziger, 2011).

While our results document a direct effect of self-selection on performance in this particular setting, we do not claim that the effect will quantitatively or qualitatively carry over to all settings. Rather, we view our results as a proof-of-concept that the op-

[3]Thereby we also join a small set of studies explicitly considering the impact of personality traits on educational outcomes or performance (e.g., Chan and Lam, 2015; Golsteyn, Non, and Zölitz, 2017). Yet, these other studies do not consider the implications of multidimensional peer effects.

[4]These studies focus on individual decisions. However, autonomy can also help improve outcomes under collective decision-making. Having the right to vote has been can affect the quality of leadership positively (e.g., Brandts, Cooper, and Weber, 2014) as well as increase the effectiveness of institutions in the presence of social dilemmas (e.g., Bó, Foster, and Putterman, 2010; Sutter, Haigner, and Kocher, 2010).

portunity to self-select peers can affect performance. Yet, the process of self-selecting peers could also be important for settings in which peer effects do not arise due to social comparisons or peer pressure, but from effort or skill complementarities (e.g., Bandiera, Barankay, and Rasul, 2010; Mas and Moretti, 2009), or setting in which peers learn from each other (e.g., Bursztyn et al., 2014; Jackson and Bruegmann, 2009). The settings across these studies differ enormously, as does the underlying mechanism. Nonetheless, all of these share the notion that the behavior or action of peers imposes an externality on the action or behavior of others and that peers can in principle also be self-selected, affecting subsequent peer interactions.

The remainder of the paper is structured as follows. The next section presents our experimental design as well as procedural details. Section 3 presents the data and describes our sample of students. We outline our empirical framework in section 4. In section 5, we analyze how self-selected peers affect performance relative to randomly assigned peers and decompose this effect in a direct effect of self-selection and an indirect effect as a result of changes in the peer composition. We then interpret the direct effect and highlight potential policy implications. Finally, section 6 concludes.

2 Experimental design

Studying the self-selection of peers and their subsequent impact on performance requires an environment in which subjects can choose peers themselves and where exogenous assignment can be implemented. Subjects must be able to compare their own performance with that of a peer in a task that lends itself to natural up- and downward comparisons. One complication in many settings is that it is difficult to isolate the person who serves as the relevant point of comparison. This is especially true if several potential peers are present at all times, among which only some constitute the set of an individual's relevant peers. As subjects might select those peers for many reasons besides their performance, it is essential not only to observe additional characteristics of all subjects, but also to collect data from an existing social group. In these groups, subjects have a clear impression of other group members and are able to select peers based on additional characteristics such as their social ties.

In this study, we used the controlled environment of a framed field experiment to overcome these challenges. We embedded our experiment in physical education classes of German secondary schools. Students from grades 7 to 10 participated in a running task, first alone and then simultaneously with a peer. Running allowed students to compare their performance with either faster or slower students, while it excluded complementaries in production between the students. Moreover, we fo-

cused on pairs as the unit of observation. This reduced the number of peers in the experimental task to a single individual and allows us to cleanly identify his or her impact. Subjects singled out specific peers by either naming them directly (in the treatment NAME) or selecting performance intervals (in PERFORMANCE). The respective treatments used these preferences to form pairs with self-selected peers or pairs were formed at random. Hence, we can compare the effect of self-selected peers with exogenously assigned ones, and can evaluate the effects of each assignment mechanism.

In the following, we present the design of our field experiment in detail and describe the implemented procedures.

2.1 Experimental design

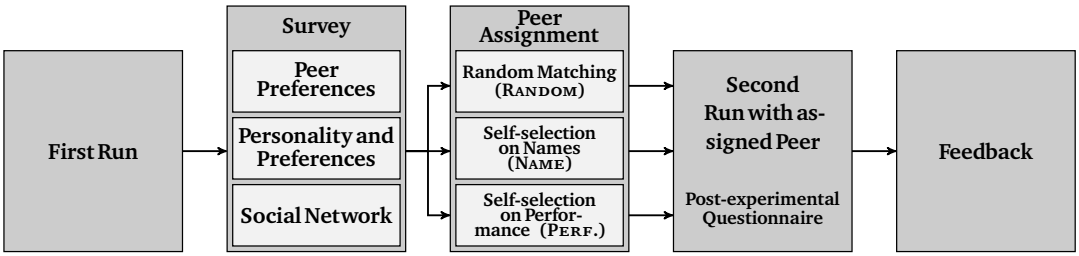
Figure 1 illustrates the experimental design. Students participated in a running task commonly known as "suicide runs", a series of short sprints to different lines of a volleyball court.[5,6] The first run – in which students ran alone – served two purposes: first, recorded times can be used as a measure of productivity and to evaluate the time improvement between the two runs; and second, we used (relative) times from the first run in combination with students' preferences to create pairs for the second run in one of the treatments described below. The second run mirrored the first one aside from the fact that students did not run alone, but rather in pairs. This means that two students performed the task simultaneously, while their times were recorded individually. Feedback about performance in both runs was only provided at the end of the experiment.

Between the two runs, students filled out a survey comprising three parts, eliciting preferences for peers, non-cognitive skills and information about the social network within each class. We elicited two kinds of preferences: first, we asked subjects to state the names of those classmates with whom they would like to perform the second

[5]The exact task is to sprint and turn at every line of the volleyball court. Subjects had to line up at the baseline. From there, they started running to the first attack line of the court (6 meters). After touching this line, they returned to the baseline again, touching the line on arrival. The next sprint took the students to the middle of the court (9 meters), the third to the second attack line (12 meters) and the last to the opposite baseline (18 meters), each time returning back to the baseline. They finished by returning to the starting point. The total distance of this task was 90 meters.

[6]The task was chosen for several reasons: (1) the task is not a typical part of the German physical education curriculum, yet it is easily understandable for the students; (2) in contrast to a pure and very familiar sprint exercise as in Gneezy and Rustichini (2004) or Sutter and Glätzle-Rützler (2015), students should only have a vague idea of their classmates' performance and cannot precisely target specific individuals in PERFORMANCE; and (3) due to the different aspects of the task (general speed, quickness in turning as well as some level of endurance or perseverance), the performance across age groups was not expected to (and did not) change dramatically.

Figure 1: Experimental design



This figure illustrates the experimental design. Peer assignment rules (RANDOM, NAME, PERFORMANCE) are randomly assigned on a classroom level.

run; and second, we asked them to state the relative performance level of their most-preferred peers. Note that we elicited all preferences irrespective of the assigned treatment and used these preferences to match students for the second run in two of the three treatments.

In addition to these preferences, the survey included socio-demographic questions and measures of personality and economic preferences: the Big Five inventory as used in the youth questionnaire of the German socio-economic panel (M. Weinhardt and Schupp, 2011), a measure of locus of control (Rotter, 1966), competitiveness[7], general risk attitude (Dohmen et al., 2011), and a short version of the INCOM scale for social comparison (Gibbons and Buunk, 1999; Schneider and Schupp, 2011). The survey concluded by eliciting the social network within every class. Subjects were asked to state up to six of their closest friends within the class.

Before and after the second run, we asked students a short set of questions about their peer and their experience during the task. Before the run, we elicited their belief about the relative performance of their peer in the first run, namely who they thought was faster. Following the second run, we asked them whether they would rather run alone or in pairs the next time, how much fun they had as well as how pressured they felt in the second run due to their peer on a five-point Likert scale.

[7]We implemented a continuous survey measure of competitiveness using a four-item scale. For this, we asked subjects about their agreement to the following four statements on a seven-point Likert scale: (i) "I am a person that likes to compete with others", (ii) "I am a person that gets motivated through competition", (iii) "I am a person who performs better when competing with somebody", and (iv) "I am a person that feels uncomfortable in competitive situations" and extracted a single principal component factor from those four items, of which the fourth item was scaled reversely.

2.2 Preference elicitation

We used the strategy method to elicit two sets of peer preferences, independent of the treatment to which a subject is assigned. The first set elicited preferences for situations in which social information is available (*name-based preferences*). Accordingly, we asked each student to state his or her six most-preferred peers from the same gender within their class, i.e., those people with whom they would like to be paired in the second run. They could select any person of the same gender, irrespective of this person's actual participation in the study or their attendance in class.[8] These classmates had to be ranked, creating a partial ranking of their potential peers.

Second, we elicited preferences solely based on the relative performance in the first run, ignoring the identities of the potential running partners (*performance-based preferences*). For this purpose, we presented subjects with ten categories comprising one-second intervals starting from $(4, 5]$ seconds slower than their own performance in the first run, to $(0, 1]$ seconds slower and $(0, 1]$ seconds faster up to $(4, 5]$ seconds faster. Appendix Figure A.1 presents a screenshot of the elicitation. We chose the range of intervals such that subjects could choose peers from a range of approximately ± 2 SD from their own performance in the first run. Subjects had to indicate from which time interval they would prefer a peer for the second run, irrespective of the potential peer's identity. Similar to the name-based preferences, we elicited a partial ranking for those performance-based preferences. Accordingly, subjects had to indicate their most-preferred relative time interval, second most-preferred relative time interval and so on.[9]

2.3 Treatments

We exogenously varied how pairs in the second run are formed by implementing one of three matching rules at the class level, where pairs are only formed within genders. The first rule matched students randomly – i.e., we employed a random matching (RANDOM) – and serves as a natural baseline treatment.

The second matching rule used the elicited name-based preferences (NAME) and the third rule formed pairs based on the elicited performance-based preferences (PERFORMANCE). Note that the problem of matching pairs constitutes a typical room-

[8] All subjects were informed that peers in the second run would always have the same gender as themselves and would also need to participate in the study.

[9] Naturally, each time interval could only be chosen once in the preference elicitation, although each interval could potentially include several peers if several subjects had similar times and thus belonged to the same interval. Similarly, some intervals may not contain any peers if no subject in the class had a corresponding time.

mate problem. We thus implemented a "stable roommate" algorithm proposed by Irving (1985) to form stable pairs using the elicited preferences.[10]

Subjects did not know the specific matching algorithm, but were only told that their preferences would be taken into account when forming pairs. Furthermore, we highlighted that the mechanism is incentive-compatible by telling students that it is in their best interest to reveal their true preferences. We informed subjects about the existence of all three matching rules in the survey to elicit both sets of preferences irrespective of the implemented treatment. Just before the second run took place, they were informed about the specific matching rule employed in their class and the resulting pairs.

In addition, we conducted an additional control treatment (NoPeer) in which students ran alone twice and which featured a shortened survey but was otherwise identical to the other treatments.[11] As the focus of this paper is the differential size of peer effects and not their existence per se, this only serves the purpose of excluding learning as a source of time improvements between the two runs. Hence, we exclude it from the main analysis and focus only on the evaluation of different peer assignment rules.

2.4 Procedures

We conducted the experiment in physical education lessons at three secondary schools in Germany.[12] All students from grades 7 to 10 (corresponding to age 12 to 16) of those schools were invited to participate in the experiment. Approximately two weeks prior to the experiment, teachers distributed parental consent forms. These forms contained a brief, very general description of the experiment. Only those students who handed in the parental consent before the study took place participated in the study.

[10]Given the mechanism proposed by Irving (1985), it is a (weakly) dominant strategy for all participants to reveal their true preferences. The matching algorithm requires a full ranking of all potential peers to implement a matching. Since we only elicited a partial ranking, we randomly filled the preferences for each student to generate a full ranking. However, in most cases subjects were assigned a peer according to one of their first three preferences. Nonetheless, if groups were small, it could be the case that subjects were not assigned one of their most-preferred peers. This is especially the case for performance-based preferences. See also the discussion in section 3.1 below.

[11]The survey asked students for their preferences for peers, socio-demographics and their social network. Moreover, in order to avoid deception, we told students in advance that they would run alone both times.

[12]Physical education lessons in most German secondary schools last for two regular lessons of 45 minutes each, thus about 90 minutes in total. At the third school, lessons only lasted 60 minutes for most classes. In order to conduct the experiment in the same manner as at the other schools, we were allowed to extend the lessons by 10 to 15 minutes, which was sufficient to complete the experiment.

The experiment started with a short explanation of the following lesson and a demonstration of the experimental task. A translation of this explanation as well as screenshots detailing the preference elicitation are presented in Appendix A.

We informed students that their teacher would receive each student's times from both runs, but no information about the pairings during the second run.[13] The students themselves did not receive any information on their performance until the completion of the experiment. Additionally, we stressed that both of their performances would be graded by their teacher – thus incentivizing both runs – and that the objective was to run as fast as possible in both runs.[14] Moreover, most students themselves were very interested in their own times. The introduction concluded with a short warm-up period. After this, the subjects were led to a location outside of the gym.

Students entered the gym individually, which ruled out any potential audience effects from classmates being present by design. Students completed the first suicide run and subsequently were handed a laptop to answer the survey. Answering the survey took place in a separate room.[15] After the completion of the survey, subjects returned the laptop to the experimenter and waited with the other students outside the gym. Upon completion of the survey by all students, they returned to the gym to receive further instructions for the second run. In particular, we reminded the students of the existence of the three matching rules, and announced which randomly assigned rule was implemented in their class as well as the resulting pairs from the matching process. Following these instructions, the entire group waited outside the gym again. Pairs were called into the gym and both students participated in the second run simultaneously on neighboring tracks.

After all pairs had finished their second suicide run, the experiment concluded with a short statement by the experimenters thanking the students for their participation. The teacher received a list of students' times in both runs and students were informed about their performance. We then asked the teacher to evaluate the general atmosphere within the class.[16]

[13] Of course, some teachers were present in the gym. In principle, they could observe the pairings and therefore reconstruct the resulting pairs. However, none of the teachers made notes about the pairings or asked for them.

[14] In order for the teacher to grade the entire set of students, the students who did not participate in the study also had to run twice. Their times were recorded for the teacher only and were never stored by us.

[15] At least one experimenter was present at all stages of the experiment to answer questions and limit communication between subjects to a minimum.

[16] Teachers indicated their agreement with three statements on a seven-point Likert scale: (1) "The class atmosphere is very good", (2) "Some students get excluded from the group", and (3) "Students stick together when it really matters".

3 Data description and manipulation check

We present summary statistics of the students in our sample in Table 1.[17] In total, 39 classes with an average class size of about 25 students participated in the experiment. On average, 73% of students within each class subsequently took part in the experiment.[18] This amounts to 627 students who participated in the treatments, with 66% being female.[19] Due to odd numbers of students within some matching groups, we randomly dropped one student in those groups to match students in pairs. Therefore, some students participated in the experiment but were only recorded once and are dropped for estimating the treatment effects in the next section. This procedure yields an estimation sample of 588 observations.

On average, female students took 27.57 seconds (SD of 2.50 seconds) in the first run. Their performance is quite stable across grades, with students from the seventh grade being somewhat slower. Male students' times improved with age: while male students in grade 7 took on average 25.33 seconds in the first run, their performance improved by about two seconds on average in grade 10. In the following, we therefore control for these effects by including gender-specific grade fixed effects in all of our regressions. Independent of their treatment assignment, males and females improved their performance in the second run by .78 seconds and .85 seconds on average, respectively.

We randomized classes into treatments within schools and grades. In Appendix Tables B.1, we check whether observable characteristics differ between our treatments. Overall, randomization seems to be successful. Any small pre-treatment difference is entirely due to our block randomization. Once we take variables used in the randomization (gender, grade, and school fixed effects) into account, the remaining differences disappear and treatments look balanced as shown in Appendix Table B.2.

[17] We focus on the students in the three main treatments, namely RANDOM, NAME and PERFORMANCE and do not include the students from the NOPEER treatment, which is discussed in Appendix E.

[18] We aimed to recruit all students from a class. However, due to numerous reasons this was not possible in every class. Normally, some students are missing on a given day due to sickness or other reasons, are injured and cannot participate in the lesson, are not allowed to take part in the study by their parents or do not want to participate. Additionally, some students simply forgot to hand in the parental consent. We do not have concerns of non-random selection into the study since students did not know in advance the exact day when the experiment was scheduled and most reasons for non-participation were rather exogenous (like injuries or sickness). Moreover, treatment randomization was at the class level within schools and therefore selection into treatments is not possible.

[19] We have more females in our sample since one school in our sample – the smallest one – was a female-only school.

