Essays in Applied Microeconomics

Inauguraldissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch

die Rechts- und Staatswissenschaftliche Fakultät der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Lukas Benedikt Robert Kießling

aus Haan (Rheinland)

Bonn, 2020

Dekan:	Prof. Dr. Jürgen von Hagen
Erstreferent:	Prof. Dr. Lorenz Goette
Zweitreferent:	Prof. Dr. Pia Pinger
Tag der mündlichen Prüfung:	10. Januar 2020

Acknowledgments

This thesis would not have been possible without the support of many people. First of all, I would like to thank my advisors Lorenz Goette and Pia Pinger for their guidance and support throughout the last years. I benefited a lot from being exposed to their different approaches to science, the confidence they put in me early on, and our insightful discussions that shaped my thinking. I would also like to thank Thomas Dohmen for his invaluable suggestions and for being part of my dissertation committee.

Moreover, I am grateful for countless hours I spent in school gyms together with my coauthors and friends Jonas Radbruch and Sebastian Schaube, for our endless discussions about both big questions and all the little details of our research projects, and for the fun we had. Special thanks also goes to my fellow students during the PhD, who made the journey joyful, and in particular to Louis Strang and Christian Zimpelmann.

I benefited tremendously from the great environment at the Bonn Graduate School of Economics, the support from the Institute for Applied Microeconomics, and the Collaborative Research Center Transregio 224. In addition, I want to thank Britta Altenburg, Simone Jost, Silke Kinzig, Vanessa Pollari, and Andrea Reykers, who helped me on various administrative matters over the years. I would also like to thank Matthias Sutter for giving me the opportunity to join his research group at the Max Planck Institute for the last year of my PhD, and the fantastic group of colleagues and friends I found there: Zvonimir Bašić, Stefania Bortolotti, Zita Green, Zwetelina Iliewa, Sofia Monteiro, Matthias Praxmarer, Shambhavi Priyam, Angelo Romano, Daniel Salicath, Ali Seyhun Saral, Stefan Schmidt, Sebastian Schneider, and Sebastian Tonke.

Finally, I want to thank my family – Katja, Michael, Linda, and Lea – for all their lifelong support, their encouragement, and for trusting me to make the right choices, and Viola for enriching my life and her love.

Contents

List of Figures

List of Tables

Introduction

What determines human behavior? This seemingly simple question is at the core of economics and the social sciences more generally. Yet, it is one of the most difficult questions to answer for at least two reasons: First, the presence and choices of others influence our own behavior and requires a systematic study of social aspects. Second, every decision relies on our subjective beliefs about unobserved states of the world, and these beliefs may reflect heterogeneous expectations about the consequences of different choices. By investigating the role of peers for behavior and by studying the systematic variation in beliefs, this thesis contributes to our understanding of human behavior and decision-making.

It is widely accepted that peers influence consumption behavior, general well-being, and performance. Yet, we do not know much about how individuals choose these peers in the first place and about the consequences of peer self-selection. The first two chapters, Chapters 1 and 2, which are joint work with Jonas Radbruch and Sebastian Schaube, therefore aim at filling this gap. We first examine how allowing individuals to choose with whom to interact affects their performance relatively to exogenously assigned peers (Chapter 1), and second, whom they actually choose as peers (Chapter 2). In order to study this self-selection of peers, we conduct a field experiment in secondary schools and allow students in two treatment arms to select their peers themselves, while in another they are randomly assigned to a peer.

Chapter 1: "Self-selection of Peers and Performance" analyzes the consequences of peer self-selection on performance. It documents that individuals, who can self-select their peers, improve their performance more than those with randomly assigned peers. In principle, these differences in performance may stem from two sources: First, individuals interact with different peers, who influence performance. Thus, accounting for differences in the peer composition may explain our findings. Second, the results could stem from a psychological effect of being able to self-select peers rather than having them assigned. Although a peer's characteristics such as his or her performance explain part of the variation in one's own performance, these peer effects cannot explain the treatment effects. Rather, our data indicate a positive effect on performance when having autonomy over peer assignments. Furthermore, the presence of peer effects in multiple dimensions has implications for the design of reassignment policies such as tracking regimes, which are based on, e.g., measures of students' ability. If policy-makers want to reassign students into classrooms or workers into teams based on peer effects in a single dimension only, they neglect the fact that reassigning rules simultaneously change other peer characteristics, giving rise to peer effects apart from the targeted dimension. These effects can counterbalance each other leading to ambiguous net effects.

Chapter 2: "Determinants of Peer Selection" studies whom individuals choose as peers and links these choices to three potential determinants. Particularly, the chapter assesses the extent to which the selection of peers depends on (i) the relative performance of peers, (ii) personality differences, and (iii) the presence of friendship ties. By quantifying the relative contributions of performance and social aspects for peer choices, we find that friendship is the most important determinant, but individuals exhibit sizable homophily both in past performance and personality. These results help to explain why previous studies often find that different groups exert peer effects of different sizes. In particular, we suggest that selective peer choices may give rise to individual-specific peer groups that result in differential peer effects.

Chapters 1 and 2 document peer effects in performance and that individuals prefer peers who are similar to themselves. Although such peer choices and peer effects yield a correlated outcomes among peers, decisions often remain highly heterogeneous in general. One potential explanation for such heterogeneities are differences in beliefs that individuals hold. The aim of Chapters 3 and 4 is to systematically study the heterogeneity of subjective beliefs and expectations in two specific contexts. First, I investigate parents' beliefs about the returns to different parenting styles and neighborhoods (Chapter 3). Second, in joint work with Pia Pinger, Philipp Seegers, and Jan Bergerhoff, I characterize gender differences in students' wage expectations and discuss potential drivers thereof (Chapter 4).

Parents are crucial for the development and success of children as they grow up. However, not much is known about how parents decide how to raise their children, and how parenting decisions depend on the environment in which a family lives. In order to answer these questions, **Chapter 3: "Understanding Parental Decision-making: Beliefs about Returns to Parenting Styles and Neighborhoods"** studies parents' beliefs about the returns to two factors affecting the development and long-term outcomes of children: (i) parenting styles defined by the extent of warmth and control parents employ in raising their children, and (ii) neighborhood quality. Based on a representative sample of over 2,000 parents in the United States, I show that parents hold well-formed beliefs: they expect large returns to the warmth dimension of parenting as well as to living in a good neighborhood. Regarding the relation of both factors, I find that parents perceive parenting as being able to compensate partly for adverse environments. Moreover, mothers expect larger returns than fathers, but there is no socioeconomic gradient in perceived returns. Parents' perceived returns are relevant for their actual decision-making in so far that they are predictive for actual parenting behavior. Hence, my results highlight that parental beliefs are an important determinant of parental decision-making, but cannot explain socioeconomic differences in parenting.

Chapter 4: "Gender Differences in Wage Expectations: Sorting, Children, and Negotiation Styles" presents evidence from a large-scale study on gender differences in wage expectations based on a sample of over 15,000 students in Germany. Studying such wage expectations before labor market entry is important as they may determine further educational or labor market choices, affect within-household bargaining or negotiations with prospective employers, and may have consequences for financial decision-making, e.g., in terms of an optimal choice of retirement and savings plans. We document a large gender gap in expected wages that amounts to approximately 500,000 EUR over the life-cycle and resembles actual wage differences. In order to understand the underlying causes and determinants, we relate these expected wages to (i) differential sorting into majors, industries, and occupations, (ii) differences in child-rearing plans, and (iii) male-female differences in negotiation styles. We show that males and females sort themselves into different majors, industries, and occupations, and follow different negotiation strategies. While child-rearing plans are comparable across genders, females expect child-penalties for giving birth to children before the age of 30.

In summary, this thesis focuses on human behavior and decision-making by investigating the role of peer influences and subjective expectations. Using data from field experiments and large-scale surveys, the four chapters provide a starting point for further analyses of social aspects and heterogeneous beliefs.

Chapter 1

Self-selection of Peers and Performance

Joint with Jonas Radbruch and Sebastian Schaube

1.1 Introduction

"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

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

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

In order to study the effects of self-selection, we conducted a framed field experiment (Harrison and List, 2004) with over 600 students (aged 12 to 16) in physical education classes of German secondary schools. Students took part in two running tasks (suicide runs) – first alone, then with a peer – and filled out a survey in between that elicited preferences for peers, personal characteristics, and the social network within each class. Our treatments exogenously varied the peer assignment in the second run using three different peer assignment rules. We implemented a random matching of pairs (Random) as well as two matching rules that used elicited preferences to implement two notions of self-selection: first, the classroom environment enabled students to state preferences for known peers (*name-based preferences*); and second, using a running task yielded direct measures of performance and thus could be used to select peers based on their relative performance in the first run (*performance-based preferences*). Using these two sets of preferences, we implemented two treatments with self-selection of peers by matching students based on either their name-based preferences (Name) or preferences over relative performance (Performance).

We find that self-selection of peers leads to an average performance improvement of 14–15 percent of a standard deviation relative to randomly assigned peers. While students in Random also improve their performance from the first to the second run, the improvements with self-selected peers almost double. Self-selection changes the peer composition, e.g., students predominantly interact with friends in Name, but tend to choose others with a similar past performance in Performance. Based on this finding, we decompose the overall treatment effect into an indirect effect that is due to the peer's altered characteristics and a direct effect of being able to self-select a peer. Although we observe substantial peer effects in multiple dimensions (e.g., in relative performance in the first run), a peer's characteristics do not explain treatment differences resulting in an indirect effect close to zero. Instead, our estimates provide evidence that there is a direct effect of peer self-selection on performance. Therefore, the process of self-selection itself increases the performance of students. Borrowing from self-determination theory (Deci and Ryan, 1985; Deci and Ryan, 2000), we interpret this direct effect as a positive effect of having autonomy: being able to self-select peers has a psychological effect that enhances intrinsic motivation and improves subsequent performance. Finally, we simulate other exogenous peer assignment rules that seek to maximize or minimize the productivity differences between students. We document that these alternative rules yield performance improvements close to those observed with randomly assigned peers and therefore lower than those with peer self-selection. These findings thus support our interpretation that self-selection of peers carries a intrinsic value beyond changes in the peer composition.

Our results have three main contributions to the literature on peer effects, social interactions, and autonomy. First, we show that self-selection changes with whom people interact and thereby affects the overall composition of the reference or peer group. Second, we present evidence that self-selection of peers affects behavioral outcomes and has a direct effect on productivity. This highlights a novel channel through which peers and their selection affect behavior and provides the first clean evidence on autonomy in a field setting. Third, we document that peer effects may be present in multiple dimensions and discuss how this limits the effects of exogenous reassignment rules.

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

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

2. In a companion paper, Kiessling, Radbruch, and Schaube (2019), we study the peer selection process in more depth and relate the selection of peers to individual-level determinants.

Moreover, our findings help to reconcile mixed evidence on the effectiveness of interventions changing class or work-group compositions to exploit peer effects (e.g., Booij, Leuven, and Oosterbeek, 2017; Carrell, Sacerdote, and West, 2013; Duflo, Dupas, and Kremer, 2011; Garlick, 2018). In our setup, the combination of two effects – the change in the peer composition and the multidimensionality of peer effects – has only a small impact on aggregate performance. More specifically, we move beyond peer effects in a single dimension and allow several characteristics such as productivity, friendship ties, and personality measures to exert peer effects.[3] Our results show that there are sizable peer effects apart from productivity. Consequently, if policy-makers reassign peers based on peer effects in a single dimension only, they neglect the fact that reassigning rules simultaneously change other peer characteristics giving rise to peer effects apart from the targeted dimension. These effects can counterbalance each other and lead to a net effect that is in our case close to zero and in general ambiguous. Hence, studies analyzing peer interactions and reassignment policies need to take into account not only a potential direct effect of self-selection, but also the multidimensionality of peer effects.

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

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

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

While the quantitative impact of different assignment mechanisms and the resulting peer composition might be specific to our setting and sample, students are a highly relevant subject group. They have not only been analyzed to study phenomena such as favoritism (Belot and Ven, 2011, and references therein), but peers during high school also have long-lasting effects on an individual's skill formation (Agostinelli, 2018) and hence on subsequent educational attainment. Moreover, 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 from effort or skill complementarities (e.g., Bandiera, Barankay, and Rasul, 2010; Mas and Moretti, 2009), or setting in which learning from peers is important (e.g., Bursztyn, Ederer, Ferman, and Yuchtman, 2014; Jackson and Bruegmann, 2009). The settings across these studies differ enormously, as does the underlying mechanism. Nonetheless, all of these share the notion that the behavior or action of peers imposes an externality on the action or behavior of others. In addition, peers can in principal also be self-selected affecting subsequent peer interactions.

The remainder of the paper is structured as follows. The next section presents our experimental design as well as procedural details. Section 1.3 presents the data and describes our sample of students. We outline our empirical framework in Section 1.4. In Section 1.5, we analyze how self-selected peers affect performance relative to randomly assigned peers and decompose this effect in a direct effect of self-selection and an indirect effect as a result of changes in the peer composition. We then interpret the direct effect and highlight potential policy implications. Finally, section 1.6 concludes.

1.2 Experimental Design

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

In this study, we used the controlled environment of a framed field experiment to overcome these challenges. We embedded our experiment in physical education classes of German secondary schools. Students from grades 7 to 10 participated in a running task, first alone and then simultaneously with a peer. Running allowed students to compare their performance with either faster or slower students, while it excluded complementaries in production between the students. Moreover, we 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 cleanly identify his or her impact. Subjects singled out specific peers by either naming them directly (in the treatment NAME) or selecting performance intervals (in PERFORMANCE). The respective treatments used these preferences to form pairs with self-selected peers or pairs were formed at random. Hence, we can compare the effect of self-selected peers with exogenously assigned ones, and can evaluate the effects of each assignment mechanism.

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

1.2.1 Experimental Design

Figure 1.1 illustrates the experimental design. Students participated in a running task commonly known as "suicide runs", a series of short sprints to different lines of a volleyball court.[5,6] The first run – in which students ran alone – served two purposes: first, recorded times can be used as a measure of productivity and to evaluate the time improvement between the two runs; and second, we used (relative) times from the first run in combination with students' preferences to create pairs for the second run in one of the treatments described below. The second run mirrored the first one aside from the fact that students did not run alone, but rather in pairs. This means that two students performed the task simultaneously, while their times were recorded individually. Feedback about performance in both runs was only provided at the end of the experiment.

Between the two runs, students filled out a survey comprising three parts, eliciting preferences for peers, non-cognitive skills and information about the social network within each class. We elicited two kinds of preferences: first, we asked sub-

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

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

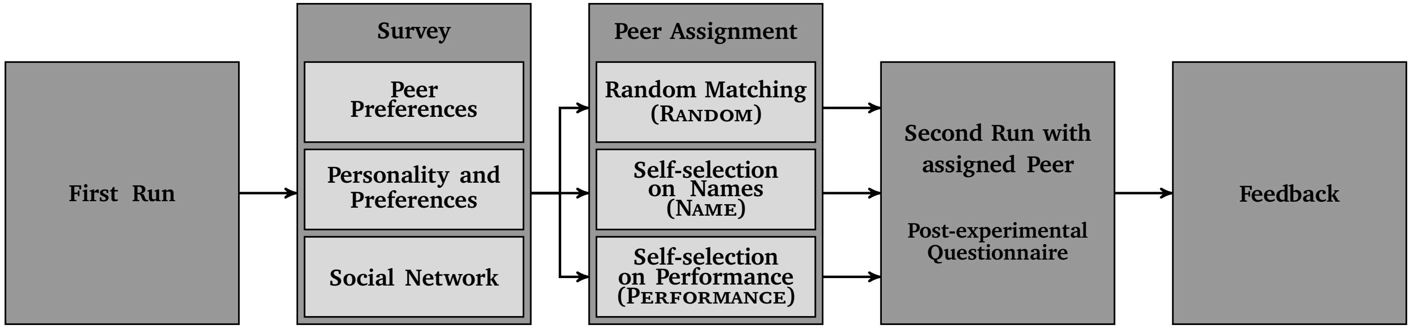


Figure 1.1. Experimental design

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

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

1.2.2 Preference Elicitation

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

Second, we elicited preferences solely based on the relative performance in the first run, ignoring the identities of the potential running partners (*performance-based preferences*). For this purpose, we presented subjects with ten categories comprising

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

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

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

1.2.3 Treatments

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

The second matching rule used the elicited name-based preferences (Name) and the third rule formed pairs based on the elicited performance-based preferences (Performance). Note that the problem of matching pairs constitutes a typical roommate problem. We thus implemented a "stable roommate" algorithm proposed by Irving (1985) to form stable pairs using the elicited preferences.[10]

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

In addition, we conducted an additional control treatment (NoPeer) in which students ran alone twice and which featured a shortened survey but was otherwise

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

10. Given the mechanism proposed by Irving (1985), it is a (weakly) dominant strategy for all participants to reveal their true preferences. The matching algorithm requires a full ranking of all potential peers to implement a matching. Since we only elicited a partial ranking, we randomly filled the preferences for each student to generate a full ranking. However, in most cases subjects were assigned a peer according to one of their first three preferences. Nonetheless, if groups were small, it could be the case that subjects were not assigned one of their most-preferred peers. This is especially the case for performance-based preferences. See also the discussion in Section 1.3.1 below.

identical to the other treatments.[11] As the focus of this paper is the differential size of peer effects and not their existence per se, this only serves the purpose of excluding learning as a source of time improvements between the two runs. Hence, we exclude it from the main analysis and focus only on the evaluation of different peer assignment rules.

1.2.4 Procedures

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

The experiment started with a short explanation of the following lesson and a demonstration of the experimental task. A translation of this explanation as well as screenshots detailing the preference elicitation are presented in Appendix 1.I.

We informed students that their teacher would receive each student's times from both runs, but no information about the pairings during the second run.[13] The students themselves did not receive any information on their performance until the completion of the experiment.

Additionally, we stressed that both of their performances would be graded by their teacher – thus incentivizing both runs – and that the objective was to run as fast as possible in both runs.[14] Moreover, most students themselves were very interested in their own times. The introduction concluded with a short warm-up period. After this, the subjects were led to a location outside of the gym.

Students entered the gym individually, which ruled out any potential audience effects from classmates being present by design. Students completed the first suicide run and subsequently were handed a laptop to answer the survey. Answering the

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

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

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

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

survey took place in a separate room.[15] After the completion of the survey, subjects returned the laptop to the experimenter and waited with the other students outside the gym. Upon completion of the survey by all students, they returned to the gym to receive further instructions for the second run. In particular, we reminded the students of the existence of the three matching rules, and announced which randomly assigned rule was implemented in their class as well as the resulting pairs from the matching process. Following these instructions, the entire group waited outside the gym again. Pairs were called into the gym and both students participated in the second run simultaneously on neighboring tracks.

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

1.3 Data Description and Manipulation Check

We present summary statistics of the students in our sample in Table 1.1.[17] In total, 39 classes with an average class size of about 25 students participated in the experiment. On average, 73% of students within each class subsequently took part in the experiment.[18] This amounts to 627 students who participated in the treatments, with 66% being female.[19] Due to odd numbers of students within some matching groups, we randomly dropped one student in those groups to match students in pairs. Therefore, some students participated in the experiment but were only recorded once and are dropped for estimating the treatment effects in the next section. This procedure yields an estimation sample of 588 observations.

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

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

17. We focus on the students in the three main treatments, namely Random, Name and Performance and do not include the students from the NoPeer treatment, which is discussed in Appendix 1.D.

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

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

Table 1.1. Summary statistics

	7th grade	8th grade	9th grade	10th grade	Total
Sociodemographic 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

Notes: 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.

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

We randomized classes into treatment and check whether observable characteristics differ between our treatments in Appendix Table 1.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 results from the randomization of classes into treatments and can be explained entirely by variation in observables. Conditional on gender-specific grade fixed effects, school fixed effects and age, these differences disappear.

1.3.1 Preferences for Peers and Manipulation Check

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

Table 1.2. Share of name-based preferences being friends

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

Notes: 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 1.2 and Figure 1.2, respectively. Two findings emerge: first, most students nominated friends as their most-preferred peer; and second, while students on average preferred to run with a slightly faster peer, there is a strong heterogeneity in this preference. We analyze these preferences in further detail in Kiessling, Radbruch, and Schaube (2019).

Figure 1.3 shows the realized match quality for all three treatments with respect to the ranking of peers in the two sets of elicited preferences. The upper panel shows the realized match quality according to name-based preferences. We observe that some people were randomly matched to someone with whom they would liked to be paired in Random and Performance. As expected, this share is rather low. While the median peer in Name corresponds to the most-preferred peer according to the elicited name-based preferences, the median peer is not part of the elicited

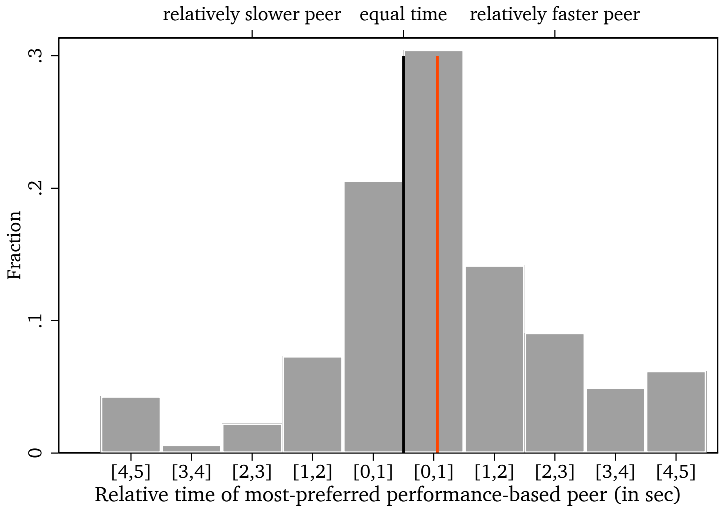


Figure 1.2. Most-preferred performance-based peer

Notes: 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).

preferences (i.e., not among the six most-preferred peers) for Random and Performance. A similar, albeit less pronounced picture arises when analyzing the match quality according to the preferences over relative performance as presented in the lower panel of Figure 1.3. We observe that students in Performance were paired with more preferred peers according to their preferences relative to the other two treatments. However, subjects might have preferred other students or relative times that were not available to them, which mechanically affects the match quality. In Appendix 1.A, we check that once we take the mechanical effect into account, the median match quality in Performance corresponds to the second most-preferred peer, i.e., we obtain a similarly pronounced pattern as in Name.

1.4 Empirical Strategy

This section outlines our empirical framework. For this purpose, we first analyze the effect of being assigned to a particular peer assignment mechanism. In a second step, we decompose this change in performance into two effects: an indirect effect stemming from a change in the peer composition and a direct effect due to self-selection. Appendix 1.B derives these estimation equations from an economic model similar to a mediation analysis described in Heckman and Pinto (2015).

The random assignment of classes into treatments allows us to estimate the average effect of peer selection on performance. Let $D^d = 1$ with $d \in \{N, P\}$ denote

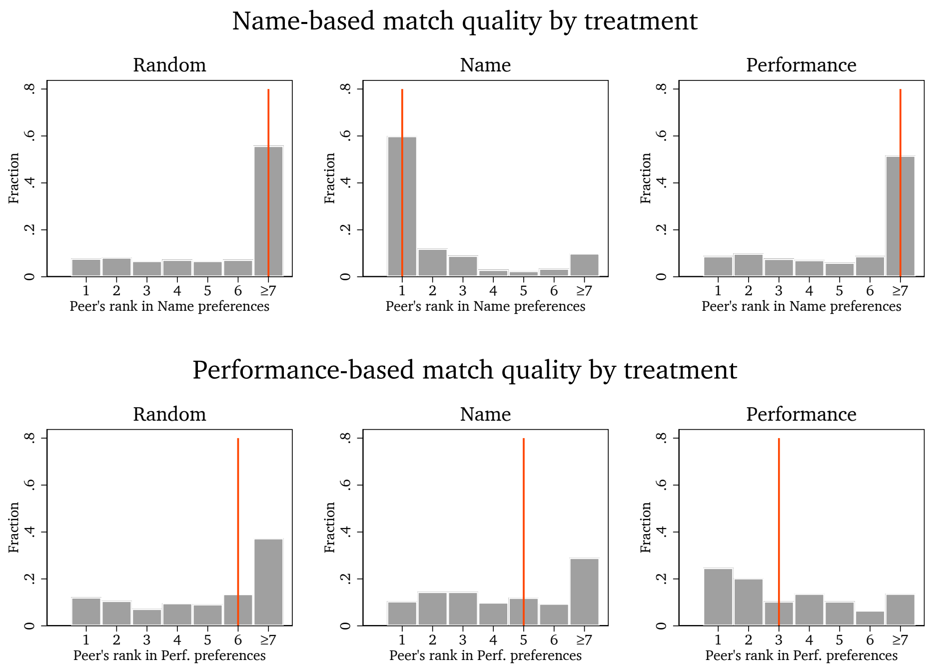


Figure 1.3. Match quality across treatments

Notes: 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.

treatment assignment to Name and Performance, respectively, and zero otherwise. We focus on percentage point improvements from the first to the second run, y_{igs}, of individual i in gender-specific grade g of school s as an outcome. Our baseline specification is then given by:

$$y_{igs} \quad = \quad \tau + \tau^N D_i^N + \tau^P D_i^P + \gamma X_i + \rho_s + \lambda_g + u_{igs} \qquad (1.1)$$

The main parameters of interest are τ^N and τ^P, the effect of being assigned to one of our treatments relative to Random. School fixed effects, ρ_s, and gender-specific grade fixed effects, λ_g, control for variation due to different schools (i.e., as a result of different locations and timing of the experiment) and variation specific to gender and grades.[20] Finally, X_i is a vector of predetermined characteristics such as 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

20. See Section 1.3 for a discussion concerning why we include gender-specific grade fixed effects rather than gender and grade fixed effects separately.

second, self-selection may change the peer composition and therefore the difference between the student's and his or her peer's characteristics. To understand the source of the average treatment effect, we decompose it into a direct effect of self-selection as well as a pure peer composition effect.[21] This takes into account the change in relative peer characteristics across treatments. We implement this decomposition using the following specification:

$$y_{igs} \; = \; \bar{\tau} + \underbrace{\bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P}_{\substack{\text{Treatments} \\ \text{(direct effects)}}} + \underbrace{\beta\theta_i(D^N, D^P)}_{\text{Peer characteristics}} + \underbrace{\gamma X_i + \rho_s + \lambda_g}_{\text{Ind. characteristics and FE}} + u_{igs} \qquad (1.2)$$

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

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

21. The direct effect mainly captures changes in performance due to being able to self-select a peer, which we interpret as an increase in autonomy (see Section 1.5.5 for a discussion of the psychological underpinnings). We acknowledge that our definition of a direct effect also captures inputs that (i) differ across treatments, and (ii) are not measured in our rich set of potential mediators (match quality, friendship ties, productivity differences, ranks and personality differences). However, we show in robustness checks that in our setting this is of minor concern only.

22. We define two indicators to measure whether the assigned peer is nominated among the first three peers for name-based preferences or falls into the three highest ranked categories for performance-based preferences. Alternative specifications are shown in Appendix 1.E.

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

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

we also include the absolute difference in these personality measures to capture potential non-linear effects.

1.5 Results

Our experimental design allows to study the causal effect of different peer assignment mechanisms on individual performance. More specifically, we compare three treatments corresponding to random matching (RANDOM), matching with self-selected peers based on name-based peer preferences (NAME) and preferences over relative performance (PERFORMANCE). As outlined in section 1.2, the random assignment of peers constitutes a natural starting point for at least two reasons: first, the pure presence of any peer might already improve performance; and second, randomly assigned peers are used to document peer effects in a wide range of settings. We contrast this baseline condition with two treatments that assign peers based on elicited preferences, i.e., in which each subject endogenously chooses her peer.

Our empirical results start by documenting average treatment effects. As introduced in Section 1.4, the average treatment effect can stem from two possible sources: if the (relative) characteristics of the peer affect performance and the treatments additionally induce a change in these characteristics, the altered peer composition might explain performance differences across treatments. Moreover, the ability to self-select a peer may directly influence the students' willingness to perform. Before we decompose each treatment effect into a *direct effect* of self-selection and an *indirect effect* due to changes in the peer composition, we establish two necessary conditions for the indirect effect to matter. First, we show that relative peer characteristics matter for individual outcomes. Second, we document that our treatments – which allow for self-selection – indeed change the relative characteristics of peers in the second run. We then decompose the average treatment effects into the two aforementioned channels. Our results conclude with an interpretation of the direct effect and a discussion of implications for peer assignment rules.

1.5.1 Average Effect of Self-selection on Performance

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

Figure 1.4 presents our first result. Subjects in RANDOM improve on average by 1.93 percentage points when paired with a random peer in the second run. However, their performance improves even more in NAME and PERFORMANCE by 3.22 and 3.58 percentage points, respectively. We present the corresponding estimates in Table 1.3.

Table 1.3. 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. PERF.	.51	.62	.38	.8	.98	.28	.8

Notes: This table presents least squares regressions according to equation (1.1) using percentage point improvements in columns (1) to (3) and times of the second run controlling for times in the first run in columns (4) to (7) as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big 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.

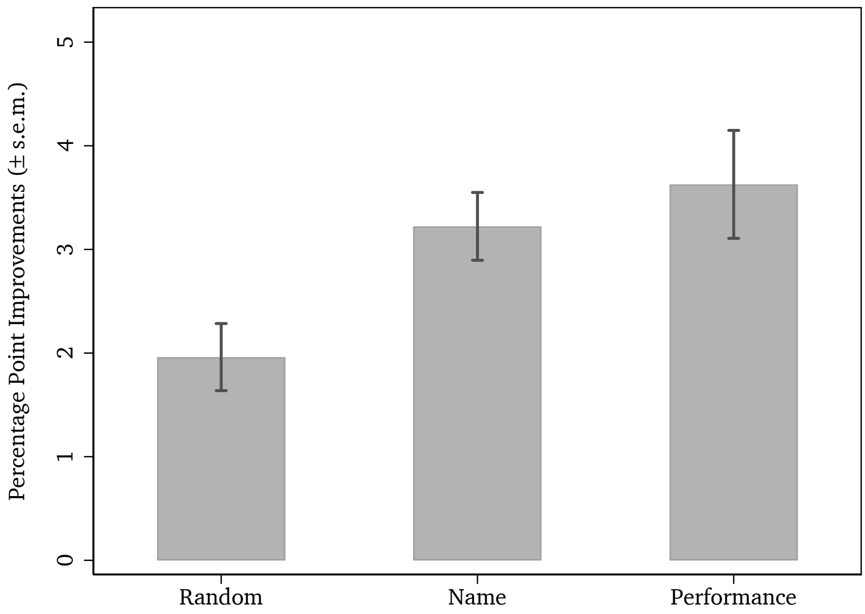


Figure 1.4. Average performance improvements

Notes: 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 1.3. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level.

Columns (1)-(3) present the estimated percentage point improvements in time according to equation (1.1). Columns (4)-(6) additionally express the results in terms of times in the second run – and standardized times in column (7) – 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 self-selected peers based on relative performance is 1.67 percentage points and thus somewhat larger, although it does not significantly differ from Name (p-value= 0.51). These effects persist when controlling for students' own personal characteristics (column (2)) as well as if we additionally control for class-level variables capturing the atmosphere within a class (column (3)). Interestingly, the average treatment effects are about the same size as the improvement in Random. On average, students are faster in the second run and this effect is nearly twice as large in Performance and Name compared to Random. Our baseline effects correspond to additional time improvements of .38 to

.41 seconds (cf. columns (4)-(6)) and account for 14% of a standard deviation in Name and 15% in Performance (cf. column (7)).[25,26]

1.5.2 Peer Characteristics Matter for Individual Improvements

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

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

In order to show the relevance of peer characteristics, we decompose the coefficient of determination, R^2, into variation that is attributable to individual characteristics and peer characteristics.[28] Note that we cannot estimate partial models to

25. Appendix 1.C presents additional robustness checks using biased-reduced linearization or group means to account for the limited number of clusters, specifications that control for outliers and reports the average treatment effects for different subgroups (by gender, grade, school). Our results are robust to all of these checks.

26. In Appendix 1.D, we document that the observed performance improvements in the three treatments described here are a result of the presence of peers and not due to learning. We present the results of an additional control treatment (NoPeer) and its implementation details. In the control treatment, subjects run twice without any peer and we find that they do not improve their time from the first to the second run; in fact, individual performance decreases. The improvements that we observe here can therefore be attributed to the presence of peers rather than learning or familiarity with the task.

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

28. As peer characteristics, we include the rank within a pair itself as well as the rank interacted with match quality with respect to both sets of preferences, friendship indicators and productivity differences. We also include personality traits of a peer and absolute differences in personality traits between peers. This corresponds to the full specification that we also use in our decomposition (col. 5 of Table 1.6).

obtain the fraction of variance explained by a set of predictors as an individual's and her peer's characteristics may be correlated (e.g., since both are from the same age group and age is related to performance, as documented in Table 1.1). We account for this interplay between different groups of explanatory variables by employing a variance decomposition based on Shapley values to calculate the marginal contribution of each group of variables (see Huettner and Sunder, 2012).

We base the variance decomposition on data from RANDOM only and estimate equation (1.2) to decompose R^2 into components attributable to individual as well as peer characteristics.[29] As Table 1.4 reports, we find that 20% of the total variation in percentage points improvements in individual performance can be attributed to characteristics of the peer, which corresponds to 78% of the explained variation. Consequently, only 6% of the total variation or 22% of the explained variation stems from individual characteristics.[30]

Table 1.4. Variance decomposition of performance improvements in RANDOM

		Variation attributable to			
Explained variation (R^2)		Peer characteristics		Individual characteristics	
.26	*(100%)*	.2	*(78%)*	.06	*(22%)*

Notes: This table presents a decomposition of the coefficient of determination, R^2, using Shapley values and is based on equation (1.2) estimated on RANDOM only.

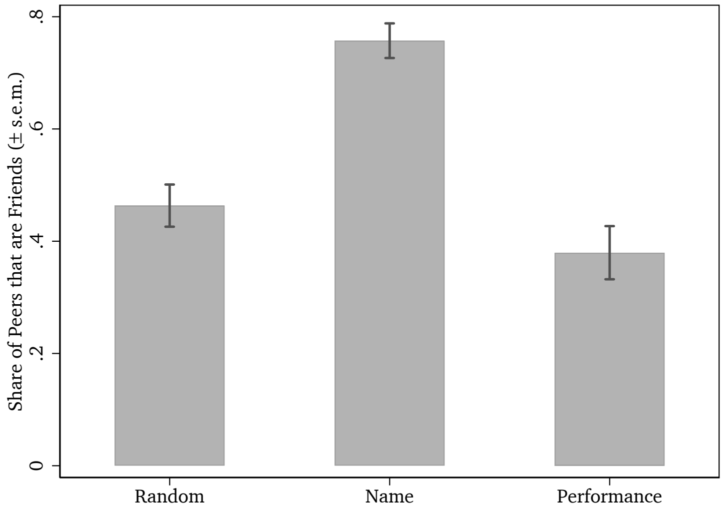
The decomposition therefore shows the importance of accounting for peer characteristics in general. Characteristics of peers are responsible for a large share of the explained variance. Hence, we need to take these peer characteristics into account for the analysis of our treatments.

1.5.3 Self-selection Changes the Peer Composition

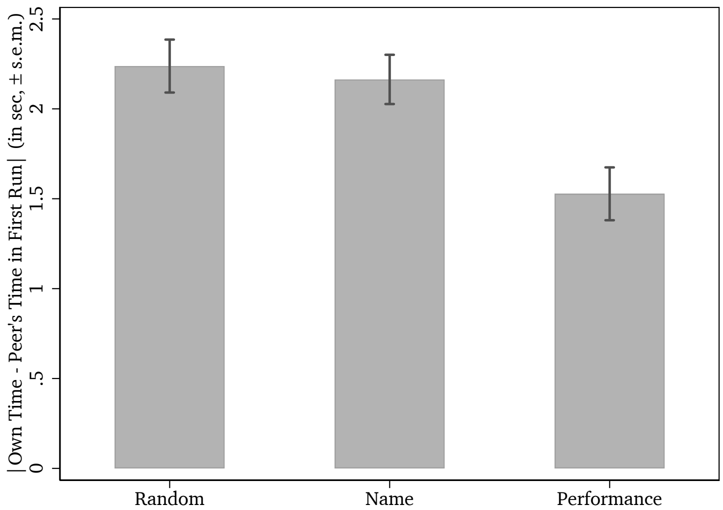
In this section, we document that treatments that allow for self-selection change with whom someone interacts. Although relative peer characteristics are important for understanding outcomes – as shown in the previous section – students also need to interact with systematically different peers when self-selecting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

29. The corresponding estimates are delegated to column (1) of Appendix Table 1.E.5.

30. Note that we explain percentage point improvements from the first to the second run and hence much of the individual-level variation is already taken out of the dependent variable. When using time in the second run as an outcome variable, individual characteristics account for approximately 54% (67% when additionally controlling for time in the first run) of the explained variation ($R^2 = 0.70$



(a) Share of Peers Being Friends



(b) Absolute Differences in Productivity

Figure 1.5. Changes in peer composition

Notes: Figure 1.5a presents the share of all students who nominated their assigned peer as a friend for each of the three treatments including standard errors. Figure 1.5b shows the average absolute within-pair difference in productivity (measured in times from the first run) and including standard errors for each treatment. We control for gender, grade and school fixed effects 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 1.A.2.

without time in the first run, $R^2 = 0.79$ with time in the first run), while peer characteristics explain the remainder of R^2. Nevertheless, the variation explained from peer characteristics remains sizable.

Figure 1.5 shows that our treatments indeed changed the peer composition with respect to two prime examples of peer characteristics, namely friendship ties and productivity differences within pairs. More specifically, Figure 1.5a shows that students are predominantly paired with friends in NAME (76% of all peers are friends), whereas the share of peers being friends in RANDOM and PERFORMANCE is 49% and 37%, respectively. As matching based on preferences over relative performance (PERFORMANCE) allows for targeting of other students with a similar or slightly higher productivity, the students' absolute time differences in the first run might change. Panel B of Figure 1.5b confirms this by showing that the average absolute difference in times from the first run is 1.53 seconds in PERFORMANCE, while it is larger than two seconds in the other two treatments (2.24 and 2.16 seconds in RANDOM and NAME). Even though students could mainly target peers along these two dimensions, we present how our treatments affect the peer composition along various other characteristics in Appendix Table 1.A.2. We find that targeting specific peers also results in systematically different peers in terms of their personality.

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

1.5.4 Decomposition into Direct and Indirect Effects of Self-selection

We now decompose the average treatment effects from Table 1.3 by taking changes in the peer composition explicitly into account. As outlined in Section 1.4, the estimated average effects potentially comprise a direct effect as a result of self-selection and an indirect effect stemming from interacting with different peers. This is the case as our treatments have two features: on the one hand, our treatments change with whom someone interacts and those peer characteristics matter as documented above; and on the other, they change the selection procedure from exogenous assignment to the self-selection of peers. The indirect effect therefore captures changes in the relative characteristics of peers (e.g., the time differences between the student and peer in the first run) due to the altered peer composition induced by being able to select them. The direct effect captures the effect of the treatment due to a change in the selection rule. The previous two subsections documented that NAME and PERFORMANCE change the peer composition relative to RANDOM and established that those relative peer characteristics are important in determining individual outcomes. The decomposition analyzes the extent to which the average treatment effects are driven by these changes in the peer composition.

The results of the decomposition based on equation (1.2) are summarized in Table 1.5 and prThe results of the decomposition based on equation (1.2) are summarized in Table 1.5 and presented in further detail in Table 1.6. In Table 1.5, we use the whole set of characteristics to decompose the average treatment effects into the direct and indirect effects. Therefore, the size of the direct effects equals the

Table 1.5. Decomposition of treatment effects

	Direct Effects		Indirect Effects	
	PP imprv.	Std. Err.	PP imprv.	Std. Err.
Name	1.24	0.50	0.13	0.24
Performance	2.21	0.68	-0.52	0.23

Notes: The table presents the resulting direct and indirect effects from a decomposition according to equation (1.2) shown in column (5) of Table 1.6. Indirect effects are defined as the changes in percentage point improvements that are explained by changes in peer characteristics relative to Random and comprises the combined effect of all peer characteristics in column (5) of Table 1.6.

coefficients of the treatment indicators in column (5) of Panel A in Table 1.6. They correspond to 1.24 percentage points in Name and 2.21 in Performance.

The decomposition shows that even though peer characteristics are highly important in understanding the variation in outcomes, the indirect effects of self-selection in the two treatments are considerably low. They correspond to only 11% of the size of the direct effect in Name and 24% in Performance.[31] In Name, we estimate a positive and insignificant indirect effect of .13 percentage point improvements (p-value = 0.59). This means that the altered peer characteristics have only a slightly positive effect on the students' performance. For Performance, we find a significant indirect effect of -.52 percentage points (p-value = 0.03). Thus, the change in the peer composition even magnifies the direct effect as it negatively rather than positively affects performance.

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

We now analyze the detailed results of the decomposition in Table 1.6. Column (1) replicates the baseline estimates from column (2) of Table 1.3 for means of comparison. In columns (2)-(4), we include different sets of peer characteristics, before we include all of them in column (5). Turning to the separate columns, we find that the size of the treatment indicators only slightly differ across specifications. Nonetheless, some of the included peer characteristics influence the individual performance in the second run. Performance-based match quality has some predictive power for

31. The indirect effect in our decomposition is induced by the impact of peer characteristics and their change through self-selection.Therefore, it corresponds to the difference in the average effect for Name and Performance and the direct effect as the direct and indirect effect add up to the average effect. The indirect effect also corresponds to multiplying the coefficients for (relative) peer characteristics from column (5) with the change in the peer composition across treatments, as described in Appendix 1.B and Appendix Table 1.A.2.

Table 1.6. Decomposition of treatment effects

	Percentage Point Improvements					
	(1) Baseline	(2) Match Quality	(3) Friend-ship ties	(4) Time Difference	(5) All	(6) Class Controls
Direct Effects						
NAME	1.37***	1.23**	1.46***	1.35***	1.24**	1.46***
	(0.50)	(0.53)	(0.49)	(0.46)	(0.50)	(0.46)
PERFORMANCE	1.69**	1.78***	1.61**	1.84***	2.21***	1.73**
	(0.65)	(0.65)	(0.64)	(0.61)	(0.68)	(0.68)
Peer Characteristics						
Faster Student × High match quality (NAME)		0.00			0.52	0.69
		(0.39)			(0.43)	(0.45)
Slower Student × High match quality (NAME)		0.31			0.46	0.62
		(0.61)			(0.66)	(0.74)
Faster Student × High match quality (PERF.)		1.17**			0.43	0.12
		(0.52)			(0.53)	(0.59)
Slower Student × High match quality (PERF.)		-2.07***			-0.71	-1.15
		(0.61)			(0.66)	(0.73)
Faster Student × Peer is Friend			-0.77*		-1.15**	-1.03**
			(0.45)		(0.53)	(0.47)
Slower Student × Peer is Friend			-0.06		0.13	0.45
			(0.53)		(0.67)	(0.79)
Faster Student × $\lvert \Delta Time\ 1 \rvert$				-0.39***	-0.35**	-0.36**
				(0.14)	(0.16)	(0.16)
Slower Student × $\lvert \Delta Time\ 1 \rvert$				1.03***	1.04***	0.84***
				(0.21)	(0.20)	(0.19)
Slower Student in Pair		3.85***	2.20***	-0.17	-0.15	0.11
		(0.44)	(0.49)	(0.45)	(0.68)	(0.76)
Abs. Diff. in Personality	No	No	No	No	Yes	Yes
Peer Characteristics	No	No	No	No	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Class-level Controls	No	No	No	No	No	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes
N	585	585	585	585	582	512
R^2	.08	.18	.15	.24	.29	.29
p-value: NAME vs. PERFORMANCE	.62	.41	.82	.43	.17	.72

Notes: This table presents least squares regressions according to equation (1.2) using percentage point improvements as the dependent variable. High match quality is an indicator that equals one if the partner was ranked within an individual's first three preferences. Personality characteristics include the Big Five, locus of control, social comparison, competitiveness, and risk attitudes. Appendix Table 1.E.6 presents the omitted coefficients of own and peer characteristics, and their absolute differences for our preferred specification in column (5). *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

performance improvements in the restricted regression in column (2). However, the effects are insignificant when controlling for all peer characteristics in column (5). Overall, the quality of the match, i.e., how well a student's preferences were satisfied by the pairing in the second run, has little to no effect on their performance. We also observe that initially faster students within a pair reduce their performance when paired with a friend, while the relatively slower students do not adjust their performance differentially for friends as peers (column (3) and (5)). In column (4), we focus on productivity differences, since faster and slower students within a pair might be affected differentially. We also allow the effect of productivity differences, $|\Delta Time1|$, to differ by the rank within a pair. We find that differences in times of the first run have a significant effect on both faster and slower students within a pair. While slower students within a pair benefit by a 1.03 percentage point improvement from running with a one second faster student, the relatively faster student's performance suffers from this productivity difference by .39 percentage points. In sum, the average performance of a pair thus improves with increasing differences in productivity.

