Kießling, Lukas; Radbruch, Jonas; Schaube, Sebastian

Working Paper

The Impact of Self-Selection on Performance

IZA Discussion Papers, No. 11365

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Kießling, Lukas; Radbruch, Jonas; Schaube, Sebastian (2018) : The Impact of Self-Selection on Performance, IZA Discussion Papers, No. 11365, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/177169

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11365

The Impact of Self-Selection on Performance

Lukas Kiessling
Jonas Radbruch
Sebastian Schaube

FEBRUARY 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11365

The Impact of Self-Selection on Performance

Lukas Kiessling
University of Bonn

Jonas Radbruch
University of Bonn and IZA

Sebastian Schaube
University of Bonn

FEBRUARY 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Impact of Self-Selection on Performance*

In many natural environments, carefully chosen peers influence individual behavior. In this paper, we examine how self-selected peers affect performance in contrast to randomly assigned ones. We conduct a field experiment in physical education classes at secondary schools. Students participate in a running task twice: first, the students run alone, then with a peer. Before the second run,we elicit preferences for peers. We experimentally vary the matching in the second run and form pairs either randomly or based on elicited preferences. Self-selected peers improve individual performance by .14-.15 SD relative to randomly assigned peers. While self-selection leads to more social ties and lower performance differences within pairs, this altered peer composition does not explain performance improvements. Rather, we provide evidence that self-selection has a direct effect on performance and provide several markers that the social interaction has changed.

JEL Classification: C93, D01, I20, J24, L23

Keywords: field experiment, self-selection, peer effects, social comparison, peer assignment

Corresponding author:
Jonas Radbruch
Institute of Labor Economics (IZA)
Schaumburg-Lippe-Straße 5-9
53113 Bonn
Germany
E-mail: radbruch@iza.org

* We thank Lorenz Götte, Sebastian Kube, Pia Pinger for their guidance and support. We also thank Viola Ackfeld, Philipp Albert, Thomas Dohmen, Ingo Isphording, Ulf Zölitz and audiences at Bonn, MBEPS 2017, VfS 2017, ESA Europe 2017, 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, and Max Planck Institute for Research on Collective Goods for helpful feedback and comments. We also thank the schools and students that participated in the experiments.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.

"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 used to motivate 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 focus on one's own work. In various dimensions of life – ranging from students in educational settings (Duflo, Dupas, and Kremer, 2011; Sacerdote, 2001) over cashiers in supermarkets (Mas and Moretti, 2009) and fruit pickers on strawberry fields (Bandiera, Barankay, and Rasul, 2005, 2009) to fighter pilots during World War II (Ager, Bursztyn, and Voth, 2016) – we look at the behavior of our peers and compare our own performance and choices with theirs.[1]

In many natural environments, the persons with whom we compare are carefully chosen rather than exogenously assigned. This peer selection may generally occur across two dimensions. In some cases, people know others well and are able to select their peers accordingly. This type of selection takes place mostly in settings where people interact frequently with each other, such as classrooms or workplaces. In other settings, selection is based on limited information, e.g., only past performance is observed or only certain characteristics are available as a basis for selection. People might consciously select into schools or workplaces comprising peers with a known ability.[2] Therefore, individuals self-select into certain environments and

[1] The influence of peers on our own behavior has long been recognized in the social sciences in general, as well as in economics more specifically. Such effects – commonly referred to as "peer effects" – are widely observed across a wide range of outcomes, not only for performance on the job or in school: indeed, other contexts include investment behavior (Bursztyn et al., 2014), consumption (Kuhn et al., 2011), program participation (Dahl, Løken, and Mogstad, 2014), propensity to exercise (Aral and Nicolaides, 2017; Babcock and Hartman, 2010) and wages in a firm (Cornelissen, Dustmann, and Schönberg, 2017). The settings across these studies differ enormously, as does the underlying mechanism (e.g., peer pressure, learning, complementarities). Nonetheless, all of these have in common that the behavior or action of peers imposes an externality on the action or behavior of others. Most of the research on peer effects takes the peer group or a single peer as given or randomly assigned.

[2] Festinger (1954) already conjectured that people tend to compare their own performance on average with slightly better performing individuals. Similarly, performance leaderboards for sales representatives are

even into specific peer groups within given environments. This is in stark contrast with environments where peers are randomly or exogenously assigned. Self-selection should therefore result in different peers, can affect subsequent behavior, and might even have a direct effect on our motivation.

In this paper, we study how different peer assignment rules – self-selection versus random assignment – affect individual performance and how self-selection itself affects interactions between peers. In a first step, we investigate how self-selected peers – based on either identity or relative performance – affect average performance in contrast to randomly assigned peers. After documenting differences in performance, we then analyze the underlying mechanisms. We explore whether self-selection leads to a different peer composition and we decompose performance improvements into a direct effect – stemming from being able to self-select a peer per se – and an indirect effects from a change in the relative peer characteristics. We provide evidence on the sources of the direct effect by documenting changes in the peer interaction and discuss which individuals tend to benefit most from these peer assignment mechanisms.

In order to study the effects of self-selection, we conducted a framed field experiment 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 matching rules. We implemented a random matching of pairs (RANDOM) as well as two matching rules that use the elicited preferences to implement self-selected peers. The specific setup of our experiment allows for two notions of self-selection, based on either social identity or the relative performance of one's classmates. First, the classroom environment enabled students to state preferences for known peers (*name-based preferences*). Second, using a running task yields direct measures of performance and thus could be used to select peers based on their relative performance in the first run (*performance-based*

widespread for motivational reasons. They allow employees to compare their performance with others despite not knowing them personally.

preferences). Utilizing these two sets of preferences, we implemented two treatments with self-selection. The treatments matched students based on name-based preferences (NAME) or preferences over relative performance (PERFORMANCE), which we elicited in the survey.

We find that both peer-assignment mechanisms with self-selected peers improve average performance by .14–.15SD relative to randomly assigned peers. Self-selection changes the peer composition, e.g., students interact predominantly with friends in NAME, while they choose students with a similar past performance in PERFORMANCE. However, this indirect effect due to changes in the peer composition cannot explain performance improvements in treatments with self-selected peers. More specifically, the indirect effect of the changed peer composition is insignificant in NAME and even negative in PERFORMANCE. Our estimates show that there is a direct effect of self-selection on performance. Therefore, this process of self-selection seems to provide an additional motivation to students. In order to investigate the sources of the direct effect, we show that students in PERFORMANCE experience more peer pressure. Furthermore, we find that only slower students within a pair improve their performance in NAME, while both the slower and faster student improve similarly in PERFORMANCE compared to students in randomly formed pairs. Both observations suggest that the within-pair interaction has changed across treatments. Finally, we examine which students in the ability distribution tend to benefit most from our peer assignment mechanisms. We find that NAME improves students across the ability distribution, while PERFORMANCE tends to favor faster students.

While the impact of a peer and the resulting quantitative effect might be specific to this setting, the underlying motive of the results are of general interest. Students have not only been successfully used to analyze phenomena like favoritism (Belot and van de Ven, 2011, and references therein), but they are also a highly relevant subject group, given that social comparisons are important drivers of effort and performance in school and consequently may affect educational attainment. The process of self-selecting peers is potentially equally important for settings in which peer effects do not arise due to social comparisons or peer pressure, but rather where effort, task or skill complementarities exist (e.g., Mas and Moretti, 2009) or where learning from peers is important (among others Bursztyn et al., 2014; Kimbrough, McGee, and Shigeoka, 2017). In these settings, peer effects originate from different mechanisms than stud-

ied in our setting, although in principle peers can also be self-selected. This may affect interactions among peers and the motivation of individuals themselves in ways similar to this study. Our results also complement the findings by Bartling, Fehr, and Herz (2014), who demonstrate that people value the opportunity to actively select relevant aspects of life, whereas we highlight the motivational benefits of subjects being able to self-select their environment (i.e., their peer).[3] As our paper shows, the direct effect of being able to self-select peers might be even more important than those induced by exogenous group assignment. Hence, studies analyzing interactions between peers and policies leveraging these insights need to take into account any selection of peers taking place within groups.

This paper relates to several strands of literature and addresses recent developments on peer effects. First, most studies have traditionally relied on (conditional) random assignment of peers (for an overview, see Herbst and Mas, 2015; Sacerdote, 2014). In order to study peer effects in performance, these studies impose – for example – that all other class members (e.g., as in Feld and Zölitz, 2017) or the entire set of friends (by leveraging social network data as in Bramoullé, Djebbari, and Fortin, 2009) serve as relevant peers. This literature builds on (conditional) random assignment to identify the existence of peer effects and circumvent statistical problems as outlined in Manski (1993). As we are interested in how self-selection actually changes peer group compositions and performance, we contrast the setting typically used in the literature (i.e., random assignment) by allowing for self-selection.

The existence of peer effects in educational settings motivated a small strand of the literature to focus on reassignment policies. Rather than assigning students randomly to classrooms, Carrell, Sacerdote, and West (2013) systematically formed classes comprising only high- and low-ability students to increase the GPA of the latter. Instead of increasing their GPA as was predicted by estimates in Carrell, Fullerton, and West (2009), the GPA actually decreased. The authors suggest that subgroups of either high- or low-ability students emerged, with little interaction between them. Therefore, the exogenous formation of classrooms changed the

[3] Similarly, having the opportunity to decide or vote has been found to positively affect the quality of leadership (e.g. Brandts, Cooper, and Weber, 2014) as well as the effectiveness of institutions (e.g. Bó, Foster, and Putterman, 2010) in the presence of social dilemmas.

class composition and thereby the set of potential peers, while within this group students self-selected their relevant peers. Booij, Leuven, and Oosterbeek (2017) also present evidence on exogenous peer group manipulations. They manipulated the group composition based on their prior ability, which led to a change in the social interaction: low-ability students were more involved in classes and reported more positive interactions within classrooms. In this paper, rather than reassigning students into classrooms as in the previous studies, we take the classes as given and focus on the peer assignment and resulting interactions within classrooms.

Researchers have recently analyzed the potentially differential effects of friends and non-friends and thus have moved away from the paradigm that all peers influence an individual's performance similarly. Lavy and Sand (2015) analyze how reciprocal – in contrast to non-reciprocal – friends affect the test scores of middle-school students. Chan and Lam (2015) further decompose the type of peers and investigate the varying effect of those types on educational attainment. In particular, they find that the specific type of peer (classmate, seatmate or friend) as well as their individual personalities matter for understanding peer effects in educational settings. In a different domain, Aral and Nicolaides (2017) suggest that only some parts of a person's social network affect exercising behavior. While Aral and Nicolaides focus on the extensive margin – i.e., the decision to exercise or not – and study who is influencing whom in a given network, we study the intensive margin – i.e., how much effort to provide – taking into account that not all people serve as relevant peers.

The closest paper to ours is Chen and Gong (2017). The authors study self-sorting of students into teams for a group task with skill complementarities. Similar to us, they find that peers are selected based on the social network and that those groups perform better than randomly formed groups. In contrast to their setting, we focus on pairs with individual production as the unit of analysis to identify a peer's effect on individual performance.