Table 1: Summary statistics

	7th grade	8th grade	9th grade	10th grade	Total
Socio-Demographic Variables					
Age	12.77	13.80	14.77	15.83	14.52
	(0.48)	(0.45)	(0.39)	(0.53)	(1.22)
Female	0.60	0.60	0.66	0.72	0.66
	(0.49)	(0.49)	(0.48)	(0.45)	(0.48)
Times (in sec)					
Time 1 (Females)	28.03	27.06	27.31	27.83	27.57
	(2.75)	(2.06)	(2.28)	(2.71)	(2.50)
Time 2 (Females)	26.98	26.46	26.47	26.94	26.72
	(1.97)	(1.74)	(2.43)	(2.37)	(2.23)
Time 1 (Males)	25.33	24.23	23.71	23.27	24.09
	(1.93)	(1.99)	(2.03)	(2.18)	(2.16)
Time 2 (Males)	24.62	23.58	22.85	22.35	23.31
	(2.01)	(1.99)	(1.70)	(1.50)	(1.98)
Class-level Variables					
# Students in class	25.54	26.00	26.25	25.03	25.68
	(2.71)	(1.96)	(2.56)	(3.17)	(2.74)
Share of participating students	0.75	0.69	0.77	0.71	0.73
	(0.11)	(0.14)	(0.16)	(0.13)	(0.14)
Share of Students in Treatments					
RANDOM	0.32	0.46	0.34	0.32	0.35
	(0.47)	(0.50)	(0.47)	(0.47)	(0.48)
NAME	0.37	0.25	0.37	0.35	0.34
	(0.48)	(0.43)	(0.49)	(0.48)	(0.47)
PERFORMANCE	0.32	0.29	0.29	0.33	0.31
	(0.47)	(0.46)	(0.46)	(0.47)	(0.46)
Observations	123	124	182	198	627

Standard deviations are presented in parentheses. Note that some students only participated in the survey in cases in which they were allowed to participate in the study but were unable to take part in the regular physical education lesson, while some others only took part in the first run if there was an odd number of students in the matching group. See the text for details.

3.1 Preferences for peers and manipulation check

Before turning to the results of the experiment, we briefly present the preferences for peers elicited in the survey. Furthermore, we show that our peer assignment based on those preferences indeed changed the actual match quality, which we define as the rank of the assigned peer in the elicited preference rankings. This means that students in the self-selected treatments had a higher probability of being matched with someone who they preferred more, i.e., who ranked higher in their name- or performance-based preferences. Hence, our experimental variation of taking the preferences into account should have an effect on the rank of the assigned peers within a subject's preferences (i.e., the quality of that match) in the respective treatment with self-selection.

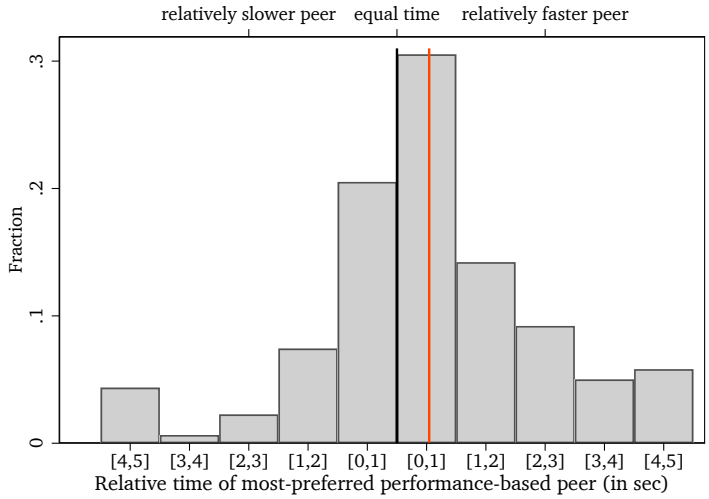
Table 2: Share of name-based preferences being friends

Name-based preference	1st	2nd	3rd	4th	5th	6th	Average
Share of peers being friends	0.89	0.79	0.73	0.60	0.49	0.41	0.65

This table presents the share of friends for each name-based preference (most-preferred peer to sixth most-preferred peer as well as pooled over all six preferences) as elicited in the survey.

We summarize the preferences for peers according to name- and performance-based preferences in Table 2 and Figure 2, respectively. Two findings emerge: first, most students nominated friends as their most-preferred peer; and second, while students on average preferred to run with a slightly faster peer, there is a strong heterogeneity in this preference. We analyze these preferences in further detail in Kiessling, Radbruch, and Schaube (2020).

Figure 2: Most-preferred performance-based peer

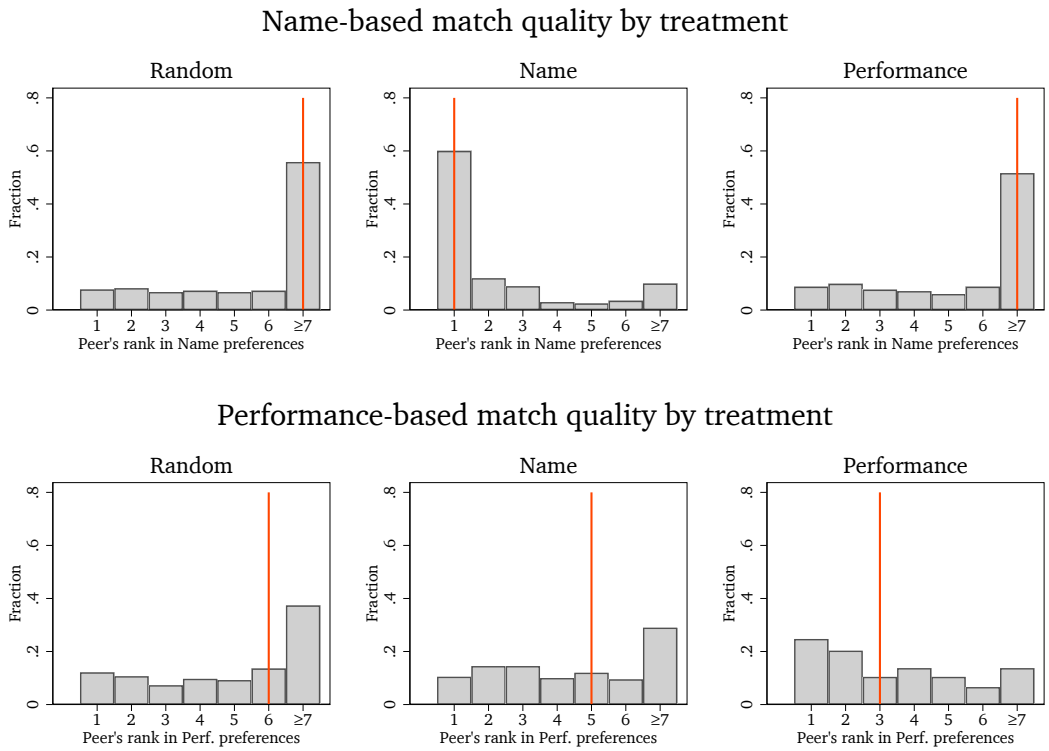


The figure presents a histogram of the peer preferences over relative performance as elicited in the survey. Vertical lines indicate own time (black line; equals zero by definition) and the mean preference of all individuals (red line; 0.56 sec faster on average, where we used the midpoint of each interval to calculate the mean).

Figure 3 shows the realized match quality for all three treatments with respect to the ranking of peers in the two sets of elicited preferences. The upper panel shows the realized match quality according to name-based preferences. We observe that some people were randomly matched to someone with whom they would liked to be paired in RANDOM and PERFORMANCE. As expected, this share is rather low. While the median peer in NAME corresponds to the most-preferred peer according to the elicited name-based preferences, the median peer is not part of the elicited preferences (i.e.,

not among the six most-preferred peers) for RANDOM and PERFORMANCE. A similar, albeit less pronounced picture arises when analyzing the match quality according to the preferences over relative performance as presented in the lower panel of Figure 3. We observe that students in PERFORMANCE were paired with more preferred peers according to their preferences relative to the other two treatments. However, subjects might have preferred other students or relative times that were not available to them, which mechanically affects the match quality. In Appendix B, we check that once we take the mechanical effect into account, the median match quality in PERFORMANCE corresponds to the second most-preferred peer, i.e., we obtain a similarly pronounced pattern as in NAME.

Figure 3: Match quality across treatments



The figure presents a histogram of match qualities for each treatment measured by the rank of the realized peer in an individual's name- (upper panel) or performance-based preferences (lower panels). Vertical red lines denote median ranks.

4 Empirical Strategy

This section outlines our empirical framework. For this purpose, we first analyze the effect of being assigned to a particular peer assignment mechanism. In a second step, we decompose this change in performance into two effects: an indirect effect stemming from a change in the peer composition and a direct effect due to self-selection. Appendix C derives these estimation equations from an economic model similar to a mediation analysis described in Heckman and Pinto (2015).

The random assignment of classes into treatments allows us to estimate the average effect of peer selection on performance. Let $D^d = 1$ with $d \in \{N, P\}$ denote treatment assignment to NAME and PERFORMANCE, respectively, and zero otherwise. We focus on percentage point improvements from the first to the second run, y_{igs}, of individual i in gender-specific grade g of school s as an outcome. Our baseline specification is then given by:

$$y_{igs} = \tau + \tau^N D_i^N + \tau^P D_i^P + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{1}$$

The main parameters of interest are τ^N and τ^P, the effect of being assigned to one of our treatments relative to RANDOM. School fixed effects, ρ_s, and gender-specific grade fixed effects, λ_g, control for variation due to different schools (i.e., as a result of different locations and timing of the experiment) and variation specific to gender and grades.[20] Finally, X_i is a vector of predetermined characteristics such as personality characteristics and – in some specifications – class-level control variables, and u_{igs} is a mean zero error term clustered at the class level.

In addition, we estimate a standard difference-in-differences specification to check the robustness of our findings using individual i's time in run t ($t = 1, 2$), who is in grade g and school s, as an outcome:

$$\begin{aligned} time_{itgs} = \tau &+ \tau_1^N D_i^N + \tau_1^P D_i^P + \tau_2 \mathbb{1}\{t = 2\} \\ &+ \tau_2^N (D_i^N \times \mathbb{1}\{t = 2\}) + \tau_2^P (D_i^P \times \mathbb{1}\{t = 2\}) \\ &+ \gamma X_i + \rho_s + \lambda_g + u_{itgs} \end{aligned} \tag{2}$$

The main parameters of interest are now the coefficients of the interaction terms τ_2^N and τ_2^P, the additional time improvement due to being assigned to one of our treatments relative to RANDOM. School fixed effects, ρ_s, and gender-specific grade fixed effects, λ_g, control for variation due to different schools (i.e., as a result of

[20] See the section 3 for a discussion concerning why we include gender-specific grade fixed effects rather than gender and grade fixed effects separately.

different locations and timing of the experiment) and variation specific to gender and grades. In addition, we control for a range of individual characteristics or present specifications with individual-level fixed effects.

Any change in outcomes can be attributed to one of two main sources: first, different peer-assignment mechanisms may affect peer interactions directly; and second, self-selection may change the peer composition and therefore the difference between the student's and his or her peer's characteristics. To understand the source of the average treatment effect, we decompose it into a direct effect of self-selection as well as a pure peer composition effect.[21] This takes into account the change in relative peer characteristics across treatments. We implement this decomposition using the following specification:

$$y_{igs} = \bar{\tau} + \underbrace{\bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P}_{\substack{\text{Treatments} \\ \text{(direct effects)}}} + \underbrace{\beta\theta_i\left(D^N, D^P\right)}_{\text{Peer characteristics}} + \underbrace{\gamma X_i + \rho_s + \lambda_g}_{\text{Ind. characteristics and FE}} + u_{igs} \tag{3}$$

We are interested in $\bar{\tau}_N$ and $\bar{\tau}_P$, the direct effects of our treatments relative to RANDOM. β denotes the influence of peer characteristics θ_i on the outcome. Changes in peer characteristics through our treatments are captured by changes in $\theta_i\left(D^N, D^P\right)$. In particular, we allow our effects to be mediated through several channels: a first set of channels capture the quality of the match measured by the rank of the peer in an individual's preferences[22], productivity differences measured by absolute differences of times in the first run, and (directed) friendship ties. We allow the effect of these to differ between the faster and slower student in a pair, given that previous research has shown that ranks affect peer interactions.[23]

[21] The direct effect mainly captures changes in performance due to being able to self-select a peer, which we interpret as an increase in autonomy (see section 5.6 for a discussion of the psychological underpinnings). We acknowledge that our definition of a direct effect also captures inputs that (i) differ across treatments, and (ii) are not measured in our rich set of potential mediators (match quality, friendship ties, productivity differences, ranks and personality differences). However, we show in robustness checks that in our setting this is of minor concern only.

[22] We define two indicators to measure whether the assigned peer is nominated among the first three peers for name-based preferences or falls into the three highest ranked categories for performance-based preferences. Alternative specifications are shown in Appendix F.

[23] For example, beginning with Murphy and F. Weinhardt (2018), several studies document the importance of ranks for subsequent outcomes when peers interact with each other (Elsner and Isphording, 2017; Gill et al., 2019). In a related manner, based on theoretical considerations, Cicala, Fryer, and Spenkuch (2018) show that individuals may select themselves into specific peer groups based on their rank within a prospective group, while Tincani (2017) sets up a model in which individuals have preferences over ranks and discusses how this can give rise to heterogeneous peer effects. Common across these studies is their emphasis on the importance of individual rank within groups for peer interactions.

While the existing literature to date has mainly concentrated on the influence of peers with respect to productivity differences and friendship ties on performance, our data allows us to go beyond this.[24] In particular, we allow for a second set of mediators based on the peer's personality and preference measures (i.e., Big Five, locus of control, competitiveness, risk attitudes, social comparison). Additionally, we also include the absolute difference in these personality measures to capture potential non-linear effects.

5 Results

Our experimental design allows to study the causal effect of different peer assignment mechanisms on individual performance. More specifically, we compare three treatments corresponding to random matching (RANDOM), matching with self-selected peers based on name-based peer preferences (NAME) and preferences over relative performance (PERFORMANCE). As outlined in section 2, the random assignment of peers constitutes a natural starting point for at least two reasons: first, the pure presence of any peer might already improve performance; and second, randomly assigned peers are used to document peer effects in a wide range of settings. We contrast this baseline condition with two treatments that assign peers based on elicited preferences, i.e., in which each subject endogenously chooses her peer.

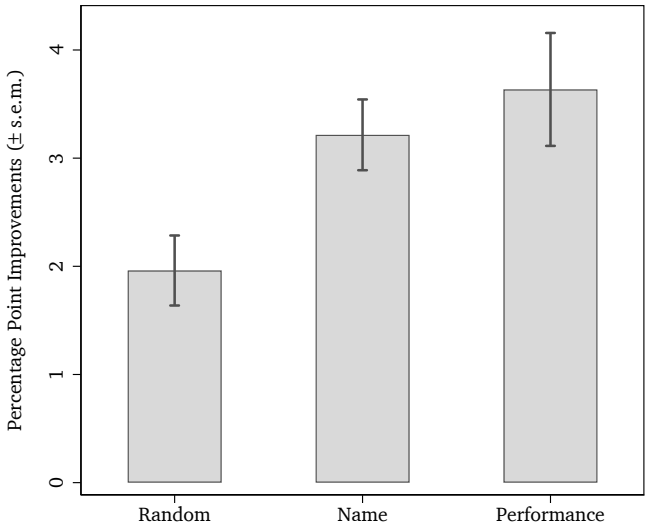
Our empirical results start by documenting average treatment effects. As introduced in section 4, the average treatment effect can stem from two possible sources: if the (relative) characteristics of the peer affect performance and the treatments additionally induce a change in these characteristics, the altered peer composition might explain performance differences across treatments. Moreover, the ability to self-select a peer may directly influence the students' willingness to perform. Before we decompose each treatment effect into a *direct effect* of self-selection and an *indirect effect* due to changes in the peer composition, we establish two necessary conditions for the indirect effect to matter. First, we show that relative peer characteristics matter for individual outcomes. Second, we document that our treatments – which allow for self-selection – indeed change the relative characteristics of peers in the second run. We then decompose the average treatment effects into the two aforementioned channels. Our results conclude with an interpretation of the direct effect and a discussion of implications for peer assignment rules.

[24] Two exceptions include Chan and Lam (2015) and Golsteyn, Non, and Zölitz (2017), who study how peer personality traits affect one's own performance.

5.1 Average effect of self-selection on performance

We analyze how average performance improvements differ between treatments. For this purpose, we use percentage point improvements as outcomes and therefore base our comparisons on the performance in the first run. This specification takes into account the notion that slower students (i.e., those with a slower time in the first run) can improve more easily by the same absolute value compared with faster students, as it is physically more difficult for the latter.

Figure 4: Average performance improvements



The figure presents percentage point improvements from the first to the second run with corresponding standard errors for the three treatments RANDOM, NAME, and PERFORMANCE corresponding to column (2) in Table 3. We control for gender-specific grade fixed effects as well as school fixed effects, and cluster standard errors at the class level.

Figure 4 presents our first result. Subjects in RANDOM improve on average by 1.93 percentage points when paired with a random peer in the second run. However, their performance improves even more in NAME and PERFORMANCE by 3.22 and 3.58 percentage points, respectively.