We control for all of these characteristics jointly in column (5), where we also add a rich set of relative peer personality characteristics. The effect of friendship ties on the initially faster students as well as the effects on productivity differences persist. More importantly, the direct effects of both NAME and PERFORMANCE remain robust, showing a direct effect of self-selection on individual performance. In order to further probe the robustness of this finding, we additionally control for proxies of the class attitude in column (6). While the estimates slightly differ in magnitude, the results are generally robust. However, as we lose some observations, our preferred specification is column (5).

Table 1.7. Variance decomposition and the role of unobservables

Panel A: Variance decomposition							
		Variation attributable to					
Explained variation (R^2)		Treatments		Peer characteristics		Individual characteristics	
0.29	*(100%)*	0.03	*(12%)*	0.21	*(72%)*	0.05	*(16%)*

Panel B: Role of unobservables			
	Oster's δ		
	$R^2_{max} = 0.50$	$R^2_{max} = 0.75$	$R^2_{max} = 1.00$
NAME	2.54	1.19	0.78
PERFORMANCE	-7.05	-3.43	-2.27

Notes: Panel A decomposes the explained variance of specification (5) of Table 1.6 in components attributable to treatments, peer and individual characteristics similar to Table 1.4. Panel B quantifies the importance of unobservables relative to observables needed for zero direct effects according to Oster (2019).

Our results, therefore, provide evidence for a direct effect of self-selection. In the remainder of this section, we provide further evidence for its robustness. First, in Panel A of Table 1.7 we replicate the variance decomposition of Table 1.4 for all three treatments and confirm the importance of the peer characteristics in terms of explaining the variation in outcomes. Second, in Panel B of Table 1.7 we address the possible concern that other characteristics for which we cannot account or control are driving the direct effect. Our results above remain relatively stable when adding different sets of peer controls, which is reassuring. A more formal approach to tackle this concern is to ask how important unobserved characteristics would have to be to explain our direct treatment effects (Altonji, Elder, and Taber, 2005; Oster, 2019). We follow Oster (2019) and calculate δ, a measurement for the relative importance of unobserved characteristics compared to observed characteristics. This measure describes how important unobserved variables would have to be relatively to observed ones to explain the direct effects, i.e., to drive down the direct effects to zero. Absolute values of δ larger than one indicate that these omitted variables have to be relatively more important than observed peer characteristics. Negative values indicate that those unobservable characteristics need to reverse the effect of observed covariates. We calculate these measures for three scenarios that differ in the maximum amount of variance that would theoretically be explained if all factors that might affect the outcomes were observed. More specifically, we calculate δ for R^2_{max} equal to 0.50, 0.75 and 1.00. In all but one extreme scenario the omitted peer characteristics are required to be more important than the observed peer characteristics. This suggests that such unobserved characteristics need to have a larger effect than productivity differences, friendship ties, match quality and all other controls – including personality traits – combined. Compared to other studies, our analysis already allows for more peer characteristics to influence subjects' behavior. Therefore, we allow for a very rich set of important characteristics and conclude that such unobserved characteristics are highly unlikely to drive the direct treatment effects.

In addition, we provide several robustness checks in the Appendix 1.E. Appendix Table 1.E.1 allows for different specifications of match quality by additionally considering the partner's match quality, an interaction between one's own and the partner's match quality, as well as feasible match quality. Appendix Table 1.E.2 considers different definitions of friendship ties apart from directed links (i.e., undirected, reciprocal, directed and reciprocal friendship ties). The results for all robustness checks remain qualitatively and quantitatively similar. Furthermore, we show in Table 1.E.3 and Appendix Figure 1.E.1 that the linear specification of productivity differences is not restrictive. Appendix Table 1.E.4 estimates the coefficients of peer characteristics on the subsample of students in Random only and imposes these coefficients on the other treatments. Furthermore, Table 1.E.5 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. All of these robustness checks support our conclusion.

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

1.5.5 Explanation of the Direct Effect

We interpret the direct effect as a positive effect of self-selection due to increased control or autonomy over the peer assignment mechanism. However, one might worry that knowledge of all three treatment conditions could lead students in Random to react negatively due to disappointment that their preferences have not been taken into account.[32] If these disappointed students drove our findings, we would falsely attribute effects to self-selection even if students in Name and Performance do not react positively.[33] If the direct effect originated from disappointment, we would expect students in Random to have less fun in the experimental task. Therefore, in column (1) of Appendix Table 1.F.1 we analyze the extent to which subjects across treatments had different perceptions regarding their fun in the second run. We find zero effects. The absence of direct effects in the fun dimension alleviates the potential concern that knowledge of all three treatments leads to disappointment when students are assigned to Random.[34]

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

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

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

We therefore conclude that the direct effects in our experiment are due to positive effects of self-selection. More specifically, we argue that the opportunity to self-select key aspects of one's environment – in our experiment having autonomy over the peer selection – has a direct effect beyond the instrumental value of changing peer characteristics. Self-determination theory provides a credible explanation through which self-selection can impact performance directly. The theory identifies autonomy as a crucial determinant of motivation: individuals who can actively select parts of their environment – most importantly their tasks in work environments – display higher intrinsic motivation (Deci and Ryan, 1985, 2000).[35] Applying this explanation to our setting suggests that not the selected peer herself increases motivation, but the mere act of selecting her. However, we do not argue that this behavioral effects stems from self-selecting any aspect, but a relevant aspect of one's environment.

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

1.5.6 The Limits of Reassignment Rules

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

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

in our context, we can use our estimates to simulate the effect of such exogenous peer assignment rules and compare their effect to outcomes under self-selection.

For this purpose, we use our estimates obtained in Section 1.5.4, using the whole set of peer characteristics (column (5) of Table 1.6). Based on these estimates, we simulate different (exogenous) assignment rules, calculate the resulting effects on performance, and compare them to performance improvements observed in our experiment. We first compare the improvements to the counterfactual of assigning the same peers in Name and Performance without the direct effect of self-selection. A comparison of other peer assignment rules with these results sheds light on the question of whether students are able to choose optimal peers. Second, we simulate the expected performance improvements under a random matching. Third, we use several assignment rules that base the assignment on one single and commonly employed peer characteristic, namely past performance. Our estimates obtained in Section 1.5.4 suggest that pairs with a higher difference in initial performances will improve their performance on average. If this is the only characteristic of a peer that affects performance, aggregate performance would be maximized as long as the sum of productivity differences within a pair is maximized.[36] In order to compare the results of self-selection against exogenous assignment rules that promise the largest aggregate improvements, we consider two matching rules that maximize these productivity differences within pairs – Equidistance and High-to-Low – that keep the distance in ranks within the class constant or pair the best-performing student with the slowest student. Additionally, we look at the effect of tracking (i.e., 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.[37]

Figure 1.6 presents the simulated average performance improvements of each assignment rule. The results show that no other peer assignment rule is able to reach similar performance improvements as those featuring self-selection. In fact, they are close to the results from our random matching, since students under those peer assignment rules do not benefit from the additional intrinsic value of self-selection. We observe that in the absence of a direct effect of self-selection, students do not experience additional improvements relative to randomly assigned peers. Compared to Equidistance and High-to-Low, students in Name (exog.) and Performance (exog.) perform worse indicating that they do not choose their peers optimally.

More surprisingly, the reassignment rules that maximize productivity differences in pairs – Equidistance and High-to-Low – do not improve average performance

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

37. We provide details on the prediction of performance improvements and the peer assignment rules in Appendix 1.H.

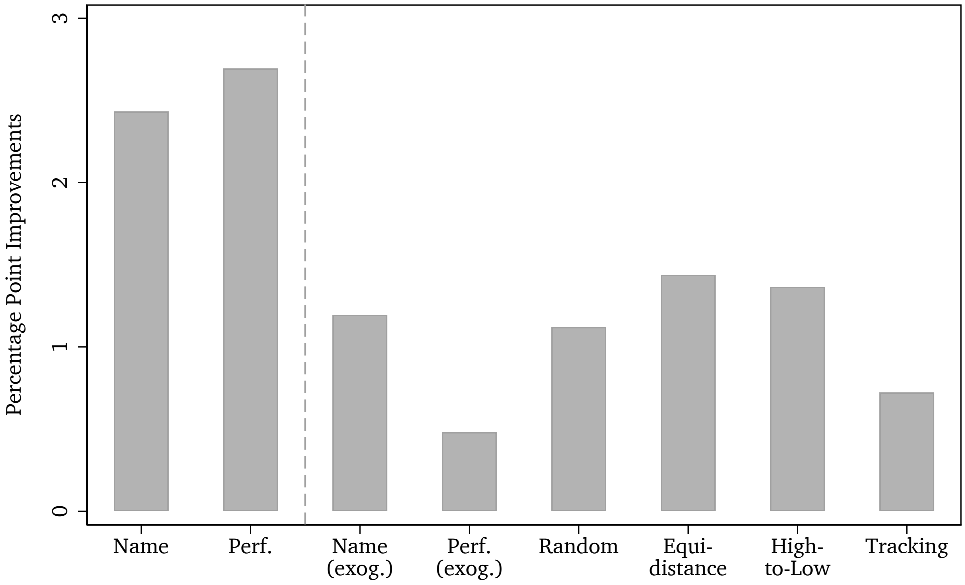


Figure 1.6. Simulation of other peer assignment rules

Notes: The figure presents predicted percentage point improvements for the two treatments (NAME, PERFORMANCE) with and without the effect of self-selection, the RANDOM-treatment as well as three simulated peer assignment rules (EQUIDISTANCE, HIGH-TO-LOW and TRACKING). We fix the personal characteristics and other covariates not at the pair level to 0, whereby effect sizes are therefore not directly comparable to treatment effects above. More details are provided in the text and Appendix 1.H.

compared to the random assignment of peers. Although both rules increase the average productivity difference in pairs by construction and affect performance through this channel, those rules also change other characteristics of the peer. The lack of any additional improvement implies that these other changes in peer characteristics offset the positive effect of increased productivity differences. This highlights important consequences of peer effects that are multidimensional if one wants to enhance overall performance.

This result suggests that reassignment rules based on specific characteristics may not work as intended given that other characteristics may affect performance at the same time. Thus, depending on the correlation structure between the characteristic used for the peer assignment rule and the omitted characteristics as well as their effect, the resulting outcomes may be either higher or lower than predicted. If peer effects are multidimensional, policy makers need to take all potential characteristics into account when reassigning students into peer groups. Consequently, designing optimal peer assignment rules might be more challenging than expected.[38] This insight further helps to understand why we observe a very small indirect effect in the decomposition of the treatment effects despite the fact that peer characteristics help to explain much of the variation in individual outcomes (cf. Table 1.6).

The simulations above suggest that self-selection of peers can be an attractive alternative compared to traditional peer assignment rules to increase individual performance. However, we want to stress that such peer assignments based on self-selection may also come at a cost. In particular, we show in Appendix Table 1.G.1 that students in PERFORMANCE experience significantly more pressure compared to the other two treatments, and individual ranks may be more perturbed between the two runs in NAME and RANDOM relative to PERFORMANCE. Hence, a policy maker might not only look at the resulting performances but also how different assignment rules affect the individuals' overall well-being.

1.6 Conclusion

Peer effects are an ever-present phenomenon discussed in a wide range of settings across the social sciences. For many situations, identifying the effect of an actively self-chosen peer is important beyond estimating peer effects in general. Our framed field experiment introduces a novel way to study the self-selection of peers in a controlled manner and is able to separate the impact of a specific peer on a subject's performance from the overall effect of self-selection. The results of our experiment provide evidence that self-selecting peers yields performance improvements of about 15% of a standard deviation relative to random assignment of peers. While peer characteristics affect the individual performance, they are not the origin of the estimated treatment effects. Rather, these improvements stem from a direct effect of self-selection. Based on self-determination theory (Deci and Ryan, 1985), we interpret this direct effect such that the ability to select one's own peer enhances a student's intrinsic motivation and subsequently increases individual performance.

Teachers or supervisors might be interested to leverage this direct effect of self-selection in addition to other forms of non-monetary incentives used in schools (Levitt et al., 2016) or workplaces (Cassar and Meier, 2018). 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 in general may be improved if individuals are given the opportunity to select their relevant comparison themselves rather than being assigned an unspecific one.

38. In general, designing optimal peer assignment rules requires an optimization taking into account all potential dimensions in which peers may exert effects. This creates a high-dimensional optimization problem that is highly difficult to solve.

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

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. In those settings, it is reasonable to assume that self-selection of peers may happen or can be implemented. For example, study groups at universities often form endogenously (Chen and Gong, 2018), researchers select their co-authors and workers in firms increasingly form self-managed work teams (Lazear and Shaw, 2007), and employees self-select with whom they work by referring others to their employer (Friebel, Heinz, Hoffman, and Zubanov, 2019; Lazear and Oyer, 2012).

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.

Appendix 1.A Randomization and Manipulation Check

Table 1.A.1 presents the randomization check of our experiment. The residual of times in the first run are constructed from a regression of times of the first run on school and grade-specific fixed effects as well as age. As can be seen the difference in times in the first run can be explained by those observables and hence are an artifact of the block randomization as classrooms rather than individuals were randomly assigned to treatments.

Table 1.A.1. Randomization check

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

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

In section 1.3.1, 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.[39] 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 1.A.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 1.3.

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 treatments affect the peer composition in other dimensions apart from the match quality in Table 1.A.2.

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

Table 1.A.2. Effects of treatments on peer composition

	Match Qual. (name)	Match Qual. (time)	Friendship Ties	Time 1	Locus of Control	Social Comparison	Competi-tiveness
Name	0.49***	0.07	0.32***	-0.08	0.12	0.00	0.03
	(0.06)	(0.04)	(0.06)	(0.19)	(0.11)	(0.10)	(0.13)
Performance	-0.06	0.24***	-0.07	-0.70***	0.46***	-0.19**	0.12
	(0.06)	(0.04)	(0.07)	(0.21)	(0.12)	(0.09)	(0.11)
Age (standardized)	-0.03	-0.12*	0.03				
	(0.04)	(0.07)	(0.08)				
N	588	588	294	294	292	293	291
p-value: Name vs. Perf.	1.0e-11	.0002	1.3e-07	.0037	.003	.079	.37
Mean in Random	.23	.3	.4	2.4	.98	1.1	1.1

	Extra-version	Agree-ableness	Conscien-tiousness	Neuro-ticism	Openness	Risk
Name	0.07	-0.14	0.09	-0.15	0.11	-0.15
	(0.11)	(0.14)	(0.09)	(0.11)	(0.13)	(0.10)
Performance	0.05	0.01	0.14	-0.20	0.28**	0.12
	(0.11)	(0.17)	(0.09)	(0.12)	(0.13)	(0.11)
N	292	292	292	292	292	292
p-value: Name vs. Perf.	.76	.19	.53	.63	.19	.031
Mean in Random	1.1	1.2	1	1.1	.98	1.1

Notes: 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.

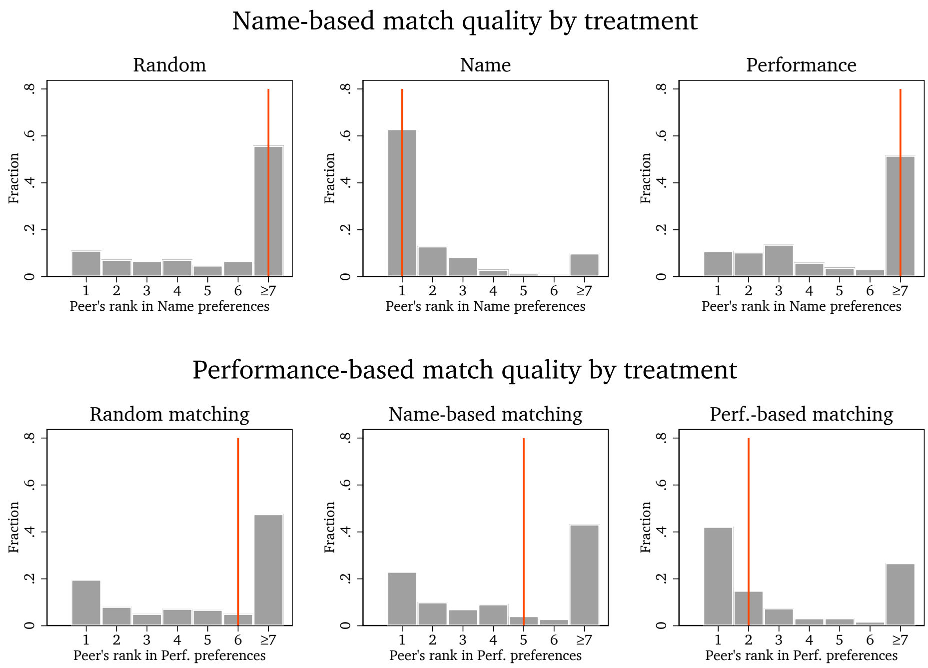


Figure 1.A.1. Feasible match quality across treatments

Notes: 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.

Appendix 1.B 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.[40]

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:

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

Let θ_d be a vector characterizing a peer's relative characteristics in treatment $d \in \{R, N, P\}$.[41] 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

$$Y_d \;=\; \alpha_d + \beta_d\theta + \gamma X + \epsilon_d \tag{1.B.2}$$

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

Potentially, there are unobserved factors in θ. We therefore split θ in a vector with the observed inputs ($\bar{\theta}$) and unobserved inputs ($\tilde{\theta}$)[42] with corresponding effects $\bar{\beta}_d$

40. 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.

41. 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).

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

and $\tilde{\beta}_d$ and can rewrite equation (1.B.2) as follows:

$$\begin{aligned} Y_d &= \alpha_d + \bar{\beta}_d\bar{\theta} + \tilde{\beta}_d\tilde{\theta} + \gamma X + \epsilon_d & (1.B.3) \\ &= \tau_d + \bar{\beta}_d\bar{\theta} + \gamma X + \tilde{\epsilon}_d & (1.B.4) \end{aligned}$$

where $\tau_d = \alpha_d + \tilde{\beta}_d E[\tilde{\theta}]$ and $\tilde{\epsilon}_d = \epsilon_d + \tilde{\beta}_d(\tilde{\theta} - E[\tilde{\theta}])$. 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 $\Delta_{R,d}$).

$$\begin{aligned} Y_d &= \tau_d + \bar{\beta}\bar{\theta} + \bar{\Delta}_{R,d}\bar{\theta} + \gamma X + \tilde{\epsilon}_d & (1.B.5) \\ &= \hat{\tau}_d + \bar{\beta}\bar{\theta} + \gamma X + \tilde{\epsilon}_d & (1.B.6) \end{aligned}$$

In what follows, we are interested in $\bar{\tau}_d = E[\hat{\tau}_d - \hat{\tau}_R]$ $(d \in \{N, P\}; \hat{\tau}_d = \tau_d + \bar{\Delta}_{R,d}\bar{\theta})$, 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 indirect effects of the mediating variables. The empirical specification of (1.B.6) is given by

$$y_{igs} = \bar{\tau} + \bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \qquad (1.B.7)$$

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}.

Appendix 1.C Robustness Checks for Average Treatment Effects

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

Table 1.C.1. Robustness checks

	Percentage Point Improvements				
	(1) Baseline	(2) BRL	(3) Group means	(4) Winsorization	(5) Truncation
NAME	1.26***	1.26**	1.13*	1.05***	0.95***
	(0.43)	(0.50)	(0.61)	(0.37)	(0.35)
PERFORMANCE	1.67**	1.67**	1.96***	1.51***	1.43***
	(0.62)	(0.72)	(0.62)	(0.51)	(0.43)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	588	588	70	588	496
R^2	.056	.056	.27	.072	.087
p-value: NAME vs. PERF.	.51	.55	.15	.37	.27

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Column (1) presents the baseline specifications as used in Table 1.3. Columns (2) uses biased-reduced linearization (BRL) to account for the limited number of clusters. Column (3) uses matching group-specific means as the unit of observation. Finally, columns (4) and (5) apply a 90% winsorization and truncation, respectively.

We further analyze the robustness of our results by looking at different subsamples. We therefore split our sample first by grades in the upper panel of Table 1.C.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 1.C.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. PERF.	.51	.21	.82	.19	.82
	(6) Female	(7) Male	(8) School 1	(9) School 2	(10) School 3
NAME	1.26*	1.21***	1.36***	1.44**	2.09***
	(0.65)	(0.44)	(0.11)	(0.65)	(0.37)
PERFORMANCE	1.68**	1.63*	1.53***	2.29***	2.22*
	(0.77)	(0.85)	(0.05)	(0.55)	(1.12)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	390	198	148	274	166
R^2	.057	.065	.065	.1	.12
p-value: NAME vs. PERF.	.53	.62	.3	.14	.88

Notes: 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 1.3. Columns (2)-(5) restrict the sample to one grade, columns (6) and (7) to each gender and columns (8)-(10) to one school.

Appendix 1.D Control Treatment to Disentangle Peer Effects from Learning

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

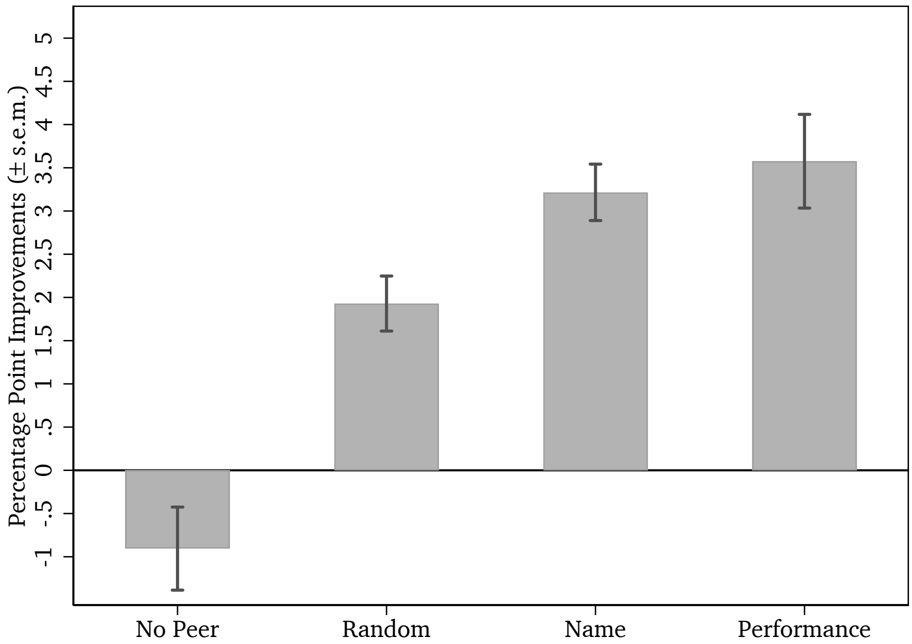


Figure 1.D.1. Average treatment effects

Notes: 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 1.D.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.

Table 1.D.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

Notes: This table presents least squares regressions using percentage point improvements in column (1) or times from the second run in columns (2) and (3) 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.

Appendix 1.E Peer Composition Robustness Checks

We run several robustness checks for the results presented in Table 1.6. First, in Table 1.E.1 we use different specifications for match quality. We consider the partner's match quality, an interaction between one's own and the partner's match quality, and feasible match quality as defined in Appendix 1.A, and find that the estimates of our direct effects are qualitatively and quantitatively the same. Second, in Table 1.E.2, we show that our results do not depend on the precise definition of friendship ties. We check whether our results change when we define friendship ties as undirected or reciprocal rather than directed. As can be seen from the table, the coefficients on the direct effects as well as on other peer characteristics remain the same. Third, we control for differences in productivity in a more flexible way in Table 1.E.3 by allowing for quartic rather than linear effects of productivity differences in column (2) (see also Figure 1.E.1 comparing linear and quartic terms graphically). In addition, we allow for a second flexible specification using fixed effects for productivity differences. More specifically, we include an indicator for each one-second interval of productivity differences between subjects within a pair. This allows for a potential non-linear influence of productivity differences on our estimates. Comparing the estimates shows that neither the quartic functional form nor the fixed effect specification is restrictive. Fourth, we estimate the influence of peer characteristics (and individual characteristics) on the sample of Random subjects only in Table 1.E.4 and use these coefficients to decompose the average effect. For this purpose, we first net out the effect of group variables such as school and gender-grade fixed effects (as well as individual characteristics) from both the outcome and independent variables such as peer characteristics according to the Frisch-Waugh-Lovell theorem using the whole sample. In a first version, we regress the outcome and peer characteristics on the fixed effects only. In a second version, we additionally net out the effect of individual characteristics from peer characteristics and the outcome. We use the residuals of those regressions to decompose the treatment effect. We then begin by estimating the influence of peer characteristics on the outcome using only subjects from Random and the residualized outcome as well as peer characteristics (column (1) and (3)). In a second step, we restrict the influence of those peer characteristics and estimate the direct treatment effects (column (2) and (4)). Finally, Table 1.E.5 restricts the control group sample to subjects with a high match quality within Random to show that the treatment effects persist for these subjects and the coefficients on peer compositional effects do not substantially change. Table 1.E.6 presents the ommited coefficients of own and peer characteristics, as well as their absolute differences, from column (5) Table 1.6 in the main text.

Table 1.E.1. Robustness Checks for match quality

	Percentage Point Improvements		
	(1) Partner's MQ	(2) Interaction	(3) Feasible
Direct Effects			
Name	1.15**	1.14*	1.19**
	(0.55)	(0.57)	(0.47)
Performance	2.23***	2.21***	2.05***
	(0.70)	(0.69)	(0.66)
Peer Characteristics			
High match quality (partner; Name)	0.28	0.18	
	(0.42)	(0.56)	
High match quality (partner; Perf.)	-0.07	0.21	
	(0.40)	(0.44)	
High match quality (own and partner; Name)		0.19	
		(0.84)	
High match quality (own and partner; Perf.)		-0.58	
		(0.94)	
Faster Student × High match quality (feasible; Name)			0.02
			(0.42)
Slower Student × High match quality (feasible; Name)			1.38*
			(0.79)
Faster Student × High match quality (feasible; Perf.)			0.83*
			(0.41)
Slower Student × High match quality (feasible; Perf.)			0.32
			(0.86)
Faster Student × Match Quality (name-based)	0.45	0.37	
	(0.40)	(0.44)	
Slower Student × Match Quality (name-based)	0.41	0.30	
	(0.65)	(0.75)	
Faster Student × Match Quality (perf.-based)	0.43	0.75	
	(0.51)	(0.65)	
Slower Student × Match Quality (perf.-based)	-0.71	-0.48	
	(0.65)	(0.61)	
Friendship Ties and Performance Differences	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	582	582	582
R^2	.29	.29	.29
p-value: Name vs. Performance	.16	.18	.24

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) adds the partner's match quality in addition to own match quality as in Table 1.6, while column (2) additionally controls for the interaction of own and partner's match quality. Finally, column (3) uses a different measure of match quality, (feasible match quality – see also Appendix 1.A), which acknowledges the fact that certain preferred peers may not be available.

Table 1.E.2. Different definitions of friendship ties

	Percentage Point Improvements			
	(1) directed	(2) undirected	(3) reciprocal	(4) dir. & rec.
Direct Effects				
Name	1.24**	1.20**	1.21**	1.14**
	(0.50)	(0.49)	(0.50)	(0.50)
Performance	2.21***	2.13***	2.21***	2.19***
	(0.68)	(0.69)	(0.68)	(0.68)
Faster Student × Peer is friend	-1.15**			-1.67*
	(0.53)			(0.85)
Slower Student × Peer is friend	0.13			-0.38
	(0.67)			(0.83)
Faster Student × Peer is friend (undirected)		-1.63***		
		(0.58)		
Slower Student × Peer is friend (undirected)		0.16		
		(0.80)		
Faster Student × Peer is friend (reciprocal)			-0.56	0.76
			(0.59)	(0.94)
Slower Student × Peer is friend (reciprocal)			0.47	0.73
			(0.53)	(0.63)
Faster Student × $\|\Delta Time\ 1\|$	-0.35**	-0.34**	-0.34**	-0.34**
	(0.16)	(0.16)	(0.16)	(0.15)
Slower Student × $\|\Delta Time\ 1\|$	1.04***	1.04***	1.05***	1.05***
	(0.20)	(0.20)	(0.20)	(0.20)
Slower Student in Pair	-0.15	-0.47	0.05	-0.18
	(0.68)	(0.74)	(0.69)	(0.68)
Match quality and performance differences	Yes	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes
N	582	582	582	582
R^2	.29	.29	.29	.29

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 1.6 for reference using directed friendship ties. Column (2) uses undirected friendship ties, column (3) reciprocal directed friendship ties, while column (4) allows for a differential effect of directed and reciprocal friendship ties.

Table 1.E.3. Robustness checks for absolute time differences

	Percentage Point Improvements		
	(1) Linear	(2) Quartic	(3) FEs
Direct Effects			
Name	1.24**	1.28**	1.20**
	(0.50)	(0.49)	(0.52)
Performance	2.21***	2.23***	2.25***
	(0.68)	(0.68)	(0.74)
Faster Student × $\|\Delta Time\ 1\|$	-0.35**	-2.70**	
	(0.16)	(1.26)	
Slower Student × $\|\Delta Time\ 1\|$	1.04***	1.27	
	(0.20)	(1.75)	
Slower Student in Pair	-0.15	-1.82*	
	(0.68)	(0.91)	
Faster Student × $\|\Delta Time\ 1\|^2$		0.90	
		(0.56)	
Slower Student × $\|\Delta Time\ 1\|^2$		-0.00	
		(0.97)	
Faster Student × $\|\Delta Time\ 1\|^3$		-0.12	
		(0.09)	
Slower Student × $\|\Delta Time\ 1\|^3$		-0.01	
		(0.18)	
Faster Student × $\|\Delta Time\ 1\|^4$		0.00	
		(0.00)	
Slower Student × $\|\Delta Time\ 1\|^4$		0.00	
		(0.01)	
Time Diff. FEs	No	No	Yes
Match Quality and Friendship Ties	Yes	Yes	Yes
Abs. Diff. in Personality	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	582	582	582
R^2	.29	.29	.3
p-value: Name vs. Performance	.17	.17	.14

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) presents the last specification of Table 1.6 for reference. Column (2) includes quartic terms of time differences in the first run (also illustrated in Appendix Figure 1.E.1) and column (3) fixed effects for every one-second difference in productivity levels of the two students.

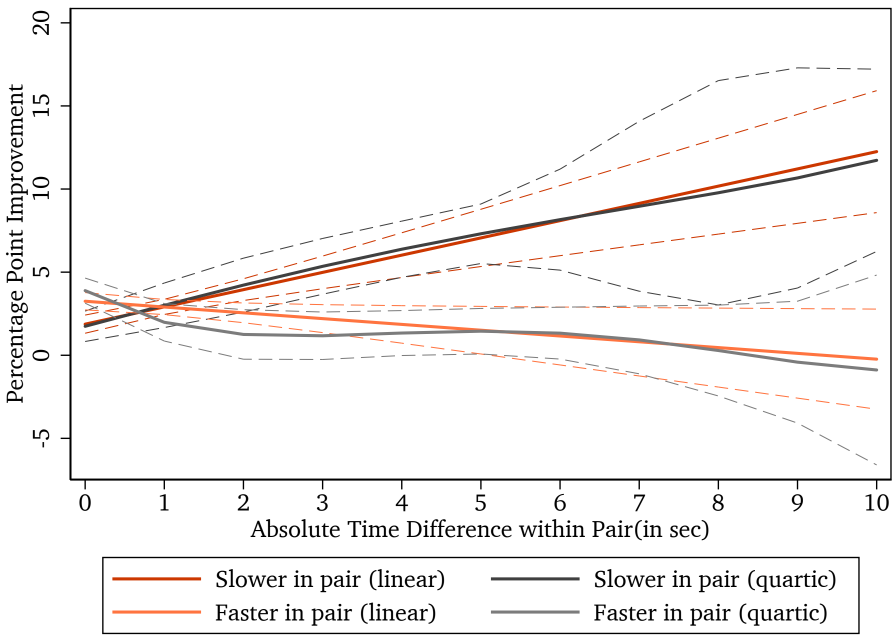


Figure 1.E.1. Robustness of linear specification in time differences

Notes: The figure presents marginal effects (solid lines) from a least squares regression using percentage point improvements as the dependent variable including 95% confidence intervals (dashed lines). It plots the linear specification (black lines) as used in the main text as well as a second specification using quartic polynomials (orange lines) of absolute time differences in the first run as regressors. We use the same set of controls as in column (5) of Table 1.6 and cluster standard errors at the class level. The corresponding regressions are presented in columns (1) and (2) of Appendix Table 1.E.3.

Table 1.E.4. Restricting coefficients of peer characteristics

	Percentage Point Improvements			
	Fixing only FEs		Fixing FEs & own char.	
	(1) only RANDOM	(2) all	(3) only RANDOM	(4) all
Direct Effects				
NAME		.77*		.79*
		(.46)		(.47)
PERFORMANCE		1.67**		1.66**
		(.67)		(.67)
Peer Characteristics				
Faster Student × High match quality (NAME)	.76	.76	.76	.76
	(.85)		(.78)	
Slower Student × High match quality (NAME)	.26	.26	.38	.38
	(1.09)		(1.01)	
Faster Student × High match quality (PERF.)	.18	.18	-.15	-.15
	(1.11)		(1.13)	
Slower Student × High match quality (PERF.)	-.41	-.41	-.14	-.14
	(1.15)		(1.2)	
Faster Student × Peer is friend	-.14	-.14	-.19	-.19
	(.66)		(.59)	
Slower Student × Peer is friend	.03	.03	-.06	-.06
	(1.28)		(1.15)	
Faster Student × $\lvert \Delta Time\ \ 1 \rvert$	-.51*	-.51	-.5	-.5
	(.3)		(.28)	
Slower Student × $\lvert \Delta Time\ \ 1 \rvert$	.78**	.78	.84**	.84
	(.32)		(.3)	
Slower Student in Pair	.13	.13	-.1	-.1
	(.99)		(.87)	
Abs. Diff. in Personality	Yes	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes	Yes
Own Characteristics	No	No	Yes	Yes
N	204	582	204	582
R^2	.24		.22	

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

Table 1.E.5. Only high match quality sample as comparison group

	Percentage Point Improvements					
	(1) RANDOM	(2) All	(3) RANDOM & NAME	(4) with Controls	(5) RANDOM & PERF.	(6) with Controls
Direct Effects						
NAME		1.24**	1.83***	1.93***		
		(0.50)	(0.55)	(0.47)		
PERFORMANCE		2.21***			2.38***	1.75**
		(0.68)			(0.71)	(0.64)
Peer Characteristics						
Faster Student	0.89	0.52				-0.47
× Match Quality (name-based)	(0.95)	(0.43)				(1.28)
Slower Student	0.15	0.46				-0.56
× Match Quality (name-based)	(1.10)	(0.66)				(1.15)
Faster Student	0.06	0.43		-0.51		
× Match Quality (perf.-based)	(1.08)	(0.53)		(0.65)		
Slower Student	-0.51	-0.71		-1.21		
× Match Quality (perf.-based)	(1.22)	(0.66)		(0.86)		
Faster Student × Peer is friend	0.10	-1.15**		-1.53		-0.98
	(0.74)	(0.53)		(1.05)		(1.87)
Slower Student × Peer is friend	0.01	0.13		-1.18		-1.38
	(1.15)	(0.67)		(1.06)		(1.13)
Faster Student × $\|\Delta Time\ 1\|$	-0.54**	-0.35**		-0.72**		-0.07
	(0.25)	(0.16)		(0.29)		(0.51)
Slower Student × $\|\Delta Time\ 1\|$	0.73**	1.04***		1.25***		1.08**
	(0.32)	(0.20)		(0.38)		(0.47)
Slower Student in Pair	0.43	-0.15		-0.44		-0.97
	(1.15)	(0.68)		(1.70)		(1.47)
Abs. Diff. in Personality	Yes	Yes	No	Yes	No	Yes
Peer Characteristics	Yes	Yes	No	Yes	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes	Yes
N	204	582	208	207	162	160
R^2	.28	.29	.16	.52	.16	.37

Notes: This table presents least squares regressions using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own and peer characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Column (1) and (2) present the last specification of Table 1.6 for RANDOM and the full sample for reference. Columns (3) to (6) show that even if we restrict the comparison group to the sample of individuals in RANDOM that received a peer with high match quality according to their name- (columns (3) and (4)) or performance-based preferences (columns (5) and (6)), respectively, the direct effects persist and the coefficients on peer compositional effects do not change much.

Table 1.E.6. Omitted Coefficients from Table 1.6 column (5)

	Own characteristics	Peer characteristics	Abs. Diff in characteristics
Agreeableness	0.12	-0.11	0.29
	(0.22)	(0.20)	(0.29)
Conscientiousness	0.01	0.13	-0.13
	(0.21)	(0.17)	(0.23)
Extraversion	0.03	0.06	-0.51**
	(0.24)	(0.20)	(0.25)
Openness to Experience	-0.49**	-0.18	0.52
	(0.19)	(0.17)	(0.33)
Neuroticism	-0.16	-0.16	-0.65**
	(0.24)	(0.19)	(0.27)
Locus of Control	0.17	0.09	-0.15
	(0.20)	(0.19)	(0.31)
Social Comparison	0.32*	0.21	-0.21
	(0.18)	(0.16)	(0.31)
Competitiveness	-0.08	-0.37	0.35
	(0.30)	(0.23)	(0.21)
Risk Attitudes	0.04	0.06	1.32
	(0.18)	(0.17)	(1.70)

Notes: This table presents omitted coefficients from Table 1.6 in the main text. Columns (1) and (2) show the coefficients on own and peer characteristics, respectively. Column (3) presents the coefficients on the absolute differences in personality measures. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Appendix 1.F Additional Material for Discussion of Direct Effects

Table 1.F.1 presents three regressions to support section 1.5.5's discussion of the psychological effect underlying the direct effects. First, we show that students in RANDOM are not disappointed by having a partner assigned. If they were disappointed, they should have less fun during the second run. As column (1) show this is not the case. Secobd, we do not find evidence that subjects with self-selected perceive winning in the second run as more important as we do not see a differential effect on fun between being faster or slower in the second run. Third, we show that prosocial students, that is individuals that score higher on agreeableness, do not show differentially direct effects. This is suggestive evidence against experimenter demand effects or other reciprocal motives driving the estimated direct effects.

Table 1.F.1. Potential psychological mechanisms for the direct effect

	Fun (std.).		PP. Imprv.
	(1)	(2)	(3)
Direct Effects			
Name	-0.01	0.01	1.24**
	(0.10)	(0.14)	(0.50)
Performance	-0.10	-0.07	2.20***
	(0.08)	(0.13)	(0.68)
Name × Slower Student in Pair (2nd Run)		-0.05	
		(0.18)	
Performance × Slower Student in Pair (2nd Run)		-0.07	
		(0.17)	
Name × Agreeableness			0.02
			(0.38)
Performance × Agreeableness			0.42
			(0.45)
Peer Characteristics			
Faster Student (2nd Run) × $\|\Delta Time\ 2\|$	-0.01	-0.01	
	(0.04)	(0.04)	
Slower Student (2nd Run) × $\|\Delta Time\ 2\|$	-0.14***	-0.14***	
	(0.04)	(0.04)	
Slower Student in Pair (2nd Run)	0.04	0.07	
	(0.18)	(0.20)	
Faster Student × $\|\Delta Time\ 1\|$			-0.35**
			(0.16)
Slower Student × $\|\Delta Time\ 1\|$			1.05***
			(0.20)
Slower Student in Pair			-0.20
			(0.66)
Match quality	Yes	Yes	Yes
Friendship indicators	Yes	Yes	Yes
Own Characteristics	Yes	Yes	Yes
Peer Characteristics	Yes	Yes	Yes
Abs. Diff. in Personality Characteristics	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes
N	582	582	582
R^2	.34	.34	.29
p-value: Name vs. Performance	.46	.63	.18

Notes: This table presents least squares regressions using a standardized measure of fun in the second run (columns (1) and (2)) or percentage point improvements (column (3)) as the dependent variable. Column (2) uses the full specification of Table 1.6 and additionally interacts the treatment indicators with one's own measure of agreeableness as a proxy of prosociality. Column (1) focuses on fun as an outcome variable that was elicited after the second run ("How much fun did you have during the second run? Please rate this on a scale from 1 – no fun at all – to 5 – a lot of fun.") and uses the full specification of Table 1.6 adapted using times and ranks from the second run. Column (2) additionally interacts treatment indicators with the final rank in the second run. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level.

Appendix 1.G Additional Material for Implications

Our treatments also have implications for individual ranks of students within a class since slower students improve more than faster ones. As ranks are important in determining subsequent outcomes (Elsner and Isphording, 2017; Gill et al., 2019; Murphy and Weinhardt, 2018), a policy maker has to take the distributional effects of peer assignment mechanisms into account.[43] 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 1.G.1 in which we regress the absolute change in percentile scores from the first to the second run on treatment indicators. The outcome variable measures the average perturbation of ranks within in a class across the two runs. The results show that Performance shuffles the ranks of students less in comparison to Random and Name. While in Random students change their position by about 15 out of 100 ranks, we find significantly less changes in the percentile score in Performance relative to Random. This change corresponds to a 27% reduction in reshuffling. However, in Name we do not find any effect compared to Random.

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

43. 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.

Table 1.G.1. Side effects of reassignment rules

	Absolute Change in Percentile Scores		Pressure (std.)
	(1) within matching group	(2) within treatment	(3)
NAME	-0.01	-0.02	0.10
	(0.01)	(0.01)	(0.18)
PERFORMANCE	-0.04**	-0.04***	0.46**
	(0.02)	(0.01)	(0.15)
Gender/Grade/School FEs, Age	Yes	Yes	Yes
Other controls	No	No	Yes
N	588	588	161
R^2	.056	.051	.32
p-value: NAME vs. PERFORMANCE	.018	.085	.17
Mean in RANDOM	.15	.14	-.16

Notes: This table presents least squares regressions using absolute change in percentile scores or a standardized measure of pressure during the second run as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Absolute changes in percentile scores within matching groups are calculated based on the change of individual ranks of students in the their class and gender from the first to the second. Percentile scores within treatment are calculated for all students within the same treatment and gender (i.e., across classrooms). Other controls include the same controls as the mediation model in Table 1.6, where we use times and ranks from the second rather than the first run as the pressure variable has been elicited after the second run. Note that information on pressure was only elicited at one of the three schools.

Appendix 1.H Simulation of Matching Rules

We simulate three matching rules and predict their impact on performance improvements using our estimates from Table 1.6. In a first step, we create artificial pairs, based on the employed matching rules described below. In a second step, we then calculate the vector θ of differences for the artificial pairs as well as the matching quality of artificial peers. Finally, we use the estimated coefficients from the column (5) of Table 1.6 to predict the performance improvements we would observe for the artificial pairs. As peer-assignment rules only change θ, we are interested in the difference in the respective sums of the indirect effect and direct effect, that is between $\bar{\tau} + \beta\theta_i^{sim}$ and $\bar{\tau} + \beta\theta_i^{obs}$ from equation (1.2), where *sim* and *obs* denote simulated and observed pair characteristics, respectively. As we consider exogenous assignment rules, we assume that the direct effect of the simulated policies equals zero as in in Random. We additionally fix the covariates X to 0 and leave out the fixed effects for the simulations and predictions. This means, we calculate the performance improvements for a particular baseline group for our treatments as well as the simulations. This enables us to compare our results of the simulations directly to the peer-assignment rules using self-selection implemented in the experiment, as we compare the performance improvements for the same group.

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

Table 1.H.1. Overview of simulated peer assignment rules

Peer assignment rule	Mean absolut productivity differences (in sec)	Predicted improvement (in pp.)	Description
Name	2.09	2.43	Self-selected peers based on names
Performance	1.41	2.69	Self-selected peers based on relative performance
Name (exog.)	2.09	1.19	Self-selected peers based on names without self-selection effect
Performance (exog.)	1.41	0.48	Self-selected peers based on relative performance without self-selection effect
Random	2.42	1.12	Randomly assigned peers
Equidistance	3.11	1.44	Same distance in ranks across pairs
High-to-Low	3.11	1.36	First to last, second to second to last etc.
Tracking	0.90	0.72	First to second, third to fourth etc.

Appendix 1.I Experimental Instructions and Protocol

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

Introduction to the Experiment

Welcome everyone to today's physical education session. As you might have already noticed, today's session is going to be different. As you already know, you will take part in a scientific study. For that purpose, you received a parental consent form and handed it back to your teacher. If you have not handed it back to your teacher, you will not take part in the study.

The study is going to be conducted by the three of us: Lukas Kiessling, Sebastian Schaube and I am Jonas Radbruch. If you have any questions throughout the study, you can address us at any point in time.

The study comprises several parts. For the first part, we would like you to do a running task called suicide runs. My colleague will shortly demonstrate this exercise.