Finally, we add to the literature on rank effects in peer interactions. Our results are in line with research documenting the importance of ranks for subsequent outcomes (Elsner and Isphording, 2017; Gill et al., 2017). If individuals have preferences over ranks, this can give rise to heterogeneous peer effects similar to our setting (Tincani, 2017). Relatedly, Cicala, Fryer,

and Spenkuch (forthcoming) also use rank-dependent preferences to build a Roy-model of social interactions, where agents can select into certain groups based on their ability to carry out different tasks. In our experiment, subjects can indirectly select their rank in the second run by choosing a specific peer or relative time (e.g., a faster or slower peer).

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 direct effects of self-selection and indirect effects of a change in the peer composition. We then discuss heterogeneous responses 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. Additionally, it might be very difficult to isolate the person who serves as a point of comparison. This is especially true if several potential peers are present at all times. Moreover, within a given group only some peers might serve as relevant comparisons. 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 use an existing social group. In these groups, subjects have a clear impression of other group members and are able to select peers on additional characteristics such as their social ties.

In this study, we used the controlled environment of a framed field experiment to overcome those 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 also excluded complementaries in production between the students. Moreover, we focused 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 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

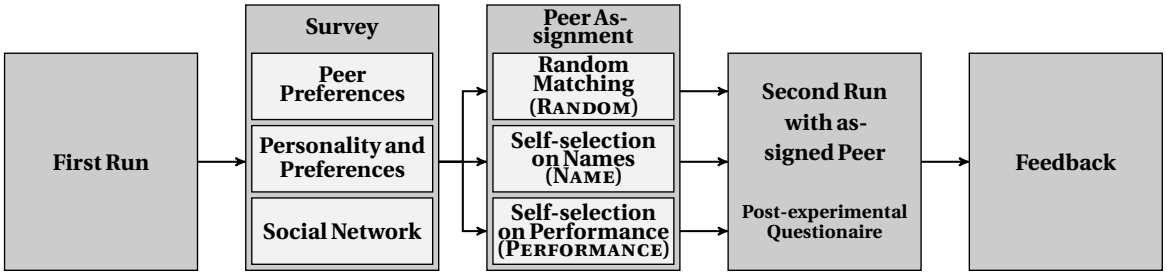
Figure 1 illustrates the experimental design. Students participated in a running task commonly known as "suicide runs", a series of short sprints to different lines of a volleyball court.[4,5] The first run – in which students run alone – served two purposes: first, recorded times can be used as a measure of ability 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 both students performed the task simultaneously, while their times were recorded

[4]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.

[5]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.

individually. Feedback about performance in both runs was provided at the end of the experiment only.

Figure 1: Experimental design



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 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 sociodemographic questions and measures of personality and preferences: the Big Five inventory as used in the youth questionnaire of the German socioeconomic panel (Weinhardt and Schupp, 2011), a measure of the locus of control (Rotter, 1966), competitiveness[6], 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 their six closest friends within the class and indicate the intensity of their friendship on a seven-point Likert scale.

[6]We implemented a novel continuous 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.

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.

2.2 Preference elicitation

We elicited two sets of peer preferences, independent of the treatment to which a subject is assigned. The first set elicited those 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.[7] 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 ten categories consisting of 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. They 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.[8]

[7] 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.

[8] Naturally, each time interval could only be chosen once in the preference elicitation, but 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.

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). This condition 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 roommate problem. We thus implemented the "stable roommate" algorithm proposed by Irving (1985) to form stable pairs using the elicited preferences.[9]

Subjects did not know the specific matching algorithm, but were only told that their preferences would be taken into account when forming pairs. 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.[10] As this only serves the purpose of excluding learning as a source of time improvements between the two runs, we exclude it from the main analysis and focus only on the evaluation of different peer assignment rules.

[9] 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.

[10] 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.

2.4 Procedures

We conducted the experiment in physical education lessons at three secondary schools in Germany.[11] 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.

The experiment started with a brief explanation of the following lesson and demonstration of the experimental task. 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.[12] The students themselves did not receive any information on their performance until the completion of the experiment. We did not incentivize students with monetary rewards. Instead, we stressed that the objective was to run as fast as possible in both runs. Moreover, teachers used the times in their own class evaluation and students themselves were also interested in their own times.[13] The introduction concluded with a short warm-up period. After this, the subjects were led to a location outside of the gym.

Students entered into the gym individually. Thus, any potential audience effects from classmates being present were ruled out 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.[14] 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

[11] Physical education lessons in most German secondary school last 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. This was sufficient to complete the experiment.

[12] 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.

[13] Note that this resembles many real-life settings with individual tasks, where individuals are not explicitly incentivized either.

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

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, announced which rule was implemented in their class and the resulting pairs from the matching process. Following these instructions, the entire group waited again outside the gym. 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.[15]

3 Data description and manipulation check

We present summary statistics of the students in our sample in Table 1.[16] In total, 627 students participated in the treatments, with 66% being female.[17,18] This corresponds to a participation rate of 73%.[19]

On average, female students took 27.57 seconds (SD of 2.50 seconds) in the first run. Their performance is quite stable across all grades, with students from the seventh grade being

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

[16]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.

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

[18]In classes with an odd number of students within a matching group, we dropped one participant randomly to match students accordingly. Therefore, some students participated in the experiment but were only recorded once and are dropped for estimating the treatment effects in the next section.

[19]We aimed at recruiting all students of 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.

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.

somewhat slower. Male students' times decreased with age: while male students in grade 7 took on average 25.33 seconds in the first run, their performance improved to 23.27 seconds on average in grade 10. In the following, we 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 treatment and check whether observable characteristics differ between our treatments in Appendix Table A.1. There are no observable differences across treatments for most variables, except for a difference in the pre-treatment times in the first run. However, this gap can be explained entirely by variation in observables. Conditional on gender-specific grade fixed effects, school fixed effects and age, these differences are no longer significant.

3.1 Preferences for peers and manipulation check

Before turning to the results of the experiment, we briefly present the preferences for peers as 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 have a higher probability of being matched with someone who they prefer more, i.e., who ranks 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.

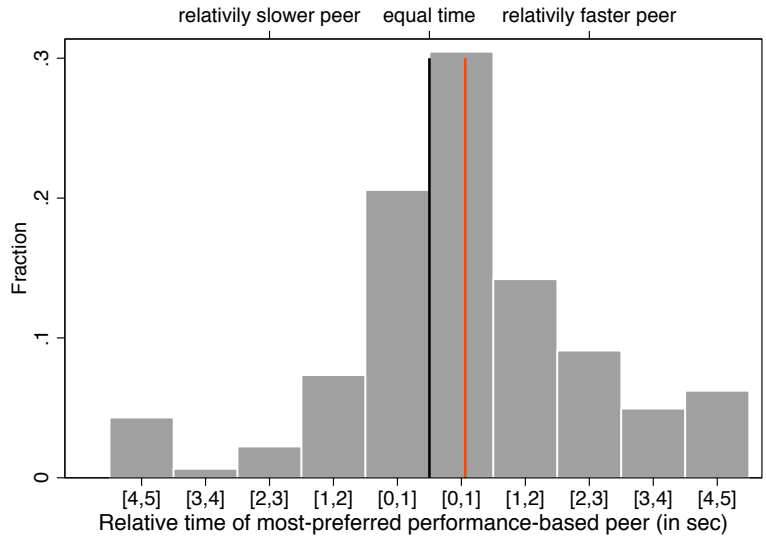
Table 2: Share of name-based preferences being friends

Name-based Preference	1st	2nd	3rd	4th	5th	6th	overall
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 nominate friends as their most-preferred peer; and second, while students prefer to run on average

Figure 2: Most-preferred performance-based peer



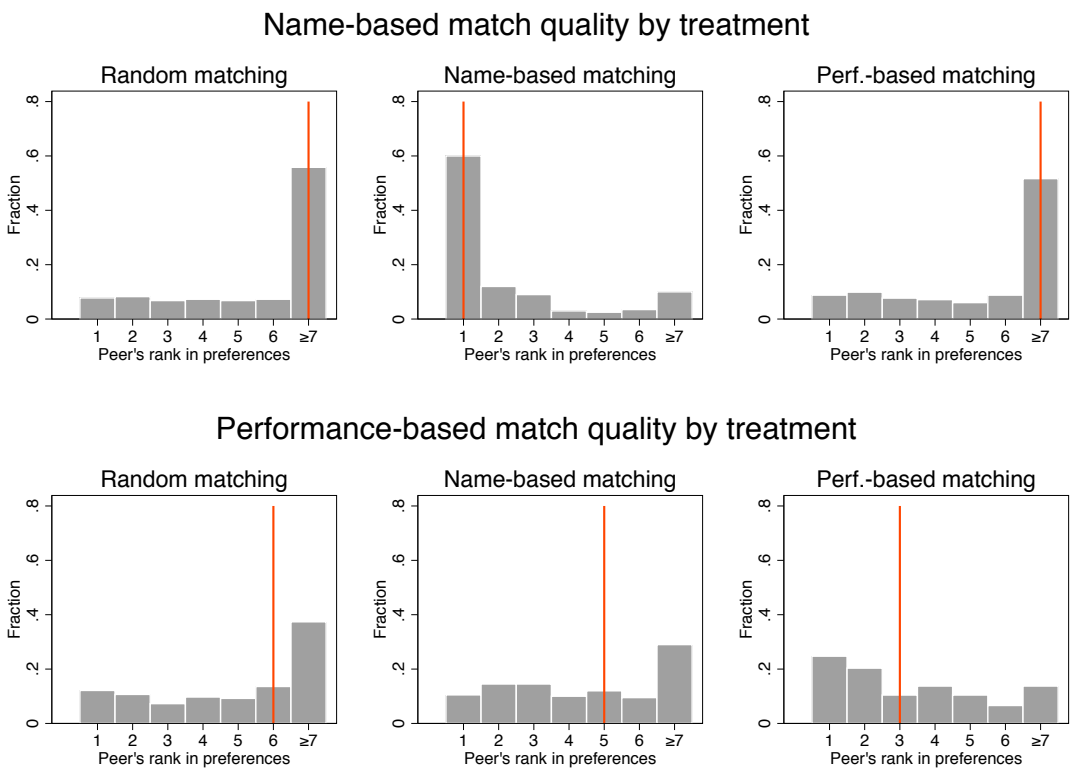
The figure presents a histogram of the peer preferences over relative performance as elicited in the survey. Vertical lines indicate own time (black line; equals zero by definition) and the mean preference of all individuals (red line; 0.56 sec faster on average, where we used the midpoint of each interval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

with a slightly faster peer, there is strong heterogeneity in this preference. We analyze the determinants of these preferences as well as how these two preference measures relate to each other in more detail in Kiessling, Radbruch, and Schaube (2018).[20]

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 are randomly matched to someone they would like to be paired with 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

[20] In Appendix B, we also show that the rankings of preferred name- and performance-based peers measure two distinct sets of preferences mitigating concerns that the two peer measures correspond to the same underlying preference.

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.

Figure 3. We observe that students in PERFORMANCE are paired with more preferred peers according to their preferences relative to the other two treatments. However, note that subjects may prefer other students or relative times that are not available to them, which mechanically affects the match quality. Moreover, to match students in PERFORMANCE, the preferences need to exhibit sufficient heterogeneity. We discuss these issues in more detail in Appendices B and C and show that sufficient heterogeneity in preferences exists to match students successfully.