We present the corresponding estimates in Table 3. Columns (1)–(3) present the estimated percentage point improvements in time according to equation (1). Columns (4)–(6) additionally express the results using a difference-in-differences specification according to equation 2 using times in the two running tasks – or standardized times in column (7) – as outcomes. A pure mean comparison of treatments reveals significant treatment effects of 1.73 and 2.19 percentage points for NAME and PERFORMANCE, respectively. Once we take the block randomization into account by including the corresponding fixed effects in column (2), assigning peers based on

Table 3: Average treatment effects

	(a) Percentage Point Imprv.			(b) Difference-in-Differences			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
NAME	1.73***	1.25***	1.36***	-0.75	-0.05		
	(0.48)	(0.43)	(0.49)	(0.58)	(0.20)		
PERFORMANCE	2.19***	1.67**	1.71**	-0.66	0.12		
	(0.68)	(0.62)	(0.65)	(0.54)	(0.22)		
Second Run				-0.49***	-0.49***	-0.49***	-0.17***
				(0.09)	(0.09)	(0.13)	(0.05)
NAME × Second Run				-0.45***	-0.45***	-0.45**	-0.16**
				(0.13)	(0.13)	(0.18)	(0.06)
PERFORMANCE × Second Run				-0.55***	-0.55***	-0.55**	-0.19**
				(0.19)	(0.19)	(0.27)	(0.09)
Own Characteristics	No	No	Yes	No	No	No	No
Gender-Grade/School FEs	No	Yes	Yes	No	Yes	No	No
Individual FEs	No	No	No	No	No	Yes	Yes
N	588	588	585	1176	1176	1176	1176
R^2	.039	.055	.078	.049	.43	.94	.94
p-value: NAME vs. PERF.	.5	.5	.59	.6	.6	.71	.71

This table presents least squares regressions according to equation (1) using percentage point improvements (panel (a)) and a difference-in-differences specification of running times (panel (b)) as the dependent variable. Own and peer characteristics include the Big 5, locus of control, social comparison, competitiveness and risk attitudes. Column (7) uses standardized times. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

name-based preferences results in an additional 1.25 percentage point improvement in performance relative to the random assignment of peers. The estimated effects for self-selected peers based on relative performance amounts to 1.67 percentage points and thus is somewhat larger, although it does not significantly differ from NAME (p-value= 0.50). These effects persist when controlling for students' own personal characteristics in column (2). Interestingly, the average treatment effects are about the same size as the improvement in RANDOM. On average, students are faster in the second run and this effect is nearly twice as large in PERFORMANCE and NAME compared to RANDOM. Our baseline effects correspond to additional time improvements of .45 to .55 seconds (cf. columns (4)–(6)) and account for 16% of a standard deviation in NAME and 19% in PERFORMANCE (cf. column (7)).[25,26]

[25]Appendix D presents additional robustness checks using biased-reduced linearization or group means to account for the limited number of clusters, specifications that control for outliers and reports the average treatment effects for different subgroups (by gender, grade, school). Our results are robust to all of these checks.

[26]In Appendix E, we document that the observed performance improvements in the three treatments described here are a result of the presence of peers and not due to learning. We present the results of an additional control treatment (NOPEER) and its implementation details. In the control treatment, subjects run twice without any peer and we find that they do not improve their time from the first to

5.2 Peer characteristics matter for individual improvements

Any decomposition of the average effect into a direct effect of self-selection and an indirect effect due to a change in the peer composition relies on two necessary conditions: first, peer characteristics need to be important for determining individual outcomes; and second, relative peer characteristics change when students can self-select their peers. We begin by providing evidence on the former condition, focusing on students in RANDOM. Therefore, we document the importance of peer characteristics by asking how much of the variation of performance improvements in RANDOM can be explained by variation in randomly assigned peer characteristics.

The intuition why peer characteristics may matter is that not all peers have the same effect on someone's performance. For example, friends who serve as a peer might influence us differently than other potential peers. Alternatively, the relative rank within a pair or productivity differences between peers may be driving individual outcomes. If some of these effects exist, then the variation in peer characteristics can explain some of the variation in the performance improvements of subjects in the data and in particular when randomly assigning those characteristics in RANDOM.[27]

In order to show the relevance of peer characteristics, we estimate equation (3) and decompose the corresponding coefficient of determination, R^2, into variation that is attributable to individual characteristics and peer characteristics.[28] We account for an interplay between different groups of explanatory variables by employing a variance decomposition based on Shapley values to calculate the marginal contribution of each group of variables (see Huettner and Sunder, 2012).

Table 4 presents this exercise for students with randomly assigned peers (RANDOM). The decomposition shows that 20% of the total variation can be attributed to characteristics of the peer, which corresponds to 79% of the explained variation. Consequently, only 6% of the total variation or 21% of the explained variation stems from individual characteristics.[29]

---

the second run; in fact, individual performance decreases. The improvements that we observe here can therefore be attributed to the presence of peers rather than learning or familiarity with the task.

[27]Only relative characteristics within a pair can help to explain differences between treatments. Since we randomize subjects into treatments, the overall distribution of peer characteristics across treatments and within classrooms remains constant. Our treatments only change with whom each student interacts within a class, and thus a peer's characteristics relative to one's own characteristics.

[28]As peer characteristics, we include the rank within a pair itself as well as the rank interacted with match quality with respect to both sets of preferences, friendship indicators and productivity differences. We also include personality traits of a peer and absolute differences in personality traits between peers. This corresponds to the full specification that we also use in our decomposition (col. 5 of Table 6).

[29]Note that we explain percentage point improvements from the first to the second run and hence much of the individual-level variation is already taken out of the dependent variable. When using time in the second run as an outcome variable, individual characteristics account for approximately 54%

Table 4: Variance decomposition of performance improvements in RANDOM

Explained variation (R^2)		Variation attributable to: Peer characteristics		Variation attributable to: Individual characteristics	
.26	*(100%)*	.2	*(79%)*	.06	*(21%)*

This table presents a decomposition of the coefficient of determination, R^2, using Shapley values and is based on equation (3) estimated on RANDOM only.

The decomposition therefore shows the importance of accounting for peer characteristics in general. Characteristics of peers are responsible for a large share of the explained variance. In addition, Appendix Table F.1 provides further evidence of peer effects in several dimensions. Hence, we need to take these peer characteristics into account for the analysis of our treatments.

5.3 Self-selection changes the peer composition

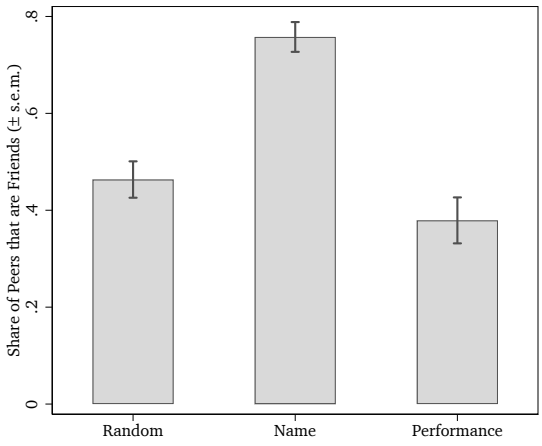
In this section, we document that treatments that allow for self-selection change with whom someone interacts. Although relative peer characteristics are important for understanding outcomes – as shown in the previous section – students also need to interact with systematically different peers when self-selecting them. A second necessary condition for the indirect effect is therefore that the relative peer characteristics have changed.

Figure 5 shows that our treatments indeed changed the peer composition with respect to two prime examples of peer characteristics, namely friendship ties and productivity differences within pairs. More specifically, Figure 5a shows that students are predominantly paired with friends in NAME (76% of all peers are friends), whereas the share of peers being friends in RANDOM and PERFORMANCE is 49% and 37%, respectively. As matching based on preferences over relative performance (PERFORMANCE) allows for targeting of other students with a similar or slightly higher productivity, the students' absolute time differences in the first run might change. Panel B of Figure 5b confirms this by showing that the average absolute difference in times from the first run is 1.53 seconds in PERFORMANCE, while it is larger than two seconds in the other two treatments (2.24 and 2.16 seconds in RANDOM and NAME). Even though students could mainly target peers along these two dimensions, we present

(67% when additionally controlling for time in the first run) of the explained variation ($R^2 = 0.70$ without time in the first run, $R^2 = 0.79$ with time in the first run), while peer characteristics explain the remainder of R^2. Nevertheless, the variation explained from peer characteristics remains sizable.

Figure 5: Changes in peer composition

(a) Share of peers being friends



(b) Absolute differences in productivity

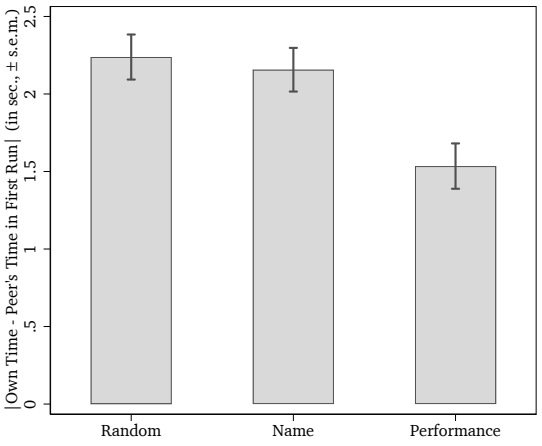


Figure 5a presents the share of all students who nominated their assigned peer as a friend for each of the three treatments including standard errors. Figure 5b shows the average absolute within-pair difference in productivity (measured in times from the first run) and including standard errors for each treatment. We control for gender, grade and school fixed effects, and cluster standard errors at the class level. We present the corresponding regressions and highlight additional compositional differences of the treatments in Appendix Table B.3.

how our treatments affect the peer composition along various other characteristics in Appendix Table B.3. We find that targeting specific peers also results in systematically different peers in terms of their personality.

This establishes that self-selection changes with whom somebody interacts. The endogenously selected peers are neither equal to random peers nor to the average peer. Their characteristics differ with respect to several important dimensions.

5.4 Decomposition into direct and indirect effect of self-selection

We now decompose the average treatment effects from Table 3 by taking changes in the peer composition explicitly into account. As outlined in section 4, the estimated average effects potentially comprise a direct effect as a result of self-selection and an indirect effect stemming from interacting with different peers. This is the case as our treatments have two features: on the one hand, our treatments change with whom someone interacts and those peer characteristics matter as documented above; and on the other, they change the selection procedure from exogenous assignment to the self-selection of peers. The indirect effect therefore captures changes in the relative characteristics of peers (e.g., the time differences between the student and peer in the first run) due to the altered peer composition induced by being able to select them. The

direct effect captures the effect of the treatment due to a change in the selection rule. The previous two subsections documented that NAME and PERFORMANCE change the peer composition relative to RANDOM and established that those relative peer characteristics are important in determining individual outcomes. The decomposition analyzes the extent to which the average treatment effects are driven by these changes in the peer composition.

Table 5: Decomposition of treatment effects

	Direct Effects		Indirect Effects	
	PP imprv.	Std. Err.	PP imprv.	Std. Err.
NAME	1.24	0.50	0.12	0.24
PERFORMANCE	2.24	0.68	-0.54	0.24

The table presents the resulting direct and indirect effects from a decomposition according to equation (3) shown in column (5) of Table 6. Indirect effects are defined as the changes in percentage point improvements that are explained by changes in peer characteristics relative to RANDOM and comprises the combined effect of all peer characteristics in column (5) of Table 6.

The results of the decomposition based on equation (3) are summarized in Table 5 and presented in further detail in Table 6. In Table 5, we use the whole set of characteristics to decompose the average treatment effects into the direct and indirect effects. Therefore, the size of the direct effects equals the coefficients of the treatment indicators in column (5) of Table 6. They correspond to 1.24 percentage points in NAME and 2.24 in PERFORMANCE.

The decomposition shows that even though peer characteristics are highly important in understanding the variation in outcomes, the indirect effects of self-selection in the two treatments are considerably low. They correspond to only 10% of the size of the direct effect in NAME and 24% in PERFORMANCE.[30] In NAME, we estimate a positive and insignificant indirect effect of .12 percentage point improvements (p-value = 0.59). This means that the altered peer characteristics have only a slightly positive effect on the students' performance. For PERFORMANCE, we find a significant indirect effect of -.54 percentage points (p-value = 0.03). Thus, the change in the peer composition even magnifies the direct effect as it negatively rather than positively affects performance.

[30]The indirect effect in our decomposition is induced by the impact of peer characteristics and their change through self-selection.Therefore, it corresponds to the difference in the average effect for NAME and PERFORMANCE and the direct effect as the direct and indirect effect add up to the average effect. The indirect effect also corresponds to multiplying the coefficients for (relative) peer characteristics from column (5) with the change in the peer composition across treatments, as described in Appendix C and Appendix Table B.3.

Therefore, our decomposition shows that while self-selection of peers indeed changes the composition of peers, these changes cannot explain the average treatment effects; rather, the additional performance improvements in NAME and PERFORMANCE stem from a direct effect of self-selection.

Table 6: Decomposition of treatment effects

	Percentage Point Improvements				
	(1) Baseline	(2) Match Quality	(3) Friend-ship ties	(4) Time Difference	(5) All
Direct Effects					
NAME	1.36***	1.22**	1.45***	1.34***	1.24**
	(0.49)	(0.53)	(0.49)	(0.46)	(0.50)
PERFORMANCE	1.71**	1.81***	1.63**	1.87***	2.24***
	(0.65)	(0.65)	(0.65)	(0.61)	(0.68)
Peer Characteristics					
Faster Student		-0.03			0.49
× High match quality (NAME)		(0.39)			(0.44)
Slower Student		0.32			0.49
× High match quality (NAME)		(0.61)			(0.65)
Faster Student		1.13**			0.42
× High match quality (PERF.)		(0.52)			(0.53)
Slower Student		-2.14***			-0.75
× High match quality (PERF.)		(0.62)			(0.66)
Faster Student			-0.79*		-1.15**
× Peer is Friend			(0.45)		(0.53)
Slower Student			-0.06		0.11
× Peer is Friend			(0.53)		(0.66)
Faster Student				-0.38**	-0.34**
× $\lvert\Delta Time\ 1\rvert$				(0.14)	(0.16)
Slower Student				1.04***	1.05***
× $\lvert\Delta Time\ 1\rvert$				(0.21)	(0.20)
Slower Student in Pair		3.82***	2.15***	-0.21	-0.19
		(0.43)	(0.48)	(0.44)	(0.67)
Abs. Diff. in Personality	No	No	No	No	Yes
Peer Characteristics	No	No	No	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes	Yes	Yes
N	585	585	585	585	582
R^2	.078	.18	.15	.24	.29
p-value: NAME vs. PERFORMANCE	.59	.38	.79	.4	.15

This table presents least squares regressions according to equation (3) using percentage point improvements as the dependent variable. High match quality is an indicator that equals one if the partner was ranked within an individual's first three preferences. Personality characteristics include the Big Five, locus of control, social comparison, competitiveness, and risk attitudes. Appendix Table F.9 presents the omitted coefficients of own and peer characteristics, and their absolute differences for our preferred specification in column (5). *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

We now analyze the detailed results of the decomposition in Table 6. Column (1) replicates the baseline estimates from column (2) of Table 3 for means of compari-

son. In columns (2)–(4), we include different sets of peer characteristics, before we include all of them in column (5). Turning to the separate columns, we find that the size of the treatment indicators only slightly differ across specifications. Nonetheless, some of the included peer characteristics influence the individual performance in the second run. Performance-based match quality has some predictive power for performance improvements in the restricted regression in column (2). However, the effects are insignificant when controlling for all peer characteristics in column (5). Overall, the quality of the match, i.e., how well a student's preferences were satisfied by the pairing in the second run, has little to no effect on their performance. We also observe that initially faster students within a pair reduce their performance when paired with a friend, while the relatively slower students do not adjust their performance differentially for friends as peers (column (3) and (5)). In column (4), we focus on productivity differences, since faster and slower students within a pair might be affected differentially. We also allow the effect of productivity differences, $|\Delta Time1|$, to differ by the rank within a pair. We find that differences in times of the first run have a significant effect on both faster and slower students within a pair. While slower students within a pair benefit by a 1.04 percentage point improvement from running with a one second faster student, the relatively faster student's performance suffers from this productivity difference by .38 percentage points. In sum, the average performance of a pair thus improves with increasing differences in productivity.

We control for all of these characteristics jointly in column (5), where we also add a rich set of relative peer personality characteristics. The effect of friendship ties on the initially faster students as well as the effects on productivity differences persist. More importantly, the direct effects of both NAME and PERFORMANCE remain robust, showing a direct effect of self-selection on individual performance.

5.5 Robustness of the decomposition

The results reported in the previous section provide evidence for a direct effect of self-selection. In this section, we provide further evidence for the existence of such an effect from several robustness checks.

Peer characteristics remain important. We replicate the variance decomposition of Table 4 for all three treatments in Panel A of Table 7 and confirm the importance of the peer characteristics in terms of explaining the variation in outcomes. In particular, 73% of the explained variation stems from variation peer characteristics.

Table 7: Variance decomposition and the role of unobservables

Panel A: Variance decomposition								
		Variation attributable to						
Explained variation (R^2)		Treatments		Peer characteristics		Individual characteristics		
0.29	*(100%)*	0.03	*(12%)*	0.21	*(73%)*	0.04	*(15%)*	

Panel B: Role of unobservables			
	Oster's δ		
	$R^2_{max} = 0.50$	$R^2_{max} = 0.75$	$R^2_{max} = 1.00$
NAME	2.59	1.22	0.79
PERFORMANCE	-6.46	-3.16	-2.09

Panel A decomposes the explained variance of specification (5) of Table 6 in components attributable to treatments, peer and individual characteristics similar to Table 4. Panel B quantifies the importance of unobservables relative to observables needed for zero direct effects according to Oster (2019).