(The following verbal explanation was accompanied with physical demonstration of the exercise)

You start at the baseline of the volleyball court and run to to this first line. You touch it with your hand and run back to the baseline. You touch the baseline with your hand and run to the next line. Touch it again, back to the baseline; touch it, and then to the third line, back to the baseline, to the fourth line and then you return to the baseline.

Everyone of you will run alone and the goal is to be as fast as possible. After this run, we will hand you a computer to fill out a survey.

After all of you have ran and filled out the survey, you will run for a second time. This time at the same time as another student. During the survey we will ask you – among other questions – with whom you would like to run. You will receive detailed information about this later on.

The goal during both runs is to be as fast as possible. We will record your running times and hand it to your teacher. Your teacher will grade your performance during both runs.

Before we start with the study, we would like to remind you again that your participation is voluntary. If anyone does not want to take part in the study, then please inform us now.

Do you have any further questions? If this is not the case, please start with the warm-up, before we start with the experiment.

(After a short warm-up, all students were asked to leave the gym and wait in an accompanying the hallway until they were called in the gym to take part in the first run. We

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

Screenshots of the Preference Elicitation during the Survey

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

	4-5 seconds slower	3-4 seconds slower	2-3 seconds slower	1-2 seconds slower	0-1 seconds slower	Own time	0-1 seconds faster	1-2 seconds faster	2-3 seconds faster	3-4 seconds faster	4-5 seconds faster
1st Preference	○	○	○	○	○	X	○	○	○	○	○
2nd Preference	○	○	○	○	○	X	○	○	○	○	○
3rd Preference	○	○	○	○	○	X	○	○	○	○	○
4th Preference	○	○	○	○	○	X	○	○	○	○	○
5th Preference	○	○	○	○	○	X	○	○	○	○	○
6th Preference	○	○	○	○	○	X	○	○	○	○	○
7th Preference	○	○	○	○	○	X	○	○	○	○	○

Figure 1.I.1. Performance-based preferences

	ID of running mate	4-5 seconds faster	3-4 seconds faster	2-3 seconds faster	1-2 seconds faster	0-1 seconds faster	Own time	0-1 seconds slower	1-2 seconds slower	2-3 seconds slower	3-4 seconds slower	4-5 seconds slower
1st Preference	---------	○	○	○	○	○	X	○	○	○	○	○
2th Preference	---------	○	○	○	○	○	X	○	○	○	○	○
3rd Preference	---------	○	○	○	○	○	X	○	○	○	○	○
4th Preference	---------	○	○	○	○	○	X	○	○	○	○	○
5th Preference	---------	○	○	○	○	○	X	○	○	○	○	○
6th Preference	---------	○	○	○	○	○	X	○	○	○	○	○

Figure 1.I.2. Name-based preferences

Introduction to the Second Run for the Whole Class

(Class was gathered for announcement)

We will shortly start with the second run. For this purpose a partner for you has been selected. In your class, the partner has been selected *randomly [based on your indication how fast you want your partner to be] [based on the classmates you nominated]*. We would like to remind you that the objective is to be as fast as possible and it is only about your own time. Your teacher will receive a list with your performance, but no information about the pairs.

(The list with pairs was read out aloud to the students and students were accompanied to the waiting zone. Students were called into the gym one pair after the other. In the gym they were led to separate, but adjacent tracks. Each student was accompanied by one experimenter, who recorded their time as well their responses to four additional questions.)

Individual Introduction Directly Before the Second Run

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

Post-run Questionnaire after the Second Run

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

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

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

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

References

Ager, Philipp, Leonardo Bursztyn, and Hans-Joachim Voth. (2016). "Killer Incentives: Status Competition and Pilot Performance during World War II." Working Paper. [5]

Agostinelli, Francesco. (2018). "Investing in Children's Skills: An Equilibrium Analysis of Social Interactions and Parental Investments." Working Paper. [9]

Aldashev, Gani, Georg Kirchsteiger, and Alexander Sebald. (2017). "Assignment Procedure Biases in Randomised Policy Experiments." *Economic Journal* 127 (602): 873–895. [32]

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. (2005). "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools." *Journal of Political Economy* 113 (1): 151–184. [31]

Aral, Sinan, and Christos Nicolaides. (2017). "Exercise Contagion in a Global Social Network." *Nature Communications* 8 (14753): [7]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2009). "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica* 77 (4): 1047–1094. [5]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2010). "Social Incentives in the Workplace." *Review of Economic Studies* 77 (2): 417–458. [5, 9]

Bartling, Björn, Ernst Fehr, and Holger Herz. (2014). "The intrinsic value of decision rights." *Econometrica* 82 (6): 2005–2039. [8, 33]

Bartling, Björn, Ernst Fehr, and Klaus M. Schmidt. (2013). "Discretion, productivity, and work satisfaction." *Journal of Institutional and Theoretical Economics* 169 (1): 4–22. [8]

Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse. (2012). "The Relationship Between Economic Preferences and Psychological Personality Measures." *Annual Review of Economics* 4 (1): 453–478. [32]

Belot, Michèle, and Jeroen van de Ven. (2011). "Friendships and Favouritism on the Schoolground – A Framed Field Experiment." *Economic Journal* 121 (557): 1228–1251. [9]

Betts, Julian R. (2011). "The Economics of Tracking in Education." In. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3, Handbook of the Economics of Education. Elsevier, 341–381. [33]

Bó, Pedro Dal, Andrew Foster, and Louis Putterman. (2010). "Institutions and Behavior: Experimental Evidence on the Effects of Democracy." *American Economic Review* 100 (5): 2205–2229. [8]

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek. (2017). "Ability Peer Effects in University: Evidence from a Randomized Experiment." *Review of Economic Studies* 84 (2): 547–578. [8]

Bradler, Christiane, Robert Dur, Susanne Neckermann, and Arjan Non. (2016). "Employee Recognition and Performance: A Field Experiment." *Management Science* 62 (11): 3085–3099. [8]

Brandts, Jordi, David Cooper, and Roberto Weber. (2014). "Legitimacy, Communication, and Leadership in the Turnaround Game." *Management Science* 61 (11): 2627–2645. [8]

Bursztyn, Leonardo, Florian Ederer, Bruno Ferman, and Noam Yuchtman. (2014). "Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions." *Econometrica* 82 (4): 1273–1301. [9]

Carrell, Scott, Bruce Sacerdote, and James West. (2013). "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81 (3): 855–882. [8, 33, 60]

Cassar, Lea, and Stephan Meier. (2018). "Nonmonetary Incentives and the Implications of Work as a Source of Meaning." *Journal of Economic Perspectives* 32 (3): 215–38. [33, 36]

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." Working Paper. [7, 8, 20]

Chen, Roy, and Jie Gong. (2018). "Can self selection create high-performing teams?" *Journal of Economic Behavior & Organization* 148: 20–33. [7, 37]

Chevalier, Judith A., M. Keith Chen, Peter E. Rossi, and Emily Oehlsen. (Forthcoming). "The value of flexible work: Evidence from uber drivers." *Journal of Political Economy*, [8]

Cicala, Steve, Roland Fryer, and Jörg Spenkuch. (2018). "Self-Selection and Comparative Advantage in Social Interactions." *Journal of the European Economic Association* 16 (4): [7, 20]

Corgnet, Brice, Joaquín Gómez-Miñambres, and Roberto Hernán-González. (2015). "Goal Setting and Monetary Incentives: When Large Stakes Are Not Enough." *Management Science* 61 (12): 2926–2944. [8]

Deci, Edward, and Richard Ryan. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer Science & Business Media. [6, 8, 33, 36]

Deci, Edward, and Richard Ryan. (2000). "The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior." *Psychological Inquiry* 11 (4): 227–268. [6, 33]

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. (2011). "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–550. [12]

Duflo, Esther, Pascaline Dupas, and Michael Kremer. (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review* 101 (5): 1739–1774. [8, 33]

Elsner, Benjamin, and Ingo Isphording. (2017). "A Big Fish in a Small Pond: Ability Rank and Human Capital Investment." *Journal of Labor Economics* 35 (3): 787–828. [20, 58]

Falk, Armin, and Michael Kosfeld. (2006). "The Hidden Costs of Control." *American Economic Review* 96 (5): 1611–1630. [8, 33]

Friebel, Guido, Matthias Heinz, Mitchell Hoffman, and Nick Zubanov. (2019). "What Do Employee Referral Programs Do?" Working Paper. [37]

Fu, Chao, and Nirav Mehta. (2018). "Ability Tracking, School and Parental Effort, and Student Achievement: A Structural Model and Estimation." *Journal of Labor Economics* 36 (4): 923–979. [33]

Garlick, Robert. (2018). "Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment." *American Economic Journal: Applied Economics* 10 (3): 345–369. [8, 33]

Gibbons, Frederick, and Bram Buunk. (1999). "Individual Differences in Social Comparison: Development of a Scale of Social Comparison Orientation." *Journal of Personality and Social Psychology* 76 (1): 129–147. [12]

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse. (2019). “First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision.” *Management Science* 65 (2): 494–507. [20, 58]

Gneezy, Uri, and Aldo Rustichini. (2004). “Gender and Competition at a Young Age.” *American Economic Review* 94 (2): 377–381. [10, 60]

Golsteyn, Bart, Arjan Non, and Ulf Zölitz. (2017). “The Impact of Peer Personality on Academic Achievement.” Working Paper. [8, 20]

Harrison, Glenn, and John List. (2004). “Field Experiments.” *Journal of Economic Literature* 42 (4): 1009–1055. [6]

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.” *Econometric Reviews* 34 (1-2): 6–31. [18, 42, 43]

Herbst, Daniel, and Alexandre Mas. (2015). “Peer Effects on Worker Output in the Laboratory Generalize to the Field.” *Science* 350 (6260): 545–549. [7]

Huettner, Frank, and Marco Sunder. (2012). “Axiomatic arguments for decomposing goodness of fit according to Shapley and Owen values.” *Electronic Journal of Statistics* 6: 1239–1250. [25]

Irving, Robert. (1985). “An Efficient Algorithm for the “Stable Roommates” Problem.” *Journal of Algorithms* 6 (4): 577–595. [13]

Jackson, C. Kirabo, and Elias Bruegmann. (2009). “Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers.” *American Economic Journal: Applied Economics* 1 (4): 85–108. [9]

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube. (2019). “Determinants of Peer Selection.” Working Paper. [7, 17]

Koch, Alexander K., and Julia Nafziger. (2011). “Self-regulation through Goal Setting*.” *Scandinavian Journal of Economics* 113 (1): 212–227. [8]

Kosfeld, Michael, and Susanne Neckermann. (2011). “Getting More Work for Nothing? Symbolic Awards and Worker Performance.” *American Economic Journal: Microeconomics* 3 (3): 86–99. [8]

Lavy, Victor, and Edith Sand. (Forthcoming). “The Effect of Social Networks on Students’ Academic and Non-cognitive Behavioural Outcomes: Evidence from Conditional Random Assignment of Friends in School.” *Economic Journal*, [7]

Lazear, Edward, and Paul Oyer. (2012). “Personnel Economics.” In *The Handbook of Organizational Economics*. Princeton University Press, 479–519. [37]

Lazear, Edward, and Kathryn Shaw. (2007). “Personnel Economics: The Economist’s View of Human Resources.” *Journal of Economic Perspectives* 21 (4): 91–114. [37]

Levitt, Steven, John List, Susanne Neckermann, and Sally Sadoff. (2016). “The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance.” *American Economic Journal: Economic Policy* 8 (4): 183–219. [8, 36]

Manski, Charles. (1993). “Identification of Endogenous Social Effects: The Reflection Problem.” *Review of Economic Studies* 60 (3): 531–542. [7]

Mas, Alexandre, and Enrico Moretti. (2009). “Peers at Work.” *American Economic Review* 99 (1): 112–145. [5, 9]

Murphy, Richard, and Felix Weinhardt. (2018). “Top of the Class: The Importance of Ordinal Rank.” Working Paper. [20, 58]

Oster, Emily. (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204. [30, 31]

Owens, David, Zachary Grossman, and Ryan Fackler. (2014). "The Control Premium: A Preference for Payoff Autonomy." *American Economic Journal: Microeconomics* 6 (4): 138–161. [8, 33]

Rotter, Julian B. (1966). "Generalized Expectancies for Internal Versus External Control of Reinforcement." *Psychological Monographs: General and Applied* 80 (1): 1–28. [12]

Sacerdote, Bruce. (2001). "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116 (2): 681–704. [5]

Sacerdote, Bruce. (2011). "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?" In. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 3, Handbook of the Economics of Education. Elsevier, 249–277. [7]

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." *DIW Data Documentation*, [12]

Sutter, Matthias, and Daniela Glätzle-Rützler. (2015). "Gender Differences in the Willingness to Compete Emerge Early in Life and Persist." *Management Science* 61 (10): 2339–2354. [10]

Sutter, Matthias, Stefan Haigner, and Martin G. Kocher. (2010). "Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations." *Review of Economic Studies* 77 (4): 1540–1566. [8]

Tincani, Michela. (2017). "Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence." Working Paper. [7, 20]

Weinhardt, Michael, and Jürgen Schupp. (2011). "Multi-Itemskalen im SOEP Jugendfragebogen." *DIW Data Documentation*, [12]

Chapter 2

Determinants of Peer Selection

Joint with Jonas Radbruch and Sebastian Schaube

2.1 Introduction

Peer effects have been documented across many different environments: skills of classmates influence grades at school (Sacerdote, 2011), co-workers affect own performance both in highly stylized settings (Falk and Ichino, 2006) as well as workplaces (Bandiera, Barankay, and Rasul, 2009; Mas and Moretti, 2009). Peers, therefore, constitute a crucial determinant of our performance. But who are these people that we select as peers? Answering this question is important for the design of policies that exploit social comparisons in educational contexts or firms. The successful implementation of such policies presupposes an understanding of the formation and composition of reference groups (e.g., Carrell, Sacerdote, and West, 2013; Kőszegi, 2014; Manski, 1993). However, we do not know much about the underlying process of peer selection. Several, sometimes conflicting determinants of peer selection are conceivable. If high-performing peers serve as a reference point, they can motivate individuals to exert more effort (e.g., Abeler, Falk, Goette, and Huffman, 2011; Koch and Nafziger, 2011). Others choose peers to compete or they select specific friends, as they might make a task more enjoyable (Park, 2019).

In this paper, we study the selection of peers and link these choices to three potential determinants. Specifically, we measure with whom individuals want to interact and analyze the extent to which these preferences for specific peers depend on (i) relative performance, (ii) personality differences, and (iii) the presence of friendship ties.[1] By studying these determinants, we can quantify the magnitudes of performance and social aspects in the peer selection process and examine their relationship. In this paper, we therefore highlight the role of individual characteris-

1. We differentiate between friends and peers as two distinct, albeit related concepts. While friends can be peers, not all friends have to be peers across all situations.

tics for reference group formation. In doing so, we provide a microfoundation for models of differential and nonlinear peer effects.

In order to study the selection of peers, we use data from a framed field experiment with over 600 students aged 12 to 16. In the experiment, students took part in two running tasks, first alone, then simultaneously with a peer. Between the two runs, we collected two different types of preferences for peers, which were subsequently used to form pairs for the second run. More specifically, we elicited students' preferences for peers by allowing them to name up to six classmates with whom they would like to be paired (*name-based preferences*) or choose their peer's relative performance (*performance-based preferences*). Moreover, we elicited personality measures and the social network within each class. Our setup thus has four crucial features to analyze peer selection in detail. First, the classroom environment enabled students to state meaningful preferences for known peers (*name-based preferences*) allowing for social aspects. Second, using a running task yields direct measures of performance. This allows us to isolate preferences over the relative performance of peers (*performance-based preferences*), creating a preference measure for peers that abstracts from social considerations. Third, our analysis relies on preference measures. This overcomes the notion that preferences for peers may not necessarily be satisfied in observed selection outcomes, e.g. due to the limited availability of peers. Fourth, by focusing on a single peer in the second run, we circumvent issues associated with multiple reference points (Kahneman, 1992) as students interact with one peer only.

Our analysis proceeds in three steps. First, we describe the heterogeneity in both preference measures, finding that friendship ties play a crucial role. About 80% of the three most-preferred name-based peers are friends. Nonetheless, this figure declines to less than 50% when considering the fifth or sixth ranked peer. Moreover, we observe that students on average prefer slightly faster peers (0.20 SD in terms of performance in the first run). However, this masks large heterogeneities in performance-based preferences. Approximately half of the students want to interact with similar (slightly faster or slower) students. The other half prefer peers who differ in their relative performance by more than one second.

In a second step, we study the determinants of peer selection based on names. In particular, we consider the extensive – whom to select – as well as the intensive margin, namely the ranking of peers. We estimate the extent to which peer selection patterns can be explained by differences in past performance, differences in personality, and the presence of friendship ties. We find that all three dimensions matter, although friendship ties are the most important determinant. If two students are friends, this increases their nomination probability (rank) by 39 percentage points (1.7 ranks). Moreover, we find substantial homophily in both past performance as well as personality. Accordingly, students select peers with whom they are similar. A one standard-deviation difference in past performance (difference in personality) reduces the probability of selecting a given classmate by approximately 6 percentage points (4.5 pp.) corresponding to 0.38 ranks (0.30 ranks). These homophily effects hold conditional on friendship ties: students select those friends as peers who are close to them with respect to personality and performance. Moreover, our results uncover heterogeneities across sub-groups. We show that the importance of these dimensions differs between males and females as well as high- and low-ability students. In particular, male subjects exhibit a stronger homophily in performance than female subjects.

In a third step, we explore the relationship between performance- and name-based preferences. Our results show that when students select peers based on names, they try to target their preferred relative performance level. This demonstrates that subjects nominate similar performing peers not only due to homophily, but also due to preferences over relative performance. The social dimensions of peer selection remain unaffected, which highlights the multidimensionality of preferences for peers.

This paper relates to the rich literature on peer effects. Although their importance is undisputed, evidence on whom people select as peers remains scarce. Yet, Manski (1993, p. 536) already noted that the "*informed specification of reference groups is a necessary prelude to [the] analysis of social effects*". This implies that studies on peer effects have to take a stance on who constitutes a reference or peer group, thus specifying who exerts potential peer effects. For example, it is common to specify the set of classmates or co-workers as reference groups on an ad-hoc basis. However, only parts of these groups may constitute relevant peers and misspecifications thereof attenuate peer effect estimates due to measurement error (Cornelissen, Dustmann, and Schönberg, 2017; Dube, Giuliano, and Leonard, 2019). In order to circumvent this problem and accommodate different peer definitions, a growing body of literature estimates peer effects for different groups separately, differentiating between genders (Beugnot, Fortin, Lacroix, and Villeval, 2019; Black, Devereux, and Salvanes, 2013; Hoxby, 2000; Lavy and Schlosser, 2011) or allowing friends and non-friends to exert different peer effects (Aral and Nicolaides, 2017; Bandiera, Barankay, and Rasul, 2009; Bandiera, Barankay, and Rasul, 2010). We document that friendship is the most important determinant for peer selection, thereby validating the use of friends as a proxy for peers.[2] Moreover, our results show that people exhibit systematic peer choice patterns. This suggests that only a subset of people serve as peers and affect behavior. In particular, this motivates the separate estimation of peer effects for different sub-groups and demographic characteristics, i.e., differential peer effects.[3] Relatedly, an individual's impact may differ across the abil-

2. A more general interpretation of this finding is that individuals prefer familiar peers and coworkers. Support for this stems from observations that only those coworkers with whom individuals have sufficiently large overlaps in working time exert peer influences (Mas and Moretti, 2009) and that familiarity between coworkers has positive effects on workers' performance more generally (Huckman, Staats, and Upton, 2009).

3. In principle, differential peer effects can be due to (i) only some individuals being relevant peers, (ii) only some individuals exerting peer effects, or a combination of both.

ity distribution: it might be large on classmates or co-workers with similar abilities, whereas for others with vastly different ability levels the effect might be small. Nonlinear peer effects implicitly incorporate these patterns of peer selection since they allow different individuals (e.g., in terms of their ability) to exert different effects (Burke and Sass, 2013; Mas and Moretti, 2009; Tan and Netessine, forthcoming).[4]

In general, individuals often self-select into workplaces or organizations based on institutional characteristics or individual traits. For example, employees select into workplaces based on latter's characteristics (e.g., incentive schemes, Dohmen and Falk, 2010; Niederle and Vesterlund, 2007), students and their parents choose schools based on the academic performance of the school (Burgess, Greaves, Vignoles, and Wilson, 2014), and individuals sort into occupations and organizations based on individual traits (e.g., prosociality, Carpenter and Myers, 2010; Friebel, Kosfeld, and Thielmann, forthcoming). We advance this literature by studying the process of peer selection within those organizations or social groups. In a similar vein, Cicala, Fryer, and Spenkuch (2018) study how students choose peer groups by sorting into specific tasks based on their comparative advantage. Our approach differs and links peer selection to social and non-social determinants to investigate how individuals weight these. By this, our paper adds to a growing literature modeling the selection of friends and the formation of social networks (see for an overview Graham, 2015; McPherson, Smith-Lovin, and Cook, 2001). However, we deliberately differentiate between friends and peers as two distinct, albeit related concepts. Friendship ties may be one factor that determines whether to choose a specific individual as a peer. Although it is quantitatively the most important factor in the peer selection process, it is neither a necessary nor sufficient indicator for actual peer choices. Methodologically, we adopt a similar framework to Girard, Hett, and Schunk (2015). Whereas they study friendship formation at a university and find homophily in several personality traits and economic preferences, we focus on peer selections within established social networks and allow – among other factors – friendship ties to affect these.

Our results help to develop a deeper understanding of the selection process for peers – or reference group formation more generally – which can be levered to design successful policy interventions. By reorganizing teams, organizations or groups, policy-makers can change the availability of potential peers and thereby channel peer interactions. By providing suitable peers, they can exploit the resulting effects (e.g., Roels and Su, 2014). The findings presented here might help to design policies and incentive contracts incorporating social interactions (Carrell, Sacerdote, and West, 2013; Kőszegi, 2014). We show that their effects potentially differ across sub-

4. Policymakers or employers can potentially exploit these nonlinearities to optimally assign peers or teams (Bhattacharya, 2009), although the consequences of such reassignments vary across studies (Booij, Leuven, and Oosterbeek, 2017; Carrell, Sacerdote, and West, 2013; Kiessling, Radbruch, and Schaube, 2018).

groups. This suggests, for example, that high-ability individuals select more similar peers compared to low-ability individuals, who place more emphasis on friendships. Therefore, separating students or workers by ability should only slightly change high-ability individuals' reference groups. However, it might have larger effects for their low-ability counterparts, affecting subsequent behavior differentially.

Our evidence on the determinants of peer selection informs the literature on the specification and formation of reference points (see for an overview O'Donoghue and Sprenger, 2018). Selected peers can serve as an "aspiration level" or goal that constitutes a reference point, as introduced by Kahneman and Tversky (1979) and used similarly for example by Brookins, Goerg, and Kube (2017) and Koch and Nafziger (2011). Some studies (e.g., Cerulli-Harms, Goette, and Sprenger, 2019; Schwerter, 2016) debate the nature and the location of reference points. We add to this literature by demonstrating that social reference points can arise endogenously through peer selection. Heterogeneous preferences over peers highlight how reference points are linked to individual characteristics.

In our experiment, we induce peer effects by allowing for social comparisons. We analyze explicitly individual preferences for social comparisons and do not focus on their consequences.[5] Only a handful of papers study to whom people compare: while some studies (Clark and Senik, 2010; Knight, Song, and Gunatilaka, 2009) find that people compare themselves to friends, co-workers or neighbors, others focus on comparisons along performance levels (Falk and Knell, 2004) or with one's own past (Senik, 2009). By contrast, ours is the first study to combine preferences along social dimensions with information about preferences for relative performance and the personality of the peer.

The remainder of the paper is structured as follows. The next section presents the data and describes our sample. Section 2.3 documents two kinds of preferences for peers, based on relative performance and names. We analyze the general determinants of the name-based preferences in Section 2.4. Finally, Section 2.5 concludes.

2.2 Data

In most environments, it is difficult to observe with whom people compare their own performance. This is especially difficult when there is not a single peer available as an objective standard but rather when several peers are observed at the same time. Additionally, peer selection may not only be based on preferences over some target performance; rather, it is potentially based on a much broader set of peers' characteristics.

5. Social comparisons may harm effort provision and work performance (Ashraf, Bandiera, and Lee, 2014; Cohn, Fehr, Herrmann, and Schneider, 2014), reduce job satisfaction (Card, Mas, Moretti, and Saez, 2012), change consumption patterns (Kuhn, Kooreman, Soetevent, and Kapteyn, 2011), and negatively affect happiness and overall well-being (Clark and Senik, 2010).

In this paper, we use the dataset of a framed field experiment studying the self-selection of peers (Kiessling, Radbruch, and Schaube, 2018) to overcome these difficulties. The experiment elicited preferences for peers in a sample of over 600 students and thus allows us to study the peer selection process. In addition to these preferences, the experiment elicited the social network and several personal characteristics.

2.2.1 Experiment

The experiment was embedded into physical education classes in German secondary schools. Subjects participated in two suicide runs, each comprising a series of short sprints along the lines of a volleyball court[6]: first, at the beginning of the experiment alone, then at the end of the experiment simultaneously with a peer. No other classmates were presented during the first or second run. For the second run, we randomly assigned classes to one of three treatment conditions, which implemented different peer assignment rules: random assignment, self-selection based on names, or self-selection based on relative performance. The treatments with self-selection of peers used the elicited preferences for peers to assign students into pairs for the second run. For this, we implemented a "stable roommate" algorithm proposed by Irving (1985) to form stable pairs. Hence, in order to be matched with their most-preferred possible peer, students had to reveal their true preferences. Students were matched within their own gender only. In order to incentivize students in both runs, we reported the individual times to teachers for grading. Moreover, students themselves were intrinsically motivated expressed by a strong interest in their individual time. Between the two runs, subjects participated in a survey. In addition to sociodemographics, the survey asked students to reveal their preferences for peers according to two dimensions and elicited several personal characteristics as well as the social network of the class. In the following, we describe each of these survey elements in more detail.

2.2.2 Preference Elicitation

The survey elicited two distinct measures for peer preferences, which were used to implement self-selected peers in the experiment. First, we elicited preferences for situations solely based on relative performance (*performance-based preferences*). Second, we asked for preferences for those settings in which social information is

6. The exact task was to sprint and turn at every line of the volleyball court. Subjects had to line up at the baseline, from where they started running to the first 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 final sprint 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.

available (*name-based preferences*). These preferences were elicited for the whole sample and independent of the treatment itself, as the treatment was only assigned after the survey took place. Note that these preferences are revealed, rather than stated preferences. In particular, there was a positive probability that these preferences were taken into account due to the random assignment of treatments after the survey.

We first discuss the elicitation of preferences for peers based on relative performance. For this purpose, the survey presented subjects with ten categories comprising one-second intervals starting from $(4, 5]$ seconds slower than their own performance in the first run, to $(0, 1]$ seconds slower and $(0, 1]$ seconds faster up to $(4, 5]$ seconds faster. We present a screenshot of the elicitation procedure in Figure 2.1. Subjects indicated from which relative performance interval they would prefer a peer for the second run, irrespective of the potential peer's identity. This means the students could not base their decision on any characteristics besides the relative performance. In the first row of the table, subjects indicated their most-preferred time interval and thereby the peer's relative performance. In the second row, they indicated their second most-preferred interval, and so forth. The preference for peers based on relative performance corresponds to the highest ranked time interval. We asked students to rank their seven most-preferred time intervals and therefore elicited a partial ranking of potential peers for performance-based preferences. Naturally, each time interval could only be chosen once, but it potentially included several peers. Similarly, some intervals might have been empty.

	4-5 seconds slower	3-4 seconds slower	2-3 seconds slower	1-2 seconds slower	0-1 seconds slower	Own time	0-1 seconds faster	1-2 seconds faster	2-3 seconds faster	3-4 seconds faster	4-5 seconds faster
1st Preference	○	○	○	○	○	X	○	○	○	○	○
2nd Preference	○	○	○	○	○	X	○	○	○	○	○
3rd Preference	○	○	○	○	○	X	○	○	○	○	○
4th Preference	○	○	○	○	○	X	○	○	○	○	○
5th Preference	○	○	○	○	○	X	○	○	○	○	○
6th Preference	○	○	○	○	○	X	○	○	○	○	○
7th Preference	○	○	○	○	○	X	○	○	○	○	○

Figure 2.1. Screenshot of the survey question on performance-based peer preferences

Notes: The figure presents a screenshot of the survey module eliciting the preferences over relative performance. In particular, it elicits a partial ranking of ten categories of relative ability ranging from 4 to 5 seconds slower to 4 to 5 seconds faster.

The second preference measure elicited preferences for situations in which selection can be based on the identity of the peer (*name-based preferences*), i.e., subjects could condition their decision on all known characteristics of their peers. We asked each student to state his or her six most-preferred peers from the same gender within their class. These classmates had to be ranked, creating a partial ranking of their peers.

When subjects nominated a student, they were asked to indicate their belief about the relative performance of the person. The belief elicitation was similar to that of the performance-based preferences described above: subjects had to indicate their beliefs about the performance of the potential peer in the first run using the same ten intervals and the same layout.

2.2.3 Personal Characteristics and Social Network

The survey also included several measures for personality traits 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[7], general risk attitude (Dohmen, Falk, Huffman, Sunde, Schupp, et al., 2011), and a short version of the INCOM scale for social comparison (Gibbons and Buunk, 1999; Schneider and Schupp, 2011). For each multiple item scale, we extracted one underlying factor with a mean of zero and a standard deviation of one.

At the end of the survey, we elicited the social network of the class.[8] The elicitation asked every student to name up to six friends in their class. Due to this constraint, we focus on undirected links. We define that friendship ties exist between person i and j if j was either nominated by student i as a friend, or j herself nominated i as a friend. This means that students can have more than six friends if they were nominated by participants who they did not nominate themselves.[9]

7. Rather than using tournament entry decisions as measures of competitiveness, we introduced a continuous measure based on a student's agreement to four items on a seven-point Likert scale. The statements were: (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" (reversely coded). We then extracted a single principal component factor from those four items

8. As preferences were elicited as the first part of the survey, this ordering induced the maximum possible time lag between the two elicitations. This makes potential spillovers between these two measures unlikely.

9. About 79% of the students nominated six friends. Thus, we were concerned that a maximum of six friends might be restrictive and accordingly define friendships as undirected rather than directed links. In robustness checks, we explore different friendship definitions, whereby our results are robust to using different definitions, such as directed and reciprocal friendships.

2.2.4 Summary statistics

We present summary statistics of our sample in Table 2.1. Overall, we have preference measures and the social network for 619 individuals from 39 classes of grades 7 to 10 (aged 12 to 16) with 66% of students being female.[10] This amounts to 73% of all students in a class participating in the experiment.[11] The average class size is about 26 and students have approximately seven friends on average, with 80% of those friends being from a student's own gender. On average, females took 27.57 seconds to finish the first run, which does not vary by age. By contrast, male performance improves with age: while the average time of males in grade 7 is 25.33 seconds, it improves to 23.21 seconds in grade 10.

Table 2.1. Summary statistics

	7th grade	8th grade	9th grade	10th grade	Total
Sociodemographic Variables					
Age	12.77	13.80	14.76	15.82	14.51
	(0.48)	(0.45)	(0.39)	(0.53)	(1.22)
Female	0.60	0.61	0.66	0.73	0.66
	(0.49)	(0.49)	(0.47)	(0.45)	(0.47)
Number of friends	6.93	7.18	7.01	6.50	6.86
	(1.35)	(1.75)	(1.57)	(1.70)	(1.63)
Share of friends of own gender	0.84	0.75	0.85	0.75	0.80
	(0.19)	(0.24)	(0.20)	(0.26)	(0.23)
Times (in sec)					
Time 1 (Females)	28.03	27.06	27.32	27.81	27.57
	(2.75)	(2.06)	(2.28)	(2.71)	(2.50)
Time 1 (Males)	25.33	24.18	23.60	23.21	24.04
	(1.93)	(2.02)	(1.82)	(2.11)	(2.11)
Class-level Variables					
# Students in class	25.54	25.97	26.29	25.01	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)
Observations	123	122	179	195	619

10. These classes are from three Germany secondary schools from the highest track, preparing students for university entry after grade 12 (*Gymnasien*). The smallest of the three schools is a female-only school, resulting in the somewhat increased share of females in our sample.

11. Only those students who submitted parental consent forms prior to the experiment, who did not choose to abstain from the study (which nobody did), and who were not absent from the physical education lesson took part in the study. Since students did not know the exact date where the study took place, we do not have any concerns about study-related absences from the classes.

2.3 Preferences for Peers

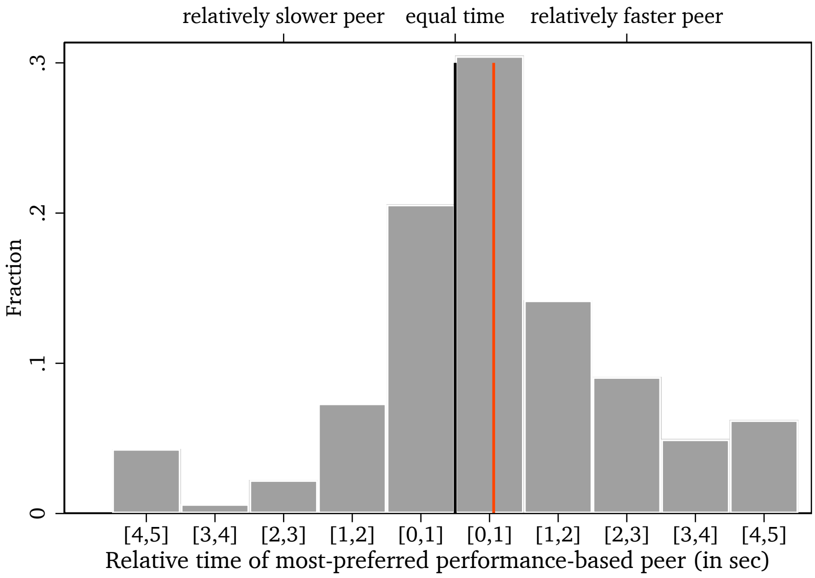
In this section, we describe two types of preferences for peers: first, students could select their most-preferred relative performance (*performance-based preference*); and second, students could select their preferred peers based on names (*name-based preferences*), allowing students to condition their peer choice on all characteristics known to them. These two distinct preference measures allow us to describe students' peer selection, namely who they prefer as a peer.

2.3.1 Performance-based Preferences

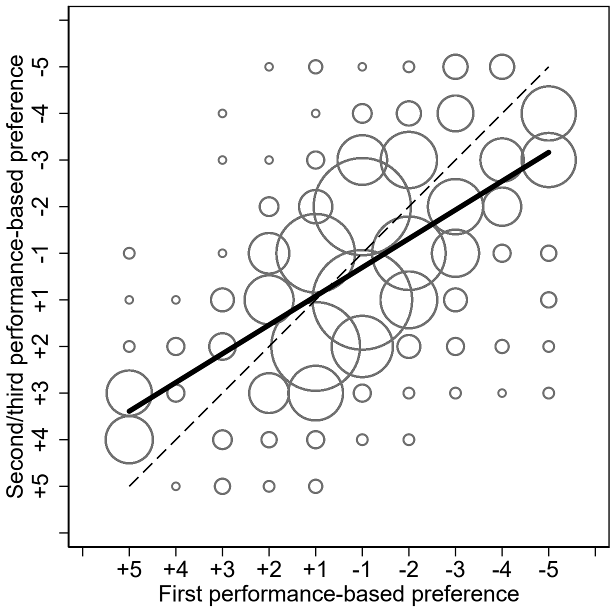
As described in section 2.2, we elicited a partial ranking over ten categories, with each category corresponding to a one-second time interval of relative performance. Figure 2.1 presents the preferences for the relative performance of peers. First, turning to the distribution of the most-preferred relative performance (Figure 2.1a), we find that students prefer performances from the entire possible set. Some students prefer peers who are 4 to 5 seconds slower, whereas others prefer peers who are up to 4 to 5 seconds faster than their own performance. Second, around half of the students prefer similar performing peers, i.e., their most-preferred peer has a performance within one second of their own performance in the first run. Finally, the majority of students prefers faster peers: the median of the distribution lies in the category with slightly faster peers and on average students prefer peers who were .56 seconds faster in the first run, corresponding to .20 SD in terms of performances in the first run. Figure 2.1b shows the relationship of the first performance-based preference with the second and third one. We observe that the second and third preference are centered around the first performance-based preference.[12] Moreover, Appendix Figures 2.A.2a and 2.A.2b reveal that the distributions across genders is similar, with males preferring somewhat faster peers than females: while males prefer peers who are .90 seconds faster (.31SD in terms of performances in the first run), females select peers who are .38 seconds (.13SD) faster.

In general, these preferences partially support the conjecture of Festinger (1954, p. 121) that people compare themselves with others who are "close to [their] own ability" and are in line with evidence from other disciplines noting tendencies to engage in upward comparisons (e.g. Huguet, Dumas, Monteil, and Genestoux, 2001).

12. In Appendix Figures 2.A.1a and 2.A.1b, we present the distributions of the second and third highest ranked interval. While the probability mass in these histograms is shifted away from an individual's own performance, this is simply an artifact of the limited number of categories, as can be seen in Figure 2.1b. The categories in which students preferred a much faster or much slower peer as the first preference naturally show a different pattern due to censoring. This explains why we do not find a perfect relationship with a slope of 1. When estimating a Tobit model accounting for censoring at the lower and upper limit, the regression coefficient on the second preferences is .97 with a standard error of .05 and we cannot reject that the coefficient equals unity.



(a) Distribution of preferences



(b) Relationship among preferences

Figure 2.1. Preferences for relative performance

Notes: Figure (a) presents a histograms of students' preferences over relative performance. The intervals used here and in the survey are one-second intervals of relative performances in the first run. Vertical lines indicate own performance (black; equals zero by definition) and mean preference (red; where we used the midpoint of each interval to calculate the mean). Figure (b) presents the relationship of the first performance-based preference and the second/third preference.

Nonetheless, this does not hold for all of our subjects. In particular, there is a sizable share of students preferring peers who do differ in ability.

2.3.2 Name-based Preferences

The second set of preferences allows students to state their preferences by selecting peers from a list of their classmates' names. In contrast to performance-based preferences, in principle students can take into account all information known to them when selecting their preferred peer.

Table 2.1. Share of name-based preferences who are friends

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

Notes: This table presents the share of nominated peers for each of the six name-based preferences elicited in the survey who are friends.

Table 2.1 presents the share of selected peers who are also friends of an individual. While 89% of all individuals select a friend as their most-preferred peer, this number decreases by about 10 percentage points for each of the following, lower ranked nominations. This pattern might be partially driven by the fact that students do not have a sufficient number friends of the same gender in the class who they can select. Nonetheless, our data shows that students have on average about seven friends, of which 78% are of their own gender, implying that students on average have 5.3 same-sex friends who they could select (see Table 2.1). Thus, this finding shows that students predominately consider their friends as peers, which is also confirmed by our more formal analysis below. However, they do not solely choose their peers based on friendship ties. Some students seem to avoid some of their friends in favor of other class members.

2.4 Determinants of Peer Selection

In order to more formally explore the underlying determinants of peer selection, we analyze how the three fundamental dimensions – performance, personality, and friendship – affect who is selected as a peer and quantify the relative importance. For our analysis, we use a nomination model similar to the social network formation literature (e.g., Girard, Hett, and Schunk, 2015). As students could nominate more than one potential peer and had to rank them, we can analyze the event that someone is nominated in the name-based preference elicitation and additionally study their rank among the selected peers. We therefore investigate the extensive and intensive margins of the selection process and highlight associated heterogeneities. In a second step, we look at the role of one determinant – the preferences for a relative performance – in greater detail. In particular, we analyze the extent to which students target their performance-based preferences.

2.4.1 Empirical Strategy

In order to analyze the determinants of peer selection in a structured way, we proceed in two steps. First, we analyze the extensive margin of peer selection. Let y_{ij} equal one if individual i nominates individual j and zero otherwise. The dataset therefore contains one observation for each possible nomination within a group. In our main analysis, we define a person to be selected as a peer if this person is part of the first three nominated name-based peers, i.e., if she is one of the three students who somebody would be most willing to be paired with in the second run.[13] We want to understand the extent to which i's nomination of j depends on three determinants: (i) differences in terms of performance in the first run ($\Delta^t(t_i, t_j)$), (ii) differences in personality ($\Delta^p(p_i, p_j)$), and (iii) the presence of friendship ties (F_{ij}). Additionally, we allow for individual-level heterogeneity in terms of observed and unobserved characteristics by including either individual characteristics ($\Omega_{ij} = \lambda X_i + \pi X_j$) or individual-level fixed effects ($\Omega_{ij} = \nu_i + \nu_j$) as well as some idiosyncratic shock (ϵ_{ij}) for each nomination. Our main specification is therefore given by:

$$y_{ij} = \underbrace{\alpha \Delta^t(t_i, t_j)}_{\substack{\text{Differences} \\ \text{in performance}}} + \underbrace{\beta \Delta^p(p_i, p_j)}_{\substack{\text{Differences} \\ \text{in personality}}} + \underbrace{\gamma F_{ij}}_{\substack{\text{Friendship} \\ \text{ties}}} + \underbrace{\Omega_{ij}}_{\substack{\text{Controls for} \\ \text{heterogeneity}}} + \epsilon_{ij} \tag{2.1}$$

In our application, we measure differences in terms of the Euclidean distance of the respective characteristic. Hence, similarity in terms of past performance is measured by the absolute distance $\Delta^t(t_i, t_j) = |t_i - t_j|$. In order to measure the difference in personality, we combine the set of standardized personality measures elicited in the survey (Big Five, locus of control, competitiveness, attitudes to engage in social comparisons and risk attitudes) to define the distance $\Delta^p(p_i, p_j) = \sqrt{\sum_k (p_{ik} - p_{jk})^2}$ with k indexing different personality measures.[14] Therefore, the coefficients α and β can be interpreted as the influence of differences in past performance and personality on the likelihood of nominating someone as a peer. Negative coefficients ($\alpha < 0$,

13. Accordingly, we define $y_{ij} = 1$ if and only if j is nominated in i's first three name-based preferences and $y_{ij} = 0$ otherwise. Given that groups were normally not very large and – as shown in Kiessling, Radbruch, and Schaube (2018) – 81% of students were matched with one of their first three preferences in the name-based matching, we consider those individuals as the most important ones. Our results are robust to this cut-off. In the Appendix, we relax this definition and consider different cut-offs. Panel A of Appendix Table 2.B.2 presents the results and shows that they qualitatively and quantitatively similar.

14. In robustness checks, we allow each of these personality measures to enter separately to explore what is driving the estimated effects. The advantage of the index is that it reduces the degrees of freedom and yields a single coefficient, which makes the impact of personality easily comparable to absolute differences in performances.

$\beta < 0$) provide evidence of homophily, namely the tendency of individuals to select others with similar characteristics (McPherson, Smith-Lovin, and Cook, 2001). Similarly, positive coefficients ($\alpha > 0$, $\beta > 0$) support heterophily, namely the tendency to avoid others who are similar.