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 – before we show how to allow for heterogeneities in the direct effect depending on the rank within a pair. In Appendix D, we show how to derive these estimation equations from an economic model similar to the mediation analysis as 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. Our baseline specification for an outcome y_{igs} of individual i in gender-specific grade g of school s is therefore 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., due to different locations and timing of the experiment) and variation specific to gender and grades.[21] Finally, X_i is a vector of predetermined characteristics such as age as well 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.

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 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-

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

selection as well as a pure peer composition effect.[22] This takes into account the change in relative peer characteristics across treatments. We implement this decomposition using the following specification:

$$y_{igs} = \bar{\tau} + \bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{2}$$

We are interested in $\bar{\tau}_N$ and $\bar{\tau}_P$, the direct effects of our treatments relative to RANDOM. Changes in peer characteristics are captured by θ_i. In particular, we allow our effects to be mediated by the quality of the match measured by the rank of the peer in an individual's preferences, ability differences and ranks withing pairs, friendship ties and a set of personality and preference measures (i.e., Big Five, locus of control, competitiveness, risk attitudes, social comparison).

Finally, we analyze the heterogeneous direct effects of ranks within pairs to analyze whether only certain individuals are reacting to our treatments using

$$\begin{aligned} y_{igs} = \bar{\tau} &+ \bar{\tau}_h^N \mathbb{1}_{\{a_i \geq a_j\}} D_i^N + \bar{\tau}_l^N \mathbb{1}_{\{a_i < a_j\}} D_i^N \\ &+ \bar{\tau}_h^P \mathbb{1}_{\{a_i \geq a_j\}} D_i^P + \bar{\tau}_l^P \mathbb{1}_{\{a_i < a_j\}} D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \end{aligned} \tag{3}$$

The indicator $\mathbb{1}_{\{a_i \geq a_j\}}$ denotes whether subject i was of higher ability (e.g., faster in the first run) than her or his peer j, and $\mathbb{1}_{\{a_i < a_j\}}$ equals one if i was of lower ability. We interact this rank indicator with the treatment indicators D_i^d ($d \in \{N, P\}$) to analyze whether the direct effect depends on the rank within a pair.

5 Results

Our experimental design allows us to study the causal effect of different peer assignment mechanisms on individual performance. Two of these assignment rules use the preferences for

[22] The direct effect mainly captures changes in motivation due to being able to self-select a peer, but also inputs that (i) differ across treatments, and (ii) are not measured in our rich set of potential mediators (match quality, friendship ties, ability differences, ranks and personality differences).

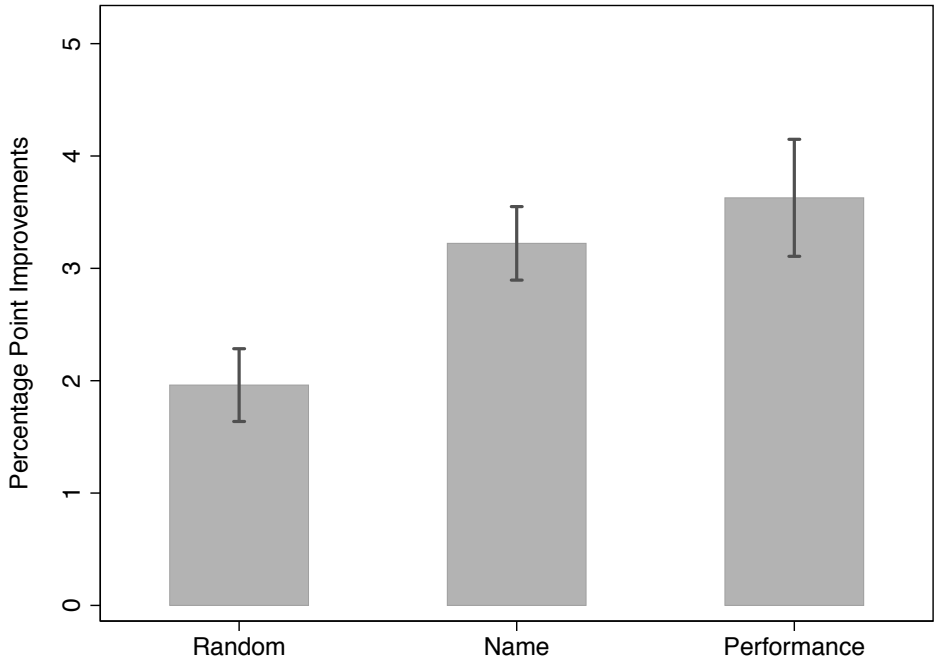
peers elicited in the survey to form pairs and therefore allow for the self-selection of peers. More specifically, the three treatments correspond to random matching (RANDOM), matching based on self-selected peers using name-based peer preferences (NAME) and using preferences over relative performance (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 (e.g., Falk and Ichino, 2006; Guryan, Kroft, and Notowidigdo, 2009). We contrast this baseline condition with two treatments that assign peers based on elicited preferences, i.e., in which subjects endogenously choose their peer.

5.1 Average effect of self-selection on performance

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

Figure 4 presents our main result. Subjects in RANDOM improve on average by 1.93 percentage points during their second run. However, 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 (3)-(7) express the results additionally in terms of (standardized) times in the second run controlling for times in the first run to confirm these effects in times rather than percentage point improvements. Assigning peers based on name-based preferences results in an additional 1.26 percentage point improvement in performance relative to the random assignment of peers. The coefficient for performance-based matching is 1.67 percentage points and thus somewhat larger, but it does not differ significantly from NAME. These effects persist when controlling for students' own personal characteristics (col-

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 (1) in Table 3. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level.

umn (2)) as well as if we additionally control for class-level variables capturing the atmosphere within a class (column (3)). Our baseline effects correspond to additional time improvements of .38 to .41 seconds and account for 14% of a standard deviation in NAME and 15% in PERFORMANCE (cf. columns (4)-(7)).[23]

In Appendix E, we show that the observed performance improvements are due to 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 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.

[23]Appendix F presents additional robustness checks using biased linear reduction standard errors, controlling for outliers, and presents the average treatment effects for different subgroups. Our results are robust to all of these checks.

Table 3: Average treatment effects

	(a) Percentage Point Imprv.			(b) Time (Second Run)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
NAME	1.26***	1.37***	1.84***	-0.38***	-0.38***	-0.48***	-0.14***
	(0.43)	(0.50)	(0.46)	(0.11)	(0.12)	(0.12)	(0.04)
PERFORMANCE	1.67**	1.69**	1.28**	-0.41***	-0.38***	-0.31**	-0.15***
	(0.62)	(0.65)	(0.60)	(0.14)	(0.14)	(0.14)	(0.05)
Time (First run)				0.69***	0.67***	0.71***	0.74***
				(0.04)	(0.04)	(0.05)	(0.04)
Class-level Controls	No	No	Yes	No	No	Yes	No
Own Characteristics	No	Yes	Yes	No	Yes	Yes	No
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	588	585	515	588	585	515	588
R^2	.056	.08	.096	.8	.81	.83	.8
p-value: NAME vs. PERFORMANCE	.51	.62	.38	.8	.98	.28	.8

This table presents least squares regressions according to equation 1 using percentage point improvements (panel (a)) and times of the second run controlling for times in the first run (panel (b)) 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 characteristics include the Big 5, locus of control, social comparison, competitiveness and risk attitudes. Class-level control variables in columns (3) and (6) include the share of participating students, three variables to capture the atmosphere within a class (missing for four classes), and indicators for the size of the matching group. Column (7) uses standardized times.

5.2 Changes in the peer composition and the direct effect of self-selection

As outlined in section 4, the estimated average treatment effects consist of a direct effect due to self-selection and an indirect effect. The latter captures changes in the relative characteristics of the peer (e.g., the time differences between the student and peer in the first run) due to the altered peer composition induced by our treatments.[24] In the following, we first document how NAME and PERFORMANCE change the peer composition relative to RANDOM, before analyzing the extent to which this change in the peer composition can explain the average treatment effect.

[24] Note that only the relative characteristics within a pair can matter for a change in the performance, given that we randomize subjects into treatments. Therefore, the overall distribution of peer characteristics across treatments is similar and constant. Our treatments only change with whom each student interacts.

It is important to check for a change in the composition and the resulting indirect effect as potentially not all peers are equally important. Suppose that only interacting and comparing yourself with a friend leads to a change in performance (e.g. Bandiera, Barankay, and Rasul, 2009) and at the same time subjects only select their friends in NAME. Alternatively, suppose that peers only matter if they have a similar performance and at the same time subjects more commonly select someone with a similar performance in PERFORMANCE. Potentially, our treatments would simply change the likelihood of interacting with such a person (i.e., change the peer composition between treatments) and these changes would explain the average treatment effect.

Figure 5 shows that our treatments indeed changed the peer composition with respect to two prime examples of relative peer characteristics, namely friendship ties and ability differences within pairs. Even though students could mainly target peers along these two dimensions, we present how our treatments affect the peer composition along various other characteristics in Appendix Table C.1. 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 other students with a similar or slightly higher ability, 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 greater than two seconds in the other two treatments (2.24 and 2.16 seconds in RANDOM and NAME).

While the existing literature to date has mainly concentrated on the influence of peers with respect to ability and friendship ties on performance, our data allows us to go beyond this.[25] In particular, we allow for a large set of different personal characteristics (competitiveness, Big Five, Locus of control, social comparison, and risk attitudes) to influence the performance.

Moreover, by having access to preferences over peers, we are able to include the match quality of a peer as a potential mediator. For this purpose, we define two indicators to measure

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

Figure 5: Changes in peer composition

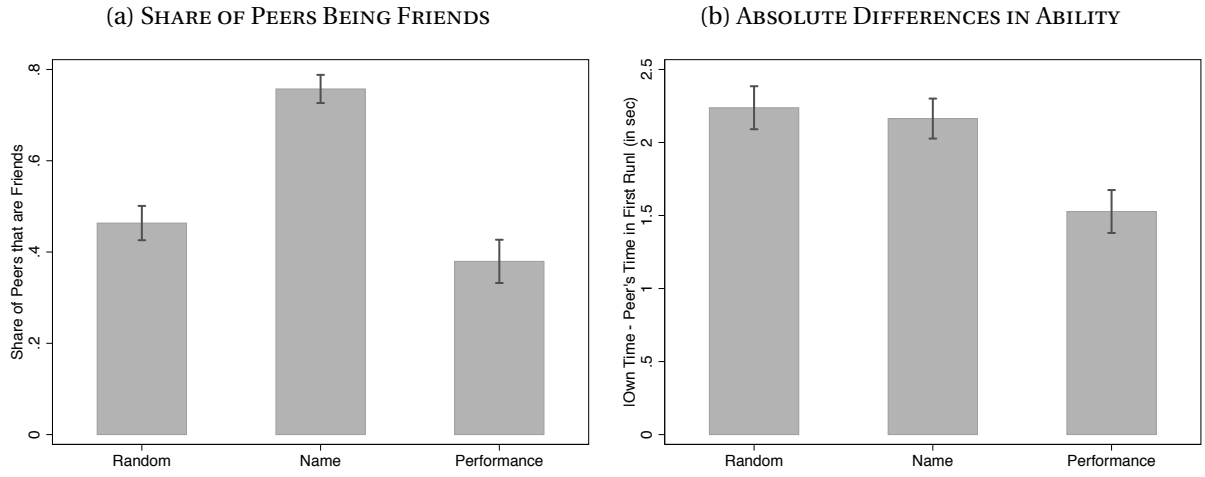


Figure 5a presents the share of all students who nominated their assigned peer as a friend for each of the three treatments including standard errors. Figure 5b shows the average absolute within-pair difference in ability (measured in times from the first run) and including standard errors for each treatment. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level. We present the corresponding regressions and highlight additional compositional differences of the treatments in Appendix Table C.1.