Omitted variables do not seem to drive our results. In Panel B of Table 7 we address the possible concern that other characteristics for which we cannot account or control are driving the direct effect. Our results above remain relatively stable when adding different sets of peer controls, which is reassuring. A more formal approach to tackle this concern is to ask how important unobserved characteristics would have to be to explain our direct treatment effects (Altonji, Elder, and Taber, 2005; Oster, 2019). We follow Oster (2019) and calculate δ, a measurement for the relative importance of unobserved characteristics compared to observed characteristics. This measure describes how important unobserved variables would have to be relatively to observed ones to explain the direct effects, i.e., to drive down the direct effects to zero. Absolute values of δ larger than one indicate that these omitted variables have to be relatively more important than observed peer characteristics. Negative values indicate that those unobservable characteristics need to reverse the effect of observed covariates. We calculate these measures for three scenarios that differ in the maximum amount of variance that would theoretically be explained if all factors that might affect the outcomes were observed. More specifically, we calculate δ for R^2_{max} equal to 0.50, 0.75 and 1.00. In all but one extreme scenario the omitted peer characteristics are required to be more important than the observed peer characteristics. This suggests that such unobserved characteristics need to have a larger effect than productivity differences, friendship ties, match quality and all other controls – including personality traits – combined. Compared to other studies, our analysis already allows for more peer characteristics to influence subjects' behavior. There-

fore, we allow for a very rich set of important characteristics and conclude that such unobserved characteristics are highly unlikely to drive the direct treatment effects.

Estimating peer effects on RANDOM only does not affect our results. One might worry that estimating peer effects on all three treatments jointly may bias our estimates. We therefore perform the following alternative estimation strategy. We estimate the peer coefficients on the subsample of students in RANDOM, resulting in unbiased estimates due to random assignment of peers. In a second step, we then impose these estimates on the two treatments featuring self-selected peers when estimating our main specification. Essentially, we are calculating the predicted performance improvements and compare them to the realized improvements to recover the direct effects of self-selection. Appendix Table F.2 shows that imposing peer effects from RANDOM on the other two treatments does not change our conclusions. In particular, the direct effects remain significant, although slightly lower for the NAME treatment.

Results are robust to different definition of key variables. In addition, we provide several robustness checks relaxing the definition of key variables used in our decomposition. Appendix Table F.3 allows for different specifications of match quality by additionally considering the partner's match quality, an interaction between one's own and the partner's match quality, as well as feasible match quality. Appendix Table F.4 considers different definitions of friendship ties apart from directed links (i.e., undirected, reciprocal, directed and reciprocal friendship ties). Appendix Table F.5 considers the time in the second run as an outcome instead of percentage point improvements. The results for all robustness checks remain qualitatively and quantitatively similar. Furthermore, we show in Table F.6 and Appendix Figure F.1 that the linear specification of productivity differences is not restrictive.

Control group comprising only high-match peers does not alter our conclusions. In Table F.7, we restrict the control group to those subjects that received one of their preferred peers by pure chance and estimate the direct effects using this control group. These matches occurred by pure chance and not due to self-selection. We find that the direct effects persist when restricting to these subgroups as a control group.

Direct effects are robust to additional class-level controls and are not an artefact of overfitting controls. In order to further probe the robustness of our findings, we additionally control for proxies of the class attitude in Appendix Table F.8. While

the estimates slightly differ in magnitude, the results are generally robust. Another concern might be that by controlling for many different own and peer variables, our results might be due to overfitting. We address this concern by adopting a post-double Lasso estimator proposed by Belloni, Chernozhukov, and Hansen (2014), which penalizes control variables but permitting valid inference on treatment effects. Our results are robust to this data-driven selection of control variables.

Taken together, our analysis shows that self-selection improves individual performance directly and not due to a change in the peer composition. This means that subjects react to observationally similar peers differently once they have chosen them actively. Characteristics of peers are important in determining outcomes, but they do not explain the average treatment effects of self-selection, which are driven by the direct effect of self-selection. Although our treatments allowed for two different notions of self-selection, it is reassuring that the estimates of the direct effects are similar across treatments.

5.6 Interpretation of the direct effect

We interpret the direct effect as a positive effect of self-selection due to increased control or autonomy over the peer assignment mechanism. However, one might worry that knowledge of all three treatment conditions could lead students in RANDOM to react negatively due to disappointment that their preferences have not been taken into account.[31] If these disappointed students drove our findings, we would falsely attribute effects to self-selection even if students in NAME and PERFORMANCE do not react positively.[32] If the direct effect originated from disappointment, we would expect students in RANDOM to have less fun in the experimental task. Therefore, in column (1) of Appendix Table G.1 we analyze the extent to which subjects across treatments had different perceptions regarding their fun in the second run. We find zero effects. The absence of direct effects in the fun dimension alleviates the potential concern

[31] This results from the fact that we elicited preferences for peers irrespective of the treatment and only announced the assignment rule after the survey, but before the second run.

[32] At the same time, this also describes a feature of many real-world settings. Imagine that a person is randomly assigned a partner from a group of available people. Even if this person has not been asked explicitly with whom she would like to interact, she still has preferences about interacting with certain people. Therefore, disappointment could also play a role in these settings. This might be true for all settings that feature exogenous assignment and overrule the underlying preferences of the involved persons.

that knowledge of all three treatments leads to disappointment when students are assigned to RANDOM.[33]

We therefore conclude that the direct effects in our experiment are due to positive effects of self-selection. More specifically, we argue that the opportunity to self-select key aspects of one's environment – in our experiment having autonomy over the peer selection – has a direct effect beyond the instrumental value of changing peer characteristics. Self-determination theory provides a credible explanation through which self-selection can impact performance directly. The theory identifies autonomy as a crucial determinant of motivation: individuals who can actively select parts of their environment – most importantly their tasks in work environments – display higher intrinsic motivation (Deci and Ryan, 1985, 2000).[34] Applying this explanation to our setting suggests that not the selected peer herself increases motivation, but the mere act of selecting her. However, we do not argue that this behavioral effects stems from self-selecting any aspect, but a relevant aspect of one's environment.

Self-determination theory and autonomy in particular have recently gained increasing attention from economists. Cassar and Meier (2018) review the economic literature on non-monetary aspects of work environments in light of self-determination theory and highlight the importance of autonomy for various behavioral outcomes. A related argument to ours also underlies the findings of Bartling, Fehr, and Herz (2014) and Owens, Grossman, and Fackler (2014). Although they do not focus on the effect of autonomy on subsequent outcomes, their studies demonstrate that people have a willingness to pay for making decisions by themselves and maintaining autonomy. Similarly, a growing body of literature demonstrates that restricting subjects choice sets and therefore restricting their autonomy and freedom can negatively influence outcomes (e.g., Falk and Kosfeld, 2006). Therefore, our results add to this literature

[33]A related issue would be that the direct effect stems from a positive effect of subjects in treatments with self-selection as they may react reciprocal towards being treated kindly (see Aldashev, Kirchsteiger, and Sebald, 2017, for an analysis how reciprocity can influence treatment effects). If students prefer to be in one of the self-selection treatments (NAME or PERFORMANCE) rather than in RANDOM and they perceive their assignment as kind, reciprocal students could respond by increasing their performance. This in turn would imply that the direct effects of our treatments are due to reciprocity or some kind of experimenter demand effects. Then prosocial students should display a stronger (direct) effect than non-reciprocal students as they are are more likely to react reciprocally. We proxy prosociality by scoring higher on the agreeableness scale of the Big Five as it is significantly correlated with reciprocity and altruism (Becker et al., 2012). Column (3) in Appendix Table G.1 reports the interaction between the agreeableness score and treatment indicators. If the above motives are the underlying causes of the direct treatment effect, we should observe a positive and statistically significant interaction between agreeableness and the treatments. However, our results do not show this relationship. We interpret this finding as evidence against reciprocal motives driving our results.

[34]Two other components of self-determination theory are relatedness and competence, referring to the need to care about something and the need to feel challenged, respectively. In our experiment, we hold these other components constant across treatments.

by highlighting the motivational benefits of autonomy and self-determination, and provide novel field evidence that having control positively affects outcomes.

5.7 The limits of peer assignment rules

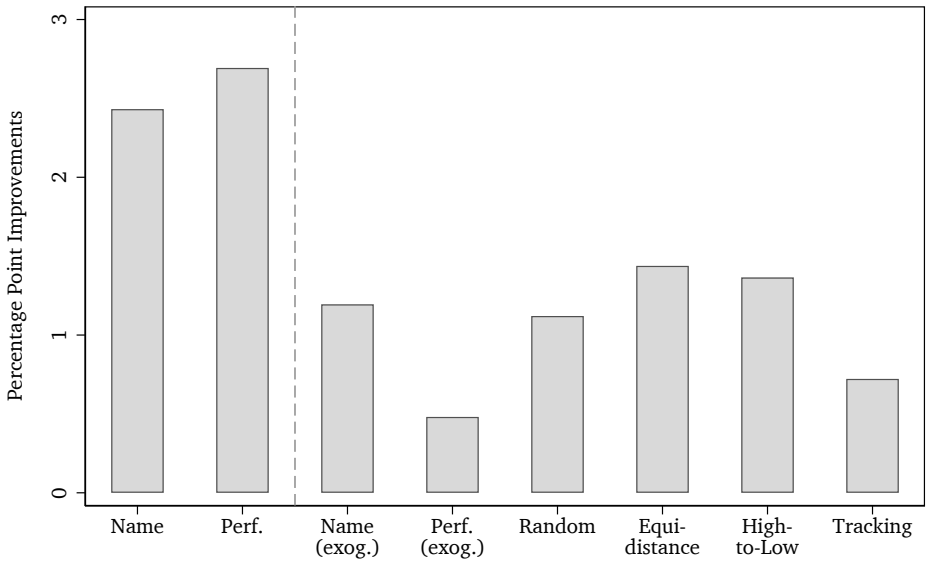
Our results show that self-selected peers lead to substantially larger performance improvements than randomly assigned peers. In practice, however, policy makers frequently do not assign peers at random. Rather, they employ a variety of peer assignment rules to help or target specific individuals. Examples include schools employing tracking (e.g., Betts, 2011; Duflo, Dupas, and Kremer, 2011; Fu and Mehta, 2018; Garlick, 2018) or pairing high-performing students with low-performing ones (e.g., Carrell, Sacerdote, and West, 2013). While we have not implemented these assignment rules in our context, we can use our estimates to simulate the effect of such exogenous peer assignment rules and compare their effect to outcomes under self-selection.

For this purpose, we use our estimates obtained in section 5.4, using the whole set of peer characteristics (column (5) of Table 6). Based on these estimates, we simulate different (exogenous) assignment rules, calculate the resulting effects on performance, and compare them to performance improvements observed in our experiment. We first compare the improvements to the counterfactual of assigning the same peers in NAME and PERFORMANCE without the direct effect of self-selection. A comparison of other peer assignment rules with these results sheds light on the question of whether students are able to choose optimal peers. Second, we simulate the expected performance improvements under a random matching. Third, we use several assignment rules that base the assignment on one single and commonly employed peer characteristic, namely past performance. Our estimates obtained in section 5.4 suggest that pairs with a higher difference in initial performances will improve their performance on average. If this is the only characteristic of a peer that affects performance, aggregate performance would be maximized as long as the sum of productivity differences within a pair is maximized.[35] In order to compare the results of self-selection against exogenous assignment rules that promise the largest aggregate improvements, we consider two matching rules that maximize these productivity differences within pairs – EQUIDISTANCE and HIGH-TO-LOW – that keep the distance in ranks within the class constant or pair the best-performing student with the slowest student. Additionally, we look at the effect of tracking (i.e., pair-

[35] Given our specification, this is true for all peer-assignment rules that match each student from the bottom half of the productivity distribution with a student from the top half.

ing the best student with the second best, third with the fourth, etc.; TRACKING). Importantly, while all of these exogeneous assignment rules are based on past performance, we take the effects of all peer characteristics into account. We compare the predicted performance improvements for those rules with our estimated performance improvements for the three assignment rules used in the experiment.[36]

Figure 6: Simulation of different peer assignment rules



The figure presents predicted percentage point improvements for the two treatments (NAME, PERFORMANCE) with and without the effect of self-selection, the RANDOM-treatment as well as three simulated peer assignment rules (EQUIDISTANCE, HIGH-TO-LOW and TRACKING). We fix the personal characteristics and other covariates not at the pair level to 0, whereby effect sizes are therefore not directly comparable to treatment effects above. More details are provided in the text and Appendix H.

Figure 6 presents the simulated average performance improvements of each assignment rule. The results show that no other peer assignment rule is able to reach similar performance improvements as those featuring self-selection. In fact, they are close to the results from our random matching, since students under those peer assignment rules do not benefit from the additional intrinsic value of self-selection. We observe that in the absence of a direct effect of self-selection, students do not experience additional improvements relative to randomly assigned peers. Compared to EQUIDISTANCE and HIGH-TO-LOW, students in NAME (EXOG.) and PERFORMANCE (EXOG.) perform worse indicating that they do not choose their peers optimally.

More surprisingly, the reassignment rules that maximize productivity differences in pairs – EQUIDISTANCE and HIGH-TO-LOW – do not improve average performance

[36]We provide details on the prediction of performance improvements and the peer assignment rules in Appendix H.

compared to the random assignment of peers. Although both rules increase the average productivity difference in pairs by construction and affect performance through this channel, those rules also change other characteristics of the peer. The lack of any additional improvement suggest that these other changes in peer characteristics offset the positive effect of increased productivity differences.

The limited effectiveness of these peer assignment rules stems from two effects: First, there exist (non-linear) peer effects in several dimensions as documented in the previous sections. Second, varying one peer characteristic simultaneously changes other observed and potentially unobserved peer characteristics that may counteract the intended effect. Together, these two effects imply that the consequences peer assignment rules are difficult to predict or even ambiguous if peer effects exist in multiple dimensions.[37] This insight further helps to understand why we observe a very small indirect effect in the decomposition of the treatment effects despite the fact that peer characteristics help to explain much of the variation in individual outcomes (cf. Table 6).

The simulations above suggest that self-selection of peers can be an attractive alternative compared to traditional peer assignment rules to increase individual performance. However, we want to stress that such peer assignments based on self-selection may also come at a cost. In particular, we show in Appendix Table H.2 that students in PERFORMANCE experience significantly more pressure compared to the other two treatments, and individual ranks may be more perturbed between the two runs in NAME and RANDOM relative to PERFORMANCE. Hence, a policy maker might not only look at the resulting performances but also how different assignment rules affect the individuals' overall well-being.

6 Conclusion

Peer effects are an ever-present phenomenon discussed in a wide range of settings across the social sciences. For many situations, identifying the effect of an actively self-chosen peer is important beyond estimating peer effects in general. Our framed field experiment introduces a novel way to study the self-selection of peers in a controlled manner and is able to separate the impact of a specific peer on a subject's performance from the overall effect of self-selection. The results of our experiment provide evidence that self-selecting peers yields performance improvements of about 16–19% of a

[37] Consequently, designing optimal peer assignment rules might be more challenging than expected as finding an optimal assignment requires an optimization taking into account all potential characteristics in which peer effects may exist. This creates a high-dimensional optimization problem that is highly difficult to solve.

standard deviation relative to random assignment of peers. While peer characteristics affect the individual performance, they are not the origin of the estimated treatment effects. Rather, these improvements stem from a direct effect of self-selection. Based on self-determination theory (Deci and Ryan, 1985), we interpret this direct effect such that the ability to select one's own peer enhances a student's intrinsic motivation and subsequently increases individual performance.

One might be eager to infer that our results give rise to a trade-off between performance improvements as a result of self-selection per se and the exogenous assignment of performance-maximizing peers. However, our simulations show that exogenous peer assignment rules, which try to lever peer effects in ability, have an impact close to zero in our case and are in general ambiguous in size and sign. This result relies on the existence of peer effects in multiple dimensions, which at least partially offset each other and in turn limit the effectiveness of exogenous reassignment rules. Hence, positive effects of peer self-selection might be performance-maximizing – even in the absence of subjects choosing "optimal" peers.

The results in this paper constitute a first proof-of-concept that self-selection of peers can directly affect performance. Yet, it is crucial to investigate whether and how this effect transfers to other situations and mechanisms of peer effects. Since our results stem from a rather specialized field setting designed to explicitly isolate the effects of self-selection, it is not clear ex-ante whether they persist over an extended period of time, in cooperative environments, or other settings. Hence, one could easily transfer our experimental design to situations in which other production functions are used or where peer effects arise via other channels, e.g., by implementing team production by reporting a function of both students' times to the teacher, or by varying the task to allow for learning or skill complementaries as sources of peer effects. Moreover, in many other settings peers can also be self-selected or such a mechanism could be implemented. For example, study groups at universities often form endogenously (Chen and Gong, 2018), researchers select their co-authors, workers in firms increasingly form self-managed work teams (Lazear and Shaw, 2007), and employees self-select with whom they work by referring others to their employer (Friebel et al., 2019; Lazear and Oyer, 2012). Hence, teachers or supervisors might be interested to explore whether leveraging such a direct effect of self-selection can improve performance in addition to other forms of non-monetary incentives used in schools (Levitt et al., 2016) or workplaces (Cassar and Meier, 2018).