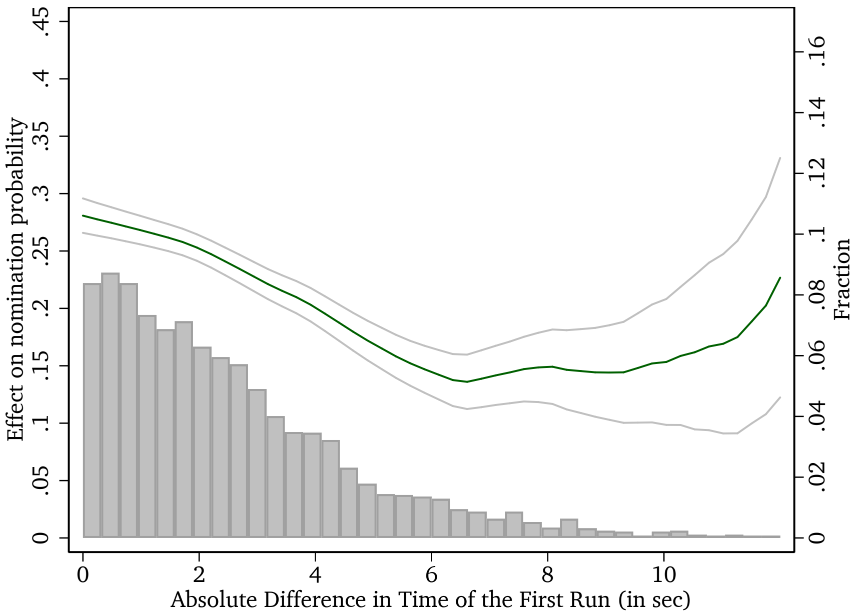
In a second step, we study the intensive margin of peer selection. We adopt the same specification as for the extensive margin (equation (2.1)), with two crucial modifications: first, we restrict the sample to all individuals who have been nominated as peers; and second, we change the dependent variable to be *j*'s rank in *i*'s preferences. For this, we define y_{ij} to the rank that individual i assigns individual j in the nomination process. The highest ranked peer receives a score of 6 and this score decreases by one with each rank in the preferences.[15]

2.4.2 Extensive Margin of Peer Selection

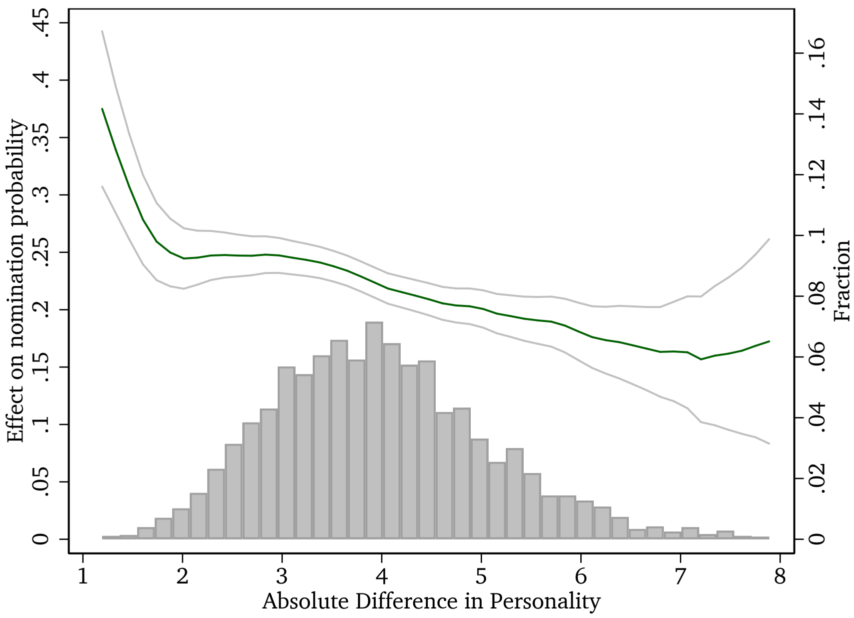
We begin our analysis by studying the extensive margin of peer selection, i.e., who individuals select as peers. Figure 2.1 provides first evidence of systematic peer selection patterns. Figure 2.1a shows that as the difference in initial performance between two individuals increases, the likelihood of nominating the other as a peer decreases. Similarly, Figure 2.1b shows a similar trend for differences in personality. Taken at face value, these relationships point towards homophily in both performance and personality. Yet, these associations could be driven by a common underlying factor (e.g., friendship ties) and potentially measure the same effect.

In order to disentangle the contribution of different factors in the peer selection process, Panel A of Table 2.1 presents a more structured analysis of the extensive margin. In particular, we estimate equation (2.1) using own and peer characteristics as well as class fixed effects in column (1), as well as individual and peer fixed effects in column (2). The results show that friendship ties are the most important determinant of peer selection. If two students are friends, this increases the nomination probability by 38 percentage points. However, we also find evidence of homophily in terms of both performances in the first run as well as personality. According to the estimates in column (2), a one-second difference in past performance or a difference of one standard deviation in personality reduces the probability of nominating a person by 3-4 percentage points. While these effects initially seem modest compared to the effect of friendship ties, it is necessary to take into account the underlying distributions of these variables. Conditional on friendship ties, increasing the absolute difference of performances in the first run by one standard deviation (2.10 sec) reduces the nomination probability by 6.3 percentage points. Similarly, increasing the difference in personality by one standard deviation reduces nomination probability

15. The exact score does not matter for our estimates, as the level is taken out by individual fixed effects. For the interpretation of the results, it is important to note that there is a difference of one between the scores.



(a) Effect of performance differences



(b) Effect of personality differences

Figure 2.1. Extensive margin of peer selection

Notes: These figures present local linear regressions of peer nominations on (a) absolute differences in initial performance and (b) absolute differences in personality including 95% confidence intervals. The underlying histograms show the distribution of the respective regressor.

by 4.5 percentage points.[16] Moreover, comparing columns (1) and (2) reveals that controlling for unobserved individual-level heterogeneity is important. Individual fixed effects allow us to capture this heterogeneity and thus controls for e.g., the popularity of students, which is otherwise unmeasured.

In order to understand the relationship between those three dimensions of peer selection, we analyze their interactions in column (3). We find that differences in performance and personality do not interact and seem to be independent, whereby the resulting coefficient is close to zero and precisely estimated. Although the coefficient of friendship ties interacted with absolute differences in personality is negative – suggesting stronger homophily in personality among friends – this effect is insignificant at conventional levels. Interestingly, we find that existing friendship ties increase the importance of differences in past performance. The homophily among friends almost doubles from 3 percentage points to 5.4 percentage points for a one-second difference in initial performance. Additional support for these results are presented in column (4). Here, we restrict the sample to the set of friends and thus ask whether the effects carry over to selection among friends. Homophily effects remain significant and even increase in magnitude. Hence, the peer selection effects estimated here are distinct from homophily that is often present in friendship formations (e.g., Girard, Hett, and Schunk, 2015; Selfhout, Burk, Branje, Denissen, Van Aken, et al., 2010).[17] Even conditional on being part of someone's social network, students select only those friends as peers who share similar characteristics.

To understand which personality facets are driving the results, we decompose aggregate impact of personality in Appendix 2.B.3 by allowing all personality measures to enter the model separately. The results show that the effect mainly stems from homophily in agreeableness, tendencies to engage in social comparisons, and – to a lesser extent – competitiveness. Importantly, the coefficients on absolute differences in performances of the first run and the presence of friendship ties remain constant, indicating that the aggregation to a single distance measure does not seem to be restrictive. Moreover, we consider different definitions of friendship ties. While in our main specification of Table 2.1 we defined friendship ties as undirected, we consider directed and reciprocal friendships in Appendix Table 2.B.4. The coefficient on the friendship indicator increases when using those alternative definitions, which arguably measure more intense friendships, although coefficients on absolute differences in performance and personality remain unaffected. This is reassuring as it alleviates the concern that the homophily terms in peer selection are mere artifacts of different friendship intensities.

16. Appendix Table 2.B.1 presents summary statistics of the absolute differences in these characteristics.

17. Appendix Table 2.B.1 documents that the average absolute difference is only slightly smaller for the sets of friends relative to the overall sample, indicating only a modest degree of homophily in *friendship* nominations.

Table 2.1. Extensive and intensive margin of peer selection

	(A) Peer Nominated				(B) Peer Nomination Ranking	
	(1)	(2)	(3)	(4)	(5)	(6)
Abs. Diff. in Time of First Run	-0.016***	-0.030***	-0.030***	-0.058***	-0.178***	
	(0.003)	(0.005)	(0.009)	(0.012)	(0.035)	
Abs. Diff. in Beliefs over Times in First Run						-0.184***
						(0.051)
Friendship Indicator	0.381***	0.392***	0.515***		1.710***	1.756***
	(0.014)	(0.017)	(0.049)		(0.115)	(0.118)
Abs. Diff. in Personality	-0.017***	-0.040***	-0.040***	-0.092***	-0.270***	-0.261***
	(0.004)	(0.009)	(0.009)	(0.024)	(0.073)	(0.079)
Abs. Diff. in Time of First Run × Abs. Diff. in Personality			0.002			
			(0.002)			
Abs. Diff. in Time of First Run × Friendship Indicator			-0.024***			
			(0.007)			
Abs. Diff. in Personality × Friendship Indicator			-0.016			
			(0.011)			
Controls for heterogeneity	Characteristics	Fixed effects	Fixed effects	Fixed effects	Fixed effects	Fixed effects
Sample	All	All	All	Friends only	All	Beliefs
Observations	6654	6646	6646	2872	2756	2756
Individuals	612	612	612	612	612	612
R^2	0.26	0.37	0.37	0.37	0.40	0.39

Notes: Panel A presents the results from the extensive margin analysis using a linear probability model according to equation (2.1) with an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable. Column (4) restricts the sample to the set of friends. Panel B presents results of the intensive margin using the ranking among those who are nominated as peers. While column (5) uses homophily in performances from the first run, we use beliefs over relative past performance rather than actual relative performance in column (6). Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

2.4.3 Intensive Margin of Peer Selection

Although the extensive margin analysis highlights whom students consider as peers, it reveals little about their relative importance. Since peers had to be ranked explicitly, we can exploit this information to more closely explore what makes a peer relatively more important. Panel B of Table 2.1 focus on this intensive margin of the peer selection process by analyzing the determinants of a peer's rank. Again, we estimate equation (2.1), but adjust the dependent variable as described in Section 2.4.1. Column (5) replicates the analysis of column (2), but uses the peer's rank as an outcome and restricts attention to all classmates that appear in the first six name-based peer preferences.[18] We find similar determinants for the ranking of peers as for the extensive margin: on average, friends are ranked 1.71 ranks higher than non-friends and students exhibit homophily in performance in the first run as well as in their personality. In particular, we find that the rank of a peer decreases by .18 ranks for each one-second time difference and by .27 ranks for each one standard-deviation difference in personalities.[19]

For the preceding analysis, we used absolute differences in past performances as a determinant of peer preferences. However, students in the experiment were not informed about their times in the first run, nor about those of their classmates. Accordingly, they had to rely on their beliefs about the relative performance of their peers when choosing them. We therefore check the robustness of our results by including the beliefs over relative performance rather than actual relative performance in column (8) and find that this does not affect our results.[20] As a second robustness check, we retain all classmates and estimate a Tobit model, in which the ranking is censored. The idea here is that all students who were not nominated have a lower rank than those who were nominated, but we do not observe their exact ranking.[21]

18. In order to analyze the ordering, we exploit the whole ranking of peers to increase power rather than analyzing the subset of the three most-preferred peers as for the extensive margin.

19. Column (2) of Table 2.B.3 in the Appendix splits up the aggregated personality measure. Similar to the extensive margin, we observe that agreeableness and the extent of engaging in social comparisons underlie the observed homophily in personality. More specifically, a one standard-deviation larger difference in agreeableness or social comparison attitudes is associated with a decrease of 0.25 and 0.16 ranks, respectively.

20. Note that we only elicited beliefs over relative performance for those students who were nominated as peers. Hence, we can only conduct this robustness check for the intensive margin and not for the extensive one. Nonetheless, as our results reveal, our conclusions neither change in a qualitatively nor quantitatively sense when using beliefs rather than actual performances. In fact, Appendix 2.C shows that beliefs and actual relative performance are strongly related to each other and validates their consistency. For this, we lever a second belief elicitation over the relative performance of the peer in the first run that was elicited just before the second run took place. This second belief measure and the one used in the elicitation of name-based preferences are indeed highly correlated, indicating that the beliefs are meaningful.

21. Since we coded the highest ranking as 6, we code all students who are not part of the six most-preferred peers as 0.

In Panel B of Appendix Table 2.B.2, we document that the results are qualitatively similar, although friendship ties become even more important than in our main specification. In summary, our results show that the results from the extensive margin analysis carry over to the intensive margin.

2.4.4 Heterogeneities in Name-based Preferences

While the previous sections have documented robust evidence of homophily in the peer selection process, different groups may choose peers differently. In order to predict the effects of different policies such as assigning students into classrooms or workers into teams, it is important to understand whether peer selection patterns differ across observable characteristics. Hence, we now shed light on the underlying heterogeneity of our estimates across sub-groups. Motivated by policies interested in promoting females or targeting low-ability students, we analyze whether males and females as well as high- and low-ability students select peers differently.

We present heterogeneities by gender and initial performance in Table 2.2. Columns (1) and (2) split the sample by gender and reveal some profound differences in the peer selection behavior of males and females. In particular, we find that males exhibit significantly stronger homophily in past performance as well as personality. By contrast, females seem to emphasize the presence of friendship ties more, although we cannot reject the hypothesis that the effect is the same across genders. In columns (4) and (5), we check for heterogeneities in ability. More specifically, we perform a median split of times in the first run within each gender and grade, and estimate equation (2.1) separately for both groups. The effect of friendship ties is more pronounced for slower students, while faster students show larger homophily effects in personality. Heterogeneities at the intensive margin are qualitatively similar as shown in Appendix Table 2.B.5.

These results highlight differential peer selection across different sub-groups. These findings have to be taken into account when thinking about peer or group assignment policies. Moreover, differences in peer selection criteria help to understand why peer effects work differently across different groups: if high-ability students exhibit strong homophily in their peer selection, they will tend to select other high-performing students as peers. Nonetheless, low-ability students choose their friends as peers, who may have low or high ability.

2.4.5 Targeting of Preferred Relative Performances

Finally, we examine the role of the preferred relative performances for the peer selection process and examine the relationship between the two sets of preferences. More specifically, we analyze the extent to which students target a relative performance level in the name-based selection process. In Appendix 2.D, we provide graphical evidence on the relation between the preferred relative performance and the selected peers. We observe that both set of preferences are positively associated with each

Table 2.2. Heterogeneities on the extensive margin of peer nominations

	Peer Nominated					
	(1) Males	(2) Females	(3) p-value	(4) Low Abil.	(5) High Abil.	(6) p-value
Abs. Diff. in Time of First Run	-0.057***	-0.027***	0.089	-0.022**	-0.042**	0.768
	(0.016)	(0.005)		(0.010)	(0.016)	
Friendship Indicator	0.348***	0.400***	0.257	0.434***	0.358***	0.004
	(0.039)	(0.020)		(0.021)	(0.022)	
Abs. Diff. in Personality	-0.105***	-0.025***	0.002	-0.027**	-0.047***	0.067
	(0.020)	(0.009)		(0.012)	(0.010)	
Controls for heterogeneity	Fixed effects	Fixed effects		Fixed effects	Fixed effects	
Observations	1408	5238		3303	3244	
Individuals	207	405		308	301	
R^2	0.39	0.35		0.44	0.43	

Notes: This table replicates Panel A, column (2) of Table 2.1 for different sub-samples. More specifically, it presents results from the extensive margin analysis using a linear probability model according to equation (2.1) with an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable. Columns (1) and (2) analyze male and female sub-samples, whereas columns (4) and (5) focus on high and low ability, defined according to the gender- and grade-specific median performance in the first run. Columns (3) and (6) present p-values of tests of equality between the two preceding columns. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

other but not perfectly related. Our preferred explanation for this imperfect relation is the fact that preferences for peers are multi-dimensional. They do not stem from a single factor, but rather are determined by the interplay of several factors.[22] Therefore, we ask whether students target peers with certain performance levels similar to their own, as indicated by the homophily documented in the previous section, or whether they try to target their preferred relative performance when selecting peers based on names.

In order to illustrate the notion that preferences for peers are indeed multi-dimensional, we enrich our previous model. In particular, we include the absolute deviation of a name-based peer's performance from the most-preferred performance in the peer selection model in equation (2.1). Table 2.3 presents the results of this exercise analogous to Table 2.1.

Focusing on the extensive margin of the selection process in Panel A, we observe that the estimated homophily in past performance is much smaller than documented in Table 2.1. Instead, there is a sizable effect of targeting one's preferred relative performance, with highly significant coefficients ranging between 1.0 and 4.9 percentage points. At the same time, the point estimate for differences in performance remains negative in all specifications and significant in some. Together, these two effects are similar in size to the homophily in performance documented in Table 2.1. Thus, students mainly select individuals who are close to their most-preferred performance (targeting of specific relative performances) but they also select peers who are close to their own performance (homophily in performances). Importantly, the other coefficients on friendship ties and personality differences remain unaffected by the inclusion of the preference for relative performance.[23] This highlights that the previous results are not a mere artifact of a preference for a specific relative performance; rather, it provides evidence that additional social dimensions are important for the peer selection process beyond mere reference points in performance.

If we concentrate on the intensive margin of the selection process in Panel B, a similar picture emerges: the absolute difference from the most-preferred relative performance is a strong predictor for the ranking among selected peers. A one-second increase in differences between the nominated peer's performance and the most-

22. A second possible explanation is that the true relation is indeed perfect and measurement error attenuates this association. Subsequently, given a true coefficient of unity, the estimated coefficients correspond to the attenuation factor λ. Using the relationship $\lambda = 1/(1+s)$ with s being the noise-to-signal ratio (Cameron and Trivedi, 2005, p. 903f.), we can calculate s. Based on the estimates in Table 2.D.1, in which we regress the preferred relative performance on a student's belief over the relative performance of her most-preferred peer, we obtain a coefficient $\hat{\beta} = 0.44$, implying $s = 1.27$. This ratio exceeds one, implying that the beliefs would need to contain more noise components than actual information. We thus conclude from this that measurement error alone is unlikely to be the sole cause for the imperfect relationship.

23. Similar to the estimates previously presented, we split the personality index in its components in Appendix Table 2.D.3

preferred relative performance leads to a decrease of .15 ranks for that peer. Again, the coefficient for homophily in performance is much smaller than before. The remaining determinants are unaffected by the inclusion of the preference for relative performance. Column (4) confirms these results using beliefs rather than actual performance. Unlike the specification with actual performance, beliefs over relative performance remain significant when including deviations from the preferred performance.

These results highlight that preferences over a peer's relative performance play a crucial role when selecting peers. While the most-preferred relative performance and the performance of the selected peer are strongly related, these measures do not coincide perfectly; rather, individuals also take into account other dimensions such as peers' similarity in terms of past performance and personality as well as existing friendship ties. By selecting peers based on their names, students can therefore condition on a richer information set. This suggests that social comparisons incorporate classical conceptualizations of reference points for effort provision, but they also depend on social factors.

2.5 Conclusion

Whom do individuals choose as peers? Answering this question is crucial to understand how peer effects work and how to design policies leveraging them. We use data from a framed field experiment and study preferences for peers to shed light on this issue. We find that individuals choose their peers predominantly, but not exclusively, along their social network. Friendship ties drive peer selections, but our sample also exhibits significant homophily in terms of individuals' performance and personality. Interestingly, among friends, similarity in performance becomes even more important for peer selection. While males choose more similar peers than females, low-performing individuals emphasize friendships more than their high-performing counterparts. By eliciting the desired relative performance of a peer, we find that most prefer peers with slightly higher but similar performance, which is in line with findings in social sciences (e.g., Blanton, Buunk, Gibbons, and Kuyper, 1999; Huguet et al., 2001). When selecting peers, individuals target a specific relative performance. Peer selection is therefore based on homophily in personality, friendship ties and a desired performance level.

Our results have important implications for estimating peer effects, designing mechanisms with social preferences and policy interventions. First, if friends are more likely to be chosen as peers, this could give rise to relatively larger impact of friends compared to non-friends. Similarly, if individuals choose peers with specific performances, these preferences may result in those peers exerting stronger effects than others. The evidence presented in this paper therefore provides a rationale for estimating models of differential (in terms of gender and friends) or nonlinear peer effects (in terms of own and peer ability). Second, by demonstrating to whom indi-

Table 2.3. Targeting of preferred relative performances

	(A) Peer Nominated			(B) Peer Nomination Ranking	
	(1)	(2)	(3)	(4)	(5)
Abs. Diff. in Time of First Run	-0.007*	-0.012*	-0.016	-0.052	
	(0.004)	(0.006)	(0.014)	(0.048)	
Abs. Diff. in Beliefs over Times in First Run					-0.149***
					(0.051)
Abs. Diff. from Perf.-based Preference	-0.010***	-0.022***	-0.049***	-0.150***	-0.175***
	(0.003)	(0.006)	(0.011)	(0.043)	(0.029)
Friendship Indicator	0.381***	0.392***		1.705***	1.722***
	(0.014)	(0.017)		(0.117)	(0.121)
Abs. Diff. in Personality	-0.017***	-0.039***	-0.094***	-0.269***	-0.260***
	(0.004)	(0.009)	(0.023)	(0.072)	(0.076)
Controls for heterogeneity	Characteristics	Fixed effects	Fixed effects	Fixed effects	Fixed effects
Sample	All	All	Friends only	All	Beliefs
Observations	6654	6646	2872	2756	2756
Individuals	612	612	612	612	612
R^2	0.26	0.37	0.37	0.40	0.41

Notes: This table presents the results of the linear probability model according to equation (2.1) using an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable and the absolute deviation from the most-preferred relative performance as an additional explanatory variable. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

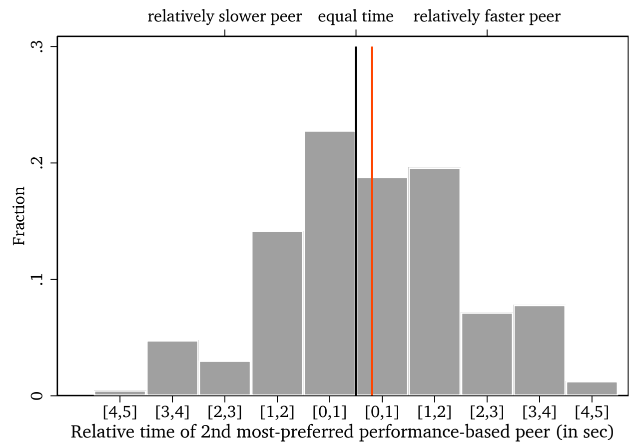
viduals compare their performance we inform theories of reference group formation. These insights in turn can be used to predict the effect of reorganizations and incentive contracts in a theoretically-disciplined manner (Ederer and Patacconi, 2010; Kőszegi, 2014). Finally, by using reassignment policies, teachers or managers influence the set of people from whom one can choose peers. On the one hand, these policies can have unintended consequences if sub-groups emerge (Carrell, Sacerdote, and West, 2013). On the other hand, policy-makers that are aware of such preferences for peers can provide suitable peers and hence indirectly affect peer selection.

The preferences for peers analyzed in this paper and their link to personal characteristics might be specific to situations where only own performance matters and with competitive components. Other peers might be selected in cooperative settings. Nevertheless, we demonstrate that the heterogeneity in social reference points and peer selection is based on systematic patterns of past performance, personality and friendship ties. These determinants are also likely to matter in other settings.

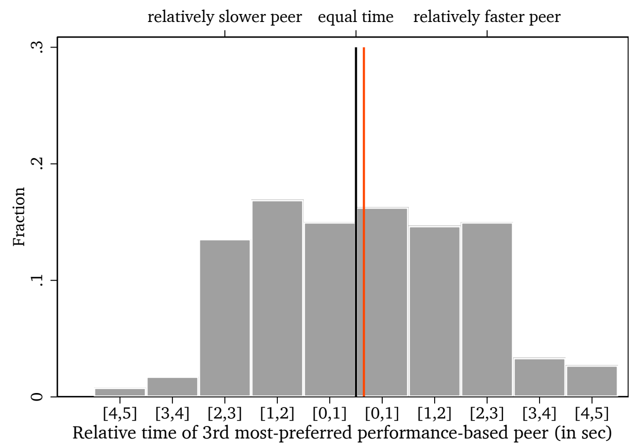
At the same time, our results open avenues for new interventions and research projects: if some peers exert positive effects on an individual's performance, can we encourage individuals to select into specific peer groups that help them to unfold their full potential? Relatedly, are students aware how their peers affect their own performance? Both of these issues raise the question whether preferences for peers would change if we provide individuals with information about peer effects or even "nudge" people to select specific peers. Our results are therefore a first step towards understanding the different aspects underlying peer choices. Future research on the interaction of personality, selection into environments and the influence of peers is needed to improve our understanding of social comparison processes, the endogenous formation of peer groups as well as their long-term consequences.

Appendix 2.A Additional Material for Performance-based Preferences

Figure 2.A.1 presents the distribution of the second and third most-preferred relative performance. We observe that these are also centered around the [0,1] second faster category but show some different pattern. Nonetheless, as reported in section 2.3.1, the differences in the distribution are due targeting the most-preferred relative performance. We thus restrict our attention to the first preference only.



(a) Second performance-based preference

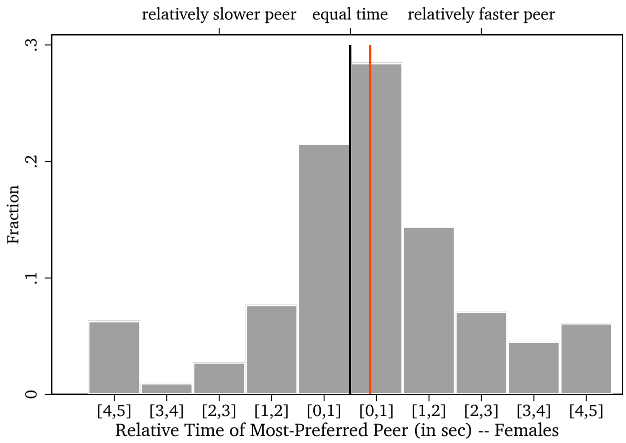


(b) Third performance-based preference

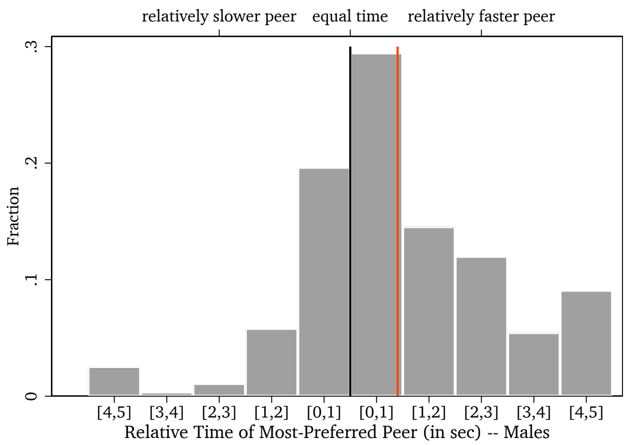
Figure 2.A.1. Distribution of second and third performance-based peer preferences

Notes: Figure (a) presents a histograms of students' preferences over relative performance. The intervals used here and in the survey are one-second intervals of relative performances in the first run. Vertical lines indicate own performances (black; equals zero by definition) and mean preference (red; where we used the mean of each interval to calculate the mean). Figure (b) presents the relationship of the first performance-based preference and the second/third preference.

In Figures 2.A.2a and 2.A.2b, we present gender splits of the most-preferred relative performance. While both distributions are relatively similar, males prefer somewhat faster peers than females. On average, females prefer peers being .38 seconds faster, whereas males prefer peers being .90 seconds faster. These correspond to 13 and 31% of a standard deviation in performances of the first run.



(a) Females



(b) Males

Figure 2.A.2. Distribution of performance-based peer preferences by gender

Notes: Figure (a) presents a histograms of students' preferences over relative performance. The intervals used here and in the survey are one-second intervals of relative performances in the first run. Vertical lines indicate own performance (black; equals zero by definition) and mean preference (red; where we used the mean of each interval to calculate the mean). Figure (b) presents the relationship of the first performance-based preference and the second/third preference.

Appendix 2.B Additional Material for Peer Selection Analysis

This appendix provides descriptive statistics and robustness checks for the analysis of the peer selection process. Table 2.B.1 provides summary statistics for the variables used in the analysis. In Table 2.B.2 Panel A we consider someone to be nominated if he is nominated at all, i.e. among the first six most-preferred peers, and zero otherwise. Similarly, Panel (B) estimates a Tobit specification using all potential peers, where we only observe the ranking for six most-preferred peers and is censored otherwise. Table 2.B.3 splits up the aggregate measure of personality and includes all dimensions separately. Table 2.B.4 uses alternative definitions of friendship to show that our results are robust with respect to the exact definition. Finally, Table 2.B.5 presents the heterogeneous effects for peer selection at the intensive margin.

Table 2.B.1. Distribution of absolute differences

	Absolute differences				
	Mean	SD	25th perc.	50th perc.	75th perc.
Full sample					
Abs. Diff. in Time of First Run	2.55	2.10	0.93	2.06	3.59
Friendship Indicator	0.46	0.50	0.00	0.00	1.00
Abs. Diff. in Personality	3.99	1.12	3.20	3.92	4.66
Abs. Diff. in Agreeableness	1.13	0.85	0.44	0.96	1.65
Abs. Diff. in Conscientiousness	1.12	0.84	0.45	0.97	1.62
Abs. Diff. in Extraversion	1.13	0.84	0.45	0.95	1.66
Abs. Diff. in Openness	1.11	0.87	0.42	0.93	1.60
Abs. Diff. in Neuroticism	1.06	0.78	0.43	0.91	1.53
Abs. Diff. in Locus of Control	1.09	0.83	0.43	0.90	1.58
Abs. Diff. in Social Comparison	1.09	0.82	0.43	0.92	1.59
Abs. Diff. in Competitiveness	1.07	0.78	0.44	0.91	1.58
Abs. Diff. in Risk Preferences	1.12	0.87	0.45	0.90	1.79
For friends only					
Abs. Diff. in Time of First Run	2.33	1.95	0.84	1.87	3.25
Abs. Diff. in Personality	3.89	1.10	3.13	3.80	4.50

Notes: This table presents summary statistics for absolute differences in several characteristics. The upper panel considers all characteristics for the whole sample, while the lower panel restricts the characteristics to friends only.

Table 2.B.2. Robustness checks: All nominated peers and censoring

	(A) Peer Nominated				(B) Peer Nomination Ranking
	(1)	(2)	(3)	(4)	(5)
Abs. Diff. in Time of First Run	-0.017***	-0.032***	-0.030***	-0.037***	-0.145***
	(0.004)	(0.006)	(0.009)	(0.009)	(0.042)
Friendship Indicator	0.495***	0.507***	0.515***		4.981***
	(0.022)	(0.024)	(0.049)		(0.311)
Abs. Diff. in Personality	-0.016***	-0.023***	-0.040***	-0.059***	-0.077*
	(0.005)	(0.007)	(0.009)	(0.017)	(0.044)
Abs. Diff. in Time of First Run × Abs. Diff. in Personality			0.002		
			(0.002)		
Abs. Diff. in Time of First Run × Friendship Indicator			-0.024***		
			(0.007)		
Abs. Diff. in Personality × Friendship Indicator			-0.016		
			(0.011)		
Controls for heterogeneity	Characteristics	Fixed effects	Fixed effects	Fixed effects	Fixed effects
Sample	All	All	All	Friends only	All
Observations	6654	6646	6646	2872	6654
Individuals	612	612	612	612	612
R^2	0.35	0.46	0.37	0.44	

Notes: Panel A presents the results from the extensive margin analysis using a linear probability model according to equation (2.1) with an indicator of being nominated as one of the sixth most-preferred name-based peers (i.e. whether a person is nominated as a peer at all) as the dependent variable. Column (4) restricts the sample to the set of friends. Panel B presents results of the intensive margin using the whole sample but allow for censoring of those classmates that were not nominated on the extensive margin. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 2.B.3. Robustness checks: Splitting up personality index

	(A) Peer Nominated	(B) Peer Nomination Ranking
	(1)	(2)
Abs. Diff. in Time of First Run	-0.029***	-0.173***
	(0.005)	(0.032)
Friendship Indicator	0.393***	1.716***
	(0.017)	(0.113)
Abs. Diff. in Agreeableness	-0.033***	-0.249***
	(0.009)	(0.054)
Abs. Diff. in Conscientiousness	-0.002	-0.054
	(0.008)	(0.053)
Abs. Diff. in Extraversion	-0.015	-0.104
	(0.010)	(0.063)
Abs. Diff. in Openness	-0.004	0.025
	(0.010)	(0.069)
Abs. Diff. in Neuroticism	-0.014	-0.141*
	(0.009)	(0.083)
Abs. Diff. in Locus of Control	-0.003	-0.056
	(0.009)	(0.079)
Abs. Diff. in Social Comparison	-0.022***	-0.163***
	(0.007)	(0.055)
Abs. Diff. in Competitiveness	-0.018*	-0.072
	(0.009)	(0.059)
Abs. Diff. in Risk Preferences	-0.005	0.006
	(0.009)	(0.086)
Controls for heterogeneity	Fixed effects	Fixed effects
Sample	All	All
Observations	6646	2756
Individuals	612	612
R^2	0.37	0.40

Notes: Panel A presents the results from the extensive margin analysis using a linear probability model according to equation (2.1) with an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable, but in which we allow for each personality measure to enter separately. Panel B presents analogous results of the intensive margin using the ranking among those who are nominated as peers. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 2.B.4. Robustness checks: Alternative definitions of friendship ties

	Peer Nominated		
	(1) Undirected	(2) Directed	(3) Reciprocal
Abs. Diff. in Time of First Run	-0.030***	-0.029***	-0.029***
	(0.005)	(0.005)	(0.005)
Abs. Diff. in Personality	-0.040***	-0.035***	-0.032***
	(0.009)	(0.008)	(0.007)
Friendship Indicator	0.392***	0.454***	0.507***
	(0.017)	(0.015)	(0.017)
Controls for heterogeneity	Fixed effects	Fixed effects	Fixed effects
Observations	6646	6646	6646
Individuals	612	612	612
R^2	0.37	0.41	0.42

Notes: This table presents the results from the extensive margin analysis using a linear probability model according to equation (2.1) with an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable for varying definitions of friendship ties. Column (1) uses undirected friendships as in the main text, column (2) defines friendship ties as directed, while column (3) only considers reciprocal links. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 2.B.5. Heterogeneities on the intensive margin of peer nominations

	Peer Nomination Ranking					
	(1) Males	(2) Females	(3) p-value	(4) Low Abil.	(5) High Abil.	(6) p-value
Abs. Diff. in Time of First Run	-0.162	-0.180***	0.884	-0.195***	-0.028	0.765
	(0.116)	(0.036)		(0.060)	(0.112)	
Friendship Indicator	1.441***	1.776***	0.234	2.174***	1.436***	0.012
	(0.245)	(0.134)		(0.140)	(0.202)	
Abs. Diff. in Personality	-0.486***	-0.200**	0.063	-0.204*	-0.304**	0.778
	(0.119)	(0.092)		(0.113)	(0.128)	
Controls for heterogeneity	Fixed effects	Fixed effects		Fixed effects	Fixed effects	
Observations	777	1979		1260	1230	
Individuals	207	405		308	301	
R^2	0.46	0.38		0.51	0.50	

Notes: This table replicates Panel B of Table 2.1 for different sub-samples. More specifically, it presents results from the intensive margin analysis using the ranking among those who are nominated as peers, in which better rankings correspond to higher values of the dependent variable (6: highest, 1: lowest). Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Appendix 2.C Relationship of Beliefs and Actual Performance

In this section, we first describe the relationship between beliefs and actual performance. Afterwards, we provide evidence that the beliefs are meaningful, which is consistent over time by leveraging a second measurement of the same belief.

Beliefs over relative performance and actual relative performance do not necessarily coincide. We therefore check how these two relate to each other. Figure 2.C.1a presents a scatter plot of the belief over relative performance of name-based peers and their actual relative performance. We observe that although the relationship is not perfect, these two are significantly related as is confirmed by the corresponding regressions in Table 2.C.1. Figure 2.C.1b displays the absolute differences between the beliefs and the actual relative performance. On average, these two have an absolute difference of 1.95 seconds.

Table 2.C.1. Relationship between beliefs over and actual relative performance

	(a) Peer's relative time (continuous)		(b) Peer is faster (binary)	
	(1)	(2)	(3)	(4)
Relative time of most-preferred name-based peer	0.25***	0.24***		
	(0.04)	(0.04)		
Preferred name-based peer is faster			0.27***	0.25***
			(0.05)	(0.05)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
Individuals	566	562	566	562
R^2	.21	.23	.16	.17

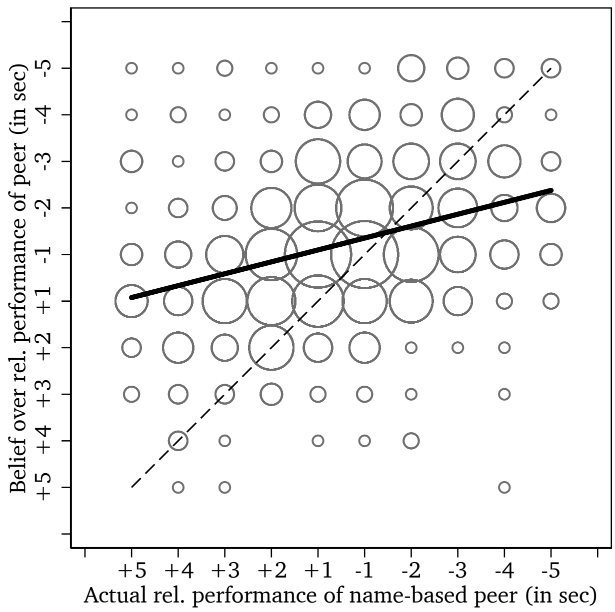
Notes: This table presents least squares regressions using a peer's relative performance according to the beliefs of the name-based preferences as the dependent variable. Figure 2.C.1 presents the results graphically. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Moreover, we are interested whether the beliefs capture pure noise or whether they are constant over time. To check for consistency of the beliefs, we lever a second (binary) belief elicited right before the second run and compare it to the beliefs elicited as part of the name-based preferences. The first two columns of Table 2.C.2 use the continuous measure of beliefs over relative performance as elicited in the name-based preferences as the dependent variable. The second set of columns uses a binary version of this indicating whether the student believed that the peer has been faster or slower. The sample is restricted to those students with peers that are nominated somewhere in the name-based preferences (i.e., for whom we have beliefs) and that are matched as a peer in the second run (i.e., only for those for whom we have a second belief measure). This naturally oversampled observations in NAME. We thus check whether the pattern differs depending on the treatment. As can be seen, the two measures are significantly related with a correlation of .58. Moreover, this correlation does not significantly vary with the assigned treatment.

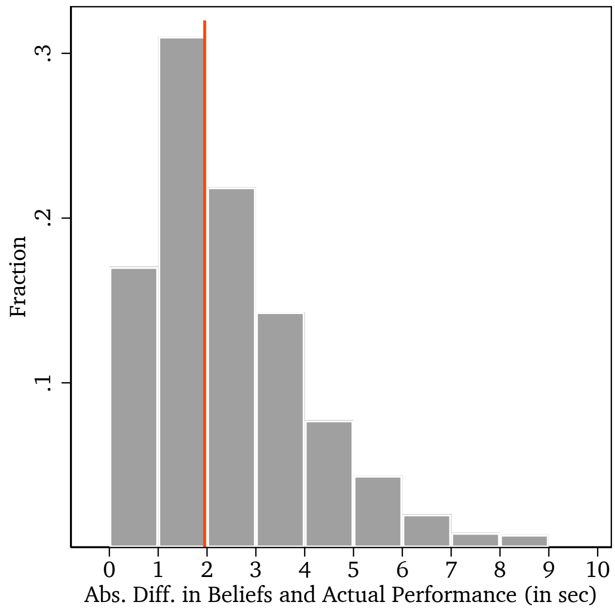
Table 2.C.2. Consistency of beliefs

	(a) Continuous belief		(b) Binary belief	
	(1)	(2)	(3)	(4)
Believe peer is faster	1.96***		0.58***	
	(0.23)		(0.05)	
RANDOM × Believe peer is faster		2.00***		0.53***
		(0.27)		(0.06)
NAME × Believe peer is faster		1.92***		0.59***
		(0.23)		(0.05)
PERFORMANCE × Believe peer is faster		2.01***		0.58***
		(0.23)		(0.05)
N	345	345	345	345
R^2	.26	.27	.3	.31

Notes: This table presents least squares regressions using the beliefs over the peer's performance as elicited in the name-based preferences as the dependent variable. The sample is restricted to those subjects with peers that are nominated in the name-based preferences and are actually matched for the second run, for which we have elicited a second (binary) belief measure. 89 observations are from students in RANDOM, 180 from NAME, and 87 from PERFORMANCE. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.



(a) Relationship



(b) Absolute difference

Figure 2.C.1. Relationship of beliefs and actual performance

Notes: Figure (a) presents the relationship beliefs over and actual relative performance of the name-based peers. The corresponding regression is presented in Table 2.C.1. Figure (b) presents a histogram of the absolute difference in beliefs and actual performance. The vertical line in (b) indicates mean absolute difference (red; where we used the mean of each interval to calculate the mean). The intervals used here and in the survey are one-second intervals of relative performances in the first run.

Appendix 2.D Additional Material for Relationship of Preferences

Figure 2.D.1 and Table 2.D.1 provide a first view on the relation of performance- and name-based preferences. More specifically, we associate preferred relative performances of each individual with beliefs over the relative performance of their peers nominated in the name-based preferences. We observe a positive relationship between the two measures as shown in Figure 2.D.1.

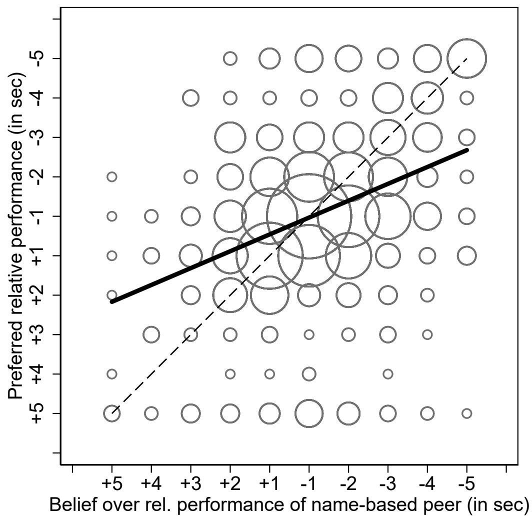


Figure 2.D.1. Relationship of performance- and name-based preferences for peers

Notes: The figures present the relationship between performance- and name-based preferences using beliefs over peer's performance. Corresponding regressions are presented in Table 2.D.1.

Table 2.D.1 quantifies this relationship: if students select a peer who they believe is one second faster, this is associated with an increase in the relative performance in the performance-based preference by .44 seconds on average (columns (1) and (2)). Similarly, we observe a significant positive relationship between binary indicators of believing that the most-preferred name-based peer is faster and choosing a faster peer in the performance-based preference in columns (3) and (4). Nonetheless, the relationship between name- and performance-based preferences is not perfect, as it would be the case if the preferences over relative performance were the only determinants of name-based preferences. If this were the case, we should observe regression coefficients of unity.

One potential explanation for the imperfect relationship between performance- and name-based preferences is measurement error. Here, we show that measurement error is unlikely to explain the imperfect association alone. Assume that we have classical measurement error and the true coefficient corresponds to one ($\beta = 1$), then by the standard attenuation bias formula (Cameron and Trivedi, 2005,

Table 2.D.1. Relationship between preferences based on names and relative performance

	Peer preference over rel. perf.			
	(a) Continuous		(b) Binary	
	(1)	(2)	(3)	(4)
Belief over peer's rel. perf.	0.44***	0.44***	0.29***	0.29***
	(0.06)	(0.06)	(0.04)	(0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
Individuals	627	623	582	578
R^2	.25	.28	.17	.2

Notes: This table presents least squares regressions using a peer's relative performance in one-second intervals or an indicator for preferring a faster peer according to the performance-based preferences as the dependent variable. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level. Figure 2.D.1 presents the results graphically.

p. 903f.), we have that if $x^* = x + v$ with v being a mean-zero error with variance σ_v^2,

$$p\lim \hat{\beta} \;=\; \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2}\beta \;=\; \lambda\beta \;=\; \lambda \tag{2.D.1}$$

as $\beta = 1$ and where λ is the attenuation factor.[24] Thus the regression coefficients in Table 2.D.2 correspond to the attenuation factors that would be needed for a perfect relationship. For a more intuitive interpretation, we rewrite the factor in terms of the noise-to-signal ratio s such that $\lambda = 1/(1+s)$. The noise-to-signal ratio tells us how much noise relative to signals the data should have if the true relationship is given by $\beta = 1$. We reproduce Table 2.D.1 here and additionally present the corresponding noise-to-signal ratios of each coefficient below the corresponding regressions. We find that all ratios exceed one, which implies that the measurements would need to have more noise components than actual information. We thus conclude that measurement error alone cannot explain the imperfect relationship.

24. For the multivariate case the formula is slightly different, but the basic idea remains the same.

Table 2.D.2. Relationship between performance- and name-based preferences

	(a) Peer's relative time (continuous)		(b) Peer is faster (binary)	
	(1)	(2)	(3)	(4)
Panel A: Using name-based beliefs				
Belief over peer's performance	0.44***	0.44***		
	(0.06)	(0.06)		
Belief over peer's performance (0/1)			0.29***	0.29***
			(0.04)	(0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
Individuals	627	623	627	623
R^2	.25	.27	.17	.2
Noise-to-signal ratio for $\beta = 1$	1.3	1.3	2.5	2.5
Panel B: Using name-based actual perf.				
Relative time of peer	0.10***	0.09***		
	(0.03)	(0.03)		
Peer is faster (0/1)			0.04	0.03
			(0.04)	(0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
Individuals	566	562	566	562
R^2	.11	.13	.095	.12
Noise-to-signal ratio for $\beta = 1$	9.2	10	26	28

Notes: This table presents least squares regressions using the most-preferred name-based peer's relative performance in one-second intervals or an indicator for preferring a faster peer according to the performance-based preferences as the dependent variable. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level. The reported signal-to-noise ratio describes the extend of measurement error needed if the true relationship is actually perfect (i.e., $\beta = 1$) rather than imperfect ($\beta < 1$). Accordingly, a noise-to-signal ratio larger than one indicates more noise than signal, equal to one corresponds to as much signal as noise and less than one more signal than noise. Figure 2.D.1 presents the results graphically.