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.[26]

The results of the decomposition based on equation 2 are presented in Table 4. Column (1) replicates the baseline estimates from column (2) of Table 3 for means of comparison. In columns (2)-(5), we include different sets of characteristics, before we allow all of them to mediate the direct effects in column (6).

Only controlling for name-based and performance-based match quality or friendship ties (column (2) and (3)) has little to no effect as the variables themselves have only small and insignificant effects on performance improvement. Hence, the estimated direct effects closely resemble the average treatment effects. In column (4), we focus on ability differences and ranks within a pair. Since faster and slower students within a pair might be affected differentially, we allow the effect of ability differences, $|\Delta Time1|$, to differ by the rank within a pair. We find that ability differences have a significant effect on both faster and slower students within

[26]Appendix Table G.3 also controls for match quality in a flexible way. The results remain qualitatively and quantitatively similar.

Table 4: Decomposition of treatment effects

	Percentage Point Improvements					
	(1) Baseline	(2) Match Qual.	(3) Friend	(4) Time Diff.	(5) Personality	(6) All
Direct Effects						
NAME	1.37***	1.36**	1.48***	1.35***	1.36***	1.26**
	(0.50)	(0.54)	(0.52)	(0.46)	(0.44)	(0.47)
PERFORMANCE	1.69**	1.74**	1.66**	1.84***	2.03***	2.18***
	(0.65)	(0.69)	(0.66)	(0.61)	(0.69)	(0.68)
Peer Characteristics						
High Match Qual. (name-based)		0.04				0.56
		(0.45)				(0.42)
High Match Qual. (perf.-based)		-0.19				-0.07
		(0.48)				(0.45)
Peer is friend			-0.38			-0.61
			(0.40)			(0.46)
Faster Student × $\lvert\Delta Time\ 1\rvert$				-0.39***		-0.35**
				(0.14)		(0.14)
Slower Student × $\lvert\Delta Time\ 1\rvert$				1.03***		1.07***
				(0.21)		(0.19)
Slower Student in Pair				-0.17		-0.14
				(0.45)		(0.46)
Abs. Diff. in Personality	No	No	No	No	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes
N	585	585	585	585	582	582
R^2	.08	.081	.082	.24	.11	.27
p-value: NAME vs. PERFORMANCE	.62	.58	.8	.43	.32	.19
Indirect Effect (NAME)						.1
Indirect Effect (PERFORMANCE)						-.49

This table presents least squares regressions according to equation 2 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. High match quality is an indicator equaling one if the partner was ranked within the first three preferences according to his or her name- or performance-based preferences. Own characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. The last two rows quantify the indirect effect for NAME and PERFORMANCE given by the combining the change in peer composition across treatments (cf. Appendix Table C.1) with the corresponding compositional effects of these characteristics in column (6). Further robustness checks are relegated to the Appendix.

a pair. On the one hand, slower students within a pair benefit strongly from running with a faster student, whereby a one-second difference in ability leads to a 1.03 percentage point improvement in the second run. On the other hand, the performance of the relatively faster student suffers from ability differences and their performance declines by .39 percentage points per second. In sum, the average performance of a pair thus improves with increased ability differences. However, the impact of ability differences does not mediate the direct treatment effects. The estimated coefficient for NAME remains stable, and the effect for PERFORMANCE even increases, implying that the indirect effect for PERFORMANCE is negative. This is partially a consequence of the smaller ability differences in PERFORMANCE relative to RANDOM as shown in Figure 5b and the overall positive impact of ability differences.

In column (5), we analyze the direct effects if we include the similarity in several personal characteristics of the two students of a pair. In contrast to ability differences and friendship ties, personality characteristics could not be targeted easily in the preference elicitation. Nonetheless, subjects could have chosen peers with certain personality characteristics indirectly in both treatments. However, the treatment effects remain stable if we control for those characteristics.

Finally, we control for all of these mediators simultaneously in column (6). The effects of the peer characteristics are in line with what we have discussed above. In the last two rows of the table, we quantify the indirect effect as the change in the coefficient of NAME and PERFORMANCE when controlling for the peer composition (column (1) vs (6)). This corresponds to multiplying the coefficients from column (6) with the change in the peer composition across treatments. We describe these changes in Appendix Table C.1.

In NAME, we estimate a positive indirect effect of .10 percentage point improvements. This means that the altered peer characteristics have only a slightly positive effect on the students' performance. The direct effect is 1.26 percentage points and therefore somewhat smaller than the average effect, but not significantly different (Wald test, p-value = 0.66). For PERFORMANCE, we observe an indirect effect of -.49 percentage points. Therefore, the change in the peer composition suppresses improvements in PERFORMANCE. The direct effect is 2.18 percentage points

and it significantly differs from the average effect (Wald test, p-value = 0.029). The magnitude of the direct effects is more than five times that of the indirect effects.[27]

Our analysis suggests 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. The direct effect could stem from an additional motivational value of self-selection, as the comparison and interaction with self-selected peers might become more important. In principle, a compositional change in unobserved characteristics – that is not measured by those included in our analysis and differs across treatments – could still account for the direct effects. However, the effect would have to be at least five times the size of the measured indirect effect.

Hence, implementing self-selection of peers has likely changed the social interaction in both treatments, either directly or by changing the influence of peer characteristics. In the next section, we present evidence that students perceive the peer interactions across treatments differently to bolster this interpretation.

5.3 Markers for changed social interactions

In this section, we study the effects of our treatments on students' experience during the tasks. Our experiment features a small post-experimental questionnaire, in which we elicited how much peer pressure students experienced and how much fun they had during the second run.[28] In order to analyze the effects of the treatments on these two variables, Table 5 presents estimates for the direct effects of our treatments based on equation 2 using standardized measures of pressure and fun as outcome variables. Here, we control for times in the second rather

[27] We present additional robustness checks in Appendix G. In Table G.1, we show that match quality itself has no influence in RANDOM. Being paired randomly with a preferred peer does not increase performance. Furthermore, Table G.2 presents the robustness of the direct effects to using only those subjects in RANDOM who are matched in line with their preferences. These matches occurred by pure chance and not due to self-selection. Finally, we document in Table G.3 that the piecewise-linear specification of ability differences and the definition of the high matching quality indicator are not restrictive by including interval fixed effects for each one-second interval of ability differences and fixed effects for each rank of the name- and performance-based preference ranking, respectively. Additionally, this table also shows that conditioning on class-level variables does not alter our results.

[28] We elicited the peer pressure measure only at one of the three schools. Therefore, we have fewer observations for this variable.

than the first run for two reasons: first, these measures are elicited after the second run; and second, a tight race could increase pressure across all treatments.

Students in PERFORMANCE experience significantly more pressure from their peer in the second run than students in RANDOM and NAME. Therefore, selecting peers based on preferences over relative performance seems to change the experience of social interactions. Note the differential effects of absolute time difference for slower and faster students within a pair on pressure: whereas slower students are always pressured to a similar degree, the pressure experienced by faster students in a pair decreases with the margin of winning.

Focusing on fun in the second run, we do not find any significant direct effects (see panel (b) of Table 5). However, we observe a significant negative effect on time differences in the second run for the slower student. Fun decreases for the slower peer with increasing distance to the peer. Combined with the zero effect of finishing second, we conclude that it is not losing per se that affects fun, but rather the margin of losing. Furthermore, the absence of direct effects alleviates a potential concern that knowledge of all three treatments leads to disappointment when students are assigned to RANDOM, namely when they are unable to select their peer themselves.[29] If those students were more disappointed, this might lead to smaller improvements by students in RANDOM compared to the two other treatments. If disappointment had driven the results, we would have expected students in RANDOM to have significantly less fun.[30]

Hence, while we find increased pressure for subjects in PERFORMANCE, we do not find any differences in fun students report across treatments. This supports the notion that the social interaction has changed at least in the pressure domain.

[29] One might also argue that this also describes a feature of real-world settings. Imagine that you are randomly assigned a partner from a group of available people. Even if you have not explicitly been asked with whom you would have liked to interact, you still have 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 people involved.

[30] A similar argument could be that our treatment effects are due to reciprocity or some kind of Hawthorne or John Henry effect, i.e., students perceive being in one or the other treatment as positive or negative. See Aldashev, Kirchsteiger, and Sebald (2017) for a discussion how this can bias treatment effects. If subjects perceive treatment assignment as being kind or unkind, we should observe some kind of reaction in the fun variable. As this is not the case, it is unlikely that the effects are due to this reason.

Table 5: Post-experimental questions

	(a) Pressure (std.).		(b) Fun (std.)	
	(1)	(2)	(3)	(4)
Direct Effects				
NAME	0.25	0.15	0.11	-0.02
	(0.21)	(0.18)	(0.08)	(0.10)
PERFORMANCE	0.36	0.51***	-0.12*	-0.10
	(0.25)	(0.15)	(0.07)	(0.08)
Peer Characteristics				
Match Quality (name-based)		0.23		0.19
		(0.18)		(0.12)
Match Quality (perf.-based)		-0.03		0.13
		(0.14)		(0.09)
Peer is friend		-0.05		0.14
		(0.27)		(0.11)
Faster Student (2nd Run) × $\lvert\Delta Time\ 2\rvert$		-0.30***		-0.01
		(0.09)		(0.04)
Slower Student (2nd Run) × $\lvert\Delta Time\ 2\rvert$		0.10		-0.13***
		(0.09)		(0.05)
Slower Student in Pair (2nd Run)		-0.27		0.02
		(0.20)		(0.10)
Gender/Grade/School FEs, Age	Yes	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Abs. Diff. in Personality	No	Yes	No	Yes
N	163	161	585	582
R^2	.098	.26	.26	.32
p-value (NAME vs. PERFORMANCE)	.65	.07	.038	.56

This table presents least squares regressions according to equation 2 using the standardized survey measure of pressure (Panel (a)) or fun (Panel (b)) 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 characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Match quality equals one if a student's peer is among his three most-preferred peers according to his name- or performance-based preferences. Note that the faster/slower student is defined according to relative times in the second run.

5.4 Do treatments change the within-pair interaction?

In order to deepen our understanding of differences across the two treatments allowing for self-selection, we estimate heterogeneous direct treatment effects with respect to the individual rank within a pair. In the previous sections, we have already shown that students with different ranks within a pair (i.e., being the faster or the slower student) react differentially in terms of both performance and how they perceive the running task. To better understand the influence of ranks and the difference of our treatments, we first focus on the heterogeneity of the direct effect with respect to the ability rank within a pair. We then proceed to look at absolute differences in times of the second run.