In this paper, we highlight that self-selecting peers can serve as a complement to other established methods such as incentives and exogenous peer assignment policies aimed at increasing individual performance. However, further research on the

interplay between endogenous group formation, social interactions and production environments remains imperative to understand how peer effects work.

References

Ager, Philipp, Leonardo Bursztyn, and Hans-Joachim Voth (2016). "Killer Incentives: Status Competition and Pilot Performance during World War II". In: NBER Working Paper Series.

Aldashev, Gani, Georg Kirchsteiger, and Alexander Sebald (2017). "Assignment Procedure Biases in Randomised Policy Experiments". In: *Economic Journal* 127.602, pp. 873–895.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber (2005). "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools". In: *Journal of Political Economy* 113.1, pp. 151–184.

Aral, Sinan and Christos Nicolaides (2017). "Exercise Contagion in a Global Social Network". In: *Nature Communications* 8.14753.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2009). "Social Connections and Incentives in the Workplace: Evidence From Personnel Data". In: *Econometrica* 77.4, pp. 1047–1094.

— (2010). "Social Incentives in the Workplace". In: *Review of Economic Studies* 77.2, pp. 417–458.

Bartling, Björn, Ernst Fehr, and Holger Herz (2014). "The intrinsic value of decision rights". In: *Econometrica* 82.6, pp. 2005–2039.

Bartling, Björn, Ernst Fehr, and Klaus M. Schmidt (2013). "Discretion, productivity, and work satisfaction". In: *Journal of Institutional and Theoretical Economics* 169.1, pp. 4–22.

Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse (2012). "The Relationship Between Economic Preferences and Psychological Personality Measures". In: *Annual Review of Economics* 4.1, pp. 453–478.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on Treatment Effects after Selection among High-Dimensional Controls†". In: *Review of Economic Studies* 81.2, pp. 608–650.

Betts, Julian R. (2011). "The Economics of Tracking in Education". In: ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3. Handbook of the Economics of Education. Elsevier, pp. 341–381.

Bó, Pedro Dal, Andrew Foster, and Louis Putterman (2010). "Institutions and Behavior: Experimental Evidence on the Effects of Democracy". In: *American Economic Review* 100.5, pp. 2205–2229.

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek (2017). "Ability Peer Effects in University: Evidence from a Randomized Experiment". In: *Review of Economic Studies* 84.2, pp. 547–578.

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non (2016). "Employee Recognition and Performance: A Field Experiment". In: *Management Science* 62.11, pp. 3085–3099.

Brandts, Jordi, David Cooper, and Roberto Weber (2014). "Legitimacy, Communication, and Leadership in the Turnaround Game". In: *Management Science* 61.11, pp. 2627–2645.

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman, and Noam Yuchtman (2014). "Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions". In: *Econometrica* 82.4, pp. 1273–1301.

Carrell, Scott, Bruce Sacerdote, and James West (2013). "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation". In: *Econometrica* 81.3, pp. 855–882.

Cassar, Lea and Stephan Meier (2018). "Nonmonetary Incentives and the Implications of Work as a Source of Meaning". In: *Journal of Economic Perspectives* 32.3, pp. 215–38.

Chan, Tszkin Julian and Chungsang Tom Lam (2015). "Type of Peers Matters: A Study of Peer Effects of Friends Studymates and Seatmates on Academic Performance".

Chen, Roy and Jie Gong (2018). "Can self selection create high-performing teams?" In: *Journal of Economic Behavior and Organization* 148, pp. 20–33.

Chevalier, Judith A., M. Keith Chen, Peter E. Rossi, and Emily Oehlsen (2019). "The Value of Flexible Work: Evidence from Uber Drivers". In: *Journal of Political Economy* 127.6, pp. 2735–2794.

Cicala, Steve, Roland Fryer, and Jörg Spenkuch (2018). "Self-Selection and Comparative Advantage in Social Interactions". In: *Journal of the European Economic Association* 16.4.

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-González (2015). "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough". In: *Management Science* 61.12, pp. 2926–2944.

Deci, Edward and Richard Ryan (1985). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media.

— (2000). "The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior". In: *Psychological Inquiry* 11.4, pp. 227–268.

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner (2011). "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences". In: *Journal of the European Economic Association* 9.3, pp. 522–550.

Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya". In: *American Economic Review* 101.5, pp. 1739–1774.

Elsner, Benjamin and Ingo Isphording (2017). “A Big Fish in a Small Pond: Ability Rank and Human Capital Investment”. In: *Journal of Labor Economics* 35.3, pp. 787–828.

Falk, Armin and Michael Kosfeld (2006). “The Hidden Costs of Control”. In: *American Economic Review* 96.5, pp. 1611–1630.

Friebel, Guido, Matthias Heinz, Mitchell Hoffman, and Nick Zubanov (2019). “What Do Employee Referral Programs Do?”

Fu, Chao and Nirav Mehta (2018). “Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation”. In: *Journal of Labor Economics* 36.4, pp. 923–979.

Garlick, Robert (2018). “Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment”. In: *American Economic Journal: Applied Economics* 10.3, pp. 345–369.

Gibbons, Frederick and Bram Buunk (1999). “Individual Differences in Social Comparison: Development of a Scale of Social Comparison Orientation.” In: *Journal of Personality and Social Psychology* 76.1, pp. 129–147.

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse (2019). “First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision”. In: *Management Science* 65.2, pp. 494–507.

Gneezy, Uri and Aldo Rustichini (2004). “Gender and Competition at a Young Age”. In: *American Economic Review* 94.2, pp. 377–381.

Golsteyn, Bart, Arjan Non, and Ulf Zölitz (2017). “The Impact of Peer Personality on Academic Achievement”.

Harrison, Glenn and John List (2004). “Field Experiments”. In: *Journal of Economic Literature* 42.4, pp. 1009–1055.

Heckman, James and Rodrigo Pinto (2015). “Econometric Mediation Analyses: Identifying the Sources of Treatment Effects from Experimentally Estimated Production Technologies with Unmeasured and Mismeasured Inputs”. In: *Econometric Reviews* 34.1-2, pp. 6–31.

Herbst, Daniel and Alexandre Mas (2015). “Peer Effects on Worker Output in the Laboratory Generalize to the Field”. In: *Science* 350.6260, pp. 545–549.

Huettner, Frank and Marco Sunder (2012). “Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values”. In: *Electronic Journal of Statistics* 6, pp. 1239–1250.

Irving, Robert (1985). “An Efficient Algorithm for the “Stable Roommates” Problem”. In: *Journal of Algorithms* 6.4, pp. 577–595.

Jackson, C. Kirabo and Elias Bruegmann (2009). “Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers”. In: *American Economic Journal: Applied Economics* 1.4, pp. 85–108.

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube (2020). "Determinants of Peer Selection".

Koch, Alexander K. and Julia Nafziger (2011). "Self-regulation through Goal Setting". In: *Scandinavian Journal of Economics* 113.1, pp. 212–227.

Kosfeld, Michael and Susanne Neckermann (2011). "Getting More Work for Nothing? Symbolic Awards and Worker Performance". In: *American Economic Journal: Microeconomics* 3.3, pp. 86–99.

Lavy, Victor and Edith Sand (2019). "The Effect of Social Networks on Students' Academic and Non-cognitive Behavioural Outcomes: Evidence from Conditional Random Assignment of Friends in School". In: *Economic Journal* 129.617, pp. 439–480.

Lazear, Edward and Paul Oyer (2012). "Personnel Economics". In: *The Handbook of Organizational Economics*. Princeton University Press, pp. 479–519.

Lazear, Edward and Kathryn Shaw (2007). "Personnel Economics: The Economist's View of Human Resources". In: *Journal of Economic Perspectives* 21.4, pp. 91–114.

Levitt, Steven, John List, Susanne Neckermann, and Sally Sadoff (2016). "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance". In: *American Economic Journal: Economic Policy* 8.4, pp. 183–219.

Manski, Charles (1993). "Identification of Endogenous Social Effects: The Reflection Problem". In: *Review of Economic Studies* 60.3, pp. 531–542.

Mas, Alexandre and Enrico Moretti (2009). "Peers at Work". In: *American Economic Review* 99.1, pp. 112–145.

Murphy, Richard and Felix Weinhardt (2018). "Top of the Class: The Importance of Ordinal Rank".

Oster, Emily (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence". In: *Journal of Business & Economic Statistics* 37.2, pp. 187–204.

Owens, David, Zachary Grossman, and Ryan Fackler (2014). "The Control Premium: A Preference for Payoff Autonomy". In: *American Economic Journal: Microeconomics* 6.4, pp. 138–161.

Rotter, Julian B. (1966). "Generalized Expectancies for Internal Versus External Control of Reinforcement". In: *Psychological Monographs: General and Applied* 80.1, pp. 1–28.

Sacerdote, Bruce (2001). "Peer Effects with Random Assignment: Results for Dartmouth Roommates". In: *Quarterly Journal of Economics* 116.2, pp. 681–704.

— (2011). "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In: ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3. Handbook of the Economics of Education. Elsevier, pp. 249–277.

Schneider, Simone and Jürgen Schupp (2011). “The Social Comparison Scale: Testing the Validity, Reliability, and Applicability of the IOWA-Netherlands Comparison Orientation Measure (INCOM) on the German Population”. In: *DIW Data Documentation*.

Sutter, Matthias and Daniela Glätzle-Rützler (2015). “Gender Differences in the Willingness to Compete Emerge Early in Life and Persist”. In: *Management Science* 61.10, pp. 2339–2354.

Sutter, Matthias, Stefan Haigner, and Martin G. Kocher (2010). “Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations”. In: *Review of Economic Studies* 77.4, pp. 1540–1566.

Tincani, Michela (2017). “Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence”.

Weinhardt, Michael and Jürgen Schupp (2011). “Multi-Itemskalen im SOEP Jugendfragebogen”. In: *DIW Data Documentation*.

Appendix – For Online Publication

A	Experimental instructions and protocol
B	Randomization and manipulation check
C	Econometric framework
D	Robustness checks for average treatment effects
E	Control treatment to disentangle peer effects from learning
F	Peer composition robustness checks
G	Additional material for discussion of direct effects
H	Simulation of matching rules and side effects

A Experimental instructions and protocol

The instructions below are translations of the German instructions for the experiment.

Introduction to the experiment

Welcome everyone to today's physical education session. As you might have already noticed, today's session is going to be different. As you already know, you will take part in a scientific study. For that purpose, you received a parental consent form and handed it back to your teacher. If you have not handed it back to your teacher, you will not take part in the study.
The study is going to be conducted by the three of us: Lukas Kiessling, Sebastian Schaube and I am Jonas Radbruch. If you have any questions throughout the study, you can address us at any point in time.
The study comprises several parts. For the first part, we would like you to do a running task called suicide runs. My colleague will shortly demonstrate this exercise.
(The following verbal explanation was accompanied with physical demonstration of the exercise)
You start at the baseline of the volleyball court and run to to this first line. You touch it with your hand and run back to the baseline. You touch the baseline with your hand and run to the next line. Touch it again, back to the baseline; touch it, and then to the third line, back to the baseline, to the fourth line and then you return to the baseline. Everyone of you will run alone and the goal is to be as fast as possible. After this run, we will hand you a computer to fill out a survey.
After all of you have ran and filled out the survey, you will run for a second time. This time at the same time as another student. During the survey we will ask you – among other questions – with whom you would like to run. You will receive detailed information about this later on.
The goal during both runs is to be as fast as possible. We will record your running times and hand it to your teacher. Your teacher will grade your performance during both runs.
Before we start with the study, we would like to remind you again that your participation is voluntary. If anyone does not want to take part in the study, then please inform us now.
Do you have any further questions? If this is not the case, please start with the warm-up, before we start with the experiment.
(Introduction was followed by short warm-up by students. After a short warm-up all students were asked to leave the gym and wait in an accompanying the hallway until

they were called in the gym to take part in the first run. We asked students whether they understood the task and, if necessary, explained the task again. Directly afterwards, they were asked to leave the gym and were led to a different room. There we asked them to complete the survey on a computer we handed them.)

Screenshots of the preferences-elicitation during the survey

(The following two screenshots, Figures A.1 and A.2, display translated elicitation screens for performance- and name-based preferences for peers.)

Figure A.1: Performance-based preferences

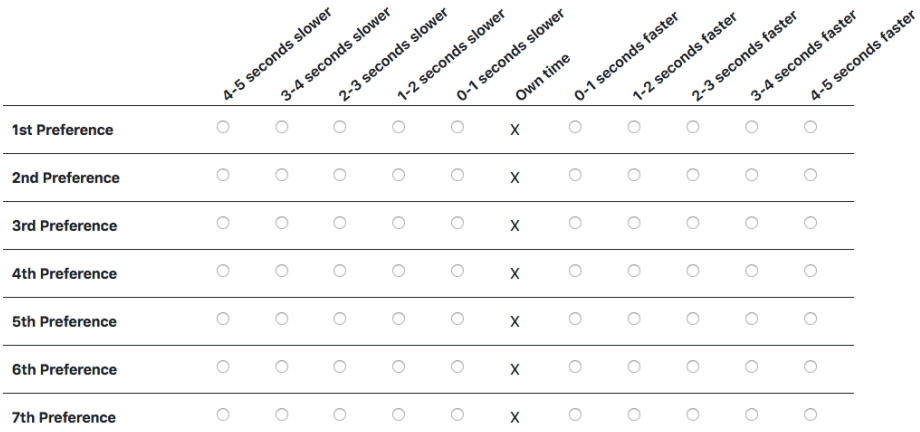


Figure A.2: Name-based preferences

Introduction to the second run for the whole class

(Class was gathered for announcement)
We will shortly start with the second run. For this purpose a partner for you has been selected. In your class, the partner has been selected *randomly [based on your indication how fast you want your partner to be] [based on the classmates you nominated]*. We would like to remind you that the objective is to be as fast as possible and it is only about your own time. Your teacher will receive a list with your performance, but no information about the pairs.
(The list with pairs was read out aloud to the students and students were accompanied to the waiting zone. Students were called into the gym one pair after the other. In the gym they were led to separate, but adjacent tracks. Each student was accompanied by one experimenter, who recorded their time as well their responses to four additional questions.)

Individual introduction directly before the second run

The two of you will now run simultaneously. Your partner has been selected randomly [based on your indication how fast you want your partner to be] [based on the classmates you nominated]. .
(We then asked each subject to assess their relative performance in the first run) Please guess, who of you two was faster during the first run?

Post-run questionnaire after the second run

(Directly after a pair participated in the second run, we asked each of the two subjects the following three questions in private)

(1) How much fun did you have during the second run? Please rate this on a scale from 1 – no fun at all – to 5 – a lot of fun

(2) If you were to run again, would you prefer to run alone or with a partner)

(3) How much pressure did you feel form your partner during the second run? Please rate this on a scale from 1 – no pressure at all – to 5 – a lot of pressure.

B Randomization and manipulation check

Table B.1 presents the randomization check of our experiment. The residual of times in the first run are constructed from a regression of times of the first run on school and grade-specific fixed effects. As can be seen the difference in times in the first run can be explained by those observables and hence are an artifact of the block randomization as classrooms rather than individuals were randomly assigned to treatments. Table B.2 shows how those small insignificant imbalances in the time in the first run (column (1)) disappear, as soon as we condition on those variables.

Table B.1: Randomization check

	Random	Name	Diff.	Performance	Diff.
Socio-Demographics					
Age	14.43	14.55	0.13	14.58	0.15
	(1.18)	(1.24)	(0.12)	(1.24)	(0.12)
Female	0.73	0.62	-0.11*	0.61	-0.12*
	(0.45)	(0.49)	(0.04)	(0.49)	(0.05)
Doing sports regularly	0.82	0.82	0.00	0.90	0.08
	(0.39)	(0.38)	(0.04)	(0.31)	(0.04)
Times (in sec)					
Time (First Run)	26.81	26.08	-0.73*	26.19	-0.62*
	(2.96)	(2.93)	(0.28)	(2.78)	(0.28)
Residual of Time (First Run)	-0.02	-0.11	-0.09	0.08	0.10
	(2.31)	(2.35)	(0.22)	(2.24)	(0.22)
Class-level Variables					
# Students in class	26.01	25.39	-0.62*	25.61	-0.41
	(2.95)	(2.02)	(0.24)	(3.11)	(0.30)
Share of participating students	0.72	0.74	0.02	0.73	0.01
	(0.16)	(0.13)	(0.01)	(0.12)	(0.01)
Grade	8.68	8.76	0.08	8.75	0.07
	(1.07)	(1.12)	(0.11)	(1.13)	(0.11)
Observations	221	213	434	193	414

*, **, and *** denote significance at the 10, 5, and 1 percent level. Standard deviations in parentheses in columns 1, 2 and 4; standard errors in column 3 and 5. Residuals of Time (First Run) are calculated as follows: We first regress all times from the first run on school, grade and gender fixed effects. We then use the residuals from this regression.