Table 2.D.3. Robustness checks: Splitting up personality index

	(A) Peer Nominated	(B) Peer Nomination Ranking
	(1)	(2)
Abs. Diff. in Time of First Run	-0.011*	-0.052
	(0.006)	(0.045)
Abs. Diff. from Perf.-based Preference	-0.022***	-0.145***
	(0.006)	(0.042)
Friendship Indicator	0.394***	1.711***
	(0.017)	(0.114)
Abs. Diff. in Agreeableness	-0.034***	-0.242***
	(0.009)	(0.057)
Abs. Diff. in Conscientiousness	-0.002	-0.055
	(0.008)	(0.052)
Abs. Diff. in Extraversion	-0.015	-0.103
	(0.010)	(0.063)
Abs. Diff. in Openness	-0.003	0.033
	(0.010)	(0.069)
Abs. Diff. in Neuroticism	-0.014	-0.136
	(0.009)	(0.082)
Abs. Diff. in Locus of Control	-0.003	-0.055
	(0.009)	(0.078)
Abs. Diff. in Social Comparison	-0.021***	-0.161***
	(0.007)	(0.054)
Abs. Diff. in Competitiveness	-0.018*	-0.072
	(0.010)	(0.057)
Abs. Diff. in Risk Preferences	-0.005	-0.006
	(0.009)	(0.085)
Controls for heterogeneity	Fixed effects	Fixed effects
Sample	All	All
Observations	6646	2756
Individuals	612	612
R^2	0.37	0.41

Notes: Panel A presents the results from the extensive margin analysis using a linear probability model according equation (2.1) with an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable, but in which we allow for each personality measure to enter separately and add absolute deviations of the most-preferred relative performance as an additional regressor. Panel B presents analogous results of the intensive margin using the ranking among those who are nominated as peers. Standard errors are shown in parentheses and clustered at the class level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. (2011). "Reference Points and Effort Provision." *American Economic Review* 101 (2): 470–492. [69]

Aral, Sinan, and Christos Nicolaides. (2017). "Exercise Contagion in a Global Social Network." *Nature Communications* 8 (14753): [71]

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. (2014). "Awards unbundled: Evidence from a natural field experiment." *Journal of Economic Behavior & Organization* 100: 44–63. [73]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2009). "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica* 77 (4): 1047–1094. [69, 71]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. (2010). "Social Incentives in the Workplace." *Review of Economic Studies* 77 (2): 417–458. [71]

Beugnot, Julie, Bernard Fortin, Guy Lacroix, and Marie Claire Villeval. (2019). "Gender and peer effects on performance in social networks." *European Economic Review* 113: 207–224. [71]

Bhattacharya, Debopam. (2009). "Inferring Optimal Peer Assignment From Experimental Data." *Journal of the American Statistical Association* 104 (486): 486–500. [72]

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. (2013). "Under Pressure? The Effect of Peers on Outcomes of Young Adults." *Journal of Labor Economics* 31 (1): 119–153. [71]

Blanton, Hart, Bram P. Buunk, Frederick X. Gibbons, and Hans Kuyper. (1999). "When Better-than-Others Compare Upward: Choice of Comparison and Comparative Evaluation as Independent Predictors of Academic Performance." *Journal of Personality and Social Psychology* 76 (3): 420–430. [90]

Booij, Adam S., Edwin Leuven, and Hessel Oosterbeek. (2017). "Ability Peer Effects in University: Evidence from a Randomized Experiment." *Review of Economic Studies* 84 (2): 547–578. [72]

Brookins, Philip, Sebastian Goerg, and Sebastian Kube. (2017). "Self-chosen goals, incentives, and effort." Working Paper. [73]

Burgess, Simon, Ellen Greaves, Anna Vignoles, and Deborah Wilson. (2014). "What Parents Want: School Preferences and School Choice." *Economic Journal* 125 (587): 1262–1289. [72]

Burke, Mary A., and Tim R. Sass. (2013). "Classroom Peer Effects and Student Achievement." *Journal of Labor Economics* 31 (1): 51–82. [72]

Cameron, Colin, and Pravin Trivedi. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press. [89, 103]

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. (2012). "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." *American Economic Review* 102 (6): 2981–3003. [73]

Carpenter, Jeffrey, and Caitlin Myers. (2010). "Why volunteer? Evidence on the role of altruism, image, and incentives." *Journal of Public Economics* 94 (11): 911–920. [72]

Carrell, Scott, Bruce Sacerdote, and James West. (2013). "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81 (3): 855–882. [69, 72, 92]

Cerulli-Harms, Annette, Lorenz Goette, and Charles Sprenger. (2019). "Randomizing Endowments: An Experimental Study of Rational Expectations and Reference-Dependent Preferences." *American Economic Journal: Microeconomics* 11 (1): 185–207. [73]

Cicala, Steve, Roland Fryer, and Jörg Spenkuch. (2018). "Self-Selection and Comparative Advantage in Social Interactions." *Journal of the European Economic Association* 16 (4): [72]

Clark, Andrew, and Claudia Senik. (2010). "Who Compares to Whom? The Anatomy of Income Comparisons in Europe." *Economic Journal* 120 (544): 573–594. [73]

Cohn, Alain, Ernst Fehr, Benedikt Herrmann, and Frédéric Schneider. (2014). "Social Comparison and Effort Provision: Evidence from a Field Experiment." *Journal of the European Economic Association* 12 (4): 877–898. [73]

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg. (2017). "Peer Effects in the Workplace." *American Economic Review* 107 (2): 425–456. [71]

Dohmen, Thomas, and Armin Falk. (2010). "You Get What You Pay For: Incentives and Selection in the Education System*." *Economic Journal* 120 (546): F256–F271. [72]

Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. (2011). "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–550. [76]

Dube, Arindrajit, Laura Giuliano, and Jonathan Leonard. (2019). "Fairness and frictions: The impact of unequal raises on quit behavior." *American Economic Review* 109 (2): 620–63. [71]

Ederer, Florian, and Andrea Patacconi. (2010). "Interpersonal Comparison, Status and Ambition in Organizations." *Journal of Economic Behavior & Organization* 75 (2): 348–363. [92]

Falk, Armin, and Andrea Ichino. (2006). "Clean Evidence on Peer Effects." *Journal of Labor Economics* 24 (1): 39–57. [69]

Falk, Armin, and Markus Knell. (2004). "Choosing the Joneses: Endogenous Goals and Reference Standards." *Scandinavian Journal of Economics* 106 (3): 417–435. [73]

Festinger, Leon. (1954). "A Theory of Social Comparison Processes." *Human Relations* 7 (2): 117–140. [78]

Friebel, Guido, Michael Kosfeld, and Gerd Thielmann. (Forthcoming). "Trust the Police? Self-Selection of Motivated Agents into the German Police Force." *American Economic Journal: Microeconomics*, [72]

Gibbons, Frederick, and Bram Buunk. (1999). "Individual Differences in Social Comparison: Development of a Scale of Social Comparison Orientation." *Journal of Personality and Social Psychology* 76 (1): 129–147. [76]

Girard, Yann, Florian Hett, and Daniel Schunk. (2015). "How Individual Characteristics Shape the Structure of Social Networks." *Journal of Economic Behavior & Organization* 115: Behavioral Economics of Education, 197–216. [72, 80, 84]

Graham, Bryan S. (2015). "Methods of Identification in Social Networks." *Annual Review of Economics* 7 (1): 465–485. [72]

Hoxby, Caroline. (2000). "Peer Effects in the Classroom: Learning from Gender and Race Variation." Working Paper. [71]

Huckman, Robert S., Bradley R. Staats, and David M. Upton. (2009). “Team Familiarity, Role Experience, and Performance: Evidence from Indian Software Services.” *Management Science* 55 (1): 85–100. [71]

Huguet, Pascal, Florence Dumas, Jean M. Monteil, and Nicolas Genestoux. (2001). “Social Comparison Choices in the Classroom: Further Evidence for Students’ upward Comparison Tendency and its Beneficial Impact on Performance.” *European Journal of Social Psychology* 31 (5): 557–578. [78, 90]

Irving, Robert. (1985). “An Efficient Algorithm for the “Stable Roommates” Problem.” *Journal of Algorithms* 6 (4): 577–595. [74]

Kahneman, Daniel. (1992). “Reference points, anchors, norms, and mixed feelings.” *Organizational Behavior and Human Decision Processes* 51 (2): 296–312. [70]

Kahneman, Daniel, and Amos Tversky. (1979). “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2): 263–292. [73]

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube. (2018). “Self-selection of peers and performance.” Working Paper. [72, 74, 81]

Knight, John, Lina Song, and Ramani Gunatilaka. (2009). “Subjective well-being and its determinants in rural China.” *China Economic Review* 20 (4): 635–649. [73]

Koch, Alexander K., and Julia Nafziger. (2011). “Self-regulation through Goal Setting*.” *Scandinavian Journal of Economics* 113 (1): 212–227. [69, 73]

Kőszegi, Botond. (2014). “Behavioral Contract Theory.” *Journal of Economic Literature* 52 (4): 1075–1118. [69, 72, 92]

Kuhn, Peter, Peter Kooreman, Adriaan Soetevent, and Arie Kapteyn. (2011). “The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery.” *American Economic Review* 101 (5): 2226–2247. [73]

Lavy, Victor, and Analía Schlosser. (2011). “Mechanisms and Impacts of Gender Peer Effects at School.” *American Economic Journal: Applied Economics* 3 (2): 1–33. [71]

Manski, Charles. (1993). “Identification of Endogenous Social Effects: The Reflection Problem.” *Review of Economic Studies* 60 (3): 531–542. [69, 71]

Mas, Alexandre, and Enrico Moretti. (2009). “Peers at Work.” *American Economic Review* 99 (1): 112–145. [69, 71, 72]

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. (2001). “Birds of a Feather: Homophily in Social Networks.” *Annual Review of Sociology* 27 (1): 415–444. [72, 82]

Niederle, Muriel, and Lise Vesterlund. (2007). “Do Women Shy Away from Competition? Do Men Compete too much?” *Quarterly Journal of Economics* 122 (3): 1067–1101. [72]

O’Donoghue, Ted, and Charles Sprenger. (2018). “Reference-dependent preferences.” In *Handbook of Behavioral Economics: Foundations and Applications*. Vol. 1, Elsevier Amsterdam. [73]

Park, Sangyoon. (2019). “Socializing at Work: Evidence from a Field Experiment with Manufacturing Workers.” *American Economic Journal: Applied Economics* 11 (3): 424–455. [69]

Roels, Guillaume, and Xuanming Su. (2014). “Optimal Design of Social Comparison Effects: Setting Reference Groups and Reference Points.” *Management Science* 60 (3): 606–627. [72]

Rotter, Julian B. (1966). “Generalized Expectancies for Internal Versus External Control of Reinforcement.” *Psychological Monographs: General and Applied* 80 (1): 1–28. [76]

Sacerdote, Bruce. (2011). “Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?” In. Edited by Eric A. Hanushek, Stephen

Machin, and Ludger Woessmann. Vol. 3, Handbook of the Economics of Education. Elsevier, 249–277. [69]

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." *DIW Data Documentation*, [76]

Schwerter, Frederik. (2016). "Social Reference Points and Risk Taking." Working Paper. [73]

Selfhout, Maarten, William Burk, Susan Branje, Jaap Denissen, Marcel Van Aken, and Wim Meeus. (2010). "Emerging late adolescent friendship networks and Big Five personality traits: A social network approach." *Journal of Personality* 78 (2): 509–538. [84]

Senik, Claudia. (2009). "Direct evidence on income comparisons and their welfare effects." *Journal of Economic Behavior & Organization* 72 (1): 408–424. [73]

Tan, Tom Fangyun, and Serguei Netessine. (Forthcoming). "When You Work with a Superman, Will You Also Fly? An Empirical Study of the Impact of Coworkers on Performance." *Management Science*, [72]

Weinhardt, Michael, and Jürgen Schupp. (2011). "Multi-Itemskalen im SOEP Jugendfragebogen." *DIW Data Documentation*, [76]

Chapter 3

Understanding Parental Decision-making: Beliefs about Returns to Parenting Styles and Neighborhoods

3.1 Introduction

Parents play a crucial role for the development and success of children, as inequalities can be traced back to early life (Francesconi and Heckman, 2016; Kalil, 2015). Yet, not much is known about the factors determining how parents decide to raise their children. In particular, evidence on the parental decision-making process and the consequences of different parenting styles remains scarce, in part due to their complexity (Attanasio, 2015). In a recent study, Doepke and Zilibotti (2017) argue that the economic environment creates incentives to engage in different forms of parenting. As parents decide where to live and how to raise their children, it is important to understand how parents perceive their environments and parenting to interact.[1]

In this paper, I study how parents perceive the returns to two factors affecting the development and long-term outcomes of children: First, I focus on parenting styles describing strategies that parents use in raising their children (Baumrind, 1967), and second, I focus on the quality of the neighborhood in which a family lives. In addition, I examine their perceived substitutability or complementarity, analyze the heterogeneity in perceived returns, and investigate the relevance of these beliefs for actual parental decision-making. Studying how parents perceive different parenting

1. In general, any observed choice may be consistent with different combinations of preferences and beliefs. Manski (2004) therefore argues that one cannot solely rely on observed behavior to infer underlying beliefs, and advocates for a direct elicitation of beliefs.

styles and neighborhoods to interact helps to predict their behavioral responses to (policy-induced) changes in the quality of neighborhoods.

In order to investigate parental beliefs, I adopt a hypothetical scenario approach used by Cunha, Elo, and Culhane (2013), Boneva and Rauh (2018), Bhalotra, Delavande, Font, and Maselko (2017), and Attanasio, Boneva, and Rauh (2019). I construct eight scenarios in which parents raise their children. Across scenarios, I vary the parenting style that parents adopt – commonly defined as different intensities of warmth and control employed in raising children (Maccoby and Martin, 1983)[2] – as well as the quality of the neighborhood families are living in. In addition, I randomize the children's age and gender across respondents. For each of these scenarios, I then elicit parental expectations about the future earnings and expected life satisfaction of the child at the age of 30.[3] This design has several noteworthy features: First, by eliciting parents' beliefs for all eight scenarios and varying one dimensions at a time, I can infer parents' perceived returns to one particular dimension while controlling for (unobserved) heterogeneity across respondents. Second, comparing scenarios that change several factors at the same time allows me to investigate the perceived substitutability or complementarity of parenting styles and neighborhoods. Third, having access to several elicited beliefs per parent, I can estimate how each parent perceives these returns and subsequently link them to their characteristics and actual parenting styles. I implement the scenarios in a survey of 2,119 parents with school-aged children in the United States, who are selected to be representative in terms of their gender, age, income, and region.

I find that parents expect considerable returns to the warmth dimension of parenting, but not to control. An increase of one standard deviation in warmth is associated with parents expecting 15.3 percent higher earnings for children at the age of 30, whereas increasing control is not perceived as yielding any returns. In addition, my estimates show that parents expect earnings to increase by 22.6 percent when raising a child in a relatively good neighborhood. When analyzing the interaction of the different factors, parents seem to adapt their expectations. Parents perceive warmth and control as complements, increasing expected earnings by an additional

2. Parenting styles have a long tradition in developmental psychology going back to Baumrind (1967). Initially, she identified three parenting styles, while Maccoby and Martin (1983) extend her original typology to four styles defined according to two dimensions – the extent of warmth, on the one hand, and control used in raising children, on the other. Depending on their intensities, these two dimensions define four distinct parenting styles: authoritative (high warmth, high control), permissive (high warmth, low control), authoritarian (low warmth, high control), and neglecting (low warmth, low control). The psychology literature often refers to these dimensions as responsiveness and demandingness instead of warmth and control.

3. This approach of eliciting future wage expectations dates back to Dominitz and Manski (1996) and has subsequently been used in a range of studies focusing on returns to human capital investments (e.g., Attanasio and Kaufmann, 2014; Hastings, Neilson, Ramirez, and Zimmerman, 2016; Jensen, 2010; Kaufmann, 2014; Nguyen, 2008).

4.6 percentage points if combining high levels of both warmth and control. Moreover, parenting is perceived as being more effective in low-quality neighborhoods. The perceived return to warmth (control) is 1.4 (1.5) percentage points higher in low-quality neighborhoods, corresponding to an increase of approximately 9 percent of the perceived return to warmth. Parents therefore expect their parenting to compensate at least in part for deprived environments. Yet, authoritative parenting styles featuring warmth and control are perceived as being more effective in high quality neighborhoods. In addition, I show that these results are not restricted to the monetary domain, but carry over to the life satisfaction domain.

How do these perceived returns vary by age and gender of the child? First, my results reveal a pronounced age gradient: high levels of warmth are perceived as more effective for younger children, while exerting control is especially important for older, teenage children living in adverse environments. I do not find differences in perceived returns by child gender. However, when focusing on parental gender, I find pronounced differences in perceived returns. Mothers expect higher returns to warmth and neighborhoods than fathers, while there are no differences in the control dimension of parenting styles. Although there is a large dispersion in perceived returns, I do not find systematic associations with other sociodemographic characteristics, which is in line with findings by Attanasio, Boneva, and Rauh (2019), but contrasts with Boneva and Rauh (2018). My findings imply that parental beliefs about returns to parenting styles and neighborhoods are similar for parents from different socioeconomic backgrounds and thus unlikely to explain socioeconomic differences in parenting behavior. Despite the absence of socioeconomic differences in perceived returns, there are systematic variations. In particular, I show that parenting values – parents' altruism and paternalism towards their own child – are strongly related to perceived returns. In particular, altruistic parents expect high payoffs for being responsive (high warmth) and living in good neighborhoods, while paternalistic parents expect larger returns to exerting control. These patterns provide empirical support for assumptions made in Doepke and Zilibotti (2017) and underline the role of parental preferences and parenting values for understanding perceived returns.

Finally, I investigate whether perceived returns are relevant for actual parenting behavior. Importantly, I find that perceived returns to both parenting dimensions are related to actual parenting behavior in the respective dimension: parents who expect larger returns to warmth (control) are more likely to raise their own children with warmth (control), highlighting that parental beliefs are consistent with actual behavior.

These results contribute to three strands of the literature. First, the paper relates to a growing literature on subjective expectations in the context of human capital formation.[4,5] It is most closely connected to studies of parental beliefs about the process of human capital formation pioneered by Cunha, Elo, and Culhane (2013). Boneva and Rauh (2018) and Attanasio, Boneva, and Rauh (2019) build on their hypothetical scenario approach to study the timing (childhood or adolescence) or type of investment (time or money), while Bhalotra et al. (2017) consider different forms of time investments (intensity of breastfeeding and child interaction). By contrast, I study a different margin by allowing the mode of interaction, i.e., the parenting style, to vary. The rationale behind this is that a time investment of one hour can have different effects, depending on the intensity of parent-child interactions and thus I pay attention to the quality rather than the quantity margin of parental investments. Apart from analyzing a new and distinct margin, I also add methodologically to this literature on subjective expectations by embedding a second belief measure to correct for measurement error. In particular, when studying the relevance of perceived returns for actual behavior, I lever two distinct measures of the same underlying factor, but measured in different domains (monetary or life satisfaction) to mitigate attenuation bias by applying a measurement error correction proposed by Gillen, Snowberg, and Yariv (2019).

Second, I contribute to a series of papers that explicitly incorporate parenting styles in addition to parental investments in their analyses. These studies analyze the development (Cobb-Clark, Salamanca, and Zhu, 2019; Cunha, 2015; Del Bono, Francesconi, Kelly, and Sacker, 2016; Ermisch, 2008; Fiorini and Keane, 2014) and intergenerational transmission of skills and preferences (Brenøe and Epper, 2019; Falk, Kosse, Pinger, Schildberg-Hörisch, and Deckers, forthcoming; Zumbuehl, Dohmen, and Pfann, 2018), a child's behavior (Dooley and Stewart, 2007) or school outcomes (Cosconati, 2012). While these papers, as well as the developmental psychology literature, are primarily concerned with the consequences of particular investments or parenting styles for child outcomes, Doepke and Zilibotti (2017) and Doepke, Sorrenti, and Zilibotti (forthcoming) choose a different approach. They focus on parental decision-making and argue that economic incentives created by

4. The literature discussed here builds on a growing literature analyzing students' subjective expectations about schooling decisions (Attanasio and Kaufmann, 2014; Giustinelli, 2016; Jensen, 2010; Kaufmann, 2014) and major choices (Arcidiacono, Hotz, and Kang, 2012; Beffy, Fougère, and Maurel, 2012; Hastings et al., 2016; Stinebrickner and Stinebrickner, 2014; Wiswall and Zafar, 2015; Zafar, 2013), or family and job preferences as well as the resulting gender differences (Kiessling, Pinger, Bergerhoff, and Seegers, 2019; Wiswall and Zafar, 2018a; Wiswall and Zafar, 2018b).

5. I focus on a literature that assesses subjective expectations as part of decision-making processes. In the context of parental beliefs, there also exist some papers (e.g., Dizon-Ross, 2019; Kinsler and Pavan, 2018) that concentrate on the accuracy of parental beliefs for outcomes such as the performance of children in school. For these outcomes, parents can learn about realizations and thus verify beliefs in principle, while for the subjective expectations considered here this is typically not the case.

the environment shape parents' parenting style choices.[6,7] The present paper complements these papers by focusing on the parental decision-making process and by presenting evidence on the perceived long-term consequences of different parenting styles in two relevant domains – earnings and life satisfaction. Moreover, my results provide support for modeling choices made in Doepke and Zilibotti (2017), namely that parental altruism and paternalism are key to understanding the choice of parenting styles.

Lastly, the paper relates to the literature showing how neighborhoods affect long-term outcomes of children (see, e.g., Chetty, Friedman, Hendren, Jones, and Porter, 2018; Chetty and Hendren, 2018a; Chetty and Hendren, 2018b; Deutscher, forthcoming, for evidence that neighborhood exposure affects a variety of social and economic outcomes) and the literature that analyzes parents' behavioral responses. Kling, Liebman, and Katz (2005), Pop-Eleches and Urquiola (2013), and Han (2019) provide evidence that parents are more involved in their children's upbringing in low-quality neighborhoods. By contrast, Patacchini and Zenou (2011) suggest that parental involvement actually increases with neighborhood quality. I contribute to this discussion by providing first evidence on parental perceptions of both neighborhood effects, as well as their interactions with parenting decisions. Moreover, my results show that parents perceive the returns to high warmth *or* high control parenting as being relatively larger in low quality neighborhoods. However, these effects are reversed when I analyze authoritative parenting styles, characterized by high levels of warmth *and* control. This indicates that conflicting findings in the literature may be due to a focus on different parenting behaviors. Collectively, these papers as well as my paper therefore suggest that the way in which parents raise their children interacts with neighborhood quality, thus pointing towards an additional mediator of neighborhood effects besides schools (e.g., Laliberté, 2018) or peers (e.g., Agostinelli, 2018).

6. In particular, they focus on inequality and occupational mobility (in terms of an incumbency premium) as two features of the environment that create such incentives. Using data from the World Value Survey, Doepke and Zilibotti (2017) provide cross-country evidence that these two measures correlate with average parenting styles in a country. Dohmen, Golsteyn, Lindahl, Pfann, and Richter (2019) provide related evidence from Sweden that the effectiveness of parenting styles indeed hinges on the economic environment. In contrast to these papers, I focus on the decision-making process of individual parents and ask whether parents think such associations exist, and study the size of and heterogeneity in those perceived returns, investigating whether these perceived returns are related to the actual decision-making of parents.

7. Relatedly, Cuellar, Jones, and Sterrett (2015) review the psychological literature on the relationship between parenting styles and neighborhoods. While a general finding in developmental psychology is that an authoritative form of parenting is most effective in raising successful children, there exists a large variety in adopted parenting styles (e.g., Chan and Koo, 2011; Dornbusch, Ritter, Leiderman, Roberts, and Fraleigh, 1987; Lamborn, Mounts, Steinberg, and Dornbusch, 1991; Steinberg, Mounts, D., and Dornbusch, 1991).

In the next section, I describe the main survey instrument as well as the data collection process. Section 3.3 documents parents' beliefs about the returns to parenting styles and neighborhoods before Section 3.4 turns to an individual-level analysis. Section 3.5 examines the relevance of individual perceived returns for parental decision-making. Finally, Section 3.6 concludes.

3.2 Survey Description and Data

My aim is to study parental beliefs about the effectiveness of different parenting styles and to analyze their interaction with the economic environment a family is living in. In order to study these beliefs, I conduct a survey with a representative sample of 2,119 parents in the United States. In this section, I describe the survey instrument and the sample for this study.

3.2.1 Hypothetical Scenario Approach

Analyzing parental beliefs is difficult for several reasons: First, inferring beliefs from observed behavior can be challenging, as different sets of preferences and beliefs can in principle rationalize a given action (Manski, 2004). Second, eliciting beliefs only about the consequences of one's own actual parenting style ignores important counterfactual beliefs that are an integral part of the decision-making process (Arcidiacono, Hotz, and Kang, 2012). Third, collecting beliefs about the parents' own behavior towards their child might trigger motivated or self-serving beliefs, resulting in over- or understating of their beliefs. In order to circumvent these issues, I adopt a hypothetical scenario approach used by Cunha, Elo, and Culhane (2013), Boneva and Rauh (2018), Bhalotra et al. (2017), as well as Attanasio, Boneva, and Rauh (2019), and elicit beliefs about the consequences of different parenting styles directly. These scenarios have the advantage of allowing me to elicit returns over different dimensions and counterfactuals by varying one dimension at a time while holding other factors constant. In addition, by asking about the consequences of a hypothetical family, I reduce the scope for self-serving beliefs.

The survey instrument consists of different scenarios varying the parenting style of parents, as well as the quality of the environment in which a family is living. I adopt the typology of parenting styles introduced by Baumrind (1967) and further specified by Maccoby and Martin (1983) and vary whether parents raise their children with high or low warmth, as well as high or low control. The combination of these two dimensions results in four distinct parenting styles: neglecting (low warmth, low control), authoritarian (low warmth, high control), permissive (high warmth, low control), and authoritative (high warmth, high control). In order to study how the effectiveness of these different parenting styles depends on the quality of the neighborhood, I elicit parents' expectations about the consequences of the four parenting styles in two different environments: one neighborhood (the "good" neighborhood) describes an environment with low unemployment and little crime, while the other has relatively high unemployment and more crime ("bad" neighborhood). This allows me to test whether parents believe that the effectiveness of different parenting styles hinges on the environment in which a family is living, as suggested in Doepke and Zilibotti (2017). Moreover, this enables me to examine whether parents perceive one parenting style as optimal, independently of the socioeconomic environment. Table 3.1 summarizes the resulting eight scenarios.

Table 3.1. Survey scenarios

Bad neighborhood (n_L)

	Low control (c_L)	High control (c_H)
Low warmth (w_L)	y_1	y_2
High warmth (w_H)	y_3	y_4

Good neighborhood (n_H)

	Low control (c_L)	High control (c_H)
Low warmth (w_L)	y_5	y_6
High warmth (w_H)	y_7	y_8

Notes: This table summarizes scenarios j ($j = 1, \ldots, 8$) in which respondents are asked to provide expected earnings for children at age 30 (y_j) for different parenting style combinations (low and high warmth/control) and neighborhoods (low or high neighborhood quality).

More specifically, I present respondents two hypothetical average American families, each having a single child whose age and gender are randomly determined, as described below. The two families differ only in the neighborhood in which they are living. One family, the "Joneses", lives in a good neighborhood that has a relatively low unemployment rate (2%), as well as a low crimes rate (10 violent crimes per 10,000 inhabitants). The other family, the "Smiths", lives in a relatively deprived neighborhood with higher unemployment (10%), as well as a higher crime rate (60 violent crimes per 10,000 inhabitants).[8] The scenarios stress that apart from living in different neighborhoods, both families have similar levels of education and income, and both families invest equal levels of time and money in their children. Across scenarios, I vary the warmth and control dimension of the parenting styles (low-low, low-high, high-low, high-high). In order to describe different parenting styles, I adopt descriptions based on established measures of parenting styles for warmth as well as control and vary the number of times parents engage in a certain

8. The underlying idea is that unemployment and crime rates correspond to measures of a latent neighborhood quality factor that potentially subsumes several other facets such as school quality or the availability of amenities. Similar proxies for neighborhood quality have been used before (e.g., Han, 2019).

type of behavior by one standard deviation of the respective distribution.[9] Appendix 3.A presents the wording of the scenarios.

Taken together, the hypothetical scenarios vary (a) the parenting style a family adopts by varying the intensity of the two dimensions warmth and control from low to high, and (b) the quality of the family's neighborhood ("good" or "bad" characterized by high or low unemployment and crime). Importantly, respondents are asked not only about one of the scenarios, but answer all of them. This feature allows me to infer the perceived returns over all three dimensions warmth, control, and neighborhood quality for each individual. By comparing individual responses across these scenarios, I am able to infer perceived returns of the three dimensions as well as their relationship in terms of their substitutability and complementarity.

3.2.2 Outcomes

The survey instrument elicits respondents' expectations for two outcomes of the hypothetical children at age 30. First, as a main outcome, I elicit parents' expectations about the expected gross yearly earnings of the children in terms of today's USD if they are working full-time. This measure allows me to calculate monetary returns over the different dimensions. In order to test whether the inferred returns carry over to other dimensions, I also elicit the expected life satisfaction at age 30 as a second outcome (measured on a scale from 1, low, to 100, high). Moreover, I can use this measure to correct for measurement error when analyzing the relation of perceived returns and actual parenting behavior. To do this, I adopt the "obviously related instrumental variable" approach proposed by Gillen, Snowberg, and Yariv (2019).

3.2.3 Randomizations

In order to analyze the extent to which parental beliefs depend on the characteristics of the child, I implement two randomizations: First, I randomly determine the gender of the child.[10] One group answers the scenarios in which both families have sons ("John" or "Simon"), while for another group, the families have daughters ("Emily" or "Sarah").[11] By comparing elicited beliefs between respondents seeing a

9. Before the actual survey, I conducted a pilot study to decide on the items in the scenarios. In particular, I chose the items that had the highest predictive power for the warmth and control dimension of parenting styles. In addition, I elicited the number of times parents engage in the respective behavior to obtain estimates of the frequency distribution. This distribution was then used to calibrate the scenarios corresponding to approximately one standard deviation difference between low and high intensities of warmth and control.

10. The randomization of gender and age is on the level of the respondent and not on the level of the hypothetical family. In other words, both families a respondent sees have either sons or daughters only, and these children have the same age.

11. These names correspond to the most popular names at the beginning of the 2000s, i.e., at a time when the hypothetical children of the scenarios were born.

son or a daughter, I can study gender differences in perceived returns. Second, the age of the child in the scenarios is randomly drawn from a uniform distribution between 6 and 16 years. The rationale for this is to analyze whether specific parenting styles are perceived more effective in certain periods as the literature on parental investments has identified periods during childhood which are crucial for skill development and long-term outcomes of children (Cunha and Heckman, 2007; Cunha, Heckman, and Schennach, 2010). Similarly, this helps to analyze whether parents perceive neighborhoods to be particularly important at certain ages.

3.2.4 Additional Survey Elements

In addition to the hypothetical scenarios described above and standard socioeconomic characteristics, the survey elicits respondents' actual parenting styles. To do this, I adopt two established measures of parenting styles as used in the German Socioeconomic Panel Study (SOEP). In particular, I use the short versions of the warmth and control dimension of parenting styles employing three- and four-items scales based on Perris, Jacobsson, Lindström, Knorring, and Perris (1980) and Schwarz, Walper, Gödde, and Jurasic (1997), respectively. Moreover, I elicit several parenting values such as the parents' belief about the malleability of their child's skills and the degree of altruism as well as paternalism towards their child.[12]

Furthermore, I ask parents to assess the quality of the neighborhood in which they are living by eliciting their agreement to the three statements (i) "My neighborhood is a good place to raise children", (ii) "I feel safe in my neighborhood", and (iii) "My child attends a school of good quality", which I use to extract a factor for subjective neighborhood quality. Additionally, based on respondents' postcodes, I can link several neighborhood characteristics provided by Chetty and Hendren (2018a,b).

3.2.5 Summary Statistics

In October and November 2018, I collected a sample of 2,119 parents in the United States in collaboration with the market research company *Research Now*. To be eligible to take part in the study, respondents have to share a household with at least one child aged between 6 to 16, and respondents were sampled to be representative in terms of their gender, age, household income, and geographic distribution.[13] Table 3.2 presents sociodemographic statistics of the final sample and the Current Population Survey (CPS): 61% of the respondents are female, with an average age of 40 years. The average household has an annual income of USD 82,644 and matches

12. These values are measured using the agreement of parents to the following statements: "I am usually willing to sacrifice my own desires to satisfy those of my child" (altruism), "As a parent, I sometimes need to be strict if my child acts against what I think is good for it" (paternalism), and "My child develops at its own pace, and there is not much I can do about that" (malleability of skills).

13. If more than one child in this age range is present in the household, one child is randomly selected, and answers to child-specific questions are elicited with regard to this child.

the geographic distribution across census regions similar to the Current Population Survey (CPS). Moreover, the sample also matches several non-targeted characteristics, such as the share of married respondents (75%) and the average number of children (2.13), but has slightly higher level of education and a lower level of employment than the CPS sample.

Table 3.2. Summary statistics

	(A) Sample		(B) CPS
	Mean	SD	Mean
Sociodemographic variables			
Female	0.61	0.49	0.57
Age	40.25	7.38	40.89
Employed	0.72	0.45	0.79
College degree	0.52	0.50	0.36
Household income (in USD)	82644	55117	78018
Family structure			
Married	0.75	0.43	0.74
Cohabitating	0.08	0.27	
Single parent	0.16	0.37	
Number of children	2.13	1.08	2.05
Share of female children	0.46	0.37	
Geographic distribution			
Northeast	0.16	0.37	0.15
Midwest	0.19	0.40	0.21
South	0.39	0.49	0.37
West	0.26	0.44	0.27
Observations	2119		

Notes: This table presents summary statistics of the sample collected for this study in Panel (A) and representative statistics of American parents based on the Current Population Survey (CPS) in Panel (B).

3.3 Parental Beliefs about the Effectiveness of Parenting Styles and Neighborhoods

In this section, I study parental beliefs about the effectiveness of different parenting styles and neighborhoods. I begin by documenting the beliefs in the scenarios elicited in the survey and estimate returns to different levels of warmth, control, and neighborhood quality. In a second step, I analyze whether the perceived returns for boys and girls, as well as younger and older children, differ from each other based on randomizations across respondents.

In order to analyze parental beliefs, I estimate the perceived returns to different parenting styles and neighborhoods by comparing an individual's beliefs in different scenarios to each other. I therefore identify returns from the within-respondent variation in beliefs. More specifically, let w_j and c_j be equal to 1 if scenario j corresponds to a parenting style with high warmth or high control, respectively, and zero otherwise. Analogously, let n_j be equal to 1 if scenario j corresponds to a high-quality neighborhood, and zero otherwise. Moreover, y_{ij} denotes respondent i's expectation over the gross yearly earnings of a child at age 30 in scenario j. My main specification is then given by

$$\begin{aligned} \log(y_{ij}) &= \beta_w w_j + \beta_c c_j + \beta_n n_j \\ &\quad + \beta_{wc}(w_j \times c_j) + \beta_{wn}(w_j \times n_j) + \beta_{cn}(c_j \times n_j) + f_i(X_i) + \epsilon_{ij}. \end{aligned} \tag{3.1}$$

The main coefficients of interest are $\beta_w, \ldots, \beta_{cn}$, which describe the parents perceptions about the returns to the different factors. While β_k with $k = w, c, n$ denote the first-order returns to warmth, control, and neighborhoods, the coefficients on the interaction terms (k=wc, wn, cn) capture whether two dimensions are complements ($\beta_k > 0$) or substitutes ($\beta_k < 0$). Positive coefficients on interaction effects therefore imply that parents expect the return of two dimensions to increases when they are paired; negative coefficients mean that the returns are jointly lower than separately. The term $f_i(X_i)$ either controls for a vector of individual-specific characteristics ($f_i(X_i) = X_i'\gamma$) or individual fixed effects ($f_i(X_i) = \delta_i$) to absorb any observed or unobserved heterogeneity across individuals, respectively. Finally, ϵ_{ij} is an idiosyncratic error term clustered on the individual level.[14]

Estimating equation (3.1) on the whole sample yields perceived returns to parenting and neighborhoods for a representative set of parents in the United States. In the following, I will also lever the individual panel dimension of the data to infer individual-level perceived returns that I can subsequently link to their determinants and actual decision-making. For this, I estimate a simplified version of equation (3.1) for each respondent separately. This recovers individual-level perceived returns denoted by $R_{warmth,i}$, $R_{control,i}$, and $R_{neighborhood,i}$ for warmth, control, and neighborhoods.[15]

14. In my main specification, I follow Boneva and Rauh (2018) and Attanasio, Boneva, and Rauh (2019), and restrict my attention to interactions of two dimensions (warmth and control, warmth and neighborhoods, control and neighborhoods). In Table 3.1, I present additional results including a triple interaction of high levels of warmth and control, as well as good neighborhoods.

15. This approach differs from Attanasio, Boneva, and Rauh (2019) and Boneva and Rauh (2018). They calculate returns to each dimension by calculating log differences between high and low characteristics and averaging over the other two dimensions. By contrast, I estimate the same specification as used for the whole sample to obtain individual-level returns and thereby control for the presence of perceived interaction effects.

3.3.1 Representative Evidence on Perceived Returns

How do parents' expectations vary over the scenarios, and what returns do they associate with different parenting styles and neighborhoods? Figure 3.1 depicts the mean parental beliefs for each of the eight scenarios from Table 3.1. Several findings emerge: First, parental beliefs for earnings of a child at age 30 vary strongly across scenarios ranging between USD 40,000 and USD 57,000, with an average of USD 47,810.[16] Second, comparing the same parenting styles across neighborhoods reveals that parents expect large returns to neighborhoods. Being raised in a relatively good neighborhood increases expected earnings by USD 7,000 to USD 8,000 on average. Third, there are sizable returns to different parenting styles. Parents expect authoritative parenting with high levels of warmth and control to compensate partly for raising children in low-quality neighborhoods. Moreover, the patterns suggest that the different dimensions interact with each other.

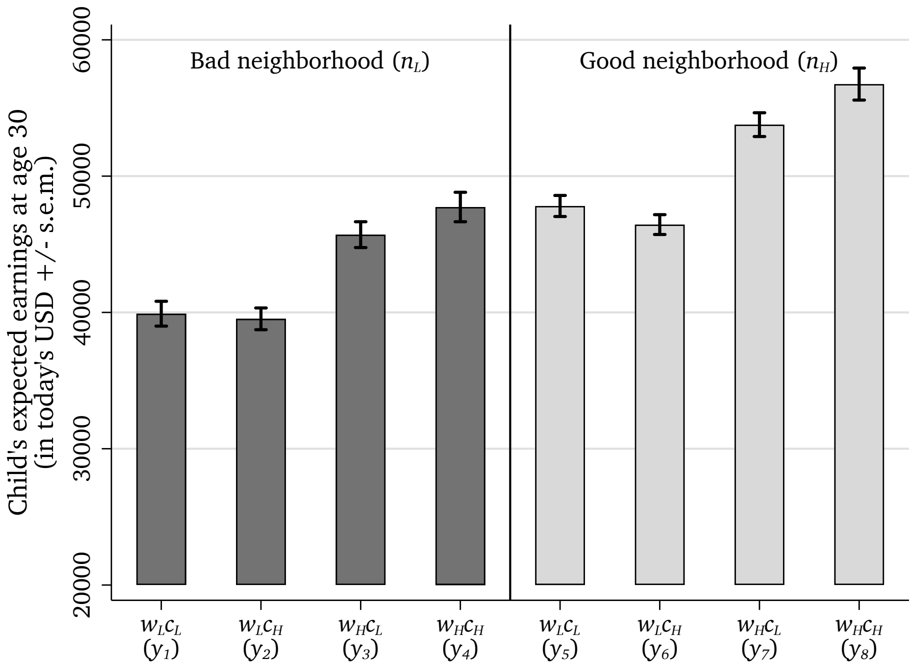


Figure 3.1. Parental beliefs about expected earnings

Notes: This figure presents parents' expectations about a child's earnings at age 30 in each of the eight scenarios. The first four bars correspond to scenarios with low neighborhood quality, while the latter four bars correspond to scenarios with high neighborhood quality. Moreover, w_jc_k (j, k = L, H) indicate different parenting styles with a low (w_L) or high level of warmth (w_H) and a low (c_L) or high level of control (c_H), respectively; cf. Table 3.1. Error bars indicate standard errors to the mean.

16. Conditional on working, respondents in the CPS earn approximately USD 46,200 at age 30.

In order to analyze these patterns in more detail, Table 3.1 presents OLS estimates as specified in equation (3.1). In columns (1) through (3), I focus on returns to primary dimensions only, while columns (4) to (6) acknowledge the presence of interactions between different dimensions of parenting styles as well as neighborhoods. Finally, column (7) investigates the interaction of all three dimensions and asks whether authoritative parenting (high warmth and high control) is more effective in good neighborhoods.

I find that parents perceive large returns to the warmth and neighborhood dimension, but no returns from exerting control. Increasing the warmth dimension of parenting by one standard deviation in column (1) increases a child's expected earnings by 16.9 percent, while the estimated perceived return to control is statistically indistinguishable from zero with a 95% confidence interval ranging from -0.2 to 1.2 percent. The perceived return to neighborhoods amounts to 21.1 percent. Neither the inclusion of sociodemographic controls in column (2) nor taking out all individual-level unobserved heterogeneity by including individual fixed effects in column (3) affects the coefficients of interest, i.e., the returns to warmth, control, and neighborhoods.

Columns (4) through (6) additionally allow for interaction effects between warmth, control, and neighborhoods. These specifications allow, for example, that the warmth and control dimensions of parenting styles are perceived as substitutes or complements, or that returns to parenting differ across neighborhoods. First, I find that the primary effects on the dimensions are similar to the previous estimates without interactions. Second, when considering interaction terms, the estimates reveal a perceived complementarity between warmth and control. Parents expect an additional return of 4.6 percentage points if children are raised with *both* high levels of warmth and control. Hence, parents expect authoritative forms of parenting (i.e., high warmth and high control) to be most effective for children's long-term success. This is similar to what has been found in the psychology literature (Baumrind, 1967; Dornbusch et al., 1987; Lamborn et al., 1991). Interestingly, there are negative interactions of good neighborhoods with warmth and control. Thus, parents perceive parenting to be *more* important in relatively adverse environments or strict parenting is less necessary if the surrounding conditions are favorable. In other words, respondents expect parenting to partly compensate for the lack of a beneficial neighborhood. This is consistent with the observation that parents become more involved in raising their children when the quality of a neighborhood decreases (e.g., Han, 2019; Kling, Liebman, and Katz, 2005; Pop-Eleches and Urquiola, 2013).[17]

Finally, column (7) introduces a triple interaction of high levels of warmth and control, as well as living in a good neighborhood, and thus measures the additional

17. For example, Kling, Liebman, and Katz (2005) provide evidence that families in high-poverty neighborhoods spend a large fraction of their time monitoring their children and keeping them safe, i.e., they exert high levels of control in raising them.

Table 3.1. Beliefs about the returns to parenting styles and neighborhoods

	log. of expected earnings at age 30 ($\log(y_{ij})$)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High warmth	0.169***	0.169***	0.169***	0.153***	0.153***	0.153***	0.163***
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.009)
High control	0.005	0.005	0.005	-0.011	-0.011	-0.011	-0.000
	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)	(0.007)	(0.008)
Good neighborhood	0.211***	0.211***	0.211***	0.226***	0.226***	0.226***	0.236***
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)	(0.009)	(0.010)
High warmth				0.046***	0.046***	0.046***	0.026**
× High control				(0.010)	(0.010)	(0.010)	(0.012)
High warmth				-0.014*	-0.014*	-0.014*	-0.035***
× Good neighborhood				(0.008)	(0.008)	(0.008)	(0.009)
High control				-0.015**	-0.015**	-0.015**	-0.036***
× Good neighborhood				(0.006)	(0.006)	(0.006)	(0.009)
High warmth × High control							0.041***
× Good neighborhood							(0.013)
Individual-level controls							
Female		-0.075***			-0.075***		
		(0.022)			(0.022)		
Age		-0.004**			-0.004**		
		(0.002)			(0.002)		
White		-0.030			-0.030		
		(0.025)			(0.025)		
College degree		0.097***			0.097***		
		(0.022)			(0.022)		
Employed		-0.052**			-0.052**		
		(0.024)			(0.024)		
log(Household income)		0.196***			0.196***		
		(0.019)			(0.019)		
Single parent		0.021			0.021		
		(0.026)			(0.026)		
Number of children		0.002			0.002		
		(0.010)			(0.010)		
Share of female children		0.014			0.014		
		(0.025)			(0.025)		
Mean exp. income (in USD)	47810	47810	47810	47810	47810	47810	47810
Controls for heterogeneity	No	Controls	FE	No	Controls	FE	FE
Observations	16952	16952	16952	16952	16952	16952	16952
Individuals	2119	2119	2119	2119	2119	2119	2119
R^2	.052	.14	.73	.052	.14	.74	.74

Notes: This table presents least squares regressions of log earnings expectations based on equation (3.1). Columns (1) through (3) focus on first-order effects. Columns (4) to (6) additionally include two-way interactions, while column (7) also adds a three-way interaction of warmth, control and neighborhoods. Standard errors clustered by respondent in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

perceived return to authoritative parenting (high warmth and high control) in good neighborhoods. While the main conclusions remain qualitatively as well as quantitatively similar to the previous results, the additional interaction shows that parents perceive the complementarity of warmth and control to be stronger in favorable neighborhoods compared to detrimental ones. Thus, parents perceive neighborhoods and intensive parenting (i.e., authoritative parenting styles) as complements. As far as these perceptions correspond to actual returns, this result suggests that increasing segregation may help to explain why the rich adopt more intensive parenting styles with higher investments, while the poor invest investments less (see also the discussion in Doepke, Sorrenti, and Zilibotti, forthcoming). Moreover, this helps to reconcile the finding of cultural complementarity in Patacchini and Zenou (2011) with my previous findings, as well as other studies documenting substitution effects between neighborhoods and parenting (e.g., Pop-Eleches and Urquiola, 2013). While parents may try to compensate for the lack of a good environment by increasing their involvement in raising children, living in a high-quality neighborhood may induce an additional complementarity for very intensive forms of parenting (e.g., authoritative parenting).