Column (1) of Table 6 replicates specification (6) of Table 4. In column (2), we allow the direct effect of our treatments to differ by rank according to equation 3. Self-selection yields a positive direct effect for all students independent of their rank in PERFORMANCE. In NAME, only slower students within a pair exhibit significant direct effects compared to RANDOM. Faster students within a pair are unaffected in NAME. This shows that selection on names motivates slower students to catch up with their faster peers. By contrast, selection on relative performance causes both students to improve their performance.

The observed within-pair interaction has direct consequences for the difference in performance levels across treatments. As the slower student within a pair drives the direct effect in NAME, we expect a decrease in the within-pair difference in levels in NAME. In Table 7, we analyze the absolute within-pair time difference in the second run. In column (1), we calculate the average treatment effect for these differences and show that they are significantly smaller for both treatments allowing for self-selection. In column (2), we decompose this effect again in a direct and indirect one using pair-level mediators, i.e., absolute time difference in the first run, friendship indicators and absolute differences in personality characteristics. We find that lower absolute differences in PERFORMANCE are an artifact of the changed peer composition and therefore due to the selection mechanism (i.e., lower absolute differences in ability), while we observe a direct convergence effect for NAME.

Table 6: Rank heterogeneity within pairs

	Percentage point imprv.	
	(1)	(2)
Direct Effects		
NAME	1.26**	0.60
	(0.47)	(0.54)
PERFORMANCE	2.18***	2.23***
	(0.68)	(0.69)
NAME × Slower Student in Pair		1.36**
		(0.55)
PERFORMANCE × Slower Student in Pair		-0.08
		(0.65)
Peer Characteristics		
Match Quality (name-based)	0.56	0.57
	(0.42)	(0.42)
Match Quality (perf.-based)	-0.07	-0.04
	(0.45)	(0.46)
Peer is friend	-0.61	-0.64
	(0.46)	(0.45)
Faster Student × $\|\Delta Time\ 1\|$	-0.35**	-0.34**
	(0.14)	(0.14)
Slower Student × $\|\Delta Time\ 1\|$	1.07***	1.06***
	(0.19)	(0.20)
Slower Student in Pair	-0.14	-0.56
	(0.46)	(0.64)
Abs. Diff. in Personality	Yes	Yes
Own Characteristics	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes
N	582	582
R^2	.27	.28
p-value (NAME vs. PERFORMANCE)	.19	.016

This table presents least squares regressions according to equation 3 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 characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Match quality equals one if a student's peer is among his three most-preferred peers according to his name- or performance-based preferences.

Table 7: Convergence of performance within pairs

	$\lvert\Delta Time2\rvert$	
	(1)	(2)
NAME	-0.48***	-0.37***
	(0.16)	(0.13)
PERFORMANCE	-0.36*	-0.20
	(0.20)	(0.21)
$\lvert\Delta Time\ 1\rvert$		0.49***
		(0.07)
Friendship Indicator		-0.44***
		(0.13)
Abs. Diff. in Personality	No	Yes
Gender/Grade/School FEs	Yes	Yes
N	294	291
R^2	.07	.52
p-value: NAME vs. PERFORMANCE	.52	.41
Mean in RANDOM	1.7	1.7

This table presents least squares regressions using absolute differences of times in 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. Additional peer composition controls include absolute differences of personality characteristics of subjects and their peers (Big Five, locus of control, social comparison, competitiveness, risk attitudes).

Although the direct effect of self-selection in both treatments is similar in sign and magnitude, the two treatments induce distinct interaction patterns within pairs. While in NAME only the slower student within a pair drives the direct effect, all students improve due to self-selection in PERFORMANCE. We also observe a similar convergence in performance levels across both treatments with self-selection. However, this result is due to the selection mechanism in PERFORMANCE and due to the interaction in NAME. In combination with the results in section 5.3, these heterogeneous effects show that our treatments work through different channels and thereby affect the subjects differently.

5.5 Implications for targeting individuals

Our results show that the process of self-selection has a heterogeneous impact on the subjects depending on the rank within a pair. However, a policy maker might not only be interested in the changed interaction within pairs, but rather they might target specific groups of individuals to improve their performance, irrespective of direct or indirect effects driving these improvements. For this purpose, we look at the heterogeneity in average treatment effects conditional on ability and simulate the effects of other rules employing exogenous peer assignment.

Figure 6 presents percentage point improvements of low-, medium- and high-ability subjects across the three assignment rules.[31] Across all treatments, the performance improvements decrease when ability increases but remain positive even for high-ability students. This mainly stems from the positive effect of ability differences for slower students within a pair and negative effect for faster ones.[32]

Although this decreasing pattern holds for all three treatments, there are some differences. Low-ability students in RANDOM show large improvements of 4.77 percentage points (p-value < 0.01), while medium- and high-ability students do not improve significantly (.96 and .36 percentage points with p-values of .30 and .31, respectively). All students across the ability distribution improve more in NAME than in RANDOM by 1.02 (p-value = 0.28), 2.00 (p-value = 0.05) and 1.01 percentage points (p-value = 0.04) for low-, medium- and high-ability students. By contrast, PERFORMANCE does not help low-ability students relative to RANDOM (.20 percentage points decrease, p-value = 0.86) but benefits students from the upper two terciles of the ability distribution by 2.57 (p-value = 0.02) and 2.16 (p-value < 0.01) percentage points. Overall, the performance improvements are more equally distributed across different levels of ability.

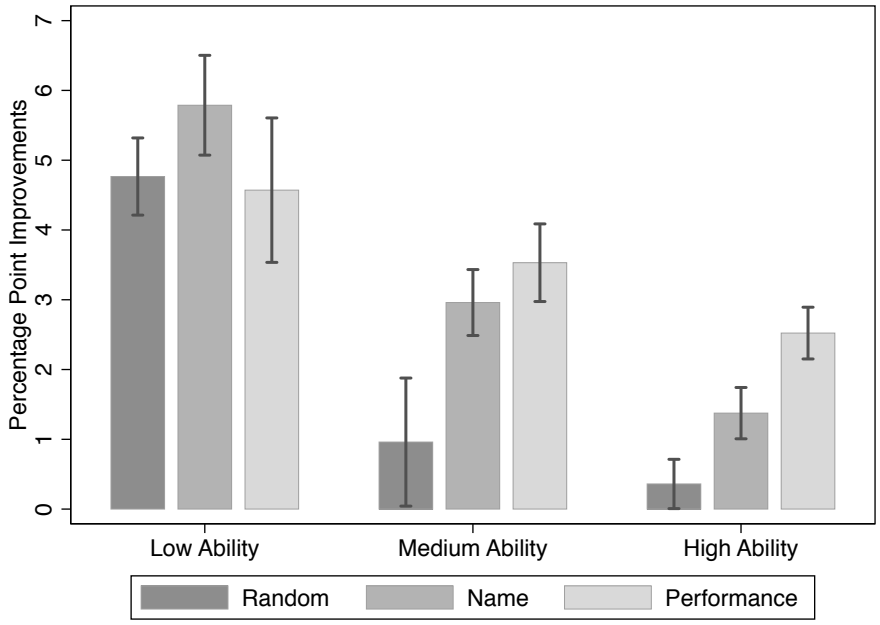
The treatments therefore target different groups of individuals. Low-ability students benefit most from name-based matching, whereas students with higher ability show the largest

[31] The corresponding regressions as well as alternative specifications are presented in Appendix Table H.1. Low, medium and high ability are defined according to terciles of times in the first run within each school, grade and gender.

[32] Table 6 shows that a one-second ability difference improves performance by 1.06 percentage points for slower students within a pair and reduces the faster students' performance by .34 percentage points.

improvements when matched using preference over relative performance. Policy makers can therefore use different peer assignment rules to benefit specific groups of individuals.

Figure 6: Heterogeneity by own ability



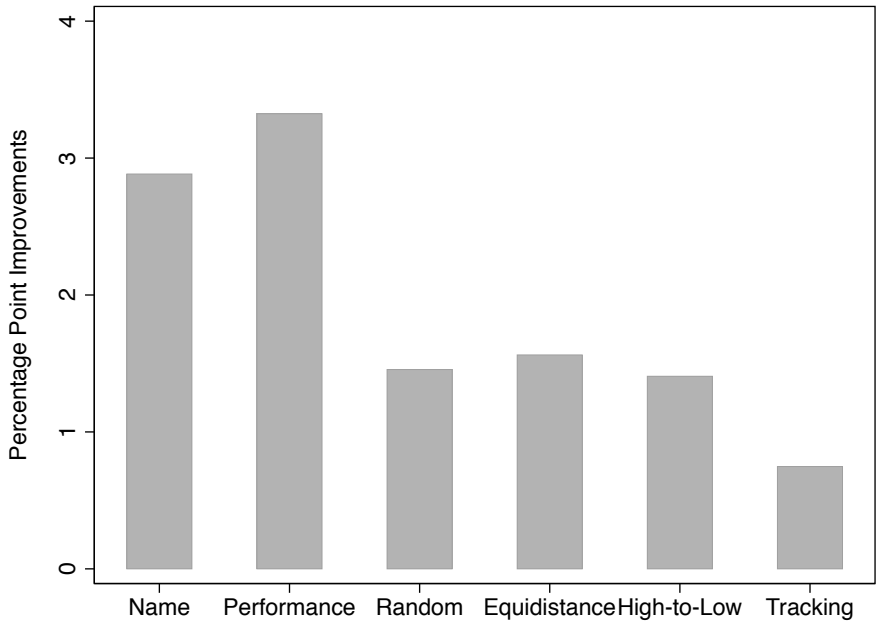
The figure presents percentage point improvements and standard errors for the three treatments RANDOM (dark gray), NAME (gray), and PERFORMANCE (light gray) by ability terciles. We control for gender, grade and school fixed effects as well as for age and cluster standard errors at the class level. The corresponding regressions are presented in Appendix Table H.1.

The previous sections and the patterns in Figure 6 imply that individual improvements are largely determined by the interplay of the peer – especially his or her relative ability – and the treatment. Table 6 shows that mainly the slower students within a pair improve in NAME, while both improve similarly in PERFORMANCE compared to the random assignment of peers. Low-ability students benefit most from this as they are more likely to be paired with faster students. This effect is amplified compared to PERFORMANCE as this treatment results in pairs with smaller ability differences relative to the other two treatments.[33] Note that this

[33]Figure 5b shows that the ability differences are indeed lower in PERFORMANCE. These lower ability differences translate into smaller indirect effects for the pair and especially for the slower peer. This is due to smaller ability differences reducing improvements for slower students in a pair, which are not compensated by the effects on faster students. See the coefficients on ability differences interacted with rank of a student in Table 6.

only results in a positive effect for low-ability students in NAME if these students choose faster students and are subsequently matched with them, a condition that is satisfied in our setting (see Appendix Table H.2). This implies that the choice of a peer by an individual carries greater weight for individual improvements in treatment NAME than in PERFORMANCE, as the former only benefits slower students in a pair whereas the latter benefits both students. This might also help to understand the absence of improvements for low-ability students in Carrell, Sacerdote, and West (2013) as students in their setting might not have chosen high-ability students as relevant peers.

Figure 7: Simulation of other peer assignment rules



The figure presents predicted percentage point improvements for the three treatments (NAME, PERFORMANCE and RANDOM) 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 I.