In section 3.1, we presented the resulting match qualities using the preferences as elicited in the survey. However, some subjects may prefer relative times, which are not available to them. For example, the fastest subject in the class might want to run with someone who is even faster, or a student wants to run with somebody else who is 1-2 seconds faster but by chance there is no one in the class with such a time. Similarly, subjects in Name may rank other students which were not present during the experiment or did not participate. We therefore present an alternative

Table B.2: Randomization check: Time in the first run

	Time (First Run; in sec.)				
	(1)	(2)	(3)	(4)	(5)
NAME	-0.75	-0.37	0.21	-0.01	-0.11
	(0.58)	(0.30)	(0.30)	(0.20)	(0.20)
PERFORMANCE	-0.66	-0.23	0.23	0.06	0.06
	(0.54)	(0.29)	(0.31)	(0.24)	(0.23)
Female		3.52***		3.20***	
		(0.26)		(0.28)	
Grade FEs	No	No	Yes	Yes	No
Gender-specific grade FEs	No	No	No	No	Yes
School FEs	No	No	Yes	Yes	Yes
N	588	588	588	588	588
R^2	.014	.34	.15	.38	.39
p-value: NAME vs. PERF.	.85	.63	.97	.74	.45

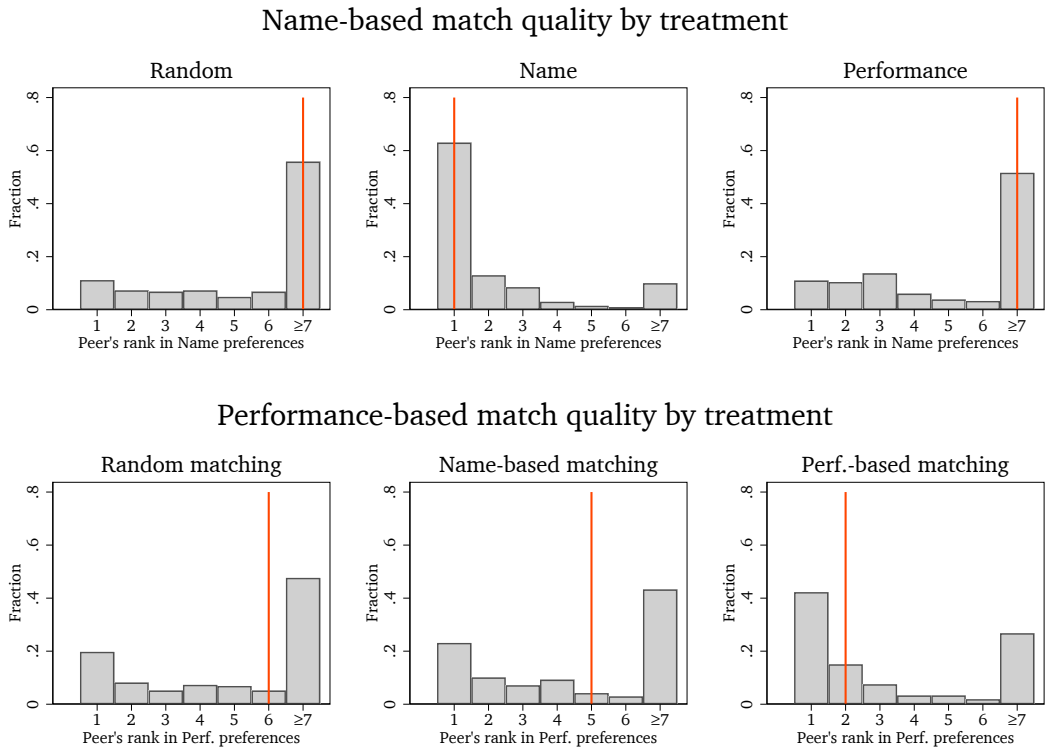
*, **, and *** denote significance at the 10, 5, and 1 percent level. Standard deviations in parentheses.

approach to evaluate the match quality by taking the availability of peers into account. This implies that the quality of a match does not correspond directly to the elicited preferences; rather, based on these preferences all available subjects (i.e., the students participating in the study) are ranked. The quality of the match is then calculated based on this new ranking and results in a realized feasible match quality.

Consequently, we determine the feasible match quality by calculating how high a classmate is ranked in a list of available classmates. [1] In NAME, this can only increase the match quality. If someone nominates another student who is not available as her most-preferred peer and she received her second highest ranked choice, this means that she is matched with her most-preferred feasible peer. Similar arguments can increase the match quality for preferences over relative performance. However, the match quality in performance can also be lower. Suppose that a student ranks the category "1-2 seconds faster" highest and there are three students in that category. However, she is only matched with her second highest ranked category. There would have been three subjects whom she would have preferred more, generating a feasible match quality of 4. We present the corresponding histograms in Figure B.1 and observe that the median of the feasible match quality is actually higher for both treatments relatively to the match qualities depicted in Figure 3.

[1] We code peers who are not ranked among the first six preferences with a match quality of 7.

Figure B.1: Feasible match quality across treatments



The figure presents a histogram of match qualities for each treatment evaluated according to either the students' name-based preferences (upper panel) or performance-based preferences (lower panel). Vertical lines denote median match qualities.

As our treatments change the peer composition, they also change the relative characteristics of peers. In order to understand which characteristics change, we analyze how our treatments affect the peer composition in other dimensions apart from the match quality in Table B.3.

Table B.3: Effects of treatments on peer composition

	Match Qual. (name)	Match Qual. (time)	Friendship Ties	Time 1	
NAME	0.49***	0.07	0.32***	-0.08	
	(0.06)	(0.04)	(0.06)	(0.19)	
PERFORMANCE	-0.06	0.23***	-0.07	-0.70***	
	(0.06)	(0.04)	(0.07)	(0.21)	
N	588	588	294	294	
R^2	.34	.075	.2	.09	
p-value: NAME vs. PERF.	8.2e-12	.00037	1.1e-07	.0037	
Mean in RANDOM	.23	.3	.4	2.4	
	Extra-version	Agree-ableness	Conscien-tiousness	Neuroticism	Openness to Experience
NAME	-0.14	0.09	-0.15	0.11	-0.15
	(0.14)	(0.09)	(0.11)	(0.13)	(0.10)
PERFORMANCE	0.01	0.14	-0.20	0.28**	0.12
	(0.17)	(0.09)	(0.12)	(0.13)	(0.11)
N	292	292	292	292	292
R^2	.05	.058	.047	.039	.03
p-value: NAME vs. PERF.	.19	.53	.63	.19	.031
Mean in RANDOM	1.2	1	1.1	.98	1.1
	Locus of Control	Social Comparison	Compe-titiveness	Risk	
NAME	0.12	0.00	0.03	0.07	
	(0.11)	(0.10)	(0.13)	(0.11)	
PERFORMANCE	0.46***	-0.19**	0.12	0.05	
	(0.12)	(0.09)	(0.11)	(0.11)	
N	292	293	291	292	
R^2	.065	.033	.03	.019	
p-value: NAME vs. PERF.	.003	.079	.37	.76	
Mean in RANDOM	.98	1.1	1.1	1.1	

This table presents least squares regressions using absolute differences in pairs' characteristics except for match quality and friendship as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. All regressions control for gender, grade and school fixed effects in regressions with individual outcomes.

C Econometric Framework

In this appendix, we outline how to interpret our estimates in light of a mediation analysis similar to Heckman and Pinto (2015). A key difference between their framework and ours is that we are interested in the direct effect of our treatments as well as indirect effects of a change in the production inputs, rather than only the latter.

In general, any observed change in outcomes of our experiment can be attributed to one of two main sources: first, different peer-assignment mechanisms may affect peer interactions directly; and second, self-selection changes the peers and therefore the difference between the student's and his or her peer's characteristics. We therefore decompose the average treatment effect into a direct effect of self-selection as well as a pure peer composition effect. This takes into account the change in relative peer characteristics across treatments.[1]

Consider the following potential outcomes framework. Let Y^P and Y^N and Y^R denote the counterfactual outcomes in the three treatments. Naturally, we only observe the outcome in one of the treatments:

$$Y = D^N Y^N + D^P Y^P + (1 - D^P)(1 - D^N) Y^R \tag{C.1}$$

Let θ_d be a vector characterizing a peer's relative characteristics in treatment $d \in \{R, N, P\}$.[2] Similar to the potential outcomes above, we can only observe the peer composition vector θ in one of the treatments and thus $\theta = D_P \theta_P + D_N \theta_N + (1 - D_P)(1 - D_N)\theta_R$ and define an intercept α analogously. The outcome in each of the treatments is therefore given by

$$Y_d = \alpha_d + \beta_d \theta + \gamma X + \epsilon_d \tag{C.2}$$

where we implicitly assume that we have a linear production function, which can be interpreted as a first-order approximation of a more complex non-linear function. The outcome depends on own characteristics X as well as treatment-specific effects of relative characteristics of the peer θ and a zero-mean error term ϵ_d, independent of X and θ.

[1] Our treatments do not change the distribution of characteristics or skills within the class or of a particular subject; rather, the treatments change with whom from the distribution a subject interacts. Due to the random assignment, we assume independence of own characteristics and the treatment.

[2] In our estimations, we include the following characteristics in θ_d: indicators whether the peer ranked high in the individual preference rankings, effects of absolute time differences for slower and faster students within pairs, the rank and presence of friendship ties within pairs, and absolute differences in personal characteristics (Big 5, locus of control, competitiveness, social comparison and risk attitudes).

Potentially, there are unobserved factors in θ. We therefore split θ in a vector with the observed inputs ($\bar{\theta}$) and unobserved inputs ($\tilde{\theta}$)[3] with corresponding effects $\bar{\beta}_d$ and $\tilde{\beta}_d$ and can rewrite equation (C.2) as follows:

$$Y_d = \alpha_d + \bar{\beta}_d\bar{\theta} + \tilde{\beta}_d\tilde{\theta} + \gamma X + \epsilon_d \tag{C.3}$$

$$= \tau_d + \bar{\beta}_d\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{C.4}$$

where $\tau_d = \alpha_d + \tilde{\beta}_d\mathbb{E}[\tilde{\theta}]$ and $\tilde{\epsilon}_d = \epsilon_d + \tilde{\beta}_d(\tilde{\theta} - \mathbb{E}[\tilde{\theta}])$. We assume $\tilde{\epsilon}_d \overset{d}{=} \epsilon$,i.e., are equal in their distribution with a zero-mean. We can express the effect of $\bar{\theta}$ in NAME and PERFORMANCE relative to the effect in RANDOM by rewriting $\beta_d = \beta + \Delta_{R,d}$. Accordingly, we rewrite the coefficients $\bar{\beta}_d$ of θ_i as the sum of the coefficients in RANDOM denoted by β and the distance of the coefficients between treatment d and RANDOM (denoted by $\Delta_{R,d}$).

$$Y_d = \tau_d + \bar{\beta}\bar{\theta} + \bar{\Delta}_{R,d}\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{C.5}$$

$$= \hat{\tau}_d + \bar{\beta}\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{C.6}$$

In what follows, we are interested in $\bar{\tau}_d = \mathbb{E}[\hat{\tau}_d - \hat{\tau}_R]$ ($d \in \{N, P\}$; $\hat{\tau}_d = \tau_d + \bar{\Delta}_{R,d}\bar{\theta}$) , i.e., the direct treatment effect of NAME and PERFORMANCE conditional on indirect effects from changes in the peer composition captured in $\bar{\theta}$. This direct effect subsumes the effect of the treatment itself ($\alpha_d - \alpha_R$), the changed impact of the same peer's observables ($\bar{\Delta}_{R,d}\bar{\theta}$), and changes in unmeasured inputs as well as their effect ($(\tilde{\beta} + \tilde{\Delta}_{R,d})\tilde{\theta}$). We interpret this direct effect in light of self-determination theory (Deci and Ryan, 1985) as an additional motivation due to being able to self-select a peer. This focus on the direct effect is a key difference compared with Heckman and Pinto (2015), who are mainly interested in the indirect effects of the mediating variables. The empirical specification of C.6 is given by

$$y_{igs} = \bar{\tau} + \bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{C.7}$$

where we are interested in $\bar{\tau}_N$ and $\bar{\tau}_P$, the direct effects of our treatments relative to RANDOM. Indirect effects are captured by $\beta\theta_i$, the effect of changed peer characteristics on the outcome y_{igs}.

[3]Furthermore, we assume that unobserved and observed inputs are independent conditional on X and D.

D Robustness checks for average treatment effects

In Table D.1, we compare the clustered standard errors with clustered standard errors using a biased-reduced linearization to account for the limited number of clusters. Comparing the first two columns, we observe that the results are robust to this alternative specification of the standard errors. In column (3), we additionally check whether looking at matching group-specific group means – i.e., the average percentage point improvement for males and females in each class – affects the estimates. While the power is reduced due to the small number of observations, the treatment effects persist and the coefficients on the treatment effects are not significantly affected. Columns (4) and (5) analyze the sensitivity of our estimates with respect to outliers. We use two different strategies. First, we apply a 90% winsorization, which replaces all observations with either a time or a percentage point improvement below or above the threshold with the value at the threshold. We replace a time of improvement below the 5th percentile with the corresponding value of the 5th percentile and all observations above the 95th percentile with the 95th percentile. Second, we truncate the data and keep only those pairs where no time or no improvement falls into the bottom 5% or top 5%. Neither winsorization nor truncation significantly changes the estimated treatment effects.

Table D.1: Robustness checks – Limited number of clusters

	Percentage Point Improvements				
	(1) Baseline	(2) BRL	(3) Group means	(4) Winsorization	(5) Truncation
NAME	1.25***	1.25**	1.15*	1.04***	0.93**
	(0.43)	(0.50)	(0.60)	(0.37)	(0.35)
PERFORMANCE	1.67**	1.67**	1.97***	1.51***	1.43***
	(0.62)	(0.72)	(0.61)	(0.52)	(0.43)
Gender-Grade/School FEs	Yes	Yes	Yes	Yes	Yes
N	588	588	70	588	496
R^2	.055	.055	.27	.071	.084
p-value: NAME vs. PERF.	.5	.54	.15	.36	.27

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Column (1) presents the baseline specifications as used in Table 3. Columns (2) uses biased-reduced linearization (BRL) to account for the limited number of clusters. Column (3) uses matching group-specific means as the unit of observation. Finally, columns (4) and (5) apply a 90% winsorization and truncation, respectively.

We further analyze the robustness of our results by looking at different subsamples. We therefore split our sample first by grades in the upper panel of Table D.2 and by schools as well as gender in the lower panel and estimate the treatment effects separately for those samples. The table shows the robustness of the estimated treatment effects as these effects persists for all subsamples with similar magnitude.

Table D.2: Robustness checks – Subsample analyses

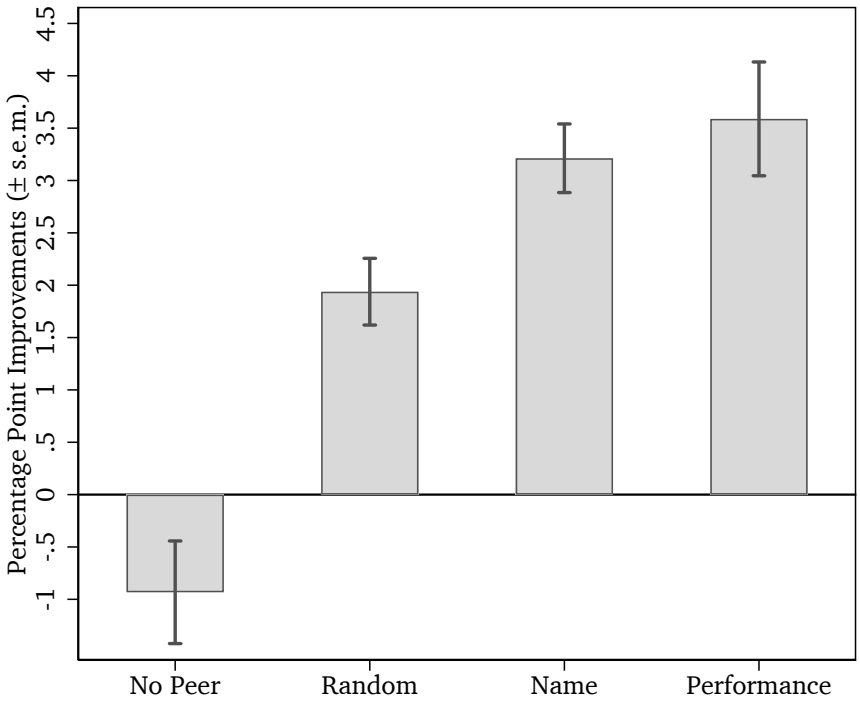
	Percentage Point Improvements				
	(1) Baseline	(2) 7th grade	(3) 8th grade	(4) 9th grade	(5) 10th grade
NAME	1.25***	1.95***	2.60***	1.50**	1.11*
	(0.43)	(0.08)	(0.34)	(0.61)	(0.62)
PERFORMANCE	1.67**	2.78***	2.50***	2.47***	1.32
	(0.62)	(0.54)	(0.15)	(0.68)	(0.89)
Gender-Grade/School FEs	Yes	Yes	Yes	Yes	Yes
N	588	116	116	174	182
R^2	.055	.073	.063	.16	.038
p-value: NAME vs. PERF.	.5	.15	.79	.21	.84
	(6) Female	(7) Male	(8) School 1	(9) School 2	(10) School 3
NAME	1.26*	1.22***	1.73***	1.43**	2.12***
	(0.65)	(0.44)	(0.16)	(0.65)	(0.38)
PERFORMANCE	1.68**	1.63*	1.33***	2.29***	2.23*
	(0.76)	(0.85)	(0.00)	(0.54)	(1.12)
Gender-Grade/School FEs	Yes	Yes	Yes	Yes	Yes
N	390	198	148	274	166
R^2	.054	.065	.036	.1	.12
p-value: NAME vs. PERF.	.53	.63	.041	.14	.89

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Column (1) presents the estimates using the whole sample as in Table 3. Columns (2)–(5) restrict the sample to one grade, columns (6) and (7) to each gender and columns (8)–(10) to one school.