3.3.2 Perceived Returns by the Child's Gender and Age

While the previous estimates are average returns across all scenarios, the design of the survey allows me to go one step further. In particular, I vary both the gender (male/female) as well as the age of the child in the scenario (6–16 years) across respondents. Table 3.2 analyzes whether parental expectations differ across these randomizations. As shown in columns (1) to (3), parents expect boys to earn more than girls when they are grown up. They expect boys to earn on average 49,492 USD and girls to earn around 7% less (46,123 USD). Despite these level differences, I do not find evidence for differences in the perceived returns across gender. Yet, there are significant changes in perceived returns when varying the age of the child. More specifically, the warmth dimension becomes less important the older the child is, according to parents' expectations. While for 6 to 9-year-old children a standard deviation increase yields a perceived return of 18.6 percent, it amounts to only 14.7 and 12.7 percent, respectively, for 10 to 12-year-old and 13 to 16-year-old children (corresponding t-tests of the difference between coefficients yield p-values of $p = 0.060$ and $p = 0.003$). In line with county exposure effects in Chetty and Hendren (2018a), I do not find evidence of perceived critical age effects, during which living in certain neighborhoods is crucial for long-run outcomes. Rather, I find that the interaction of the control dimension of parenting and neighborhoods is perceived to be of particular importance for older children. More specifically, parents associate control to yield a 2.9 percentage point return in adverse environments for the oldest age group in my sample. By contrast, there is no such effect for the youngest age group (test of the difference between coefficients: $p = 0.042$). Thus, parents adapt their return expectations to characteristics of children, such as their age.

3.3.3 Robustness Checks Using Different Sample Restrictions

In Table 3.3, I check the robustness of the main findings by restricting the sample in various ways. First, I restrict the sample in column (1) to those respondents who report being one of the main caregivers of the child. Second, after eliciting expectations in the scenarios, I asked how certain parents were about their responses and

Table 3.2. Perceived returns by child's gender and age

	$\log(y_{ij})$		p-value	$\log(y_{ij})$			p-values		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Boys	Girls	(1)-(2)	6-9 years	10-12 years	13-16 years	(4)-(5)	(4)-(6)	(5)-(6)
High warmth	0.163***	0.142***	0.209	0.186***	0.147***	0.127***	0.060	0.003	0.325
	(0.012)	(0.012)		(0.014)	(0.015)	(0.014)			
High control	-0.018*	-0.003	0.266	-0.018	-0.008	-0.005	0.557	0.434	0.832
	(0.010)	(0.010)		(0.013)	(0.012)	(0.012)			
Good neighborhood	0.219***	0.233***	0.462	0.241***	0.223***	0.214***	0.439	0.248	0.711
	(0.013)	(0.013)		(0.017)	(0.016)	(0.016)			
High warmth × High control	0.052***	0.040***	0.547	0.037**	0.052***	0.050***	0.510	0.575	0.924
	(0.013)	(0.014)		(0.017)	(0.016)	(0.016)			
High warmth × Good neighborhood	-0.023**	-0.006	0.255	-0.019	-0.019	-0.006	0.974	0.447	0.521
	(0.011)	(0.011)		(0.012)	(0.015)	(0.013)			
High control × Good neighborhood	-0.011	-0.020**	0.468	0.001	-0.017	-0.029***	0.237	0.042	0.454
	(0.009)	(0.008)		(0.010)	(0.012)	(0.011)			
Mean exp. income (in USD)	49492	46123		48373	46999	47915			
Controls for heterogeneity	FE	FE		FE	FE	FE			
Observations	8528	8416		5888	4896	6168			
Individuals	1066	1052		736	612	771			
R^2	.75	.71		.75	.74	.72			

Notes: This table presents least squares regressions of log earnings expectations based on equation (3.1) for different sample splits according to the child's gender (columns 1 and 2) and age group (columns 4-6). Reported p-values stem from t-tests of interaction terms in fully interacted regression models. Standard errors clustered by respondent in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

exclude in column (2) those who report being uncertain or very uncertain. Third, it is possible that respondents either pay little attention and quickly click through the survey or simply perform other activities besides answering the survey. I therefore exclude respondents with the 5% lowest and highest response times in column (3). Finally, I focus on those respondents who have children similar to those in the scenarios and potentially hold more accurate beliefs. Thus, I restrict the sample to those who have children of the same gender (column 4), the same age group (column 5), or both the same gender and age group (column 6). As shown in Table 3.3, neither excluding non-main caregivers, focusing on certain respondents only, or removing respondents with very short or long response times affects the estimates in columns (1) through (3). When restricting the sample to those respondents who answer scenarios with hypothetical children sharing their own children's characteristics, the estimates remain robust, although they lose some precision due to smaller samples.

Table 3.3. Robustness of perceived returns for different samples

	log. of expected earnings at age 30 ($\log(y_{ij})$)					
	(1) Main caregivers	(2) Certain response	(3) Response time	(4) Same sex	(5) Same age	(6) Same sex+age
High warmth	0.154***	0.155***	0.159***	0.152***	0.143***	0.138***
	(0.009)	(0.010)	(0.009)	(0.010)	(0.013)	(0.017)
High control	-0.012*	-0.012	-0.006	-0.011	-0.007	0.002
	(0.007)	(0.008)	(0.007)	(0.008)	(0.011)	(0.015)
Good neighborhood	0.226***	0.217***	0.235***	0.217***	0.227***	0.220***
	(0.010)	(0.011)	(0.010)	(0.011)	(0.015)	(0.021)
High warmth × High control	0.046***	0.051***	0.041***	0.040***	0.048***	0.037**
	(0.010)	(0.011)	(0.010)	(0.011)	(0.015)	(0.019)
High warmth × Good neighborhood	-0.014*	-0.009	-0.018**	-0.014	-0.009	-0.004
	(0.008)	(0.009)	(0.008)	(0.009)	(0.012)	(0.016)
High control × Good neighborhood	-0.014**	-0.017**	-0.015**	-0.013*	-0.012	-0.010
	(0.006)	(0.007)	(0.006)	(0.008)	(0.011)	(0.015)
Mean exp. income (in USD)	47835	48932	47040	48802	48390	49278
Controls for heterogeneity	FE	FE	FE	FE	FE	FE
Observations	16384	12792	15272	12312	7376	4000
Individuals	2048	1599	1909	1539	922	500
R^2	.74	.74	.73	.73	.73	.74

Notes: This table presents least squares regressions of log earnings expectations based on equation (3.1). Column (1) restricts the sample to respondents who are main caregivers to their children. Column (2) excludes parents who report being uncertain about their responses. Column (3) excludes respondents with the 5% highest and lowest response times. Columns (4) to (6) restricts the sample to parents whose children and the child in the scenario have the same characteristics in terms of gender (column 4), age group (column 5), and gender, as well as age group (column 6). Standard errors clustered by respondent in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

3.3.4 Relationship of Returns in the Earnings and Life Satisfaction Domain

The previous results stem from scenarios in which parents were asked about their expectations for children's earnings at age 30, who are raised with a particular parenting style and in a specific neighborhood. Although monetary returns are appealing for their ease of interpretation, one potential concern with them is that parents may not perceive expected earnings at age 30 as the relevant outcome to evaluate the consequences of different parenting styles. Parents may perceive non-monetary outcomes such as well-being or life satisfaction as more important. In order to test whether the results from the monetary domain are comparable to those from other domains, I study a second outcome measure, expected life satisfaction of children at age 30 (measured on a scale from 1 to 100), which parents may have in mind when deciding about the adoption of different parenting styles.

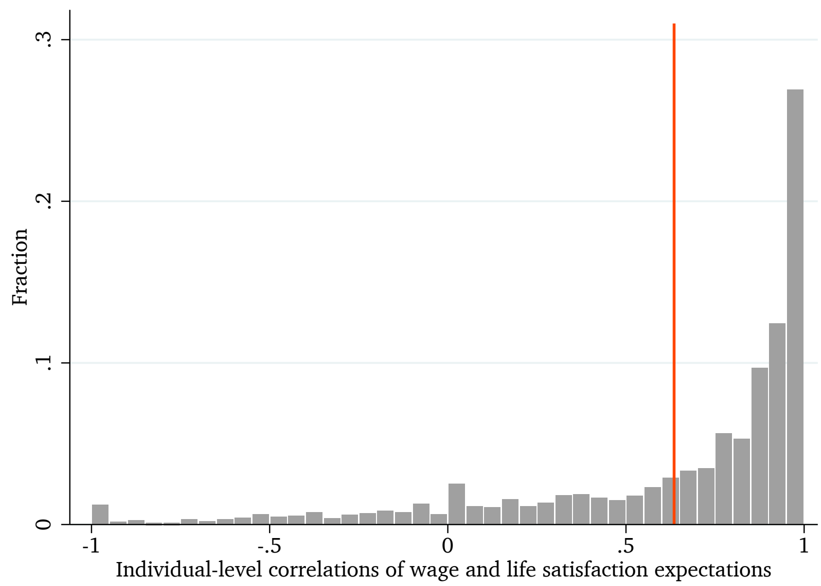


Figure 3.2. Correlations of earnings and life satisfaction expectations

Notes: This figure presents the distribution of individual-level correlations of earnings and life satisfaction expectations. The red line indicates the mean correlation across respondents of .63.

In Figure 3.2, I examine the relationship between expectations in the earnings and life satisfaction domain. More specifically, the figure displays the distribution of individual-level correlations between expectations across the two domains. For each individual, I calculate the correlation of their expectations for earnings and life satisfaction across the eight scenarios. As depicted, most correlations exceed 0.50

with a mean correlation of 0.63.[18] When analyzing the correlation of individual-level returns rather than levels, I also find strong correlations between returns in the monetary and returns in the life satisfaction domain, as shown in Appendix Table 3.B.1. Furthermore, Appendix Table 3.B.2 replicates Table 3.1 by using expected life satisfaction instead of expected earnings as an outcome. The results are both qualitatively and quantitatively similar. This implies that responses in terms of expected earnings are sensible outcomes, capturing returns that not only apply to a monetary domain. In the following, I therefore restrict my attention to monetary returns.

3.3.5 Accuracy of Beliefs and Perceived Returns

How accurate are the beliefs parents report in the scenarios? In this section, I briefly discuss their accuracy. As reported in Table 3.1, the average expected earnings across all eight scenarios is USD 47,810, which is similar to the mean annual earnings in the CPS (approx. USD 46,200 for individuals aged 30 and working). The coefficients on individual-level controls in columns (2) and (5) also reveal patterns consistent with findings from the literature on subjective wage expectations (e.g., Kaufmann, 2014): Females expect lower earnings, while college educated individuals as well as those with higher household incomes report higher earnings expectations. Moreover, similar to findings from the psychology literature (e.g., Chan and Koo, 2011; Dornbusch et al., 1987; Lamborn et al., 1991), parents associate neglecting parenting (low warmth and control) with low outcomes, and authoritative parenting (high warmth and control) with high future outcomes.

In order to compare the perceived returns to actual returns, I conduct two comparisons.[19] First, I compare perceived returns from my sample to average marginal effects of intensive parenting styles from Falk et al. (forthcoming). They estimate how children's skills develop as a function of intensive parenting styles. While they do not consider different dimensions of parenting styles (i.e., warmth and control), they construct a latent factor based on similar survey items. Falk et al. find marginal effects ranging from 0.313 to 0.424, which are somewhat higher than the combined effects of warmth and control reported in Table 3.1.[20] Second, I exploit the fact that

18. Appendix Figure 3.B.1 presents the distribution using rank correlations. These have the advantage of merely requiring an ordinal rather than a cardinal scaling for life satisfaction. The figure reveals that the individual-level correlations are even higher when relying on ranks rather than levels.

19. Boneva and Rauh (2018) show in a related setting, in which they analyze the perceived returns to parental investments at different ages, that the hypothetical scenario approach adopted in this paper yields perceived returns similar to actual returns.

20. Note that the outcomes I am interested in here are long-term outcomes at age 30. In contrast, Falk et al. (forthcoming) are interested in the development of skills during childhood. Since these skills translate only imperfectly into earnings, these higher returns are consistent with the perceived returns reported here.

respondents were asked to state their beliefs for children of *average* American families. I draw on data from the National Longitudinal Survey of Youth 1997 (NLSY97), in which children aged 12-17 in 1997 evaluate both their mothers' and their fathers' parenting style. Regressing the log earnings of respondents in 2013, when they were on average 30 years old, on indicators for warmth and control, as well as their interaction (see Appendix Table 3.C.1) reveals returns similar to the average perceived returns in my sample: The return to mother's warmth and control is .104 and .020, respectively, while the coefficient on the interaction is .026, indicating returns both quantitatively and qualitatively consistent with those in Table 3.1. Using their fathers' parenting styles yields similar results.[21] Taken together, the perceived returns in my dataset seem to be consistent with actual returns from other settings.

3.4 Heterogeneity in Individual-level Returns

The previous section documented perceived returns to different parenting styles and neighborhoods. Yet, these returns depict only average patterns. Hence, I additionally estimate equation (3.1) individual by individual to recover each parent's perceived returns, and subsequently link these to individual determinants.[22]

Figure 3.1 and Table 3.1 present the distributions of returns to the three dimensions warmth, control, and neighborhood. Several findings emerge. First, there is large heterogeneity in perceived returns. The majority of respondents expect positive returns to all three dimensions, with less than 20% of the sample expecting negative returns to warmth and neighborhoods. This number amounts to approximately 40% for control. Second, there is a sizable fraction of parents who do not expect parenting styles or neighborhoods to matter, with shares of 14% for neighborhoods to 32% in the control dimension. Third, correlations of returns across the three dimensions are positive, though not perfect, indicating that the different dimensions are related, but capture distinct concepts.[23] Taken together, most parents expect that parenting can pay off for children's long-term outcomes.

To what extent is the heterogeneity in the distribution of perceived returns systematic? One point of departure is to investigate potential differences in the perceived returns by parental gender. In particular, there is evidence that mothers spend

21. Note that these estimates are correlations and should not be interpreted as causal. Yet, respondents in the survey were asked to state their beliefs over the outcomes of children of *average* American families. Hence, looking at these basic regressions is informative, despite not accounting for measurement error, the endogeneity of parenting styles, and other confounding factors. In addition to monetary returns, Appendix Table 3.C.1 also presents results from the NLSY on children's high school GPA with similar patterns: The warmth dimension of parenting has large positive returns, while control has smaller, albeit positive returns.

22. To avoid results being driven by outliers, I winsorize perceived returns at the 1% and 99% level.

23. Moreover, expecting zero returns is highly correlated across the different dimensions; see Appendix Table 3.D.1.

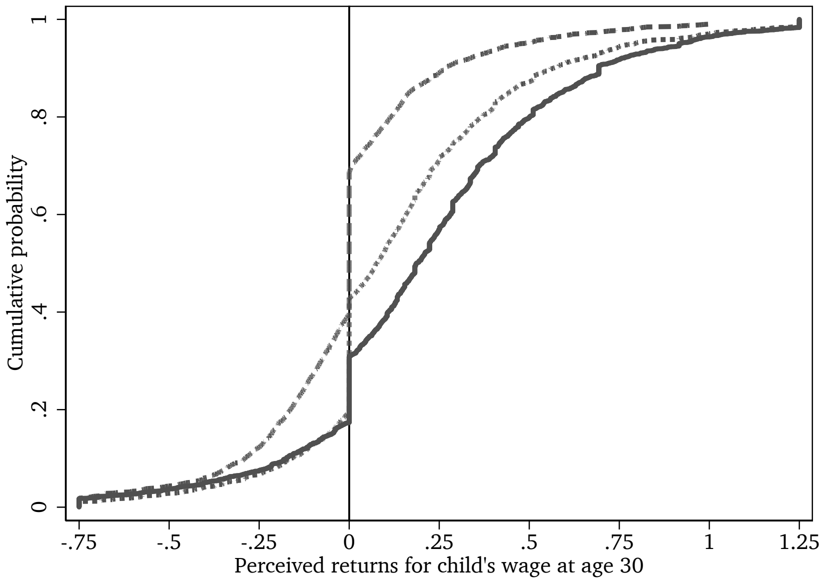


Figure 3.1. Distribution of individual-level perceived returns

Notes: This figure presents the distributions of individual-level perceived returns based on equation (3.1) for the dimensions warmth ($R_{warmth,i}$; dotted), control ($R_{control,i}$; dashed), and neighborhood ($R_{neighb.,i}$; solid).

Table 3.1. Correlations of individual-level perceived returns

	$R_{warmth,i}$	$R_{control,i}$	$R_{neighb.,i}$
$R_{warmth,i}$	1.000		
$R_{control,i}$	0.254***	1.000	
$R_{neighb.,i}$	0.290***	0.255***	1.000

Notes: This table presents correlations of individual-level returns across the three dimensions warmth ($R_{warmth,i}$), control ($R_{control,i}$) and neighborhood ($R_{neighb,i}$). *, **, and *** denote significance at the 10, 5, and 1 percent level.

about twice as much time on child-rearing activities as fathers (Guryan, Hurst, and Kearney, 2008).[24] I therefore analyze differences in the distribution of perceived returns between fathers (dark, dashed lines) and mothers (light, solid lines) in Figures 3.2a–3.2c to the three dimensions warmth, control, and neighborhood. The figures reveal significant gender differences in parental perceptions: Mothers expect larger returns than fathers in the warmth (t-test of equality of means: $p < 0.001$;

24. Moreover, data from the National Longitudinal Survey of Youth 1997 suggest that mothers are approximately 5.7 and 1.8 percentage points more likely to adopt parenting styles featuring high levels of warmth and control, respectively (see Appendix Table 3.C.2).

Kolmogorov-Smirnov tests of equality of distributions: $p < 0.001$) and neighborhood dimensions (t-tests: $p < 0.001$, KS-test: $p = 0.004$), while there are no significant differences in the control dimension (t-test: $p = 0.291$, KS-test: $p = 0.150$). Moreover, mothers' higher perceived returns seem to be relatively uniform across the distribution.

In the following, I analyze whether perceived returns are related to other parental characteristics besides gender. For this, I estimate

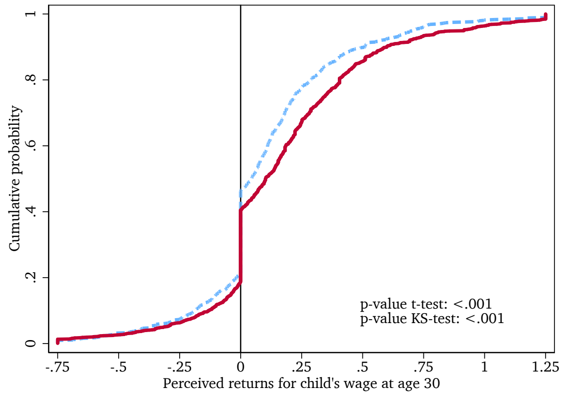
$$R_{k,i} \;=\; \alpha_0 \,+\, \alpha_1 X_i \,+\, \eta_{k,i}, \tag{3.1}$$

in which $R_{k,i}$ denotes the perceived return of individual i to dimension $k \in$ {warmth, control, neighborhood}, estimated based on equation (3.1), X_i is a vector of parental characteristics, and $\eta_{k,i}$ denotes idiosyncratic noise. I consider two sets of variables: First, I employ sociodemographic characteristics such as gender, age, and education; second, I associate returns with a respondent's parenting values (malleability of skills, altruism, and paternalism towards a child).

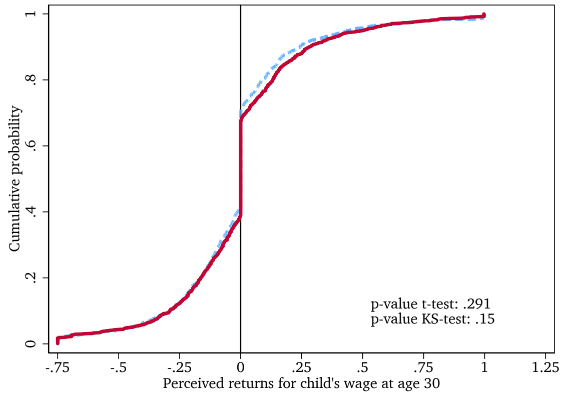
Table 3.2 presents estimates based on equation (3.1) for each return measure separately. Panel A focuses on sociodemographic determinants of perceived returns. Interestingly, apart from gender differences in the warmth (+6.2%) and neighborhood dimensions (+6.3%) as shown in Figure 3.2, almost no other characteristics seem to be systematically associated with perceived returns. In particular, I cannot reject the hypothesis that all other sociodemographic coefficients jointly equal zero in each of the three specifications regarding warmth (F-test: $p = 0.108$), control (F-test: $p = 0.935$), and neighborhoods (F-test: $p = 0.300$) in columns (1)–(3), respectively. The absence of a relationship is surprising, given that Boneva and Rauh (2018) find systematic associations for some characteristics, but it is in line with other studies (e.g., Attanasio, Boneva, and Rauh, 2019), which do not find associations either.[25] Thus, there are sizable differences in perceived returns by parental gender, but no differences along variables capturing differences in socioeconomic status. Moreover, these perceived returns are highly predictive for actual parenting styles, as I will show in the next section. The absence of associations between sociodemographics and returns therefore indicates that these beliefs capture an important aspect of parental decision-making that is distinct from standard individual characteristics and constraints.

When analyzing the effects of parenting values on returns in Panel B of Table 3.2, some interesting patterns emerge. All three return measures are significantly related to parents' beliefs about the malleability of skills, similar to Attanasio, Boneva, and Rauh (2019) and Boneva and Rauh (2018). In particular, those parents who believe

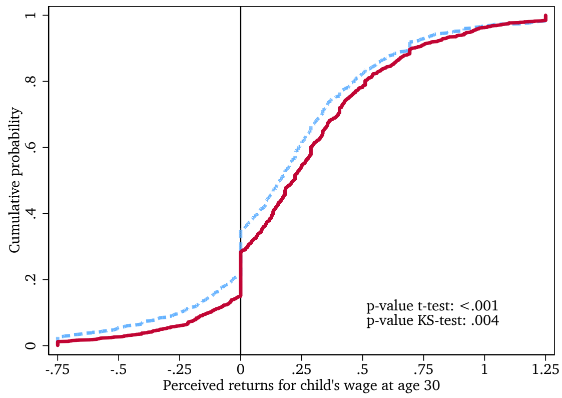
25. One explanation for these differences is that Boneva and Rauh (2018) and Attanasio, Boneva, and Rauh (2019) study families in the United Kingdom and only the latter study employs a representative sample of parents.



(a) Warmth ($R_{warmth,i}$)



(b) Control ($R_{control,i}$)



(c) Neighborhood ($R_{neighb.,i}$)

Figure 3.2. Distribution of individual-level perceived returns by parental gender

Notes: These figures present the distributions of individual-level perceived returns based on equation (3.1) for the dimensions warmth ($R_{warmth,i}$; Figure 3.2a), control ($R_{control,i}$; Figure 3.2b) and neighborhood ($R_{neighb.,i}$; Figure 3.2c) for mothers (solid, red) and fathers (dashed, blue) separately.

Table 3.2. Determinants of individual-level perceived returns

	(A) only sociodemographics			(B) incl. parenting values		
	(1)	(2)	(3)	(4)	(5)	(6)
	$R_{warmth,i}$	$R_{control,i}$	$R_{neighb.,i}$	$R_{warmth,i}$	$R_{control,i}$	$R_{neighb.,i}$
Sociodemographic characteristics						
Female	0.062***	0.015	0.063***	0.056***	0.014	0.052***
	(0.017)	(0.014)	(0.020)	(0.017)	(0.014)	(0.020)
Age	-0.000	-0.000	0.003***	-0.001	-0.001	0.002*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
White	0.047**	0.016	0.004	0.044**	0.017	0.005
	(0.019)	(0.015)	(0.022)	(0.019)	(0.015)	(0.022)
College degree	-0.005	-0.007	-0.004	-0.002	-0.003	0.001
	(0.019)	(0.015)	(0.020)	(0.019)	(0.015)	(0.020)
Employed	-0.021	-0.003	-0.016	-0.022	-0.004	-0.017
	(0.020)	(0.015)	(0.023)	(0.020)	(0.015)	(0.023)
log(Household income)	0.019	0.006	-0.003	0.018	0.005	-0.009
	(0.014)	(0.012)	(0.016)	(0.014)	(0.012)	(0.016)
Single parent	0.047*	-0.005	0.019	0.045*	-0.001	0.012
	(0.025)	(0.020)	(0.027)	(0.025)	(0.020)	(0.027)
Number of children	0.009	-0.005	0.002	0.009	-0.005	0.001
	(0.008)	(0.006)	(0.009)	(0.008)	(0.006)	(0.009)
Share of female children	-0.006	-0.000	0.009	-0.005	0.003	0.006
	(0.021)	(0.017)	(0.022)	(0.021)	(0.017)	(0.022)
Parenting values						
Altruism towards child (std.)				0.018**	-0.005	0.028***
				(0.007)	(0.006)	(0.009)
Paternalism towards child (std.)				0.001	0.013*	0.002
				(0.008)	(0.007)	(0.009)
Malleability of skills (std.)				0.030***	0.011*	0.050***
				(0.008)	(0.007)	(0.009)
Average return	.15	-.01	.23	.15	-.01	.23
Individuals	2119	2119	2119	2109	2109	2109
R^2	.015	.0017	.011	.023	.0053	.028

Notes: This table presents regressions of individual-level perceived returns to warmth ($R_{warmth,i}$; columns 1 and 4), control ($R_{control,i}$; columns 2 and 5) as well as neighborhood ($R_{neighb.,i}$; columns 3 and 6) on sociodemographic characteristics and parenting values according to equation (3.1). Individual-level perceived returns are estimated based on equation (3.1) for each individual separately. Robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

that skills are malleable perceive returns to be higher. In other words, those parents who do not share this belief react less to differences across scenarios. Moreover, returns in the warmth and neighborhood dimensions are related to the parents' altruism towards their children, whereas returns in the control dimension are associated with parental paternalism. This supports theoretical results by Doepke and Zilibotti (2017), who show that sufficiently paternalistic parents adopt parenting styles with more control, i.e., authoritarian or authoritative parenting styles in which parents exert effort to mold their children's preferences.[26] Parents' altruism and paternalism are two key parameters in their model, leading to different parenting styles.

In Appendix 3.D, I present a different approach to analyze the determinants of perceived returns. There, I rely on a measurement error correction based on two different measures of perceived returns: One measure in the monetary domain, as used throughout the paper, and another one using perceived returns in the life satisfaction domain. I adopt "obviously related instrumental variables" (Gillen, Snowberg, and Yariv, 2019), as outlined in the next section. In particular, I check whether perceived returns predict specific individual characteristics. The results in Appendix Table 3.D.3 confirm the previous patterns: Females expect larger returns to warmth as well as neighborhoods, and parenting values show the same associations as reported above.

3.5 Relevance of Perceived Returns for Actual Behavior

To what extent do perceived returns, as described above, map into actual parental decision-making? Although establishing causality without shifting parental beliefs is difficult, I can analyze the association of perceived returns with actual parenting behavior. Hence, I focus on the predictive power of returns for actual parenting styles. Remember that perceived returns capture some aspects of parenting that are not related to sociodemographic characteristics, but at the same time vary systematically with parenting behavior in the hypothetical scenarios. If perceived returns translated into actual parental decision-making, their relevance would be even higher in light of the lacking relationship to sociodemographic characteristics.

In order to examine the relevance of perceived returns, I relate the perceived returns from the hypothetical scenarios to the parents' actual parenting style elicited in the survey by estimating

$$PS_{k,i} \;=\; \delta_0 + \delta_1 R_{k,i} + \delta_2 X_i + \nu_{k,i}, \tag{3.1}$$

in which $PS_{k,i}$ denotes a standardized measure of the actual parenting style and $R_{k,i}$ correspond to the standardized individual-level perceived return estimated according to equation (3.1) in the warmth (k = warmth) or control (k = control)

26. Appendix Table 3.D.2 shows that respondents with more children, females, paternalistic parents, and those who believe that skills are malleable are less likely to expect zero returns to parenting. Accounting for respondents reporting zero returns does not change the results reported in this section.

dimensions, X_i is the same set of sociodemographic controls as before, and $\nu_{k,i}$ is an error term.[27,28]

In Table 3.1, I examine the relevance of perceived returns for actual parenting styles based on equation (3.1). Panel (A) focuses on the warmth dimension by relating estimated returns in the earnings (column 1) and life satisfaction domain (column 2) to warm parenting. The estimates reveal that returns in both domains are significantly related to parenting behavior. An increase of one standard deviation in perceived returns is associated with a .043 standard deviation increase in the warmth dimension of parenting styles. Although these individual-level returns are subject to measurement error, as they are estimated only on eight observations per respondent, they capture a similar underlying factor. I therefore lever the two different perceived return measures and implement the "obviously related instrumental variables" (ORIV) estimator proposed by Gillen, Snowberg, and Yariv (2019).[29] Applying this measurement error correcting in columns (3) and (4), I find even larger associations of .084-.088 standard deviations for an increase of one standard deviation in perceived returns that even hold when simultaneously controlling for perceived returns in the control dimension.

A similar picture arises when analyzing the role of perceived returns to control for the control dimension of parenting. While the perceived returns in the monetary domain are positive but insignificant at conventional levels ($p = 0.161$), accounting for measurement error using ORIV reveals significant associations even if controlling for return to warmth.

Finally, Appendix 3.F presents additional results linking perceived returns to neighborhoods to actual neighborhood characteristics. The results show that although there is only limited evidence for perceived returns to predict subjective neighborhoods assessments or economic conditions, perceived returns are strongly related to living in areas less segregation. In sum, the results from this section suggest that parents do not only adapt their expectations when faced with scenarios of varying parenting styles, but the corresponding returns are also relevant for their actual parenting behavior.

27. Appendix Figure 3.E.1 and Appendix Table 3.E.1 show that an exploratory factor analysis indeed recovers two factors corresponding to warmth and control from the set of survey items used to elicit a respondent's parenting style.

28. While the individual-level returns are estimated both for the first-order returns ($R_{k,i}$ for k = warmth, control, neighborhood) as well as interactions ($R_{k,i}$ for $k = wc, wn, cn$), I restrict attention to first-order returns as they dominate over interactions effects in size, as shown in Table 3.1.

29. Their estimation procedure involves duplicating each observation and use each measure once as a regressor and once as an instrument. In order to account for the larger number of observations, I follow Gillen, Snowberg, and Yariv (2019) and bootstrap standard errors.

Table 3.1. Relevance of perceived returns for actual parenting styles

	(A) Parenting Style – Warmth				(B) Parenting Style – Control			
	(1) Expected earnings	(2) Expected Life Satis.	(3) ORIV	(4) ORIV	(5) Expected earnings	(6) Expected Life Satis.	(7) ORIV	(8) ORIV
$R_{warmth,i}$	0.043**	0.044**	0.084***	0.088***				-0.039
	(0.017)	(0.019)	(0.021)	(0.025)				(0.026)
$R_{control,i}$				-0.021	0.026	0.055***	0.092***	0.102***
				(0.032)	(0.018)	(0.018)	(0.029)	(0.032)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2119	2119	4238	4238	2119	2119	4238	4238
Individuals	2119	2119	2119	2119	2119	2119	2119	2119
R^2	.046	.046	.044	.044	.036	.04	.045	.045

Notes: This table examines the relevance of perceived returns for actual parenting styles by estimating equation (3.1). Panel (A) presents results for the warmth dimension using (standardized) perceived returns to warmth, while Panel (B) presents corresponding results in the control dimension. Columns (1) and (5) use returns in the earnings domain, while columns (2) and (6) employ returns in the life satisfaction domain. Columns (3) and (7) implement ORIV estimators (ORIV, Gillen, Snowberg, and Yariv, 2019) to correct for measurement error in perceived returns using the two return measures as instruments for each other. Columns (4) and (8) additionally include (instrumented) perceived returns to control and warmth, respectively. All specifications include controls for sociodemographic characteristics as in Table 3.1. Robust standard errors in columns (1), (2), (5), and (6); bootstrapped standard errors in columns (3), (4), (7), and (8) in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

3.6 Conclusion

While parenting crucially affects the development of children, parenting itself remains a "mystifying subject" (Bornstein, 2002). In order to better understand how parents decide, I focus on parents' beliefs constituting an inherent part of their decision-making process. I conduct a survey that is among the first to investigate parental beliefs of a representative sample of parents. In the main part of the survey, I elicit beliefs using a hypothetical scenario approach that varies two factors with importance for the development of children and, hence, their long-term outcomes: first, the parenting style defined by the levels of warmth and control parents employ in raising their children, and, second, the quality of the neighborhood a family is living in. This allows me to infer parents' perceived returns to and sheds light on the perceived substitutability or complementarity of the different dimensions.

My analysis shows that parents expect large returns to high levels of warmth and beneficial neighborhoods. Parenting styles with high levels of control are only associated with positive returns if they are paired with warmth suggesting that these two dimensions are perceived as complements. Moreover, I show that parents expect parenting and neighborhoods to interact. In particular, they believe that parenting can partly compensate for living in deprived neighborhoods. Yet, the perceived return to authoritative parenting (i.e., high levels of warmth *and* control) is higher in good neighborhoods. This latter result indicates one potential explanation for divergent parenting practices along socio-economic groups: increasing segregation may lead to divergent parenting practices due to different perceived returns to intensive parenting (as also suggested by Doepke, Sorrenti, and Zilibotti, forthcoming), which may translate into different actual returns.

When studying perceived returns on the individual level, my estimates reveal some profound gender differences: mothers expect significantly larger returns than fathers in the warmth and neighborhood dimension, while parental perceptions are similar for the control dimension. Perhaps surprisingly, other sociodemographic characteristics of these perceived returns are not related to parental beliefs. The absence of a socio-economic gradient in perceived returns suggests that they are an unlikely candidate to explain socioeconomic differences in parenting behavior. Rather, the interaction between parenting and neighborhoods could provide an explanation for persistent differences in parenting across sociodemographic groups which might increase as neighborhoods become more homogeneous over time (Putnam, 2016). To the extent that some form of "optimal parenting" exists, my results suggest that the optimal parenting behavior may be environment-specific. Moreover, parenting values show distinct patterns: while paternalistic parents expect larger returns to control, altruistic ones perceive larger returns to warmth and neighborhoods. These findings extend previous research on the determinants of parental beliefs (Attanasio, Boneva, and Rauh, 2019; Boneva and Rauh, 2018) and lend empirical support for assumptions made in Doepke and Zilibotti (2017).

Importantly, the perceived returns I recover are relevant for actual parenting behavior. Hence, they capture an important determinant of parental decision-making, but cannot be proxied by standard socio-economic variables. This highlights the value of studying beliefs to understand parental decision-making processes.

The results of this paper open at least two avenues for further research. First, since the returns to parental investments hinge on the parenting style (Cunha, 2015), it would be interesting to analyze the relationship between the quality margin of parenting considered in this paper and the quantity margin as in the previous literature (Attanasio, Boneva, and Rauh, 2019; Bhalotra et al., 2017; Boneva and Rauh, 2018). Second, as beliefs about returns to parenting hinges on the quality of neighborhoods, this calls for a deeper understanding of the human capital formation process and the relationship between parenting and a family's environment.

Appendix 3.A Wording of Hypothetical Scenarios

In the following, I present the wording of the main survey instrument containing the hypothetical scenarios. Both the age (6-16 years) as well as the gender of the child in question (male/female) are randomized, resulting in male names (John and Simon) or female names (Sarah and Emily) for the children in the scenarios.

We are interested in your opinion about how important different parenting styles are for the future of children.

For this purpose, we would like to ask you to imagine two average American families, the Joneses and the Smiths, who make decisions how to raise their children. More specifically, we will show you different scenarios, and ask what you think the likely yearly earnings and life satisfaction of their children at age 30 will be. There are no clear right or wrong answers, and we know these questions are difficult. Please try to consider each scenario carefully and tell us what you believe the likely outcome will be.

Mr and Mrs Jones have one son (daughter), John (Sarah). John (Sarah) is 6 (7-16) years old. The Joneses live in a good neighborhood with little crime (10 violent crimes per 10,000 inhabitants) and low unemployment (2%). Now let's think about the future of John (Sarah). Assuming John (Sarah) is working full-time, what do you expect his (her) gross yearly earnings (in today's USD) to be when he (she) is 30 years old in each of the following scenarios? What do you expect his (her) life satisfaction to be at age 30 on a scale from 1 (low) to 100 (high)?

Scenario 1: *John (Sarah)'s parents show him (her) once per week that they like him (her). At the same time, they tell him (her) every other day that he (she) has to obey their decisions.*

Scenario 2: *John (Sarah)'s parents show him (her) once per week that they like him (her). At the same time, they tell him (her) once per week that he (she) has to obey their decisions.*

Scenario 3: *John (Sarah)'s parents show him (her) every other day that they like him (her). At the same time, they tell him (her) every other day that he (she) has to obey their decisions.*

Scenario 4: *John (Sarah)'s parents show him (her) every other day that they like him (her). At the same time, they tell him (her) once per week that he (she) has to obey their decisions.*

Now imagine a different family, the Smiths. In many respects, the Smiths are very similar to the Joneses. For example, Mr and Mrs Smith have one son (daughter), Simon (Emily), who is also 6 (7-16) years old and as smart as John (Sarah). Mr and Mrs Smith also have similar levels of income and education as Mr and Mrs Jones and *spend as much time and money on raising their child. However, there is one difference. Unlike the Joneses, the Smiths live in a bad neighborhood with much crime (60 violent crimes per 10,000 inhabitants per year) and high unemployment (10%). Assuming Simon (Emily) is working full-time, what do you expect his (her) gross yearly earnings (in today's USD) to be when he (she) is 30 years old in each of the following scenarios? What do you expect his (her) life satisfaction to be at age 30 on a scale from 1 (low) to 100 (high)?*

Scenario 5: *Simon (Emily)'s parents show him (her) once per week that they like him (her). At the same time, they tell him (her) every other day that he (she) has to obey their decisions.*

Scenario 6: *Simon (Emily)'s parents show him (her) once per week that they like him (her). At the same time, they tell him (her) once per week that he (she) has to obey their decisions.*

Scenario 7: *Simon (Emily)'s parents show him (her) every other day that they like him (her). At the same time, they tell him (her) every other day that he (she) has to obey their decisions.*

Scenario 8: *Simon (Emily)'s parents show him (her) every other day that they like him (her). At the same time, they tell him (her) once per week that he (she) has to obey their decisions.*

Appendix 3.B Relationship of Perceived Returns Across Domains

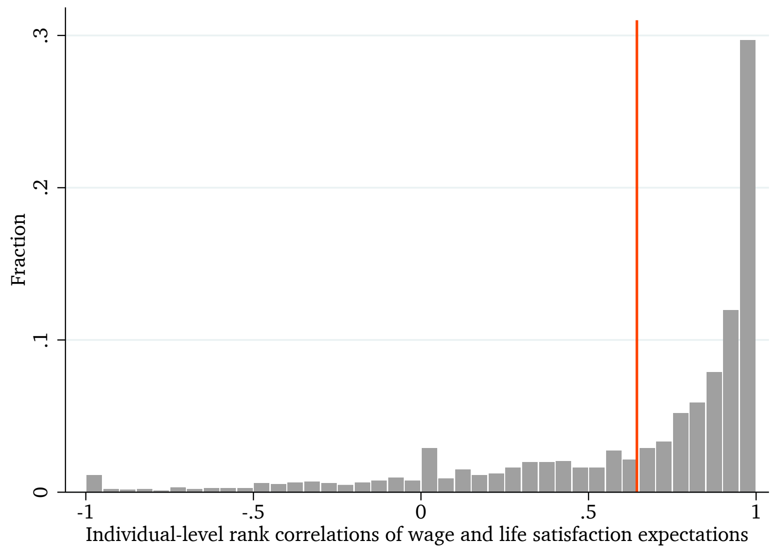


Figure 3.B.1. Rank correlation of earnings and life satisfaction expectations

Notes: This figure presents the distribution of individual-level rank correlations of earnings and life satisfaction expectations. The red line indicates the mean rank correlation across respondents.

Table 3.B.1. Relationship of perceived returns in earnings and life satisfaction domain

	$R^{LS}_{warmth,i}$		$R^{LS}_{control,i}$		$R^{LS}_{neighb.,i}$	
	(1)	(2)	(3)	(4)	(5)	(6)
$R_{warmth,i}$	0.652***	0.651***				
	(0.036)	(0.036)				
$R_{control,i}$			0.534***	0.533***		
			(0.040)	(0.040)		
$R_{neighb.,i}$					0.376***	0.377***
					(0.035)	(0.035)
Controls	No	Yes	No	Yes	No	Yes
Individuals	2119	2119	2119	2119	2119	2119
R^2	.28	.28	.2	.2	.11	.11

Notes: This table presents regressions of individual-level perceived returns in the life satisfaction domain ($R^{LS}_{k,i}$) on perceived returns in the monetary domain ($R_{k,i}$) for k = warmth, control, neighborhood. Returns are calculated from estimating equation (3.1) for each individual using either expected earnings ($R_{k,i}$) or expected life satisfaction ($R^{LS}_{k,i}$) at age 30 as an outcome. Controls include sociodemographic characteristics as in Table 3.1. Robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 3.B.2. Parental beliefs about perceived returns in the life satisfaction domain

	log. of expected life satisfaction at age 30 ($\log(ls_{ij})$)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High warmth	0.201***	0.201***	0.201***	0.180***	0.180***	0.180***	0.183***
	(0.009)	(0.009)	(0.009)	(0.011)	(0.011)	(0.011)	(0.011)
High control	-0.004	-0.004	-0.004	-0.033***	-0.033***	-0.033***	-0.030***
	(0.005)	(0.005)	(0.005)	(0.008)	(0.008)	(0.008)	(0.009)
Good neighborhood	0.145***	0.145***	0.145***	0.159***	0.159***	0.159***	0.162***
	(0.010)	(0.010)	(0.010)	(0.012)	(0.012)	(0.012)	(0.012)
High warmth × High control				0.063***	0.063***	0.063***	0.057***
				(0.012)	(0.012)	(0.012)	(0.014)
High warmth × Good neighborhood				-0.022**	-0.022**	-0.022**	-0.028***
				(0.009)	(0.009)	(0.009)	(0.011)
High control × Good neighborhood				-0.006	-0.006	-0.006	-0.011
				(0.008)	(0.008)	(0.008)	(0.011)
High warmth × High control × Good neighborhood							0.011
							(0.015)
Individual-level controls							
Female		0.000			0.000		
		(0.036)			(0.036)		
Age		0.003			0.003		
		(0.002)			(0.002)		
White		-0.026			-0.026		
		(0.040)			(0.040)		
College degree		-0.022			-0.022		
		(0.039)			(0.039)		
Employed		0.062			0.062		
		(0.042)			(0.042)		
log(Household income)		0.100***			0.100***		
		(0.031)			(0.031)		
Single parent		-0.034			-0.034		
		(0.050)			(0.050)		
Number of children		-0.009			-0.009		
		(0.019)			(0.019)		
Share of female children		0.057			0.057		
		(0.044)			(0.044)		
Mean exp. life satis. (0-100)	53	53	53	53	53	53	53
Controls for heterogeneity	No	Controls	FE	No	Controls	FE	FE
Observations	16952	16952	16952	16952	16952	16952	16952
Individuals	2119	2119	2119	2119	2119	2119	2119
R^2	.021	.034	.8	.021	.034	.8	.8

Notes: This table presents least squares regressions of log life satisfaction expectations based on equation (3.1). Columns (1) through (3) focus on first-order effects, while columns (4) to (6) add interactions. Standard errors clustered by respondent in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Appendix 3.C Parenting in National Longitudinal Survey of Youth 1997 (NLSY97)

Table 3.C.1. Gender differences in parenting styles (NLSY97)

	(A) Mother's PS		(B) Father's PS	
	(1) log(earnings)	(2) HS GPA	(3) log(earnings)	(4) HS GPA
Warmth	0.104**	0.271***	0.147***	0.294***
	(0.045)	(0.043)	(0.042)	(0.040)
Control	0.020	0.121**	0.076*	0.088**
	(0.051)	(0.048)	(0.046)	(0.043)
Warmth × Control	0.026	-0.018	-0.021	0.002
	(0.060)	(0.057)	(0.057)	(0.055)
Observations	5046	5832	4873	5645
R^2	.0037	.017	.0061	.023

Notes: This table uses data from the National Longitudinal Survey of Youth 1997 and regresses the child's log earnings in 2013 (i.e., when they are on average 30 years old) on the child's reports of each of its parents' parenting style. Columns (1) and (2) focus on the mother's warmth and control, while columns (3) and (4) report analogous regressions for fathers. Robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 3.C.2. Gender differences in parenting styles (NLSY97)

	(A) PS Warmth		(B) PS Control	
	(1)	(2)	(3)	(4)
Mother	0.057***	0.058***	0.018***	0.017***
	(0.004)	(0.004)	(0.003)	(0.004)
Mean of dependent variable	.65	.65	.55	.55
Observations	16968	12310	16968	12310
R^2	.0036	.035	.00032	.027

Notes: This table uses data from the National Longitudinal Survey of Youth 1997 and regresses the child's report of each of its parents' parenting style (measured by binary indicators) on an indicator for mothers. Columns (1) and (2) focus on the warmth dimension, while columns (3) and (4) focus on control. Control variables include the age and gender of the child, the parent's education, the log household income, and an indicator for whether both parents are present at home. Standard errors clustered on child-level in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Appendix 3.D Further Results on Determinants of Perceived Returns

To what extent are the associations reported in Section 3.4 driven by zero responses as shown in Figure 3.1? Table 3.D.1 shows that respondents who perceive no returns in one dimension are also more likely to also report zero returns in another. This pattern is especially pronounced for both parenting dimensions, suggesting that these individuals do not expect parenting to matter for long-term outcomes of children. Panel A of Table 3.D.2 shows that fathers, older respondents, as well as those with fewer children and who do not believe that skills are malleable are more likely to report zero responses in the parenting domains. Panel B shows how the results in Table 3.2 would change once I restrict the sample to respondents perceiving non-zero returns. The patterns are qualitatively and quantitatively similar to the whole sample.