While we have shown that self-selected peers improve aggregate performance compared to randomly assigned ones, in many situations peers are not assigned at random but rather in line with a specific matching rule. Schools employ tracking (e.g., Duflo, Dupas, and Kremer, 2011) or pair high-ability students with low-ability ones (e.g., Carrell, Sacerdote, and West, 2013). We

can use our estimates to simulate the effect of such peer-assignment rules and compare their effect to the outcomes under self-selection. From our estimates obtained in section 5.2, we know that pairs with a higher difference in ability will improve their performance. If this is the only characteristic of a peer that affects performance, aggregate performance would be maximized as long as the sum of ability differences within a pair is maximized.[34] 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 ability differences within pairs (EQUIDISTANCE and HIGH-TO-LOW). Additionally, we look at the effect of tracking (i.e., pairing the best student with the second best, third with the fourth, etc.; TRACKING). We compare the predicted performance improvements for those rules with our estimated performance improvements for the three assignment rules used in the experiment.[35]

Figure 7 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 these students under those peer assignment rules do not benefit from the additional motivational value of self-selection. More surprisingly, the reassignment rules that maximize ability differences in pairs – EQUIDISTANCE and HIGH-TO-LOW – do not improve average performance compared to the random assignment of peers. Although both rules increase the average ability 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 implies that these other changes in peer characteristics offset the positive effect of increased ability differences.

In general, depending on the objectives such as targeting specific groups of individuals, a policy maker such as a teacher might want to implement different peer assignment mechanisms. While our treatments allowing for self-selected peers seem to induce similar performance improvements on average, they affect different individuals. Compared to RANDOM, we

[34]This holds true for all peer-assignment rules that match each student from the bottom half of the ability distribution with a student from the top half.

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

observe performance improvements across the entire ability distribution in NAME, but only for higher-ability students in PERFORMANCE. Nonetheless, such peer assignments may come at a cost, such as increased pressure in PERFORMANCE (as documented in section 5.3) or a large perturbation of individual ranks in NAME.[36] Hence, a policy maker might not only look at the resulting outcomes 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 .14-.15 SD. These cannot be explained by indirect effects of a differing peer composition; rather, they stem from a direct effect, corresponding to a changed social interaction since students are able to select their partner themselves. This implies that self-selected peers can serve as a substantial motivator to improve performance.

Teachers or supervisors might be interested to leverage this direct effect of self-selection. They may allow students to choose their study group themselves or introduce flexible seating patterns in offices such that employees can self-select their seat mates, office partners or colleagues. Since our results suggest that self-selecting peers improves outcomes, the effectiveness of social comparison interventions (as, e.g., in Allcott and Kessler, 2015) more generally may be improved if individuals are given the opportunity to select their relevant comparison themselves rather than being assigned an unspecific one.

The results reported in this paper are also in line with earlier studies, which indicate that being paired with high-ability peers leads on average to higher performance (e.g., Carrell,

[36] We document this perturbation of ranks in the Appendix Table H.3.

Fullerton, and West, 2009). Combined with the process of self-selecting high- or low-ability peers, this can set ex-ante similar individuals on divergent trajectories in classrooms and organizations. Repeatedly choosing higher-ability peers can lead to continuous improvements, whereas selecting lower-ability peers may stall individual development.

In general, our findings give rise to a trade-off between the additional motivation due to self-selection and the exogenous assignment of performance-maximizing peers. On the one hand, giving subjects discretion over the peer choice enhances motivation and thereby increases performance. On the other hand, the resulting pairs are not necessarily performance-maximizing or optimal, as also described in Carrell, Sacerdote, and West (2013). It is therefore interesting to ask whether it is possible to overcome this trade-off: How do subjects' choices and subsequent performance change once they are informed how different peers affect their performance or are nudged to select stronger peers? However, some students may prefer slower peers; for example, to avoid pressure or due to status concerns. Hence, faster peers might not be a superior choice for all individuals.

Our experimental design can easily be transferred to situations in which other production functions are used or where peer effects arise via other channels, e.g., implementing team production by reporting a function of both students' times to the teacher, or varying the task to allow for learning or skill complementaries as sources of peer effects. Self-selection of peers can often be observed in those settings. For example, study groups at universities often form endogenously (Chen and Gong, 2017), researchers select their co-authors and workers in firms increasingly form self-managed work teams (Lazear and Shaw, 2007).

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 (Dec. 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.

Allcott, Hunt and Judd Kessler (Oct. 2015). "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons". In: *NBER Working Paper Series*.

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

Babcock, Philip and John Hartman (Dec. 2010). "Networks and Workouts: Treatment Size and Status Specific Peer Effects in a Randomized Field Experiment". In: *NBER Working Paper Series*.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005). "Social Preferences and the Response to Incentives: Evidence from Personnel Data". In: *Quarterly Journal of Economics* 120.3, pp. 917–962.

— (2009). "Social Connections and Incentives in the Workplace: Evidence From Personnel Data". In: *Econometrica* 77.4, pp. 1047–1094.

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

Belot, Michèle and Jeroen van de Ven (2011). "Friendships and Favouritism on the Schoolground – A Framed Field Experiment". In: *Economic Journal* 121.557, pp. 1228–1251.

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.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009). "Identification of Peer Effects through Social Networks". In: *Journal of Econometrics* 150, pp. 41–55.

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, Richard Fullerton, and James West (2009). "Does Your Cohort Matter? Measuring Peer Effects in College Achievement". In: *Journal of Labor Economics* 27.3, pp. 439–464.

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.

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 (2017). "Can self selection create high-performing teams?"

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

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg (Feb. 2017). "Peer Effects in the Workplace". In: *American Economic Review* 107.2, pp. 425–456.

Dahl, Gordon B., Katrine V. Løken, and Magne Mogstad (July 2014). "Peer Effects in Program Participation". In: *American Economic Review* 104.7, pp. 2049–2074.

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 Andrea Ichino (2006). "Clean Evidence on Peer Effects". In: *Journal of Labor Economics* 24.1, pp. 39–57.

Feld, Jan and Ulf Zölitz (2017). "Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects". In: *Journal of Labor Economics* 35.2, pp. 387–428.

Festinger, Leon (1954). "A Theory of Social Comparison Processes". In: *Human Relations* 7.2, pp. 117–140.

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 (2017). "First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision".

Gneezy, Uri and Aldo Rustichini (May 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".

Guryan, Jonathan, Kory Kroft, and Matthew Notowidigdo (Oct. 2009). "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments". In: *American Economic Journal: Applied Economics* 1.4, pp. 34–68.

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.

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

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube (2018). "To whom may you compare: Preferences for peers".

Kimbrough, Erik, Andrew McGee, and Hitoshi Shigeoka (2017). "How Do Peers Impact Learning? An Experimental Investigation of Peer-to-Peer Teaching and Ability Tracking". In: *IZA Discussion Paper Series*.

Kuhn, Peter, Peter Kooreman, Adriaan Soetevent, and Arie Kapteyn (Aug. 2011). "The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery". In: *American Economic Review* 101.5, pp. 2226–2247.

Lavy, Victor and Edith Sand (2015). "The Effect of Social Networks on Student's Academic and Non-Cognitive Behavioral Outcomes: Evidence from Conditional Random Assignment of Friends in School".

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

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 (Mar. 2009). "Peers at Work". In: *American Economic Review* 99.1, pp. 112–145.

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.

— (2014). "Experimental and quasi-experimental analysis of peer effects: two steps forward?" In: *Annual Review of Economics* 6.1, pp. 253–272.

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.

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 Randomization check

Table A.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)	0.25	-0.00	-0.25	-0.00	-0.25
	(2.96)	(2.93)	(0.28)	(2.78)	(0.28)
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 on school, grade and gender fixed effects as well as an indicator for the first or second run. We then use the residuals from this regression.

B Description and comparison of peer preferences

In this section, we briefly describe the preferences elicited in the survey and then compare preference over relative performance and based on names. Suppose that all subjects want to be paired with a faster peer. Subsequently, we may not be able to match this most-preferred peer to half of the sample. This implies that we need a sufficient amount of heterogeneity in performance-based preferences to match pairs optimally given their preferences. Figure B.1a presents a histogram of the most-preferred relative performance of a peer. It shows that – although subjects prefer a similar or slightly faster peer on average – preferences are still heterogeneous mitigating the concern that we are unable to provide subjects with peers according to their preferences. Moreover, Appendix C presents a manipulation check of our treatments and shows that we are indeed able to form pairs based on these elicited preferences.

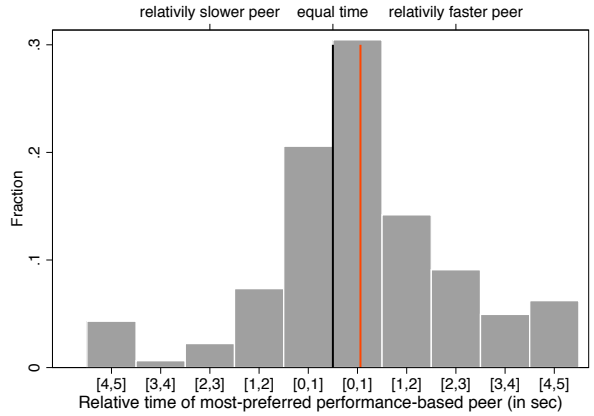
In Figure B.1b, we present the corresponding histogram using the relative times of the most-preferred name-based peers. On average, students choose similar peers of similar ability, but the dispersion of preferences is much larger than for performance-based preferences. Moreover, the most-preferred name-based peer is a friend in 89% of all cases (see Table 2).

In order to show the difference between name- and performance-based preferences, we make use of the elicited beliefs over the relative performance of peers nominated in the name-based preferences. As the elicitation procedure of those beliefs is identical to that of the preferences over relative performance, we can therefore check if subjects want to choose the same kind of peer in terms of relative performance. If only relative performance matters as a criterion for the selection process, subjects should choose a peer, which they believe has the same relative performance as they choose in the performance-based selection process. At least, this difference should be very small.[1] Since subjects beliefs might be noisy, we can repeat this exercise with the actual performance differences in the first run. Figure B.2 shows that although on average subjects choose somebody with a similar performance (based on their belief or actual

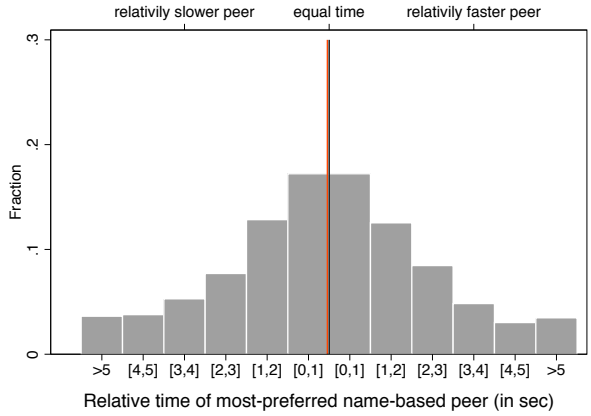
[1] This holds as long as subjects believe that there exists at least one class member with their most-preferred time. Across all three treatments, 67% of all students nominate someone in their name-based preferences whom they belief has the same relative time as their most-preferred performance-based peer. Note that this constitutes a lower bound as we can only check this for the six most-preferred name-based peers (for which we have the beliefs over relative performance) and not for the remaining class members.