E Control treatment to disentangle peer effects from learning

Table E.1 and Figure E.1 present the estimated average treatment effects and the margins including an additional control treatment. The NoPeer treatment featured the same design as all other treatments. The only difference was that students participated in the running task twice without a peer. Moreover, we shortened the survey for this treatment by removing the questionnaires on personal characteristics. The control treatment was conducted to show that the observed performance improvements are not due to learning. If learning drives our effects, we should observe performance improvements in NoPeer, which is not the case. Even if this control treatment had yielded performance improvements, this would not affect any of our results. To see this, note that we are interested in a between treatment comparison of performance improvements. Learning effects between the runs should therefore be constant across treatments.

Figure E.1: Average treatment effects



The figure presents percentage point improvements from the first to the second run with corresponding standard errors for the three treatments Random, Name, and Performance and an additional control treatment, where students run two times without a peer (NoPeer). See column (1) in Table E.1 for the corresponding regression. We control for gender, grade and school fixed effects, and cluster standard errors at the class level.

Table E.1: Average treatment effects including NoPeer-Treatment

	(a) PP. Imprv.	(b) Diff.-in-Diff.		
	(1)	(2)	(3)	(4)
Name	1.27***	-0.03		
	(0.42)	(0.20)		
Performance	1.65**	0.14		
	(0.62)	(0.23)		
NoPeer	-2.87***	0.23		
	(0.62)	(0.30)		
Second run		-0.49***	-0.49***	-0.17***
		(0.09)	(0.13)	(0.05)
Name × Second Run		-0.45***	-0.45**	-0.16**
		(0.13)	(0.18)	(0.06)
Performance × Second Run		-0.55***	-0.55**	-0.19**
		(0.19)	(0.27)	(0.09)
NoPeer × Second Run		0.73***	0.73***	0.25***
		(0.12)	(0.17)	(0.06)
Gender-Grade/School FEs	Yes	Yes	No	No
Individual FEs	No	No	Yes	Yes
N	715	1430	1430	1430
R^2	.14	.41	.94	.94

This table presents least squares regressions using percentage point improvements as the dependent variables (Panel (a)) or a difference-in-differences specification using times from both runs (Panel (b)). Column (4) presents estimates using standardized times. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

F Peer composition robustness checks

In addition to the variance decomposition presented in Table 4, Table F.1 provides evidence on peer effects in several dimensions using data from RANDOM only. In particular, we observe that there exist significant and non-linear peer effects in performance differences as well as several personality traits.

We run several robustness checks for the results presented in Table 6. First, we estimate the influence of peer characteristics (and individual characteristics) on the sample of RANDOM subjects only in Table F.2 and use these coefficients to decompose the average effect. For this purpose, we first net out the effect of group variables such as school and gender-grade fixed effects (as well as individual characteristics) from both the outcome and independent variables such as peer characteristics according to the Frisch-Waugh-Lovell theorem using the whole sample. In a first version, we regress the outcome and peer characteristics on the fixed effects only. In a second version, we additionally net out the effect of individual characteristics from peer characteristics and the outcome. We use the residuals of those regressions to decompose the treatment effect. We then begin by estimating the influence of peer characteristics on the outcome using only subjects from RANDOM and the residualized outcome as well as peer characteristics (column (1) and (3)). In a second step, we restrict the influence of those peer characteristics and estimate the direct treatment effects (column (2) and (4)).

Second, in Table F.3 we use different specifications for match quality. We consider the partner's match quality, an interaction between one's own and the partner's match quality, and feasible match quality as defined in Appendix B, and find that the estimates of our direct effects are qualitatively and quantitatively the same.

Third, in Table F.4, we show that our results do not depend on the precise definition of friendship ties. We check whether our results change when we define friendship ties as undirected or reciprocal rather than directed. As can be seen from the table, the coefficients on the direct effects as well as on other peer characteristics remain the same.

Fourth, in Table F.5, we use the time in the second run as outcome, instead of percentage point improvements. Our results of the decomposition remain robust for this outcome variable.

Fifth, we control for differences in productivity in a more flexible way in Table F.6 by allowing for quartic rather than linear effects of productivity differences in column (2) (see also Figure F.1 comparing linear and quartic terms graphically). In addition, we allow for a second flexible specification using fixed effects for productivity dif-

Table F.1: Peer effects in RANDOM

	(1) Match Qual.	(2) Friend	(3) Time Diff.	(4) Personality	(5) All
Faster Student	0.72				0.66
× High match quality (NAME)	(0.87)				(0.85)
Slower Student	-0.32				-0.02
× High match quality (NAME)	(1.39)				(1.10)
Faster Student	0.96				0.11
× High match quality (PERF.)	(1.17)				(1.09)
Slower Student	-1.99				-0.76
× High match quality (PERF.)	(1.48)				(1.23)
Faster Student		-0.11			0.12
× Peer is Friend		(0.82)			(0.87)
Slower Student		-0.11			0.14
× Peer is Friend		(1.05)			(1.24)
Faster Student			-0.45**		-0.59**
× $\lvert\Delta Time\ 1\rvert$			(0.18)		(0.26)
Slower Student			0.84**		0.68*
× $\lvert\Delta Time\ 1\rvert$			(0.34)		(0.33)
|Δ Agreeableness|				-0.93	-0.82
				(1.00)	(0.99)
|Δ Conscientiousness|				0.09	0.14
				(0.71)	(0.75)
|Δ Extraversion|				-0.19	-0.19
				(0.40)	(0.47)
|Δ Openness|				-1.37**	-1.31*
				(0.61)	(0.67)
|Δ Neuroticism|				1.38	1.44
				(0.92)	(0.94)
|Δ Locus of Control|				-0.14	-0.17
				(0.88)	(0.83)
|Δ Social Comp.|				-0.56	-0.55
				(0.70)	(0.72)
|Δ Competitiveness|				-0.47	-0.46
				(0.55)	(0.56)
|Δ Risk attitudes|				0.80*	0.77
				(0.41)	(0.48)
Slower Student in Pair	3.89***	2.68***	-0.35	2.77***	0.32
	(0.61)	(0.56)	(0.89)	(0.58)	(1.03)
Peer Characteristics	No	No	No	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes	Yes	Yes
N	205	205	205	204	204
R^2	.12	.11	.17	.2	.27

This table presents least squares regressions according to equation (3) using percentage point improvements as the dependent variable on RANDOM only. High match quality is an indicator that equals one if the partner was ranked within an individual's first three preferences. Personality characteristics include the Big Five, locus of control, social comparison, competitiveness, and risk attitudes. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

ferences. More specifically, we include an indicator for each one-second interval of productivity differences between subjects within a pair. This allows for a potential non-linear influence of productivity differences on our estimates. Comparing the estimates shows that neither the quartic functional form nor the fixed effect specification is restrictive.

Sixth, Table F.7 restricts the control group sample to subjects with a high match quality within RANDOM to show that the treatment effects persist for these subjects and the coefficients on peer compositional effects do not substantially change.

Finally, Table F.8 provides results from robustness checks that additional include a set of variables capturing the atmosphere within each class as reported by the teachers. In addition, one might be worried that adding a series of own and peer characteristics results in overfitting driving our results. In the second column of Table F.8 we therefore adopt a post-double selection (PDS) Lasso method described by Belloni, Chernozhukov, and Hansen (2014) that penalizes control variables, but allows for inference on treatment indicators. Both of these tests show that our results are robust to these additional checks. Table F.9 presents the omitted coefficients of own and peer characteristics, as well as their absolute differences, from column (5) Table 6 in the main text.

Table F.2: Restricting coefficients of peer characteristics

	Percentage Point Improvements			
	Fixing only FEs		Fixing FEs & own char.	
	(1) only RANDOM	(2) all	(3) only RANDOM	(4) all
Direct Effects				
NAME		0.79*		0.83*
		(0.46)		(0.46)
PERFORMANCE		1.73**		1.72**
		(0.67)		(0.67)
Peer Characteristics				
Faster Student × High match quality (NAME)	0.64	0.64	0.62	0.62
	(0.87)		(0.79)	
Slower Student × High match quality (NAME)	0.19	0.19	0.32	0.32
	(1.09)		(1.00)	
Faster Student × High match quality (PERF.)	0.19	0.19	-0.13	-0.13
	(1.11)		(1.11)	
Slower Student × High match quality (PERF.)	-0.60	-0.60	-0.31	-0.31
	(1.15)		(1.22)	
Faster Student × Peer is friend	-0.12	-0.12	-0.19	-0.19
	(0.72)		(0.64)	
Slower Student × Peer is friend	0.08	0.08	-0.02	-0.02
	(1.28)		(1.15)	
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.52*	-0.52	-0.51	-0.51
	(0.31)		(0.29)	
Slower Student × $\lvert\Delta Time\ 1\rvert$	0.78**	0.78	0.83**	0.83
	(0.32)		(0.30)	
Slower Student in Pair	0.07	0.07	-0.17	-0.17
	(0.95)		(0.83)	
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes	Yes
Own Characteristics	No	No	Yes	Yes
N	204	582	204	582
R^2	.23		.21	

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. We use residualized dependent and independent variables, where we take out the variation of individual-specific variables. The first two columns take out the variation of the set of fixed effects, while the last two columns additionally take out variation of own characteristics. Columns (1) and (3) present least squares regressions in RANDOM only, while columns (2) and (4) use all three treatments, but restrict the coefficients to equal the preceding columns.

Table F.3: Robustness Checks for match quality

	Percentage Point Improvements		
	(1) Partner's MQ	(2) Interaction	(3) Feasible
Direct Effects			
NAME	1.16**	1.14**	1.20**
	(0.54)	(0.56)	(0.47)
PERFORMANCE	2.26***	2.24***	2.09***
	(0.71)	(0.69)	(0.66)
Peer Characteristics			
High match quality (partner; NAME)	0.25	0.13	
	(0.43)	(0.55)	
High match quality (partner; PERF.)	-0.05	0.24	
	(0.40)	(0.44)	
High match quality (own and partner; NAME)		0.23	
		(0.83)	
High match quality (own and partner; PERF.)		-0.60	
		(0.95)	
Faster Student × High match quality (feasible; NAME)			0.01
			(0.41)
Slower Student × High match quality (feasible; NAME)			1.37*
			(0.77)
Faster Student × High match quality (feasible; PERF.)			0.77*
			(0.41)
Slower Student × High match quality (feasible; PERF.)			0.30
			(0.86)
Faster Student × Match Quality (name-based)	0.43	0.33	
	(0.40)	(0.45)	
Slower Student × Match Quality (name-based)	0.45	0.31	
	(0.64)	(0.75)	
Faster Student × Match Quality (perf.-based)	0.41	0.75	
	(0.51)	(0.66)	
Slower Student × Match Quality (perf.-based)	-0.75	-0.51	
	(0.65)	(0.61)	
Friendship Ties and Performance Differences	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes
N	582	582	582
R^2	.29	.29	.29
p-value: NAME vs. PERFORMANCE	.15	.16	.22

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) adds the partner's match quality in addition to own match quality as in Table 6, while column (2) additionally controls for the interaction of own and partner's match quality. Finally, column (3) uses a different measure of match quality, (feasible match quality – see also Appendix B), which acknowledges the fact that certain preferred peers may not be available.

Table F.4: Different definitions of friendship ties

	Percentage Point Improvements			
	(1) directed	(2) undirected	(3) reciprocal	(4) dir. & rec.
Direct Effects				
NAME	1.24**	1.20**	1.22**	1.15**
	(0.50)	(0.49)	(0.50)	(0.50)
PERFORMANCE	2.24***	2.16***	2.24***	2.23***
	(0.68)	(0.68)	(0.68)	(0.68)
Faster Student × Peer is friend	-1.15**			-1.63*
	(0.53)			(0.83)
Slower Student × Peer is friend	0.11			-0.37
	(0.66)			(0.85)
Faster Student × Peer is friend (undirected)		-1.62***		
		(0.58)		
Slower Student × Peer is friend (undirected)		0.13		
		(0.79)		
Faster Student × Peer is friend (reciprocal)			-0.59	0.71
			(0.61)	(0.94)
Slower Student × Peer is friend (reciprocal)			0.44	0.68
			(0.53)	(0.66)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.34**	-0.33**	-0.33**	-0.33**
	(0.16)	(0.16)	(0.16)	(0.15)
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.05***	1.05***	1.06***	1.06***
	(0.20)	(0.20)	(0.20)	(0.20)
Slower Student in Pair	-0.19	-0.49	0.02	-0.21
	(0.67)	(0.73)	(0.68)	(0.68)
Match quality and performance differences	Yes	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes	Yes
N	582	582	582	582
R^2	.29	.29	.28	.29

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 6 for reference using directed friendship ties. Column (2) uses undirected friendship ties, column (3) reciprocal directed friendship ties, while column (4) allows for a differential effect of directed and reciprocal friendship ties.

Table F.5: Decomposition of treatment effects using time in second run

	Time (Second Run; in sec.)				
	(1) Baseline	(2) Match Quality	(3) Friend-ship ties	(4) Time Difference	(5) All
Direct Effects					
NAME	-0.38***	-0.36**	-0.41***	-0.38***	-0.36**
	(0.12)	(0.14)	(0.13)	(0.12)	(0.14)
PERFORMANCE	-0.38***	-0.39**	-0.37**	-0.42***	-0.51***
	(0.14)	(0.14)	(0.15)	(0.14)	(0.16)
Peer Characteristics					
Faster Student		0.07			-0.07
× High match quality (NAME)		(0.11)			(0.13)
Slower Student		-0.08			-0.11
× High match quality (NAME)		(0.18)			(0.18)
Faster Student		-0.27*			-0.13
× High match quality (PERF.)		(0.14)			(0.14)
Slower Student		0.29			0.18
× High match quality (PERF.)		(0.18)			(0.19)
Faster Student			0.23*		0.31*
× Peer is Friend			(0.12)		(0.15)
Slower Student			0.02		-0.01
× Peer is Friend			(0.17)		(0.19)
Faster Student				0.08*	0.06
× $\lvert\Delta Time\ 1\rvert$				(0.04)	(0.04)
Slower Student				-0.16**	-0.15*
× $\lvert\Delta Time\ 1\rvert$				(0.07)	(0.08)
Slower Student in Pair		-0.39**	-0.08	0.14	0.14
		(0.16)	(0.13)	(0.13)	(0.21)
Own Characteristics					
Time (First Run)	0.67***	0.70***	0.69***	0.75***	0.75***
	(0.04)	(0.05)	(0.05)	(0.06)	(0.06)
Abs. Diff. in Personality	No	No	No	No	Yes
Peer Characteristics	No	No	No	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	585	585	585	585	582
R^2	.81	.81	.81	.81	.83
p-value: NAME vs. PERFORMANCE	.98	.87	.83	.77	.35