Table 3.D.1. Correlations of zero perceived returns

	Warmth	Control	Neighb.
Warmth	1.000		
Control	0.823***	1.000	
Neighb.	0.365***	0.385***	1.000

Notes: This table presents correlations of indicators for whether a respondent expects zero returns to warmth, control, or neighborhoods. *, **, and *** denote significance at the 10, 5, and 1 percent level.

The perceived returns analyzed here are subject to measurement error as they are inferred from eight observations only. In order to mitigate the role of measurement error, I lever two distinct measures of the same underlying return measure. More specifically, I lever the perceived return measure constructed from parental beliefs in the life satisfaction domain to isolate the common variation in both measures. I adopt the "obviously related instrumental variables" (ORIV) estimator proposed by Gillen, Snowberg, and Yariv (2019). This estimator uses two measures of the same underlying dimension and instruments one measure with the other. Implementing the estimator requires that the (instrumented) perceived returns is an explanatory variable rather than the dependent variable as in Table 3.2. Hence, I perform the following exercise similar to the analysis in Section 3.5, in which I examine the relevance of perceived returns. I duplicate all observations and check whether perceived returns can predict a specific characteristic conditional on all other characteristics by estimating

$$x_i \;=\; \delta_0 + \delta_1 R_{k,i} + \delta_2 X_{i,-x_i} + \nu_i. \tag{3.D.1}$$

Here, $\delta_0 = (\delta_{0,m}, \delta_{0,ls})$ denote the constants corresponding to the original and duplicated observations, $R_{k,i}$ denotes the (standardized) perceived return in dimension k

Table 3.D.2. Perceived returns accounting for zero responses

	(A) Zero returns			(B) Returns excluding zeros		
	(1) Parenting	(2) Neighb.	(3) All	(4) $R_{warmth,i}$	(5) $R_{control,i}$	(6) $R_{neighb.,i}$
Sociodemographic characteristics						
Female	-0.037*	-0.001	-0.015	0.067***	0.020	0.060***
	(0.021)	(0.016)	(0.013)	(0.021)	(0.018)	(0.022)
Age	0.004***	0.001	0.001	0.000	-0.001	0.003**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
White	0.010	0.038**	0.024*	0.061***	0.022	0.020
	(0.022)	(0.017)	(0.013)	(0.024)	(0.019)	(0.025)
College degree	-0.028	-0.001	-0.011	-0.010	-0.003	-0.000
	(0.021)	(0.018)	(0.014)	(0.023)	(0.019)	(0.022)
Employed	-0.035	-0.015	-0.016	-0.039	-0.005	-0.029
	(0.023)	(0.018)	(0.015)	(0.025)	(0.020)	(0.025)
log(Household income)	0.003	0.005	0.002	0.025	0.006	-0.008
	(0.015)	(0.012)	(0.010)	(0.017)	(0.015)	(0.018)
Single parent	0.009	0.005	0.007	0.064**	-0.003	0.016
	(0.027)	(0.022)	(0.018)	(0.031)	(0.025)	(0.030)
Number of children	-0.029***	-0.005	-0.015***	0.005	-0.005	0.001
	(0.009)	(0.008)	(0.005)	(0.009)	(0.008)	(0.010)
Share of female children	0.009	0.022	0.022	-0.008	0.003	0.017
	(0.026)	(0.021)	(0.017)	(0.027)	(0.023)	(0.025)
Parenting values						
Altruism towards child (std.)	-0.007	-0.001	-0.007	0.023**	-0.007	0.033***
	(0.010)	(0.008)	(0.007)	(0.009)	(0.008)	(0.010)
Paternalism towards child (std.)	-0.024**	-0.007	-0.013*	-0.004	0.018**	-0.000
	(0.010)	(0.008)	(0.007)	(0.011)	(0.009)	(0.011)
Malleability of skills (std.)	-0.020**	-0.015*	-0.012*	0.034***	0.015*	0.054***
	(0.010)	(0.008)	(0.006)	(0.010)	(0.009)	(0.010)
Mean of dependent variable	.23	.14	.08	.21	0	.24
Individuals	2109	2109	2109	1626	1626	1821
R^2	.019	.0063	.012	.028	.0074	.033

Notes: This table presents regressions of an indicator of zero perceived returns (Panel A) or individual-level perceived returns excluding those with zero returns (Panel B) on sociodemographic characteristics and parenting values according to equation 3.1. Individual-level perceived returns are estimated based on equation (3.1) for each individual separately. The dependent variable in column (1) corresponds to an indicator equal to one if returns to both warmth and control are perceived to be zero, while column (2) focuses on zero perceived returns in the neighborhood dimension. Column (3) checks for all three dimensions simultaneously. Columns (4) to (6) correspond to columns (4) to (6) of Table 3.2, but exclude individuals that report zero perceived returns according to column (1) and (2), respectively. Robust standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

(k = warmth, control, neighborhood) in the monetary domain for the original observations (life satisfaction domain for the duplications) that is instrumented with the return in the life satisfaction domain (monetary domain for duplications), and

$X_{i,-x_i}$ corresponds to a vector of all characteristics excluding x_i.[30] In order to account for the duplications, I follow Gillen, Snowberg, and Yariv (2019) and bootstrap standard errors using 100 replications.

Table 3.D.3 presents the results of this exercise. Each cell corresponds to a coefficient from a regression of equation (3.D.1): An increase of one standard deviation in perceived returns in the warmth or neighborhood dimension is associated with a 3.6-4.0 percentage point increase in the probability of being female and parenting values show similar patterns as before.

30. Note that this is equivalent to estimating seemingly unrelated regressions of these characteristics on perceived returns in both dimensions separately, for which the coefficient on the perceived returns is restricted to being equal across specifications.

Table 3.D.3. Determinants of individual-level perceived returns using ORIVs

	Coefficients on perc. returns		
	(1)	(2)	(3)
	$R_{warmth,i}$	$R_{control,i}$	$R_{neighb.,i}$
Sociodemographic characteristics			
Female	0.036***	0.022	0.040*
	(0.012)	(0.016)	(0.024)
Age	-0.186	-0.193	0.775**
	(0.188)	(0.254)	(0.372)
White	0.040***	0.013	0.004
	(0.012)	(0.014)	(0.021)
College degree	-0.014	-0.018	-0.011
	(0.014)	(0.014)	(0.019)
Employed	-0.026	-0.012	-0.025
	(0.013)	(0.015)	(0.020)
log(Household income)	0.015	0.028	0.026
	(0.020)	(0.026)	(0.028)
Single parent	0.011	-0.002	0.015
	(0.010)	(0.014)	(0.016)
Number of children	0.068*	-0.028	0.040
	(0.038)	(0.036)	(0.055)
Share of female children	-0.012	0.001	-0.017
	(0.011)	(0.015)	(0.015)
Parenting values			
Altruism towards child (std.)	0.088***	-0.027	0.120**
	(0.030)	(0.037)	(0.049)
Paternalism towards child (std.)	0.035	0.128***	0.022
	(0.028)	(0.038)	(0.053)
Malleability of skills (std.)	0.137***	0.043	0.249***
	(0.028)	(0.036)	(0.050)

Notes: This table presents regressions of a respondent's characteristic x_i on the instrumented perceived return and all other individual characteristics based on equation (3.D.1). Each cell reports the coefficient of the perceived returns from a separate regression with the characteristics on the left as the dependent variable. Column (1) uses perceived returns to warmth, column (2) perceived returns to control, and column (3) perceived returns to neighborhoods as the regressor of interest. Bootstrapped standard errors in parentheses. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Appendix 3.E Exploratory Factor Analysis

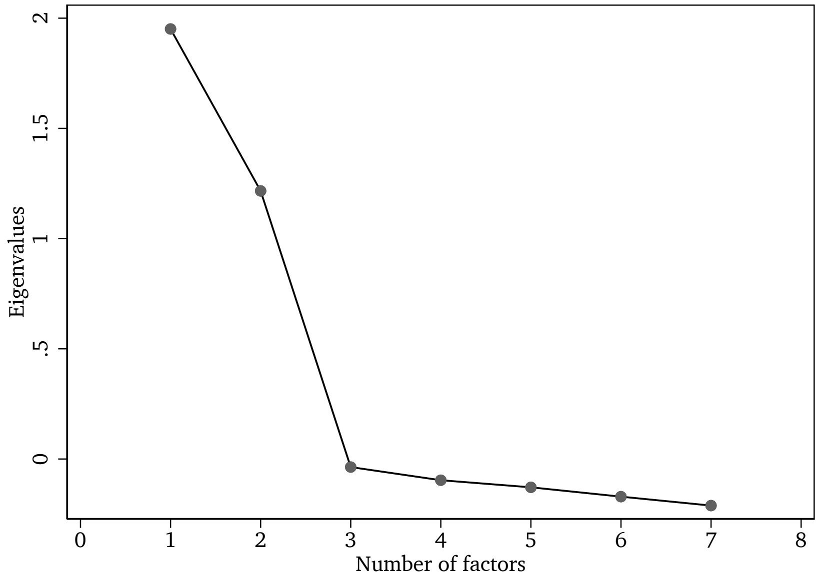


Figure 3.E.1. Scree plot of parenting style items

Notes: This figure presents a scree plot of the eigenvalues from an exploratory factor analysis using seven items based on Perris et al. (1980) and Schwarz et al. (1997) to measure parenting styles in the warmth and control dimensions, respectively.

Table 3.E.1. Rotated factor loadings of actual parenting styles

	Rotated factor loadings	
	(1) Warmth	(2) Control
Warmth measures (Perris et al., 1980)		
(1) I show my son/daughter with words and gestures that I like him/her	0.72	0.06
(2) I cheer up my son/daughter when he/she is sad	0.74	0.09
(3) I praise my son/daughter	0.75	0.07
Control measures (Schwarz et al., 1997)		
(4) I tend to be a strict parent	0.08	0.57
(5) If my son/daughter does something against my will, I punish him/her	0.06	0.68
(6) I make it clear to my son/daughter that he/she is not to break the rules or question my decisions	0.12	0.67
(7) I never waive from my rules	0.07	0.51

Notes: This table presents rotated factor loadings from an exploratory factor analysis using seven items based on Perris et al. (1980) and Schwarz et al. (1997) to measure parenting styles in the warmth and control dimensions, respectively.

Appendix 3.F Relevance of Perceived Returns for Neighborhood Characteristics

In this section, I examine whether estimated returns in the neighborhood dimension are related to the quality of the neighborhood a family is living in. I use two approaches to answer this question. First, the survey elicits the parents' agreement to three statements: (i) "My neighborhood is a good place to raise children", (ii) "I feel safe in my neighborhood", and (iii) "My child attends a school of good quality" on a 5-point scale. I extract a factor from these statements as a measure of the subjective neighborhood quality. Second, linking neighborhood characteristics from Chetty and Hendren (2018a,b) to respondents in my survey, I perform a second factor analysis that reveals two factors: a first factor capturing economic conditions in a neighborhood (NQ 1), and a second factor (NQ 2) related to measures of segregation and urbanization. Table 3.F.1 presents analogous estimates to Table 3.1 using both the subjective assessment or objective measures of neighborhood quality as outcome variables.[31]

I find that only perceived returns in the monetary domain are significantly associated with the subjectively assessed quality of a neighborhood. Returns in the life satisfaction domain or ORIVs do not reveal a significant association. Yet, when looking at objective measures of the neighborhood quality in columns (4) and (5), I find that higher perceived returns to neighborhoods are associated positively, but not significantly with economic conditions of a neighborhood ($p = 0.169$). They are, however, negatively related with its segregation. Although parental beliefs do not predict subjective neighborhood assessments, they are related to objective measures of the neighborhood quality. This suggests that respondents are not necessarily aware how their environment shapes their own assessments of the return to neighborhoods. Taken together, parental beliefs are not only systematically related to actual parenting styles, but also associated with characteristics of the parents' place of residence. This supports the conjecture that these parental beliefs are a fundamental part of parental decision-making processes.

31. One caveat of this approach to keep in mind is that some neighborhood characteristics are historical data and thus may have changed over time. Yet, Chetty, Friedman, et al. (2018) document that these characteristics are relatively stable over time and good predictors of today's conditions.

Table 3.F.1. Relevance of perceived returns for neighborhood quality

	(A) Subjective NQ			(B) NQ 1	(C) NQ 2
	(1) Expected earnings	(2) Expected Life Satis.	(3) ORIV	(4) ORIV	(5) ORIV
$R_{neighb.,i}$	0.043**	-0.005	0.058	0.062	-0.161***
	(0.019)	(0.019)	(0.043)	(0.045)	(0.038)
Controls	Yes	Yes	Yes	Yes	Yes
Observations	2104	2104	4208	4164	4164
Individuals	2104	2104	2104	2082	2082
R^2	.11	.11	.104	.135	.113

Notes: This table examines associations of perceived returns and measures of the actual neighborhood quality. Columns (1) to (3) present the results for the respondents' subjective assessments of the quality of their neighborhood. The outcome variable is a factor constructed from agreement to the three statements (i) "My neighborhood is a good place to raise children", (ii) "I feel safe in my neighborhood", and (iii) "My child attends a school of good quality". Columns (1) and (2) use returns in the expected earnings and expected life satisfaction domains, while column (3) implements the obviously related instrumental variables estimator (ORIV, Gillen, Snowberg, and Yariv, 2019) to correct for measurement error in perceived returns using the two return measures as instruments for each other. Columns (4) and (5) present corresponding results for objective measures of a neighborhood's quality based on respondents' postcodes using ORIVs. NQ 1 refers to a factor capturing economic conditions in an area, while NQ 2 is related to measures of segregation and urbanization. All specifications include controls for sociodemographic characteristics. Robust standard errors in columns (1) and (2) or bootstrapped standard errors in columns (3) to (5) in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% level.

References

Agostinelli, Francesco. (2018). “Investing in Children’s Skills: An Equilibrium Analysis of Social Interactions and Parental Investments.” Working Paper. [115]

Arcidiacono, Peter, V. Joseph Hotz, and Songman Kang. (2012). “Modeling college major choices using elicited measures of expectations and counterfactuals.” *Journal of Econometrics* 166 (1): 3–16. [114, 116]

Attanasio, Orazio P. (2015). “The Determinants of Human Capital Formation during the Early Years of Life: Theory, Measurement, and Policies.” *Journal of the European Economic Association* 13 (6): 949–997. [111]

Attanasio, Orazio P., Teodora Boneva, and Christopher Rauh. (2019). “Parental Beliefs about Returns to Different Types of Investments in School Children.” Working Paper. [112–114, 116, 121, 132, 138, 139]

Attanasio, Orazio P., and Katja M. Kaufmann. (2014). “Education choices and returns to schooling: Mothers’ and youths’ subjective expectations and their role by gender.” *Journal of Development Economics* 109: 203–216. [112, 114]

Baumrind, Diana. (1967). “Child Care Practices Anteceding Three Patterns of Preschool Behavior.” *Genetic Psychology Monographs* 75: (1), 43–88. [111, 112, 116, 123]

Beffy, Magali, Denis Fougère, and Arnaud Maurel. (2012). “Choosing the Field of Study in Postsecondary Education: Do Expected Earnings Matter?” *Review of Economics and Statistics* 94 (1): 334–347. [114]

Bhalotra, Sonia, Adeline Delavande, Paulino Font, and Joanna Maselko. (2017). “Perceived returns from early-life investments and maternal investments in children.” Working Paper. [112, 114, 116, 139]

Boneva, Teodora, and Christopher Rauh. (2018). “Parental Beliefs about Returns to Educational Investments—The Later the Better?” *Journal of the European Economic Association* 16 (6): 1669–1711. [112–114, 116, 121, 129, 132, 138, 139]

Bornstein, Marc H. (2002). “Preface.” In *Handbook of Parenting. Practical Issues in Parenting*. Edited by M. H. Bornstein. Vol. 5, Lawrence Erlbaum Associates, xi–xiv. [138]

Brenøe, Anne Ardila, and Thomas Epper. (2019). “Parenting Values Moderate the Intergenerational Transmission of Time Preferences.” Working Paper. [114]

Chan, Tak Wing, and Anita Koo. (2011). “Parenting Style and Youth Outcomes in the UK.” *European Sociological Review* 27 (3): 385–399. [115, 129]

Chetty, Raj, John N. Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. (2018). “The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility.” Working Paper. [115, 152]

Chetty, Raj, and Nathaniel Hendren. (2018a). “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects.” *Quarterly Journal of Economics* 133 (3): 1107–1162. [115, 119, 125, 152]

Chetty, Raj, and Nathaniel Hendren. (2018b). “The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates.” *Quarterly Journal of Economics* 133 (3): 1163–1228. [115, 119, 152]

Cobb-Clark, Deborah A., Nicolás Salamanca, and Anna Zhu. (2019). “Parenting style as an investment in human development.” *Journal of Population Economics* 32 (4): 1315–1352. [114]

Cosconati, Marco. (2012). “Parenting Style and the Development of Human Capital in Children.” Working Paper. [114]

Cuellar, Jessica, Deborah J. Jones, and Emma Sterrett. (2015). "Examining Parenting in the Neighborhood Context: A Review." *Journal of Child and Family Studies* 24 (1): 195–219. [115]

Cunha, Flavio. (2015). "Subjective Rationality, Parenting Styles, and Investments in Children." In *National symposium on family issues. Diverging Destinies: Families in an Era of Increasing Inequality*. Edited by P. Amato, A. Booth, S. McHale, and J. Van Hook. Vol. 5, Springer. Chapter 6, 83–94. [114, 139]

Cunha, Flavio, Irma Elo, and Jennifer Culhane. (2013). "Eliciting Maternal Subjective Expectations about the Technology of Cognitive Skill Formation." Working Paper. [112, 114, 116]

Cunha, Flavio, and James J. Heckman. (2007). "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47. [119]

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. (2010). "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931. [119]

Del Bono, Emilia, Marco Francesconi, Yvonne Kelly, and Amanda Sacker. (2016). "Early Maternal Time Investment and Early Child Outcomes." *Economic Journal* 126 (596): F96–F135. [114]

Deutscher, Nathan. (Forthcoming). "Place, Peers, and the Teenage Years: Long-run Neighborhood Effects in Australia." *American Economic Journal: Applied Economics*, [115]

Dizon-Ross, Rebecca. (2019). "Parents' Beliefs About Their Children's Academic Ability: Implications for Educational Investments." *American Economic Review* 109 (8): 2728–2765. [114]

Doepke, Matthias, Giuseppe Sorrenti, and Fabrizio Zilibotti. (Forthcoming). "The Economics of Parenting." *Annual Review of Economics*, [114, 125, 138]

Doepke, Matthias, and Fabrizio Zilibotti. (2017). "Parenting With Style: Altruism and Paternalism in Intergenerational Preference Transmission." *Econometrica* 85 (5): 1331–1371. [111, 113–115, 117, 134, 138]

Dohmen, Thomas, Bart Golsteyn, Lena Lindahl, Gerard Pfann, and André Richter. (2019). "Teach Your Children Well – Determinants and Consequences of Parenting Styles." Working Paper. [115]

Dominitz, Jeff, and Charles F. Manski. (1996). "Eliciting Student Expectations of the Returns to Schooling." *Journal of Human Resources* 31 (1): 1–26. [112]

Dooley, Martin, and Jennifer Stewart. (2007). "Family income, parenting styles and child behavioural–emotional outcomes." *Health Economics* 16 (2): 145–162. [114]

Dornbusch, Sanford M., Philip L. Ritter, P. Herbert Leiderman, Donald F. Roberts, and Michael J. Fraleigh. (1987). "The Relation of Parenting Style to Adolescent School Performance." *Child Development* 58 (5): 1244–1257. [115, 123, 129]

Ermisch, John. (2008). "Origins of Social Immobility and Inequality: Parenting and Early Child Development." *National Institute Economic Review* 205 (1): 62–71. [114]

Falk, Armin, Fabian Kosse, Pia Pinger, Hannah Schildberg-Hörisch, and Thomas Deckers. (Forthcoming). "Socio-Economic Status and Inequalities in Children's IQ and Economic Preferences." *Journal of Political Economy*, [114, 129]

Fiorini, Mario, and Michael P. Keane. (2014). "How the Allocation of Children's Time Affects Cognitive and Noncognitive Development." *Journal of Labor Economics* 32 (4): 787–836. [114]

Francesconi, Marco, and James J. Heckman. (2016). "Child Development and Parental Investment: Introduction." *Economic Journal* 126 (596): F1–F27. [111]

Gillen, Ben, Eric Snowberg, and Leeat Yariv. (2019). "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 124 (4): 1826–1863. [114, 118, 135–137, 146, 148, 153]

Giustinelli, Pamela. (2016). "Group Decision Making with Uncertain Outcomes: Unpacking Child-Parent Choice of the High School Track." *International Economic Review* 57 (2): 573–602. [114]

Guryan, Jonathan, Erik Hurst, and Melissa Kearney. (2008). "Parental Education and Parental Time with Children." *Journal of Economic Perspectives* 22 (3): 23–46. [131]

Han, Joal Kaiyuan. (2019). "Parental Involvement and Neighborhood Quality: Evidence from Public Housing Demolitions in Chicago." Working Paper. [115, 117, 123]

Hastings, Justine S., Christopher A. Neilson, Anely Ramirez, and Seth D. Zimmerman. (2016). "(Un)informed college and major choice: Evidence from linked survey and administrative data." *Economics of Education Review* 51: 136–151. [112, 114]

Jensen, Robert. (2010). "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics* 125 (2): 515–548. [112, 114]

Kalil, Ariel. (2015). "Inequality Begins at Home: The Role of Parenting in the Diverging Destinies of Rich and Poor Children." In *National symposium on family issues. Diverging Destinies: Families in an Era of Increasing Inequality*. Edited by P. Amato, A. Booth, S. McHale, and J. Van Hook. Springer. Chapter 5, 63–82. [111]

Kaufmann, Katja M. (2014). "Understanding the income gradient in college attendance in Mexico: The role of heterogeneity in expected returns." *Quantitative Economics* 5 (3): 583–630. [112, 114, 129]

Kiessling, Lukas, Pia Pinger, Jan Bergerhoff, and Philipp Seegers. (2019). "Gender Differences in Wage Expectations: Sorting, Children, and Negotiation Styles." Working Paper. [114]

Kinsler, Josh, and Ronni Pavan. (2018). "Local Distortions in Parental Beliefs over Child Skill." Working Paper. [114]

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. (2005). "Bullets Don't Got No Name: Consequences of Fear in the Ghetto." In. *Discovering Successful Pathways in Children's Development: New Methods in the Study of Childhood and Family Life*. Edited by T.S. Weisner. University of Chicago Press, 243–282. [115, 123]

Laliberté, Jean-William. (2018). "Long-term Contextual Effects in Education: Schools and Neighborhoods." Working Paper. [115]

Lamborn, Susie D., Nina S. Mounts, Laurence Steinberg, and Sanford M. Dornbusch. (1991). "Patterns of Competence and Adjustment among Adolescents from Authoritative, Authoritarian, Indulgent, and Neglectful Families." *Child Development* 62 (5): 1049–1065. [115, 123, 129]

Maccoby, Eleanor E., and John A. Martin. (1983). "Socialization in the context of the family: Parent-child interaction." In *Handbook of Child Psychology*. Edited by P. Mussen. Vol. 4, Wiley. [112, 116]

Manski, Charles F. (2004). "Measuring Expectations." *Econometrica* 72 (5): 1329–1376. [111, 116]

Nguyen, Trang. (2008). "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar." Working Paper. [112]

Patacchini, Eleonora, and Yves Zenou. (2011). "Neighborhood Effects and Parental Involvement in the Intergenerational Transmission of Education." *Journal of Regional Science* 51 (5): 987–1013. [115, 125]

Perris, Carlo, Lars Jacobsson, Hakån Lindström, Lars von Knorring, and Hjördis Perris. (1980). "Development of a new inventory for assessing memories of parental rearing behaviour." *Acta Psychiatrica Scandinavica* 61 (4): 265–274. [119, 150, 151]

Pop-Eleches, Cristian, and Miguel Urquiola. (2013). "Going to a Better School: Effects and Behavioral Responses." *American Economic Review* 103 (4): 1289–1324. [115, 123, 125]

Putnam, Robert D. (2016). *Our kids: The American dream in crisis*. Simon, and Schuster. [138]

Schwarz, Beate, Sabine Walper, Mechthild Gödde, and Stephanie Jurasic. (1997). "Dokumentation der Erhebungsinstrumente der 1. Haupterhebung (überarb. Version)." Reports of the research group "Familienentwicklung nach der Trennung" (Vol. 14). [119, 150, 151]

Steinberg, Laurence, Nina S. Mounts, Lamborn Susie D., and Sanford M. Dornbusch. (1991). "Authoritative Parenting and Adolescent Adjustment across Vaaried Ecological Niches." *Journal of Research on Adolescence* 1 (1): 19–36. [115]

Stinebrickner, Ralph, and Todd R. Stinebrickner. (2014). "A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout." *Review of Economic Studies* 81 (1): 426–472. [114]

Wiswall, Matthew, and Basit Zafar. (2015). "Determinants of College Major Choice: Identification using an Information Experiment." *Review of Economic Studies* 82 (2): 791–824. [114]

Wiswall, Matthew, and Basit Zafar. (2018a). "Human Capital Investments and Expectations about Career and Family." Working Paper. [114]

Wiswall, Matthew, and Basit Zafar. (2018b). "Preference for the Workplace, Investment in Human Capital, and Gender." *Quarterly Journal of Economics* 133 (1): 457–507. [114]

Zafar, Basit. (2013). "College Major Choice and the Gender Gap." *Journal of Human Resources* 48 (3): 545–595. [114]

Zumbuehl, Maria, Thomas Dohmen, and Gerard Pfann. (2018). "Parental Involvement and the Intergenerational Transmission of Economic Preferences and Attitudes." Working Paper. [114]

Chapter 4

Gender Differences in Wage Expectations: Sorting, Children, and Negotiation Styles

Joint with Pia Pinger, Philipp Seegers, and Jan Bergerhoff

4.1 Introduction

The gender gap in labor earnings ranges among the best documented facts in the empirical economic literature and is subject to regular policy debates[1]. Overall, the unconditional gap ranges from 5 to 35% across different OECD countries and in both absolute and relative terms it tends to be particularly large for individuals with a college degree or higher (OECD, 2015). Moreover, convergence in male-female wages remains slow despite sustained efforts towards achieving gender-based equality of opportunity.

A closely-related gender gap is the gap in *ex-ante* wage *expectations*, i.e., male-female differences in expectations about labor market returns *before* entering the labor market (see, e.g., Blau and Ferber (1991) and Brunello, Lucifora, and Winter-Ebmer (2004) for initial and Reuben, Wiswall, and Zafar (2017) for more recent evidence). Such male-female gaps in labor market expectations are important as they potentially determine education and labor market choices, household bargaining, and wage setting. They are also an important component in financial decision-making, e.g., regarding the optimal choice of retirement and savings plans. Moreover, there may exist important feedback effects whereby expected wages drive actual wage differences (e.g., through wage negotiations), and actual observable wage disparities affect expectations, thus providing a rationale for persistent gender wage gaps.

1. For a recent summary of the literature, see Blau and Kahn (2017) and Kunze (2018).

The aim of this paper is to provide encompassing and large-scale evidence on gender wage expectations, as well as investigating how they are affected by a substantial number of different factors using a single dataset and coherent framework. For this purpose, we have elicited wage expectations for counterfactual study trajectories among more than 15,000 German students from all regions, universities, study fields and over the entire prospective working life. In addition, the data contain elicited expectations about future labor force participation, working hours, childrearing plans, and wage negotiations, as well as information on perceived and actual ability, personality, IQ, beliefs and preferences.

We provide two sets of results. In a first instance, we document a range of stylized facts about male-female wage expectations, including population-wide and subgroup-specific gaps in expected wages, distributional differences in ranks and levels, and differences in expected life-cycle wage trajectories. We show that the gender gap in expected wages is significant and large across all subgroups and along the entire distribution. Moreover, it is similar to the observed actual wage gap among recent graduates.[2] In terms of life-cycle wage developments, females expect flatter wage trajectories, with an initial gap of 14 percent increasing to 27 percent at the age of 55. The accumulated life-cycle gap in expected wages hence amounts to eighteen distributional ranks, or more than 500,000 Euros. In terms of magnitude, this "perceived return to being male" is close to the actual return of obtaining a university degree. In the second part, we provide comprehensive evidence on its determinants, both along the expected wage distribution and regarding expected life-cycle wage trajectories. In line with previous literature on expected and actual wage differences (Blau and Kahn, 2017; Bütikofer, Jensen, and Salvanes, 2018; Francesconi and Parey, 2018), we find that a large portion of the overall gap in expected wages relates to academic and occupational sorting patterns and a much smaller part to IQ, perceived or actual ability and personality traits. Contrary to the evidence for actual wages (Bertrand, Goldin, and Katz, 2010; Daniel, Lacuesta, and Rodriguez-Planas, 2013; Goldin and Katz, 2016; Kleven, Landais, and Sogaard, forthcoming) but in line with Kuziemko, Pan, Shen, and Washington (2018), child-related labor force interruptions prove largely unimportant for wage expectations, although a perceived wage penalty seems to exist for having children before the age of 30. Moreover, we provide first empirical evidence on the relationship between expected wages, initial wage claims, reservation wages and a novel measure of expected negotiation styles. While initial wage claims closely relate to expected wage outcomes, females envisage substantially less scope for wage negotiations than males. Differences in anticipated negotiation styles explain 13-14% of the gender gap and thus hold similar importance as differences in major choice or occupational sorting. Finally, we

2. Among German college graduates, the gender wage gap is 20% overall and reduces to 5-10% after accounting for a large number of controls (Destatis, 2014; Destatis, 2017b; Francesconi and Parey, 2018). It is thus comparatively large.

provide suggestive evidence that wage expectations are prospective- or preference-based rather than adaptive, as personal experiences of actual gender gaps in different labor markets or student jobs no not translate into relative wage expectations.

Our study thus contributes to a buoyant literature on wage expectations, which, pioneered by Manski (Dominitz and Manski, 1997; Manski, 2004), has repeatedly documented the importance of elicited expectations and beliefs for explaining education choices and labor market behaviors (e.g., Arcidiacono, Hotz, and Kang, 2012; Jensen, 2010; Kaufmann, 2014; Stinebrickner and Stinebrickner, 2014; Zafar, 2011). It also relates to a range of prior studies documenting the existence of a gender gap in ex-ante wage expectations in a number of specialized samples, i.e., containing information from students enrolled in particular colleges/universities or fields of study. These studies have separately identified several potential drivers of the gender gap in wage expectations, including differences in major choice, personality traits, and economic preferences (Reuben, Wiswall, and Zafar, 2017; Zambre, 2018).

In this paper, we move beyond the existing evidence in at least three respects. First, we present the first large-scale study on gender wage expectations, both in terms of sample size and scope, and regarding the range of available measures. The considerable size and diversity of our sample allows us to make claims about the overall magnitude of the gender gap in wage expectations, as well as exploring heterogeneities across study fields, aspired occupations, regional labor markets, and numerous background characteristics. Moreover, by asking about expected wages at three points in the future and for different study scenarios, we construct within-individual life-cycle wage trajectories to obtain expected differences in growth rates, relative ranks, and expected lifetime labor earnings. Second, our comprehensive data allow us to relate gender gaps in expected wages to a vast array of potential determinants in one coherent framework. Potential drivers include sorting into study fields and occupations, personality traits, perceived and actual ability, economics preferences, child-rearing plans and labor supply. Third, information about prospective wage negotiations permits us to document the importance of gender differences in anticipated wage negotiations and relate wage claims and negotiation strategies to expected wage outcomes. To the extent that wage negotiations are an important component of the wage-setting process, our results thus provide an important link between expected and actual wages, as well as an explanation why the gender gap in expected wages mirrors the gender gap in actual wages.

The remainder of the paper is organized as follows. In the next section, we discuss the sample, questionnaire measures and construction of life-cycle wage trajectories. Section 4.3 documents male-female differences in wage expectations both for starting wages and over the life cycle. This section also shows that differences in expected wages relate to differences in actual wages. Section 4.4 then presents evidence on gender differences in a number of dimensions that have been shown to explain large parts of the variation in actual wage gaps. Most notably, we account for sorting into study fields and occupations, expectations about child-rearing responsibilities, and differences in negotiation patterns. Decomposition analyses assess the relative importance of these factors. Finally, section 4.5 concludes.

4.2 Data

This section reports on our sample and questionnaire measures. We start out by describing our sample and questionnaire measures of expected wages, labor supply and children, initial wage claims and reservation wages, sorting, and background characteristics. Then, we explain how we construct expected wage trajectories and measures of negotiation styles.

4.2.1 Sample

Our sample comprises 15,348 students and 1,155 recent graduates (since our focus is on student expectations, we will henceforth use the word "students"). All individuals were recruited as part of the German student study "Fachkraft 2030", surveyed in the second half of March 2015 (Seegers, Bergerhoff, Hartmann, and Knappe, 2016). In addition, a subsample of 12,734 students (82.97%) completed a supplementary psychological questionnaire comprising measures of personality traits, economic preferences, and IQ.

Students were contacted via the mailing list of a popular nationwide job board.[3] They were contacted via email and took part in an online questionnaire.[4] The sample closely compares to the overall population of German students in terms of region, university type, study fields, and likelihood to hold a student job (Seegers et al., 2016).

4.2.2 Measures

Individuals answered a comprehensive questionnaire regarding their own background and university enrollment, expectations about their course of studies, labor market expectations, expectations about child-rearing, and wage negotiation plans. They also provided information about expected future employment and student jobs. Finally, part of the sample completed a short IQ test, as well as a questionnaire about personality traits and preferences.

Wage expectations and realized wages. We asked subjects to indicate their expected yearly labor earnings in current Euros before taxes and at different points over the life cycle: (i) in their first job after graduation ($w^s_{i,st}$), (ii) at the age of

3. The job board jobmensa.de is operated by Studitemps GmbH and is the largest platform for student jobs.

4. The questionnaire was filled in by 8% of contacted students. Participation was incentivized using Amazon vouchers amounting to 5,000 EUR (1 x 1,000 EUR, 4 x 250 EUR, 10 x 100 EUR, 40 x 50 EUR vouchers).

40 ($w^s_{i,40}$), and (iii) at the age of 55 ($w^s_{i,55}$). We chose these time points for several reasons. First, starting wages are likely to be a natural reference point for many students and most related to their expected labor market negotiations. Starting wages are also most often elicited in the literature on wage expectations (Arcidiacono, Hotz, and Kang, 2012; Webbink and Hartog, 2004). Second, the age of 40 is the time when individuals will have likely completed their prospective family planning, such that child-related differences in expected wage trajectories should become apparent at this point. Third, the age of 55 is close to the time where wages peak but before early retirement sets in (Piopiunik, Kugler, and Wößmann, 2017).

We asked students to state these expected wages under three different scenarios, regarding their course of studies: (a) if they complete their current (*first*) studies ($w^f_{i,t}$), (b) if they change to their second most preferred *alternative* field of study ($w^a_{i,t}$), and (c) if they *dropout* and do not complete any further educational degree ($w^d_{i,t}$). Thus, given three scenarios (a)-(c), denoted by s, and three points over the life cycle (i)-(iii), denoted by t, we elicit a total of nine expected wages ($w^s_{i,t}$). In addition, we ask all individuals to state the probability of each of the respective scenarios materializing ($p^s_{i,t}$).

Assuming these scenarios to be mutually exclusive, i.e., that students either finish, change study fields or drop out, we can use the above information to construct our measure of *overall expected wages* as follows:

$$w_{i,t} \;=\; p^f_{i,t}w^f_{i,t} + p^a_{i,t}w^a_{i,t} + p^d_{i,t}w^d_{i,t} \quad \forall t \;\in\; \{st, 40, 55\}. \tag{4.1}$$

We reweight probabilities in cases where the stated probabilities add up to more than one hundred percent (7 percent). Moreover, we exclude individuals (less than 1%) who indicated implausible large expected wages of more than 1,000,000 EUR per year.

Our measure of *realized wages* are actual labor earnings before taxes reported by the graduates in our sample. All expected and actual labor earnings variables were winsorized at the 1% and 99% level. The mean level of expected starting wages in our student sample is 35,870 EUR per year (SD=16,093). The mean realized wage in the graduate sample amounts to 35,961 EUR per year (SD=25,093).

Labor supply and children. Our data contain several measures of expected labor supply and child-related career breaks. First, expected labor supply is captured by the expected number of weekly working hours. To match the information about expected wages, we asked for the expected number of weekly working hours at the same points in time, i.e., right after graduation ($h^s_{i,st}$), at the age of 40 ($h^s_{i,40}$), and at the age of 55 ($h^s_{i,55}$) for each of the three scenarios $s = f, a, d$.[5] Second, we elicited whether the students in our sample already have children and, if not, at what age

5. We also elicit the subjective probability of involuntary unemployment. However, similar to what has been found in the literature (e.g., Baker, Bettinger, Jacob, and Marinescu, 2018), we the

they expect the birth of their first child. Third, we asked how many children students expect to have in total and how many months they are planning to stay home with each child.

Initial wage claims, reservation wages, and discrimination. Respondents were asked about the initial salary students would demand as they enter a wage negotiation (initial wage claim, $w_{i,I}$).[6] We also inquired about the lowest wage rate at which a student would be willing to accept a job after finishing her studies (reservation wage, $w_{i,R}$). Based on initial wage claims and reservation wages, we construct a measure of negotiation style (see Section 4.2.4). Moreover, respondents stated whether they would expect to earn the same wage if they were a member of this opposite sex but with identical skills, characteristics, traits, and qualifications. If the answer is "no", we interpret this as an indicator of perceived gender discrimination.

Major and occupational sorting. Students in Germany are required to enroll for a particular field of studies when they first enter a teaching college or university. Hence, at the time of the survey, students have already selected study fields in line with their academic interests and occupational preferences. We elicited the current study field as a choice out of a list of fifteen majors. In addition, we asked respondents for their career aspirations. They could choose out of 429 pre-defined occupations or make use of a free text field. All indicated occupations were subsequently classified in terms of the ISCO-08 occupational classification reflecting job tasks as well as skills and occupational hierarchies.[7]

Personality traits, economic preferences, beliefs about ability, and IQ. Research in personality psychology and economics shows that males and females display substantial differences in personality traits, economic and social preferences, and beliefs about one's own ability (Bertrand, 2011; Bian, Leslie, and Cimpian, 2017; Borghans, Heckman, Golsteyn, and Meijers, 2009; Croson and Gneezy, 2009; Schmitt, Realo, Voracek, and Allik, 2008). Our data allow us to systematically account for these differences. In order to elicit beliefs about own ability, respondents marked their

reported expected probability of being unemployed is implausibly large in our sample for both males (25 percent at start and 15 percent at the age of 40) and females (32 percent at start and 19 percent at the age of 40) compared to employment rates of 93% for recent university graduates in Germany (Eurostat, 2018). We, therefore, do not use this variable in main part of the paper, acknowledging that this might lead to conservative estimates of the gender wage gap, as males report a 7 percent lower probability of involuntary unemployment at employment start and a 4 percent lower probability of involuntary unemployment at the age of 40.

6. While not all jobs require wage negotiations, Hall and Krueger (2012) show that the incidence of wage negotiations is much higher for highly-educated individuals with college degrees compared to the general population. Moreover, it is common in Germany to state an initial wage claim when applying for a position.

7. For evidence on the importance of tasks for the gender wage gap, see Stinebrickner, Stinebrickner, and Sullivan (2019).

relative position in the distribution of students regarding their (a) perceived academic ability and (b) perceived work-related ability on a scale from 0 to 100. Four fifth of the sample additionally participated in a survey on personality, economic preferences, and IQ. First, we measured IQ based on ten items from a Raven-type Matrices IQ test (Raven and Court, 1998). Second, a student's Big Five personality traits (agreeableness, conscientiousness, emotional stability, extraversion and openness) were assessed using the 50 item IPIP test (Goldberg, Johnson, Eber, Hogan, Ashton, et al., 2006). Finally, to elicit altruism, impatience, positive and negative reciprocity, risk aversion and trust, we employed an experimentally-validated survey module (Falk, Becker, Dohmen, Huffman, and Sunde, 2018). In the following, we use the term "Perceived/actual ability & personality" to refer to the set of these measures.

4.2.3 Wage Trajectories and Life-time Labor Earnings

We use elicited wage expectations to approximate lifetime wage trajectories as well as total lifetime labor earnings. For this purpose, we assume a standard Mincer-type earnings function where log-normally distributed wages are a quadratic function of potential experience:

$$\ln w_{i,t} = \alpha_i + \beta_i exp_{i,t} + \gamma_i exp_{i,t}^2. \tag{4.2}$$

Using the elicited information about wage expectations at three different points in time ($w_{i,st}$, $w_{i,40}$ and $w_{i,55}$), we can use equation (4.2) to determine the parameters α, β and γ for each individual separately. We then use the above relationship to calculate individual-specific expected wages for each year ($\widehat{w}_{i,t}\ \forall t \notin \{st, 40, 55\}$).[8]

Based on these expected wage calculations for each year of an individual's working life, lifetime earnings can be calculated as the sum of expected yearly earnings, i.e.,

$$\widehat{w}_{i,life} = \sum_{st}^{65} \widehat{w}_{i,t}, \tag{4.3}$$

where all expected wages are given in current Euros and we assume an average retirement age of 65 years.