Figure B.1: Heterogeneity in preferences

(a) Heterogeneity in perf.-based preferences



(b) Heterogeneity in name-based preferences



The figures present histograms of the most-preferred relative performances of the students in Performance (Panel (a), same as in Figure 2) and the relative time of the most-preferred name-based peers (Panel (b)). The intervals used here and in the survey are one-second intervals of relative times in the first run. Vertical lines indicate own time (black; equals zero by definition) and mean preference (red; 0.56 sec faster for performance-based preferences where we used the mean of each interval to calculate the mean, and 0.05 sec slower for name-based preferences).

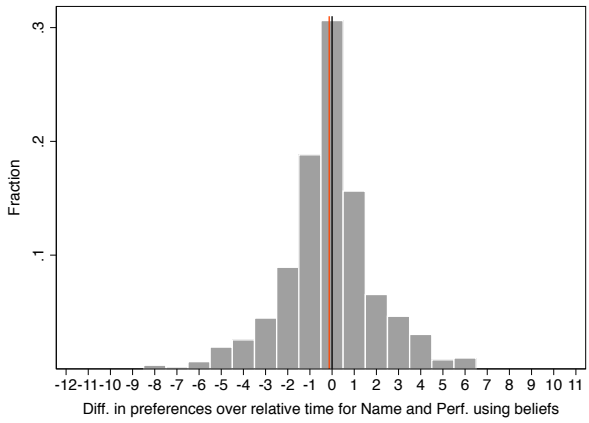
times), there is a lot of variation in those preferences.[2] Therefore, we can conclude that the two sets of preferences are distinct preferences and that not only relative performance matters for the name-based selection process.[3]

[2]The correlation between beliefs over the peer's performance and his or her actual performance is .55, indicating that subjects' beliefs are relatively accurate. The share of subjects with absolute differences less or equal than one second is 65% and 42%, and the mean differences are -.13 and .57 seconds for beliefs and actual times, respectively.

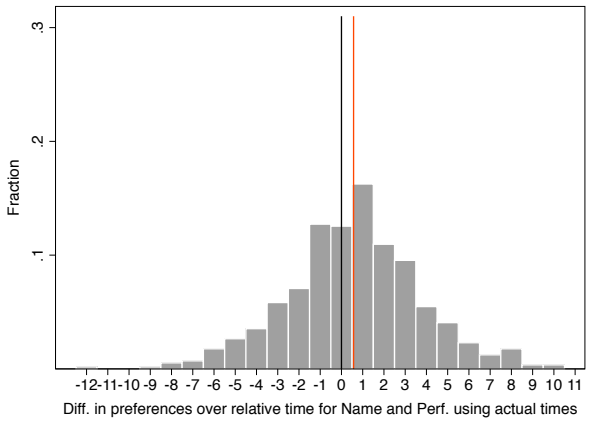
[3]Note that even if the differences were zero, the name-based preferences would be informative as there may be several class members with relative times similar to the performance-based preferences.

Figure B.2: Dissimilarity of preferences

(a) Dissimilarity of preferences using beliefs



(b) Dissimilarity of preferences using times



We plot the difference between the first preference for relative performance and the relative performance of the first preference for name-based preferences. Vertical red lines indicate the mean differences. In panel (a) we use subjects' beliefs over relative performance, while panel (b) uses actual relative times. If subjects choose someone in the same category for name- and performance-based preferences, this difference is zero.

C Manipulation checks

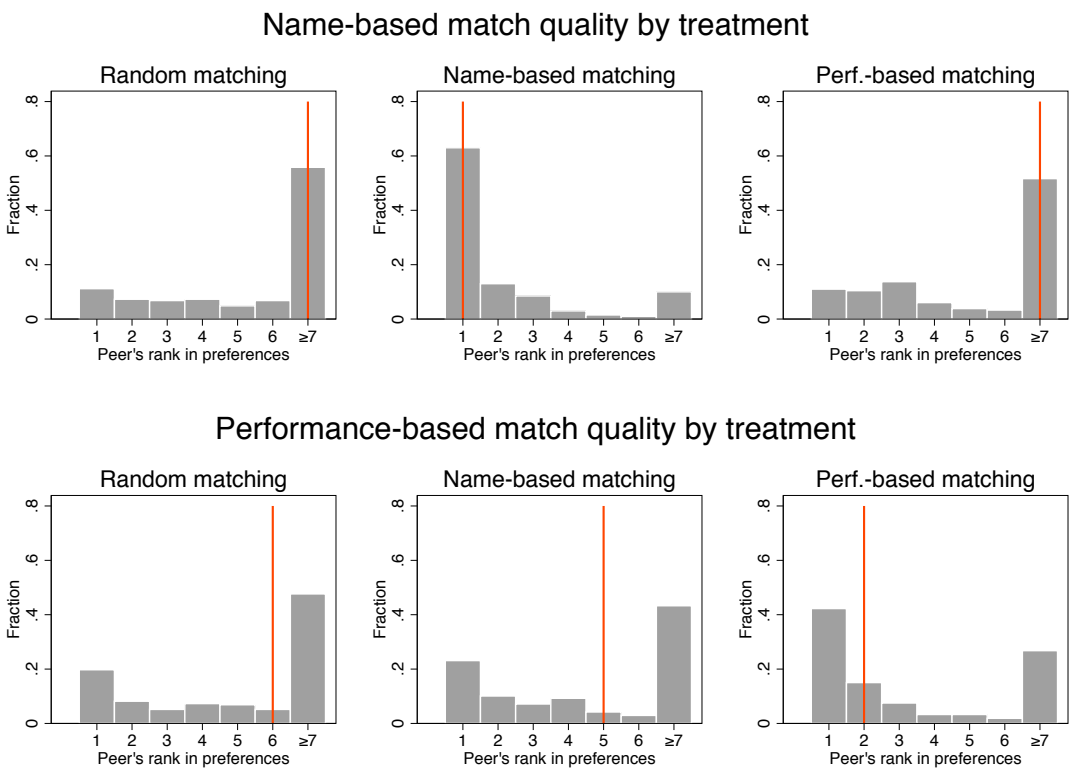
In section 3.1, we presented the resulting match qualities using the preferences as elicited in the survey. However, some subjects may prefer relative times, which are not available to them. For example, the fastest subject in the class might want to run with someone who is even faster, or a student 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 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 C.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.

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 treat-

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

Figure C.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.

ments affect the peer composition in other dimensions apart from the match quality in Table C.1.

Table C.1: Effects of treatments on peer composition

	Match Qual. (name)	Match Qual. (time)	Friendship Ties	Time 1	
Name	0.49***	0.07	0.27***	-0.08	
	(0.06)	(0.04)	(0.06)	(0.19)	
Performance	-0.06	0.24***	-0.12*	-0.70***	
	(0.06)	(0.04)	(0.07)	(0.21)	
N	588	588	294	294	
R^2	.34	.083	.19	.09	
p-value: Name vs. Performance	1.0e-11	.0002	3.4e-07	.0037	
Mean in Random	.23	.3	.43	2.4	
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
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. Performance	.19	.53	.63	.19	.031
Mean in Random	1.2	1	1.1	.98	1.1
	Locus of Control	Social Comparison	Competitiveness	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. Performance	.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 as well as age in regressions with individual outcomes.

D 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:

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

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

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

[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).

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.

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 D.2 as follows:

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

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

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 $\Delta_{R,d}$).

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

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

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 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

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

indirect effects of the mediating variables. The empirical specification of D.6 is given by

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

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}.

Finally, we analyze heterogeneous direct effects of ranks within pairs using equation D.8:

$$\begin{aligned} y_{igs} = \bar{\tau} &+ \bar{\tau}_h^N \mathbb{1}_{\{a_i \geq a_j\}} D_i^N + \bar{\tau}_l^N \mathbb{1}_{\{a_i < a_j\}} D_i^N \\ &+ \bar{\tau}_h^P \mathbb{1}_{\{a_i \geq a_j\}} D_i^P + \bar{\tau}_l^P \mathbb{1}_{\{a_i < a_j\}} D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \end{aligned} \tag{D.8}$$

The indicator $\mathbb{1}_{\{a_i \geq a_j\}}$ denotes if subject i was of higher ability (e.g., faster in the first run) than her or his peer j, and $\mathbb{1}_{\{a_i < a_j\}}$ equals one if i was of lower ability. We interact this rank indicator with the treatment indicators D_i^d ($d \in \{N, P\}$) to analyze whether the direct effect depends on the rank within a pair.

E Control treatment to disentangle peer effects from learning

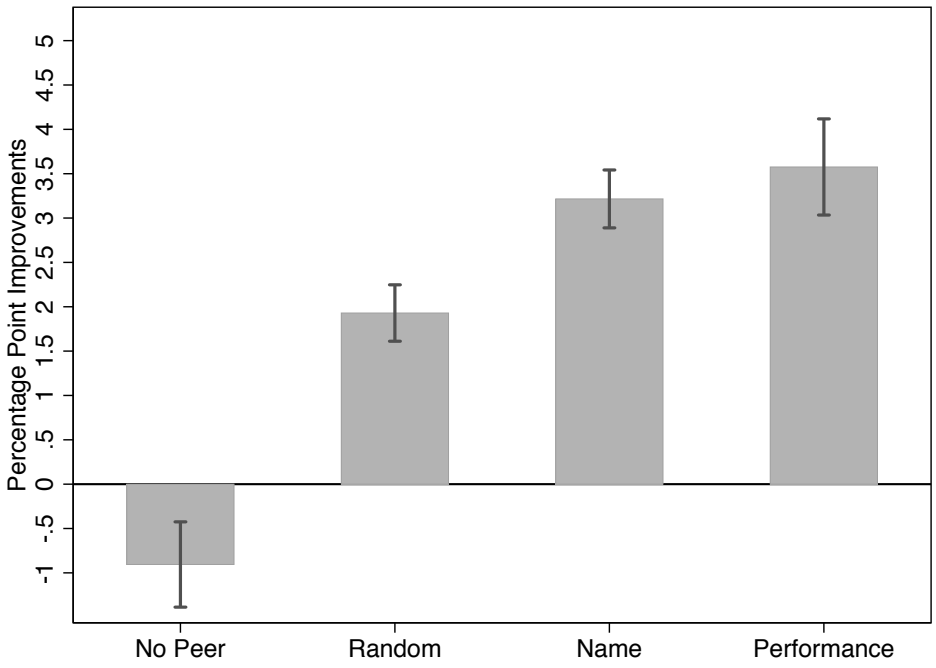
Table E.1 and Figure E.1 present the estimated average treatment effects and the margins including an additional control treatment. The NoPeer treatment featured the same design as all other treatments. The only difference was that students participated in the running task twice without a peer. Moreover, 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.

Table E.1: Robustness checks

	(a) PP. Imprv.	(b) Time (Second Run)	
	(1)	(2)	(3)
Name	1.29***	-0.37***	-0.14***
	(0.42)	(0.11)	(0.04)
Performance	1.65**	-0.40***	-0.15***
	(0.62)	(0.14)	(0.05)
NoPeer	-2.84***	0.82***	0.31***
	(0.61)	(0.16)	(0.06)
Controlling for Time (First Run)	No	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	715	715	715
R^2	.14	.81	.81

This table presents least squares regressions using percentage point improvements (Panel (a)) or times from the second run (Panel (b)) as the dependent variables. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

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 as well as age and cluster standard errors at the class level.