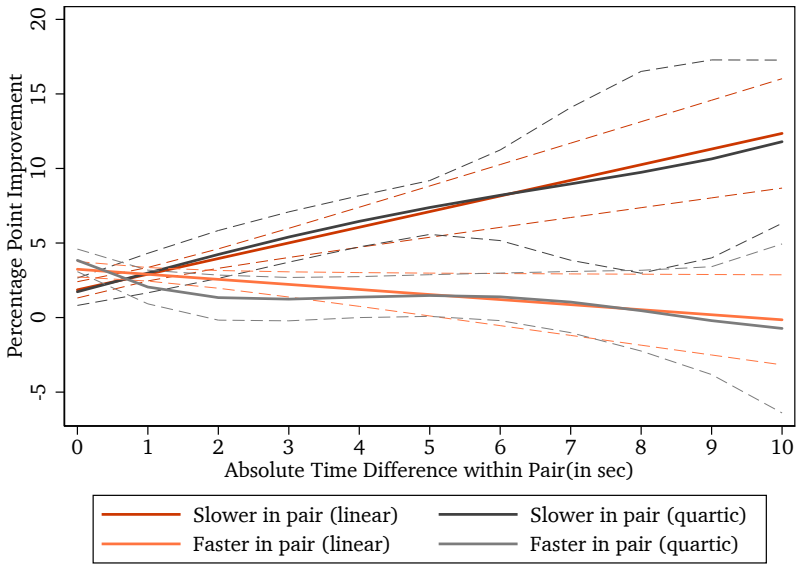
This table presents least squares regressions according to equation (3) using time in the second run as the dependent variable. High match quality is an indicator that equals one if the partner was ranked within an individual's first three preferences. Personality characteristics include the Big Five, locus of control, social comparison, competitiveness, and risk attitudes. Appendix Table F.9 presents the omitted coefficients of own and peer characteristics, and their absolute differences for our preferred specification in column (5). *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Table F.6: Robustness checks for absolute time differences

	Percentage Point Improvements		
	(1) Linear	(2) Quartic	(3) FEs
Direct Effects			
NAME	1.24**	1.28**	1.19**
	(0.50)	(0.48)	(0.51)
PERFORMANCE	2.24***	2.27***	2.30***
	(0.68)	(0.68)	(0.74)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.34**	-2.51*	
	(0.16)	(1.26)	
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.05***	1.23	
	(0.20)	(1.74)	
Slower Student in Pair	-0.19	-1.77*	
	(0.67)	(0.90)	
Faster Student × $\lvert\Delta Time\ 1\rvert^2$		0.82	
		(0.55)	
Slower Student × $\lvert\Delta Time\ 1\rvert^2$		0.04	
		(0.97)	
Faster Student × $\lvert\Delta Time\ 1\rvert^3$		-0.10	
		(0.09)	
Slower Student × $\lvert\Delta Time\ 1\rvert^3$		-0.02	
		(0.19)	
Faster Student × $\lvert\Delta Time\ 1\rvert^4$		0.00	
		(0.00)	
Slower Student × $\lvert\Delta Time\ 1\rvert^4$		0.00	
		(0.01)	
Time Diff. FEs	No	No	Yes
Match Quality and Friendship Ties	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes
N	582	582	582
R^2	.29	.29	.3
p-value: NAME vs. PERFORMANCE	.15	.15	.13

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 6 for reference. Column (2) includes quartic terms of time differences in the first run (also illustrated in Appendix Figure F.1) and column (3) fixed effects for every one-second difference in productivity levels of the two students.

Figure F.1: Robustness of linear specification in time differences



The figure presents marginal effects (solid lines) from a least squares regression using percentage point improvements as the dependent variable including 95% confidence intervals (dashed lines). It plots the linear specification (black lines) as used in the main text as well as a second specification using quartic polynomials (orange lines) of absolute time differences in the first run as regressors. We use the same set of controls as in column (5) of Table 6 and cluster standard errors at the class level. The corresponding regressions are presented in columns (1) and (2) of Appendix Table F.6.

Table F.7: Only high match quality sample as comparison group

	Percentage Point Improvements				
	(1) All	(2) Random & Name	(3) with Controls	(4) Random & Perf.	(5) with Controls
Direct Effects					
Name	1.24**	1.83***	1.93***		
	(0.50)	(0.56)	(0.48)		
Performance	2.24***			2.46***	1.78***
	(0.68)			(0.75)	(0.63)
Peer Characteristics					
Faster Student × Match Quality (name-based)	0.49				-0.49
	(0.44)				(1.27)
Slower Student × Match Quality (name-based)	0.49				-0.50
	(0.65)				(1.15)
Faster Student × Match Quality (perf.-based)	0.42		-0.52		
	(0.53)		(0.64)		
Slower Student × Match Quality (perf.-based)	-0.75		-1.22		
	(0.66)		(0.85)		
Faster Student × Peer is friend	-1.15**		-1.53		-1.07
	(0.53)		(1.05)		(1.78)
Slower Student × Peer is friend	0.11		-1.20		-1.38
	(0.66)		(1.06)		(1.11)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.34**		-0.72**		-0.08
	(0.16)		(0.29)		(0.51)
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.05***		1.25***		1.13**
	(0.20)		(0.38)		(0.44)
Slower Student in Pair	-0.19		-0.43		-1.13
	(0.67)		(1.68)		(1.34)
Abs. Diff. in Personality	Yes	No	Yes	No	Yes
Peer Characteristics	Yes	No	Yes	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes	Yes	Yes
N	582	208	207	162	160
R^2	.29	.16	.52	.16	.37

This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 6 for reference. Columns (2) to (5) show that even if we restrict the comparison group to the sample of individuals in Random that received a peer with high match quality according to their name- (columns (2) and (3)) or performance-based preferences (columns (4) and (5)), respectively, the direct effects persist and the coefficients on peer compositional effects do not change much.

Table F.8: Additional robustness checks using class-level controls and post-double selection Lasso

	(1) Class Con.	(2) PDS Lasso
Direct Effects		
NAME	1.48***	1.04**
	(0.45)	(0.47)
PERFORMANCE	1.75**	2.00***
	(0.68)	(0.60)
Peer Characteristics		
Faster Student	0.63	0.15
× High match quality (NAME)	(0.45)	(0.38)
Slower Student	0.66	0.67
× High match quality (NAME)	(0.73)	(0.68)
Faster Student	0.08	
× High match quality (PERF.)	(0.59)	
Slower Student	-1.20	-0.66
× High match quality (PERF.)	(0.74)	(0.58)
Faster Student	-1.00**	
× Peer is Friend	(0.48)	
Slower Student	0.39	0.28
× Peer is Friend	(0.78)	(0.63)
Faster Student	-0.35**	-0.36**
× $\lvert\Delta Time\ 1\rvert$	(0.16)	(0.14)
Slower Student	0.85***	0.98***
× $\lvert\Delta Time\ 1\rvert$	(0.19)	(0.21)
Slower Student in Pair	0.07	-0.05
	(0.75)	(0.55)
Abs. Diff. in Personality	Yes	No
Peer Characteristics	Yes	Yes
Own Characteristics	Yes	Yes
Class-level Controls	Yes	No
Gender-Grade/School FEs	Yes	Yes
N	512	582
R^2	.29	
p-value: NAME vs. PERFORMANCE	.72	.16

This table presents least squares regressions using percentage point improvements as the dependent variable and a set of class-level controls capturing the atmosphere within a class (missing for some classes) and results from the post-double selection (PDS) Lasso method by Belloni, Chernozhukov, and Hansen (2014). The PDS Lasso always includes the treatment indicators and gender-grade as well as school fixed effects and penalizes the remaining control variables to avoid overfitting. Standard errors in this specification are only valid for the treatment indicators and fixed effects. See Belloni, Chernozhukov, and Hansen (2014) for further details. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those (note that the PDS Lasso forces some coefficients such as all absolute differences in personality measures to zero). *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Table F.9: Omitted Coefficients from Table 6 column (5)

	Own characteristics	Peer characteristics	Abs. Diff in characteristics
Agreeableness	0.12	-0.11	0.31
	(0.22)	(0.20)	(0.29)
Conscientiousness	-0.01	0.14	-0.13
	(0.20)	(0.17)	(0.23)
Extraversion	0.03	0.05	-0.50*
	(0.24)	(0.20)	(0.25)
Openness to Experience	-0.47**	-0.18	0.52
	(0.19)	(0.17)	(0.33)
Neuroticism	-0.16	-0.15	-0.66**
	(0.24)	(0.19)	(0.28)
Locus of Control	0.15	0.09	-0.14
	(0.20)	(0.19)	(0.31)
Social Comparison	0.32*	0.20	-0.23
	(0.18)	(0.16)	(0.31)
Competitiveness	-0.09	-0.36	0.34
	(0.29)	(0.23)	(0.21)
Risk Attitudes	0.04	0.07	0.51
	(0.18)	(0.17)	(1.43)

This table presents omitted coefficients from Table 6 in the main text. Columns (1) and (2) show the coefficients on own and peer characteristics, respectively. Column (3) presents the coefficients on the absolute differences in personality measures. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

G Additional material for discussion of direct effects

Table G.1 presents three regressions to support section 5.6's discussion of the psychological effect underlying the direct effects. First, we show that students in RANDOM are not disappointed by having a partner assigned. If they were disappointed, they should have less fun during the second run. As column (1) show this is not the case. Second, we do not find evidence that subjects with self-selected perceive winning in the second run as more important as we do not see a differential effect on fun between being faster or slower in the second run. Third, we show that prosocial students, that is individuals that score higher on agreeableness, do not show differentially direct effects. This is suggestive evidence against experimenter demand effects or other reciprocal motives driving the estimated direct effects.

Table G.1: Potential psychological mechanisms for the direct effect

	Fun (std.).		PP. Imprv.
	(1)	(2)	(3)
Direct Effects			
NAME	-0.01	0.01	1.24**
	(0.10)	(0.14)	(0.50)
PERFORMANCE	-0.10	-0.06	2.24***
	(0.08)	(0.13)	(0.68)
NAME × Slower Student in Pair (2nd Run)		-0.05	
		(0.18)	
PERFORMANCE × Slower Student in Pair (2nd Run)		-0.07	
		(0.17)	
NAME × Agreeableness			0.02
			(0.39)
PERFORMANCE × Agreeableness			0.41
			(0.47)
Peer Characteristics			
Faster Student (2nd Run) × $\lvert\Delta Time\ \ 2\rvert$	-0.01	-0.00	
	(0.05)	(0.05)	
Slower Student (2nd Run) × $\lvert\Delta Time\ \ 2\rvert$	-0.14***	-0.14***	
	(0.04)	(0.04)	
Slower Student in Pair (2nd Run)	0.04	0.07	
	(0.18)	(0.20)	
Faster Student × $\lvert\Delta Time\ \ 1\rvert$			-0.34**
			(0.16)
Slower Student × $\lvert\Delta Time\ \ 1\rvert$			1.06***
			(0.20)
Slower Student in Pair			-0.24
			(0.66)
Match quality	Yes	Yes	Yes
Friendship indicators	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Abs. Diff. in Personality Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs	Yes	Yes	Yes
N	582	582	582
R^2	.34	.34	.29
p-value: NAME vs. PERFORMANCE	.5	.67	.16

This table presents least squares regressions using a standardized measure of fun in the second run (columns (1) and (2)) or percentage point improvements (column (3)) as the dependent variable. Column (2) uses the full specification of Table 6 and additionally interacts the treatment indicators with one's own measure of agreeableness as a proxy of prosociality. Column (1) focuses on fun as an outcome variable that was elicited after the second run ("How much fun did you have during the second run? Please rate this on a scale from 1 – no fun at all – to 5 – a lot of fun.") and uses the full specification of Table 6 adapted using times and ranks from the second run. Column (2) additionally interacts treatment indicators with the final rank in the second run. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

H Simulation of matching rules

We simulate three matching rules and predict their impact on performance improvements using our estimates from Table 6. In a first step, we create artificial pairs, based on the employed matching rules described below. In a second step, we then calculate the vector θ of differences for the artificial pairs as well as the matching quality of artificial peers. Finally, we use the estimated coefficients from the column (5) of Table 6 to predict the performance improvements we would observe for the artificial pairs. As peer-assignment rules only change θ, we are interested in the difference in the respective sums of the indirect effect and direct effect, that is between $\bar{\tau} + \beta\theta_i^{sim}$ and $\bar{\tau} + \beta\theta_i^{obs}$ from equation (3), where sim and obs denote simulated and observed pair characteristics, respectively. As we consider exogenous assignment rules, we assume that the direct effect of the simulated policies equals zero as in in RANDOM. We additionally fix the covariates X to 0 and leave out the fixed effects for the simulations and predictions. This means, we calculate the performance improvements for a particular baseline group for our treatments as well as the simulations. This enables us to compare our results of the simulations directly to the peer-assignment rules using self-selection implemented in the experiment, as we compare the performance improvements for the same group.

In addition to our three treatments, we simulate four types of peer assignment rules. First, we simulate two settings in which we assign the self-selected peers exogenously (NAME (EXOG.) and PERFORMANCE (EXOG.)). Hence, the resulting pairs are the same as in the self-selection treatment, but we exclude the direct effect of self-selection. Second, we implement an ability tracking assignment rule, TRACKING, in the spirit of the matching also employed in Gneezy and Rustichini (2004). Students are matched in pairs, starting with the two fastest students in a matching group and moving down the ranking subsequently. This rule minimizes the absolute distance in pairs. Third, we employ a peer assignment rule that fixes the distance in ranks for all pairs (EQUIDISTANCE). We rank all students in a matching group and match the first student with the one in the middle and so forth. More specifically, if G denotes the group size, the distance in ranks is $G/2 - 1$ for all pairs. This rule is one way to maximize the sum of absolute differences in pairs, but keeps the distance across pairs similarly. Fourth, we match the highest ranked student with the lowest one, the second highest ranked with the second lowest one and so forth (HIGH-TO-LOW). This is similar to Carrell, Sacerdote, and West (2013), who match low-ability students with those students from whom they would benefit the most (i.e., the fastest students). Again, this assignment rule maximizes the sum of absolute differences in pairs. Ta-

ble H.1 summarizes initial performance differences within pairs of the experimental treatments as well as the simulated assignment rules and the predicted performance improvements.

Table H.1: Overview of simulated peer assignment rules

Peer assignment rule	Mean absolut productivity differences (in sec)	Predicted improvement (in pp.)	Description
Self-selection of peers			
NAME	2.09	2.43	Self-selected peers based on names
PERFORMANCE	1.41	2.69	Self-selected peers based on relative performance
Exogeneous peer assignment			
NAME (EXOG.)	2.09	1.19	Self-selected peers based on names without self-selection effect
PERFORMANCE (EXOG.)	1.41	0.48	Self-selected peers based on relative performance without self-selection effect
RANDOM	2.42	1.12	Randomly assigned peers
EQUIDISTANCE	3.11	1.44	Same distance in ranks across pairs
HIGH-TO-LOW	3.11	1.36	First to last, second to second to last etc.
TRACKING	0.90	0.72	First to second, third to fourth etc.

Our treatments also have implications for individual ranks of students within a class since slower students improve more than faster ones. As ranks are important in determining subsequent outcomes (Elsner and Isphording, 2017; Gill et al., 2019; Murphy and F. Weinhardt, 2018), a policy maker has to take the distributional effects of peer assignment mechanisms into account.[1] Since low-ability students improve relatively more than high-ability students in NAME and RANDOM, these treatments yield potentially large changes of a student's rank within the class between the two runs. By contrast, PERFORMANCE will tend to preserve the ranking of the first run as improvements are distributed more equally relative to the two other treatments. We confirm this intuition in Table H.2 in which we regress the absolute change in percentile scores from the first to the second run on treatment indicators. The outcome variable measures the average perturbation of ranks within in a class across the two runs. The results show that PERFORMANCE shuffles the ranks of students less in comparison to RANDOM and NAME. While in RANDOM students change their position by about 15 out of 100 ranks, we find significantly less changes in the percentile score in PERFORMANCE relative to RANDOM. This change corresponds to a 27% reduction in reshuffling. However, in NAME we do not find any effect compared to RANDOM.

[1] Suppose that a policy maker wants to establish a rank distribution (ranks based on times in the second run) that mirrors the ability distribution (ranks based on times in the first run) due to some underlying fairness ideal (e.g., she wants to shift the distribution holding constant individual ranks). In other words, she might want to implement a peer assignment mechanism that preserves individual ranks rather than shuffle them.

As another side effect we consider the pressure students experienced during the second run due to their peer. We find that students in PERFORMANCE experience significantly more pressure than students in the other two treatments.

Table H.2: Side effects of reassignment rules

	Absolute Change in Percentile Scores		Pressure (std.)
	(1) within matching group	(2) within treatment	(3)
NAME	-0.00	-0.02	0.10
	(0.01)	(0.01)	(0.20)
PERFORMANCE	-0.04**	-0.04***	0.46**
	(0.02)	(0.01)	(0.15)
Gender/Grade/School FEs	Yes	Yes	Yes
Other controls	No	No	Yes
N	588	588	161
R^2	.048	.048	.32
p-value: NAME vs. PERFORMANCE	.013	.077	.2
Mean in RANDOM	.15	.14	-.16

This table presents least squares regressions using absolute change in percentile scores or a standardized measure of pressure during the second run as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Absolute changes in percentile scores within matching groups are calculated based on the change of individual ranks of students in the their class and gender from the first to the second. Percentile scores within treatment are calculated for all students within the same treatment and gender (i.e., across classrooms). Other controls include the same controls as the mediation model in Table 6, where we use times and ranks from the second rather than the first run as the pressure variable has been elicited after the second run. Note that information on pressure was only elicited at one of the three schools.