F Robustness checks for average treatment effects

In Table F.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.

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 F.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 F.1: Robustness checks

	Percentage Point Improvements				
	(1) Baseline	(2) BRL	(3) Group means	(4) Win.	(5) Trunc.
NAME	1.26***	1.26**	1.15*	1.05***	0.95***
	(0.43)	(0.50)	(0.58)	(0.37)	(0.35)
PERFORMANCE	1.67**	1.67**	2.12***	1.51***	1.43***
	(0.62)	(0.72)	(0.60)	(0.51)	(0.43)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	588	588	70	588	496
R^2	.056	.056	.33	.072	.087
p-value: NAME vs. PERFORMANCE	.51	.55	.088	.37	.27

This table presents least squares regressions using times (Panel (a)) or percentage point improvements (Panel (b)) 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 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.

Table F.2: Robustness checks – Subsample analyses

	Percentage Point Improvements				
	(1) Baseline	(2) 7th grade	(3) 8th grade	(4) 9th grade	(5) 10th grade
NAME	1.26***	1.95***	2.60***	1.53**	1.08*
	(0.43)	(0.08)	(0.35)	(0.59)	(0.61)
PERFORMANCE	1.67**	2.78***	2.51***	2.53***	1.32
	(0.62)	(0.63)	(0.15)	(0.62)	(0.88)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	588	116	116	174	182
R^2	.056	.073	.064	.16	.039
p-value: NAME vs. PERFORMANCE	.51	.21	.82	.19	.82
	(6) Female	(7) Male	(8) School 1	(9) School 2	(10) School 3
NAME	1.36***	1.44**	2.09***	1.26*	1.21***
	(0.11)	(0.65)	(0.37)	(0.65)	(0.44)
PERFORMANCE	1.53***	2.29***	2.22*	1.68**	1.63*
	(0.05)	(0.55)	(1.12)	(0.77)	(0.85)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	148	274	166	390	198
R^2	.065	.1	.12	.057	.065
p-value: NAME vs. PERFORMANCE	.3	.14	.88	.53	.62

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.

G Peer composition robustness checks

We run three robustness checks for the results presented in Table 4. First, to provide further evidence that it is not the quality of the match itself that drives our results, we estimate the effect of match quality within RANDOM (cf. Table G.1). As subjects in RANDOM are matched with someone they prefer by pure chance, this allows us to estimate the impact of match quality itself. The estimates show that match quality itself has no significant effect on the performance in RANDOM. Second, in Table G.2 we restrict our estimation sample to subjects with a high match quality only to show that the treatment effects persist for these subjects and the coefficients on peer compositional effects do not substantially change. Third, we control for differences in ability and matching quality in a more flexible way in Table G.3 by including interval fixed effects for ability differences and fixed effects for every rank of the preferences. More specifically, we include an indicator for each one-second interval of ability differences between subjects within a pair. Similarly, we include indicators for each rank in the two sets of preferences to check whether the high match quality indicators are restrictive. This allows for a potential non-linear influence of ability differences and match quality on our estimates. Comparing the estimates shows that neither the piecewise-linear functional form of ability differences nor using high match quality indicators is restrictive. Finally, this table shows that the decomposition presented in Table 4 is robust to the inclusion of additional class-level controls.

Table G.1: Effect of match quality within RANDOM

	Percentage Point Improvements			
	(1) Name MQ.	(2) Name MQ. with Controls	(3) Perf MQ.	(4) Perf MQ. with Controls
High Match Qual. (name-based)	0.23	0.71		
	(0.91)	(0.86)		
Faster Student × $\lvert\Delta Time\ 1\rvert$		-0.44**		-0.46*
		(0.19)		(0.23)
Slower Student × $\lvert\Delta Time\ 1\rvert$		0.81*		0.80*
		(0.38)		(0.38)
Slower Student in Pair		-0.04		-0.10
		(1.03)		(1.01)
Peer is friend		-0.41		-0.19
		(0.83)		(0.71)
High Match Qual. (perf.-based)			-0.25	0.03
			(1.09)	(0.88)
Abs. Diff. in Personality	No	Yes	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes
N	205	204	205	204
R^2	.055	.26	.055	.25

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 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 only observations within RANDOM. If we restrict the sample to students in RANDOM, the explanatory power of the match quality (MQ) is not significant.

Table G.2: Only high match quality sample as comparison group

		Only high Matching Quality			
	(1) All	(2) Random&Name	(3) with Controls	(4) Random&Perf.	(5) with Controls
Name	1.26**	1.83***	1.67***		
	(0.47)	(0.55)	(0.47)		
Performance	2.18***			2.38***	1.98***
	(0.68)			(0.71)	(0.63)
High Match Qual. (name-based)	0.56				
	(0.42)				
High Match Qual. (perf.-based)	-0.07				
	(0.45)				
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.35**		-0.63**		-0.33
	(0.14)		(0.27)		(0.46)
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.07***		1.35***		1.25**
	(0.19)		(0.35)		(0.51)
Slower Student in Pair	-0.14		-0.61		-1.71**
	(0.46)		(0.68)		(0.82)
Peer is friend	-0.61		-1.13		-1.44*
	(0.46)		(0.74)		(0.78)
Abs. Diff. in Personality	Yes	No	Yes	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	582	208	207	162	160
R^2	.27	.16	.49	.16	.33

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 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 4 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 name- (columns (3) and (4)) or performance-based preferences (columns (5) and (6)), respectively, our treatment effects persist and the coefficients on peer compositional effects do not change much.

Table G.3: Robustness Check

	(1) Linear	(2) Time Int. FE	(3) Match Qual. FE	(4) Class Controls
Direct Effects				
NAME	1.26**	1.22**	1.06**	1.44***
	(0.47)	(0.49)	(0.51)	(0.44)
PERFORMANCE	2.18***	2.23***	2.33***	1.70**
	(0.68)	(0.74)	(0.71)	(0.70)
Peer Characteristics				
High Match Qual. (name-based)	0.56	0.57		0.77
	(0.42)	(0.44)		(0.46)
High Match Qual. (perf.-based)	-0.07	-0.06		-0.40
	(0.45)	(0.40)		(0.52)
Faster Student × $\lvert\Delta Time\ 1\rvert$	-0.35**		-0.29**	-0.36**
	(0.14)		(0.14)	(0.15)
Slower Student × $\lvert\Delta Time\ 1\rvert$	1.07***		1.09***	0.87***
	(0.19)		(0.21)	(0.18)
Slower Student in Pair	-0.14		-0.07	0.20
	(0.46)		(0.43)	(0.47)
Peer is friend	-0.61	-0.60	-1.06**	-0.44
	(0.46)	(0.47)	(0.51)	(0.49)
Time Diff. FEs	No	Yes	No	No
Match Qual. FEs	No	No	Yes	No
Class-level Controls	No	No	No	Yes
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes
N	582	582	582	512
R^2	.27	.29	.3	.27
p-value: NAME vs. PERFORMANCE	.19	.17	.089	.72

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 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 4 for reference. Column (2) includes fixed effects for every one-second difference in ability levels of the two students. Column (3) includes an indicator for each rank within the two sets of preference rankings. Finally, column (4) includes additional class-level controls.

H Additional material for implications

Table H.1 shows the regressions underlying Figure 6. In particular, in column (1) we estimate equation 1 but interact treatment indicators with ability terciles (low, medium, or high). Ability terciles are defined according to tercile splits of times in the first run within each school, grade and gender. Column (2) repeats the exercise using quintiles rather than terciles to show that the pattern holds for finer splits.

As argued in section 5.5, low-ability students in NAME need to prefer to be and subsequently are matched with faster students on average. We present the shares of students in NAME who prefer a faster student (based on their name-based preferences) and who are matched to a faster student for the three ability terciles defined above in Table H.2. Indeed, low-ability students in NAME are more likely to prefer a faster peer and on average are matched to faster peers than students of higher ability.

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., 2017), 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.3, where 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

[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.

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.

Table H.1: Heterogeneous treatment effects by own ability

	Percentage Point Improvements	
	(1) Ability Terciles	(2) Ability Quintiles
Low Ability	3.21***	4.49***
	(1.02)	(1.52)
Medium-Low Ability		0.43
		(1.31)
Medium Ability	-0.59	-0.45
	(1.18)	(1.41)
Medium-High Ability		-0.53
		(0.98)
High Ability	-1.19	-1.54
	(0.88)	(1.00)
Name × Low Ability	1.02	1.27
	(0.92)	(1.53)
Name × Medium-Low Ability		1.47
		(1.11)
Name × Medium Ability	2.00*	1.65
	(1.00)	(1.21)
Name × Medium-High Ability		1.28
		(0.77)
Name × High Ability	1.01**	0.90*
	(0.48)	(0.53)
Performance × Low Ability	-0.20	-0.65
	(1.18)	(1.97)
Performance × Medium-Low Ability		1.77
		(1.23)
Performance × Medium Ability	2.57**	1.94
	(1.03)	(1.25)
Performance × Medium-High Ability		2.25***
		(0.67)
Performance × High Ability	2.16***	2.15***
	(0.49)	(0.63)
Gender-Grade/School FEs, Age	Yes	Yes
N	588	588
R^2	.39	.41

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) assigns one of three ability levels – low, medium or high – according to tercile splits of times in the first run within each school, grade and gender and presents the underlying regression for Figure 6. Column (2) uses quintiles rather than terciles to show that the pattern is robust to other definitions of ability quantiles.

Table H.2: Share of students preferring and receiving a faster peer in NAME

	Ability Tercile		
	Low	Medium	High
Preferred name-based peer is faster	0.75	0.60	0.25
Realized name-based peer is faster	0.75	0.58	0.21

This table presents the share of students preferring a faster peer in NAME and the realized share. Ability terciles – low, medium or high – are assigned according to tercile splits of times in the first run within each school, grade and gender.

Table H.3: Absolute change in percentile scores

	Absolute Change in Percentile Scores	
	within matching group	within treatment
NAME	-0.01	-0.02
	(0.01)	(0.01)
PERFORMANCE	-0.04**	-0.04***
	(0.02)	(0.01)
Gender/Grade/School FEs, Age	Yes	Yes
N	588	588
R^2	.056	.051
p-value: NAME vs. PERFORMANCE	.018	.085
Mean in RANDOM	.15	.14

This table presents least squares regressions using absolute change in percentile scores 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).

I Simulation of matching rules

We simulate three matching rules and predict their impact on performance improvements using our estimates from Table 4. 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 (6) of Table 4 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 2, where sim and obs denote simulated and observed pair characteristics, respectively. Furthermore, we assume that the direct effect of the simulated policies equals the one 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.

We simulate the following three peer assignment rules. First, 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. Second, 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. Third, 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. Table I.1 summarizes the distance in ability of the experimental treatments as well as the simulated assignment rules.

Table I.1: Overview of simulated peer assignment rules

Peer Assignment Rule	Simulated?	Mean Ability Distance (in sec)	Description
Name	No	2.09	Self-selected peers based on names
Performance	No	1.41	Self-selected peers based on relative performance
Random	No	2.42	Randomly assigned peers
Equidistance	Yes	3.11	Same distance in ranks across pairs
High-to-Low	Yes	3.11	First to last, second to second to last etc.
Tracking	Yes	0.90	First to second, third to fourth etc.