4.2.4 Distributional Differences and Negotiation Styles

Apart from analyzing gender differences at the mean, we investigate the *gender gap in terms of levels and ranks* along the entire expected wage distribution. While

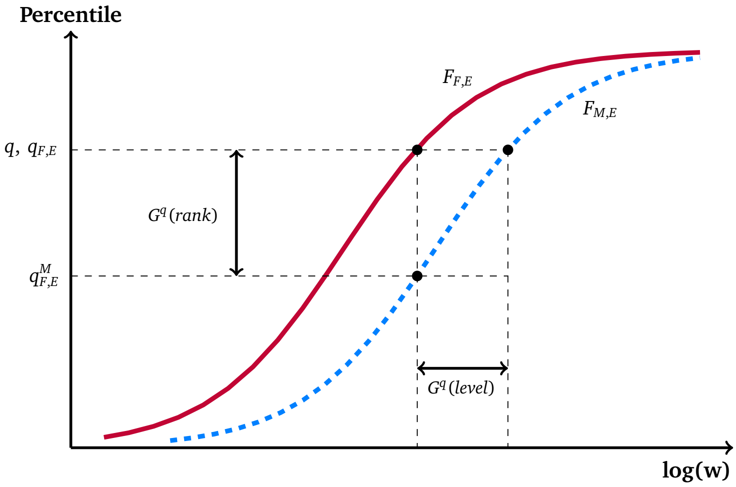
8. Note that the expected starting year ($t = st$) differs across individuals. Since we know each individual's expected year of graduation as well as their age, we calculate $\widehat{w}_{i,t}$ for all years $t > st$. This implies that our sample changes during the initial prospective working period, i.e., up to the point where all students in our sample expect to have graduated (see also footnote 15).

level differences are commonly analyzed (e.g., Francesconi and Parey, 2018), to the best of our knowledge gender gaps expressed in ranks have not been analyzed so far. Intuitively, a respective female takes up a different position (and rank) in the female expected wage distribution than in the distribution of male expected wages, whereby we compare the difference between these two rank measures. Accordingly, for each quantile $q_{F,E}$ of the female (log) expected wage distribution $F_{F,E}$, we compute the rank $q^M_{F,E}$ in the male distribution $F_{M,E}$ that corresponds to the same (log) wage level. The rank gap for a given quantile is then given by $G^q(rank) = q_{F,E} - q^M_{F,E}$ and the corresponding level gap by $G^q(level) = F_M^{-1}(q) - F_F^{-1}(q)$ (see also Bayer and Charles, 2018, for details on this methodology). We thus express male and female wages on the same underlying scale, namely in terms of the expected wage distribution of males. Panel (a) of Figure 4.1 illustrates both measures of the gender gap.

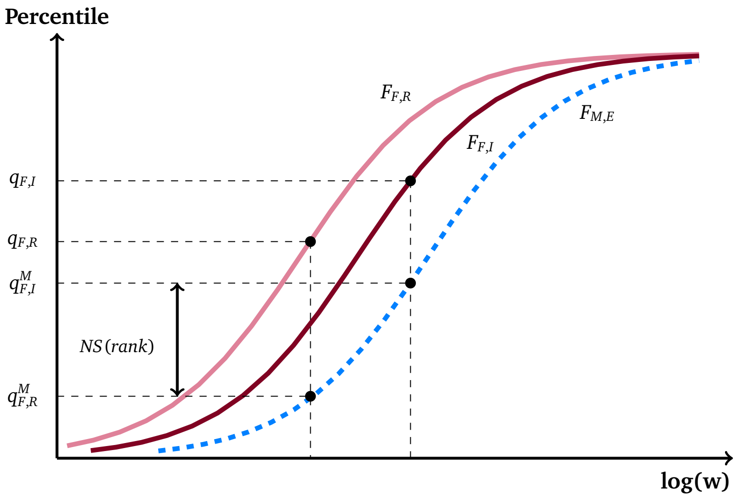
Expanding on this idea, we construct a measure of *negotiation styles* that is well defined and comparable across genders. Such comparisons across different distributions are not trivial as they require some form of anchoring. To provide such an anchor, we express initial wage claims and reservation wages of both genders in terms of ranks of the male wage distribution (see panel (b) of Figure 4.1). Thus, given that the initial wage claim (reservation wage) of a given female in our sample lies on a certain quantile $q_{F,I}$ ($q_{F,R}$), we calculate the corresponding quantile in the male expected wage distribution F_M. Using this, we then determine the corresponding rank of initial wage claims and reservation wages with respect to the male wage distribution ($q^M_{F,I}$ and $q^M_{F,R}$). Next, we proceed analogously with the initial wage claims and reservation wages of males. In a second step, we then define the negotiation style of individual i as the difference between her transformed rank of initial wage claims and reservation wages, i.e., $NS_i(rank) = q^M_{i,g,I} - q^M_{i,g,R}$ with $g = F, M$ for females and males, respectively. Our measure of negotiation styles thus captures "boldness" in wage negotiations, namely how much more a respective individual is willing to ask for, when compared to her minimum acceptable wage. Note that despite being based on initial wage claims, this measure likely captures a general willingness to ask for a relatively higher wage, both initially and in later wage negotiations.

4.3 Gender Differences in Wage Expectations

This section first documents the gender gap in wage expectations across scenarios (current major, alternative major, dropout) and over the life cycle (starting, age 40, age 55). We also present the overall gap (weighted by scenario probabilities), distribution-wide differences, and differences in individual life-cycle wage trajectories. Finally, we provide evidence that wage expectations tend to be accurate on average, suggesting that the wage gap in expectations maps into actual wage differences.



(a) Expected wages



(b) Negotiation styles

Figure 4.1. Calculation of ranks in expected wages, initial wage claims, and reservation wages

Notes: Figure 4.1a illustrates the decomposition of the gender gap in terms of ranks and levels. For a given quantile in the female expected wage distribution F_F (red, solid), the rank gap is defined as the difference between a given quantile and the quantile position that a respective female would assume in the male distribution $F_{M,E}$ (blue, dashed): $G^q(rank) = q_{F,E} - q^M_{F,E}$. Similarly, the level gap is defined as the expected wage difference between a male and a female both evaluated at the same quantile ($G^q(level) = F_M^{-1}(q) - F_F^{-1}(q)$). Fgure 4.1b illustrates how ranks of initial wage claims (dark red, solid) and reservation wages (light red, solid) of females are calculated using the male (log) expected wage distribution $F_{M,E}$ (blue, dashed). Our measure of negotiation styles for individual i is given by the difference in ranks between her initial wage claim ($q^M_{i,g,I}$) and reservation wage ($q^M_{i,g,R}$): $NS_i(rank) = q^M_{i,g,I} - q^M_{i,g,R}$ with $g = F, M$ depending on individual i's gender.

4.3.1 The Male-Female Gap in Wage Expectations

Panel A of Table 4.1 presents mean expected wages for each of the different scenarios (graduating in one's major, graduating with an alternative major, or dropping out) and at three points over the prospective working life. It shows that regardless of the scenario or age, all male-female differences in expected wages are statistically different from zero and substantial in size. Thus for example, while male students expect to earn on average 40,582 EUR after graduating from their current major, females expect a mere 85% of this amount (34,331 EUR). Moreover, the wage gap increases at higher prospective ages and is more pronounced for the current major choice, where the lifetime gap in expected wages cumulates to almost 600,000 Euros. Besides, for both males and females, expected wages conditional on finishing the current major are higher compared to the starting wages of the alternative major or for dropping out of university.[9]

To simplify the analysis, we henceforth focus on overall expected wages, i.e., by taking into account the notion that with a certain probability students change majors or drop out as shown in equation (4.1). The resulting overall expected wage rates are presented in panel B of Table 4.1 and their respective distributions in Figures 4.1a to 4.1c. Again, the male-female gap in overall expected wages is statistically significant and large. At the beginning of their careers, male students expect to earn on average 39,076 EUR, while female students expect 33,434 EUR (86%). The difference in expectations increases until the age of 40, when most children will be born, and rises further until the age of 55, when wage trajectories tend to peak. Male students expect to earn 58,301 EUR at the age of 40 and 70,518 EUR at the age of 55, whereas females report wage expectations of 45,765 EUR (78%) and 51,291 EUR (73%). Over the life cycle, this gap in expectations cumulates to an average of more than half a million Euros. To put this number into perspective, the 525,969 EUR lifetime "expected return to being male" is close to the average lifetime return to obtaining a university degree (Piopiunik, Kugler, and Wößmann, 2017).[10]

When looking at gender gaps in expectations by major, a similar pattern emerges. While substantial heterogeneity exists in terms of levels – humanities majors on average expect the lowest starting wages, while law students expect the highest – female students always expect to earn substantially less than their male counterparts and the gap in expected wages increases over the life cycle. However, the expected wage

9. This finding is consistent with recent evidence that students select into majors according to their perceived comparative advantage (Kirkeboen, Leuven, and Mogstad, 2016).

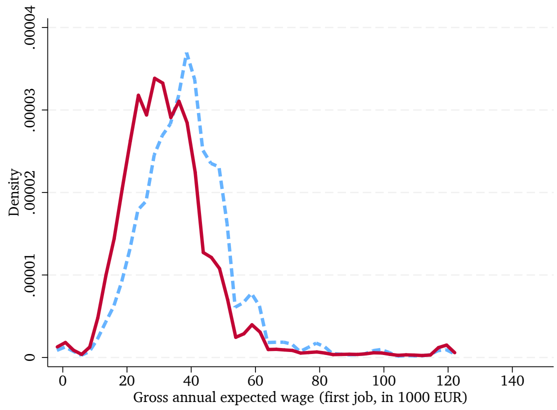
10. Lifetime returns in Piopiunik, Kugler, and Wößmann (2017) are discounted using a net discount rate of 1.5%. We thus approximate gross returns as 3568 EUR x 12 months x 37 years - 1891 EUR x 12 months x 45 years = 563,052 EUR using the numbers reported in Table 1 of their paper. Alternatively, we can apply the same discount rate of 1.5% to yearly expected incomes in our sample. Doing so results in a discounted expected lifetime earnings of 366,464 EUR compared to 387,431 EUR for the return to obtaining a university degree.

Table 4.1. Descriptive statistics of expected and actual gross annual wages in current Euros

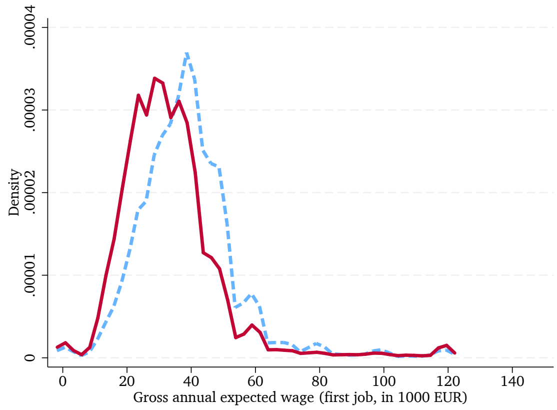
	Summary statistics				
	Males	Females	Diff.	Ratio	N
A. By scenario (expected wages)					
Current major					
Starting	40,582	34,331	6,252	0.85	15,348
Age 40	61,475	47,514	13,961	0.77	15,348
Age 55	74,698	53,361	21,337	0.71	15,348
Lifetime	2,482,233	1,895,315	586,919	0.76	12,734
Probability to finish major	81	84	-3	1.04	15,348
Alternative major					
Starting	38,156	33,685	4,471	0.88	15,348
Age 40	53,225	43,665	9,559	0.82	15,348
Age 55	64,048	48,434	15,614	0.76	15,348
Lifetime	2,165,761	1,744,971	420,790	0.81	12,828
Probability to major change	9	7	1	0.86	15,348
Dropout					
Starting	27,017	24,326	2,690	0.90	15,348
Age 40	34,296	27,980	6,316	0.82	15,348
Age 55	38,892	30,276	8,616	0.78	15,348
Lifetime	1,369,630	1,132,489	237,141	0.83	12,828
Probability of college dropout	11	9	2	0.82	15,348
B. Overall (expected wages)					
Starting	39,076	33,434	5,642	0.86	15,348
Age 40	58,301	45,765	12,536	0.78	15,348
Age 55	70,518	51,291	19,227	0.73	15,348
Lifetime	2,356,291	1,830,322	525,969	0.78	12,734
C. Actual wages (graduates)					
Starting	38,728	33,945	4,783	0.88	1,155
Lifetime	2,621,885	1,904,946	716,939	0.73	825

Notes: Ratio refers to the ratio of female to male expected wages/probabilities. Lifetime wages are constructed based on equations (4.2) and (4.3). Lifetime wages of graduates are based on actual starting wages and wage expectations at the age of 40 and 55. All wages are winsorized at the 1% and 99% level.

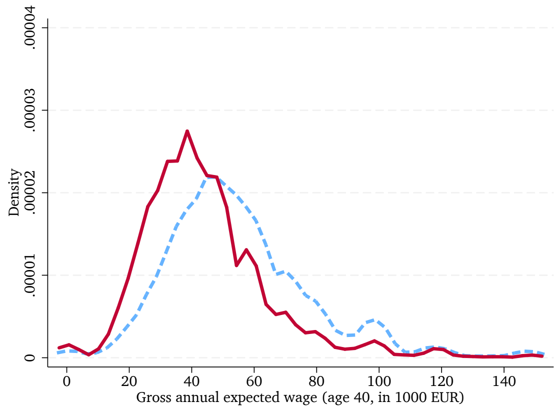
gap tends to be smaller in majors with a larger share of females (e.g., medical/health sciences, humanities) relative to majors mostly chosen by males (e.g., STEM, economics/business; see section 4.C for details). Consistent with Goldin (2014), we also observe smaller gender differences for occupations that are characterized by a linear hours-earnings relationship (e.g., teachers) compared to occupations with



(a) Expected wage – First job



(b) Expected wage – Age 40



(c) Expected wage – Age 55

Figure 4.1. Expected yearly gross wages

Notes: Figure 4.1a–4.1c present kernel densities of expected overall wages upon graduation (4.1a), at the age of 40 (4.1b), and at the age of 55 (4.1c) of female (red, solid) and male (blue, dashed) students in our sample. All expected wages are winsorized at the 1% and 99% level.

nonlinear/convex hours-earnings profiles (e.g., lawyers), see section 4.B and Table 4.B.2 for results.

4.3.2 Gender Gaps along the Expected Wage Distribution in Levels and Ranks

In the previous section, we described the gender gap at the mean. However, there might also be important distributional heterogeneities if, e.g., most of the gap was driven by differences at the very top or bottom of the distribution. Regarding actual wages, distributional differences are indeed heterogeneous. In Germany, the actual gender gap varies across the wage distribution, and decreases for university graduates with rising wage levels (Antonczyk, Fitzenberger, and Sommerfeld, 2010; Francesconi and Parey, 2018).[11] In the following, we characterize the gap in wage expectations at different points of the expected wage distribution using quantile regressions in terms of both log levels and ranks.[12]. The analyses of levels and ranks correspond to two different thought experiments. First, level differences are informative about the absolute (percentage) gain in wages that a female at a certain quantile could expect to receive if she were male. Second, rank differences reveal how much lower a respective female ranks on the male wage distribution given her respective expected wage. In other terms, if the labor market was a competition with wages as a prize, then rank differences inform us about how much worse a female would expect to perform in that competition due to her gender.

Table 4.2 describes the gender gap at five points along the expected wage distribution, namely the 10th, 25th, 50th, 75th and 90th percentiles. The estimates in the first row of panel A show that the gender gap in levels for lower quantiles is larger than for higher quantiles, decreasing from about 24 to 11 percentage points. The gap in expectations thus mirrors the actual distributional wage gap among students Francesconi and Parey (see Figure 4 in 2018). Panel B characterizes the gap using ranks as introduced in Section 4.2.4, revealing a somewhat larger, hump-shaped difference.[13] While the difference between males and females is on average five ranks at the 10th percentile, it increases to 21 ranks at the median and decreases again to nine ranks at the 90th percentile. However, the smaller rank difference at the lower end of the wage distribution reflects a lack of mass in lower tail of the male wage distribution. We thus conclude that both level and rank differences indicate a somewhat smaller gap at the top end of the distribution compared to the rest. Apart from

11. These findings for Germany contrast evidence from Sweden and the United States, where gender gaps are more pronounced at the upper part of the wage distribution, and thus overall larger among college graduates (Albrecht, Björklund, and Vroman, 2003; Bertrand, Goldin, and Katz, 2010).

12. Again, we use ranks of wages as measured in the male log wage distribution, following the approach introduced by Bayer and Charles (2018).

13. This is in line with findings from Bayer and Charles (2018), who find that black-white gaps in earnings are more pronounced when analyzing them in terms of ranks rather than levels.

Table 4.2. Level and rank gaps

	Quantiles				
	10th	25th	50th	75th	90th
A. Level gap					
Female	-0.236	-0.221	-0.238	-0.138	-0.108
	(0.012)	(0.003)	(0.009)	(0.005)	(0.011)
Including controls					
+ Majors	-0.178	-0.148	-0.129	-0.137	-0.121
	(0.012)	(0.010)	(0.006)	(0.009)	(0.012)
+ IQ and personality	-0.156	-0.114	-0.103	-0.091	-0.071
	(0.017)	(0.012)	(0.009)	(0.010)	(0.014)
+ Perceived ability	-0.154	-0.108	-0.098	-0.082	-0.077
	(0.018)	(0.012)	(0.009)	(0.009)	(0.015)
B. Rank gap					
Female	-5.2	-12.6	-20.6	-19.1	-8.5
	(0.3)	(0.5)	(0.7)	(0.8)	(0.8)
Including controls					
+ Majors	-4.0	-8.1	-12.4	-13.7	-7.0
	(0.4)	(0.6)	(0.7)	(0.9)	(1.0)
+ IQ and personality	-3.9	-6.3	-10.3	-10.3	-5.3
	(0.5)	(0.7)	(0.9)	(1.1)	(1.3)
+ Perceived ability	-3.9	-6.4	-9.8	-9.5	-5.1
	(0.6)	(0.7)	(0.9)	(1.0)	(1.3)

Notes: Each cell of this table reports the female coefficient that characterizes the gender differences for different quantiles. Panel A uses log expected wages as an outcome and thus reports level gaps, while panel B uses percentile ranks of expected wages measured in the expected wage distribution of males and therefore reports rank gaps as outlined in Section 4.2.4. Ability measures comprise IQ and personality traits and perceived ability comprises the subjective position in the distribution of academic and job-related skills, respectively. Log gross annual wages are winsorized at the 1% and 99% level.

heterogeneities in sorting, this finding might suggest that women at the middle and lower end of the distribution are less confident regarding their perceived or actual abilities. Indeed, after major choice as well perceived and actual ability (IQ, preferences, and personality) are accounted for, the gender gap in wage expectations becomes much more similar across quantiles.[14] The remaining gap is thus seems to accrue to male-female differences that exist along the entire distribution. Examples of such differences are child-rearing demands and negotiation preferences. Later in this paper, we will determine the relative importance of these factors.

14. See Table 4.A.1 for major-specific heterogeneities.

4.3.3 Life-cycle Trajectories in Expected Wages

The evidence presented in section 4.3.1 indicates that the gender gap in wage expectations increases with potential experience. To investigate the magnitude and relative importance of rising expected wage gaps over time, we use the three wage expectations (after graduation, at the age of 40, and at the age of 55) to fit individual-specific Mincerian wage trajectories as described in section 4.2.3. Figure 4.2a presents how male and female graduates expect earning trajectories to evolve over their respective lifetimes.[15] The figure reveals that the gender gap increases over time and this increase accelerates in the early-thirties when individuals start a family. Moreover, it increases until the age of 50 and stabilizes at 72% (i.e., females expect to earn 72% of the male wage at the age of 50). Expressed in terms of labor market experience, females need about nine years of prospective experience (from the age of 25 to 34) to reach the wage level that males expect to receive upon graduation (approx. 40,000 EUR). Males in turn expect to earn on average 49,000 EUR after nine years of experience, which is almost as high as the highest average wage level that females expect to earn throughout their entire careers (51,000 EUR at the age of 50).

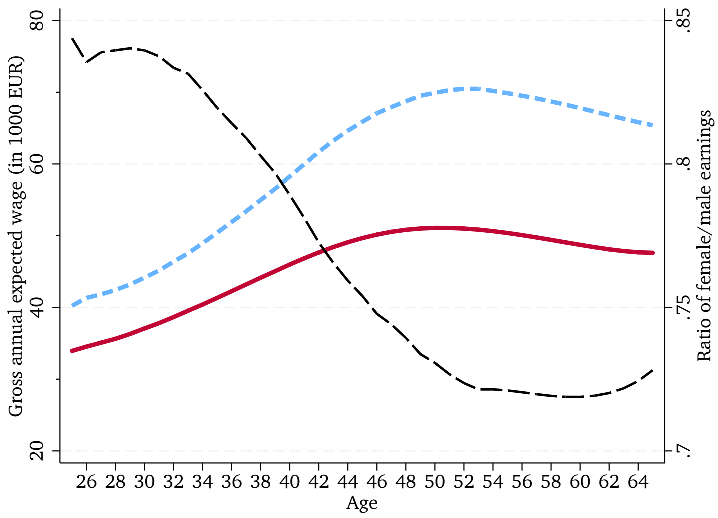
Figure 4.2b illustrates the distribution of annual wage growth by growth category ($<$2%, 2-4%, 4-6%, 6-8%, 8-10%, $\geq$10%). It shows that the vast majority of students expect annual wage growth rates of less than 4%. However, male students are more likely than females to expect larger growth rates. Thus, almost half of all female students expect their yearly wages to grow by less than 2%, compared to 35% of males. Moreover, students who expected high starting wages expect lower growth rates, and this pattern is more pronounced for females. Taken together, these patterns imply that expected wage trajectories of male and female students diverge over the life cycle. Nonetheless, while overall the gap in expected wages widens over the prospective life cycle at all parts of the expected starting wage distribution (see Figure 4.3a), rank differences persist or increase only slightly (Figure 4.3b).[16]

4.3.4 Comparing Expected Wages to Actual Wages

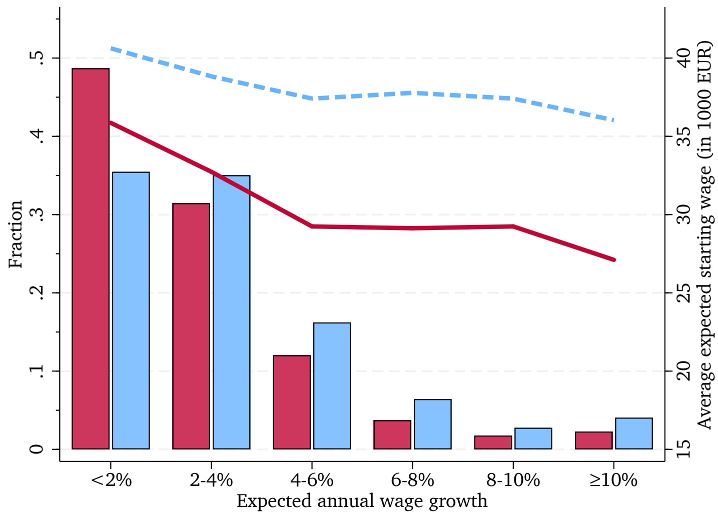
The above-described gender gap in wage expectations might translate into male-female differences in career decisions or family planning. Nonetheless, in terms of distributional concerns, fairness, and policy-making, its empirical relevance also de-

15. Note that Figure 4.2a expresses all expected wages in terms of a respondent's age while Table 4.1 presents expected starting wages irrespective of age. As there are students who graduate in their late-twenties or early-thirties, the sample used for this figure thus changes at initial ages. At the age of 25, approximately 39% of all students expect to have graduated from university. At the age of 28, 72%, at 30 this share amounts to 85% and at the age of 32 to 92%. Approximately 98% of all students expect to have graduated from university by the age of 35.

16. Figure 4.A.1 in the Appendix also confirms that the ranks in the starting wage distribution are highly correlated with ranks at the age of 40 and 55.



(a) Expected wages over the life-cycle



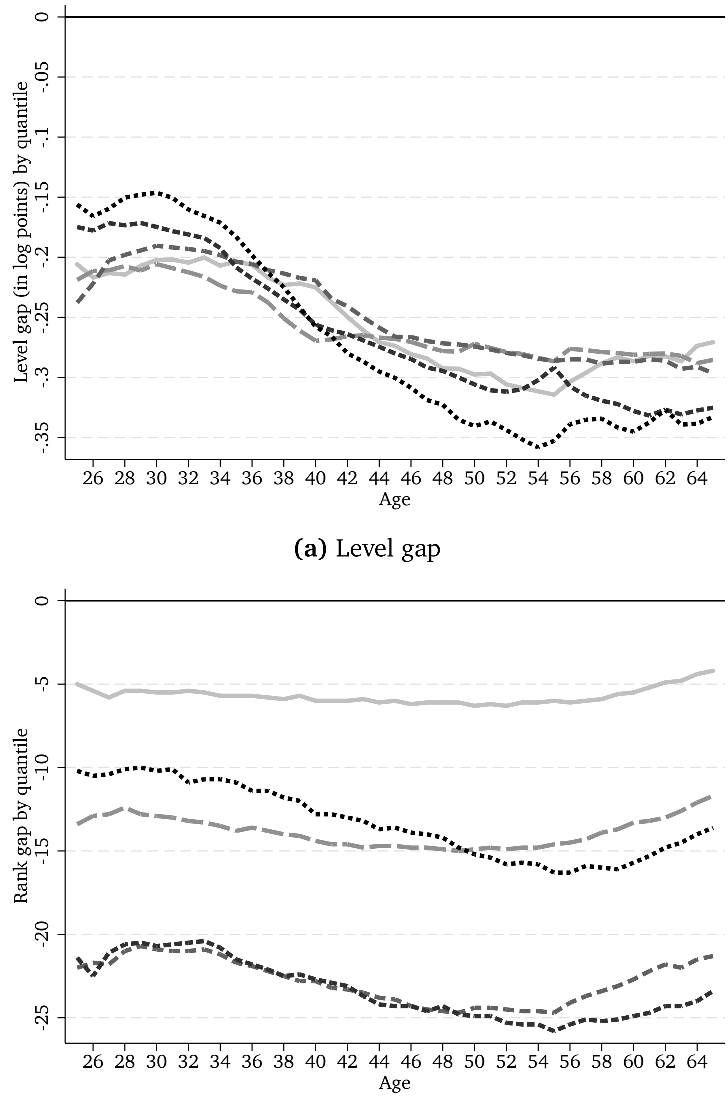
(b) Expected annual wage growth until age 40

Figure 4.2. Life-cycle wage trajectories and wage growth

Notes: Figure 4.2a shows the evolution of wages over the life cycle (females: red, solid; males: blue, dashed), including the female-male ratio (black, long-dashed). Figure 4.2b presents the expected annual wage growth until the age of 40 (bars measured on the left axis) and average expected starting wages (lines measured on the right axis) in each wage growth category separately for female (red, left bars) and male (blue, right bars) students in our sample. All wages are winsorized at the 1% and 99% level.

pends on the extent to which these expectations translate into actual gender wage differences.

Several pieces of evidence suggest that this is indeed the case. First, follow-up surveys on graduates who were initially surveyed about their wage expectations



(a) Level gap

(b) Rank gap

Figure 4.3. Rank and level gaps over the life-cycle for different initial quantiles

Notes: This figure presents the evolution of the wage gap measured in levels ranks for females starting at the 10th (very light, solid), 25th (light, long-dashed), 50th (medium, dashed), 75th (dark, short-dashed), and 90th (very dark, dotted) percentile of their wage distribution over the life cycle. Gaps are estimated using quantile regressions at each age, similar to Table 4.2.

during college show a close relation between the expectations and later realizations (Webbink and Hartog, 2004; Wiswall and Zafar, 2018a).[17] Second, the wage

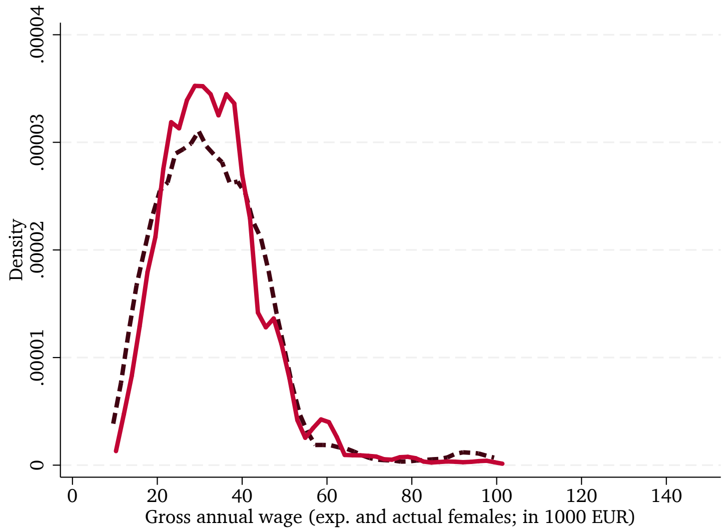
17. See also Attanasio and Kaufmann (2014), Filippin and Ichino (2005), and Schweri and Hartog (2017) for evidence that expectations predict subsequent real-life outcomes.

gap in expectations that we observe mimics the actual (conditional and unconditional) wage gap in Germany, as well as the fact that women experience much flatter life-cycle wage profiles (Destatis, 2017b; Francesconi and Parey, 2018). Thus, for example Francesconi and Parey (2018) report an overall actual gap among recent university graduates in Germany of 19.1%, while we find one of 15.5% in expectations. Besides, they report an actual gap of 10.5% among economics majors, whereas the gap in expectations among economics majors in our sample amounts to 10.45%. Third, the gender gap in starting wage expectations and the gender gap among recent graduates in our data are almost identical, and the same holds true for respective wage levels. Male recent graduates earn 38,728 EUR on average, and students in our sample expect to earn 39,076 EUR upon entry into the labor market (see Table 4.1). The corresponding values for female graduates and students are 33,945 EUR and 33,434 EUR, respectively. Finally, we find that the respective distributions overlap (see Figure 4.4), aside from slightly more mass at the lower end of the distribution among recent graduates. By comparing log (expected) wages of graduates and students in a regression framework (see Appendix Table 4.C.1), we can show that any of the observed differences stem from non-standard employment relationships (e.g., internships, part-time work). After controlling for gender, field of study, and working hours, there are no differences between expected and actual wages.

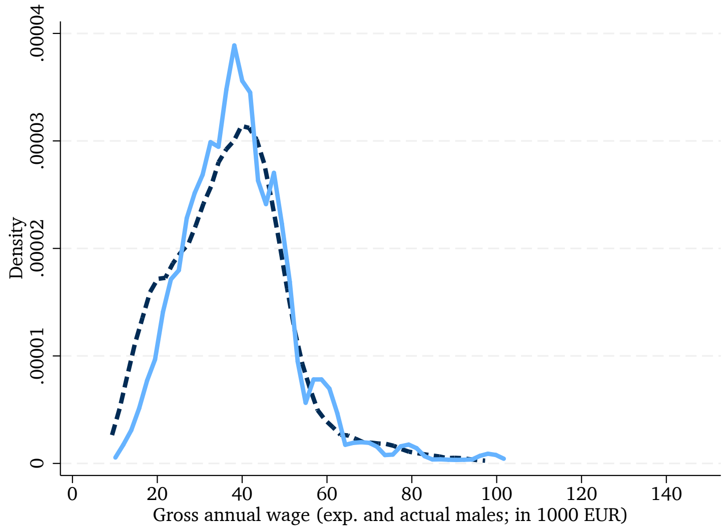
The empirical similarity of wage expectations and actual wages thus suggests that expectations reflect the expected outcome of (future) wage setting (Table 4.A.2 shows compelling evidence that this is indeed the case) and that women tend to anticipate lower wages mostly due to factors related to their gender. In the following, we will investigate this claim by shedding particular light on the relative importance of preference-based occupational sorting, child-related career breaks, and wage negotiation styles.

4.4 Explaining the Gender Gap in Wage Expectations

In this section, we examine the relative importance of several potential drivers of the gender gap in wage expectations. Alongside differential sorting into majors and occupations as well as differences in perceived/actual ability and personality traits or economic preferences, we focus in particular on the respective roles of anticipated child-rearing responsibilities and expected negotiation styles. For these factors to drive the gender gap in wage expectations documented in Section 4.3.1, two conditions need to be met: first, they need to differ across genders, and second, they need to matter for wage expectations. Sections 4.4.1 and 4.4.2 thus proceed by documenting male-female differences in child-rearing plans and wage negotiation patterns, respectively. Finally, section 4.4.3 presents regression and decomposition analyses to explore the relative importance of these and other potential divers for the gap in wage expectations.



(a) Females



(b) Males

Figure 4.4. Comparison of expected and actual wages

Notes: These figures present kernel densities of expected overall wages of female (red, solid; Figure 4.4a) and male (blue, solid; Figure 4.4b) students in our sample as well as the same distributions for actual wages of graduates (darker colors, dashed). All wages are winsorized at the 1% and 99% level.

4.4.1 Expected Child-rearing Responsibilities

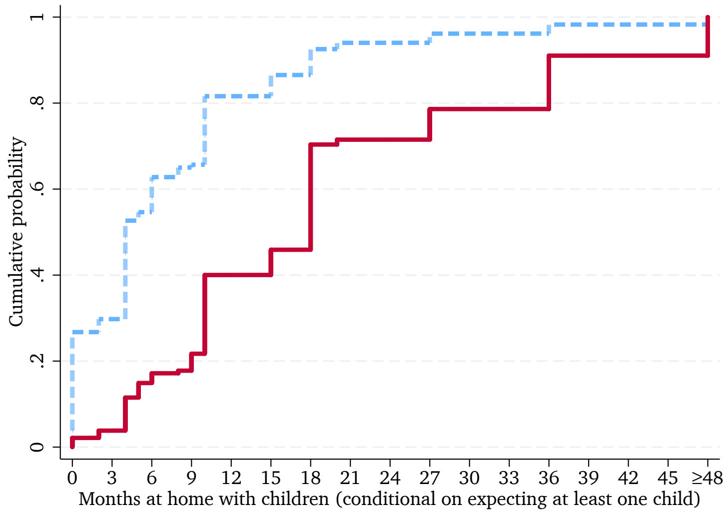
Biological and social differences in child-bearing and -rearing responsibilities are an important factor in explaining male-female differences in labor market outcomes (Bertrand, Goldin, and Katz, 2010; Daniel, Lacuesta, and Rodriguez-Planas, 2013; Goldin and Katz, 2016; Kleven, Landais, and Sogaard, forthcoming). First, women who intend to have children may select into occupations with flatter earnings profiles or linear pay structures, i.e., in anticipation of child-related wage penalties (Blau and Ferber, 1991; Goldin and Katz, 2016). Moreover, different fertility preferences, for example if women wanted more children or children at an earlier point in time, may affect a woman's household bargaining position regarding her child-rearing responsibilities and prospective labor market attachment. Second, career breaks in the form of parental leave may lead to a reduction in human capital, work-related networks, and experience, inducing females with children to earn lower relative (expected) wages afterwards (Albrecht, Edin, Sundström, and Vroman, 1999). Third, reduced working hours among women with children may exert an additional penalty in (expected) female wages, especially if long hours relate to promotions or increasing marginal returns (Angelov, Johansson, and Lindahl, 2016; Goldin, 2014).

Table 4.1. Summary statistics on family planning

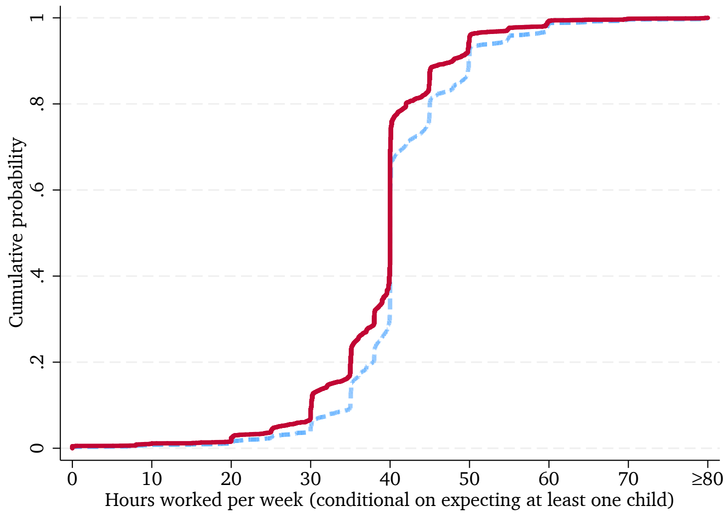
	Males	Females	Diff.	N
A. All respondents				
Wants to have children	0.88	0.87	0.02	15,348
Already has at least one child	0.03	0.02	0.01	15,256
Exp. working hours per week (age 40)	41.04	39.20	1.85	15,348
B. Conditional on wanting at least one child				
Age at birth of first child	30.59	29.38	1.21	13,370
Early parent (before age 30)	0.54	0.71	-0.16	13,427
Exp. number of children	2.27	2.20	0.07	13,427
Expected months at home per child	4.87	9.65	-4.78	11,666
Exp. working hours per week (age 40)	41.04	39.01	2.03	13,427

Notes: Panel A presents information on family planning and labor supply for all students in the sample, while panel B conditions on those respondents who want to have at least one child.

Table 4.1 summarizes male-female differences in fertility preferences, expected child-related career breaks, and expected weekly working hours. Regarding fertility preferences, the differences across genders are minor. 87% of females and 88% of males want to have children and conditional on parenthood, whereby both genders prefer to have on average around 2.2 children. However, women expect to have children about one year earlier than men and a much larger fraction (71% versus 54%) would like to have children before turning 30 years old. This age difference matches reality to the extent that males tend to be at least one year older in three quarters of all couples (German microcensus, 2010). Larger differences emerge when it comes to child-related career breaks. Males expect to stay home for around 5 months per child as opposed to females, who estimate that they will stay home for around 10 months with each child (see also Figure 4.1b). Expected differences in working hours at the age of 40 are again minor. The average expected number of working hours at the age of 40 among all individuals (panel A of Table 4.1) is almost identical to that for individuals who expect to have children (panel B of Table 4.1 and Figure 4.1b) and there is no significant difference if we restrict the sample to individuals with and without (expectant) children. Arguably, the age of 40 might be too late to capture a reduction in working times among individuals who expect to have children in their late-twenties. However, even among individuals who plan to have children in their late-thirties we do not find significant differences.



(a) Expected time home with kids



(b) Expected hours worked per week

Figure 4.1. CDFs of expected time at home with kids and working hours

Notes: This figure presents cumulative distribution functions of (a) time spent at home with children (career break) and (b) hours worked per week at the age of 40 conditional on expecting at least one child for both female (red, solid) and male (blue, dashed) students in our sample.

Figure 4.2 reveals that both males and females who expect to have children early, i.e., before the age of 30, expect longer career breaks and are also planning to work fewer hours. Young prospective parents thus seem to (rationally) anticipate less time-consuming careers. Nonetheless, as can be seen in panel (c) of Figure 4.2, females expect a wage penalty of 1,514 EUR for early parenthood (p-value < 0.01), while for males there is no difference (premium of 324 EUR, p-value $= 0.42$).

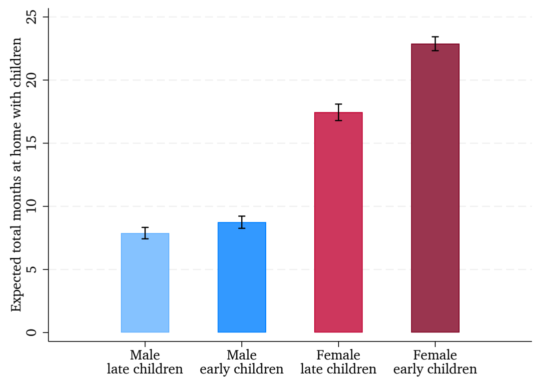
Some of the above expectations regarding fertility and time with children diverge from what we observe for current cohorts. Thus, for example, while only 13% of the women in our sample expect to remain childless, we see in current cohorts that 28% of women with an academic degree have no children at the age of 45 (Destatis, 2013). Moreover, women plan to interrupt their careers on average for 9.7 months for each child, but most expect to work full-time again at the age of 40. Among current cohorts, we observe that women with academic degrees interrupt their careers on average for 19 months (4 months for males) and only 32% of college-educated women with children under the age of 18 years work full time (Destatis, 2017a; Fabian, Rehn, Brandt, and Briedis, 2013). By contrast, males expect and realize almost no child-related interruptions or working-time reductions. In this sense, our findings are much in line with recent evidence presented in Kuziemko et al. (2018) showing that women underestimate the impact of motherhood on their future labor supply. Nonetheless, given that the students in our sample represent *future* cohorts of parents, it is somewhat difficult to distinguish false expectations from fundamental changes in child-rearing choices.

4.4.2 Negotiation Patterns

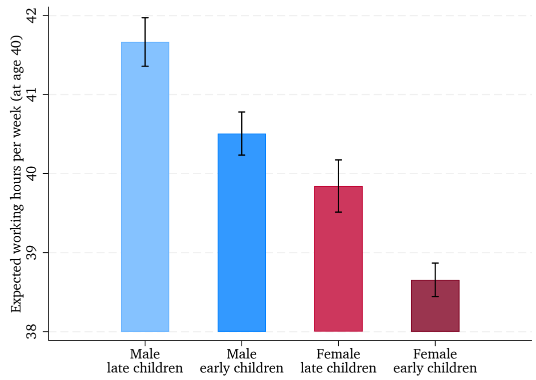
Wage negotiation strategies as well as initial wage claims and reservation wages may explain why a strong link exists between expected and actual wages (see section 4.3.4). For example, male-female differences in expected and actual wages may emerge if males are bolder in their initial wage claims or if females are more easily negotiated down towards their reservation wages.[18]

Table 4.2 presents initial wage claims, expected wages, and reservation wages of males and females first in Euros (panel A) and then in terms of ranks in the expected wage distribution of males (panel B). Expected wages on average lie between the initial wage claim and the reservation wage, indicating that most individuals expect to start a wage negotiation by claiming salaries above what they expect to receive. Similarly, they expect to settle on expected wages that lie above their respective

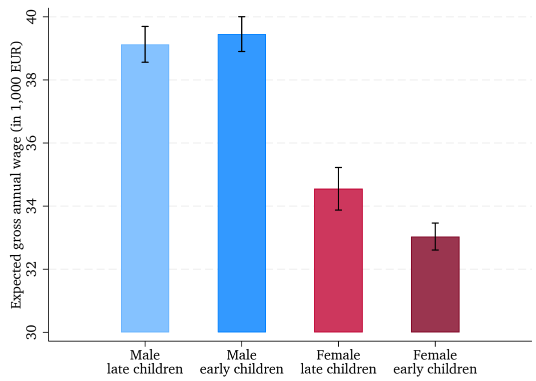
18. University graduates are usually asked to state their initial wage claim when applying for a position, such that there is little room for women to shy away from initiating a negotiation (Babcock and Laschever, 2009; Small, Gelfand, Babcock, and Gettman, 2007).



(a) Effect on time at home with children



(b) Effect on hours worked



(c) Effect on expected wages

Figure 4.2. Expected time at home with children, expected working hours, and expected wages for younger and older parents

Notes: This figure presents bar graphs of (a) time spent at home with children (career break) and of (b) hours worked per week at the age of 40 and (c) expected wages of younger and older parents conditional on expecting at least one child for both female (red) and male (blue, dark lines) students in our sample including 95% confidence intervals. Lighter colors indicate that females or males expect their first child after the age of 30, darker colors before the age of 30.

Table 4.2. Summary statistics on negotiation patterns

	Negotiation patterns			
	Males	Females	Diff.	N
A. Expressed in levels/Euro				
Initial wage claim	41,789	33,714	8,075	15,348
Expected wage	39,076	33,434	5,642	15,348
Reservation wage	34,355	28,002	6,352	15,348
B. Expressed in ranks				
Initial wage claim	58	40	18	15,348
Expected wage	50	37	14	15,346
Reservation wage	42	28	14	15,348
Negotiation style	16	13	3	15,348

Notes: Panel A reports mean initial wage claims, expected and reservation wages in Euro for both males and females. Panel B expresses these in ranks measured on the male expected wage distribution. See Section 4.2.4 for a description of how to calculate these ranks.

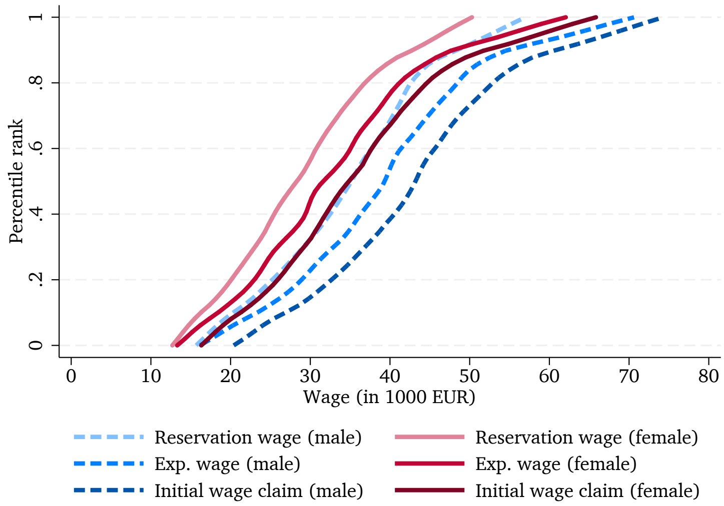
reservation wages.[19] Moreover, as shown in Figure 4.3a, this is true for both males and females and along the entire expected wage distribution.[20]

Males consistently expect to enter wage negotiations with a higher wage claim and reservation wage than females. The difference is substantial and the distribution of male reservation wages matches the distribution of female initial wage claims (compare Figure 4.3a). When expressing wage claims and reservation wages relative to an individual's expected wage, we also uncover that men tend to be bolder in their wage claims. As Figure 4.3b illustrates, males intend to claim a larger initial wage for every expected wage and they also expect to settle on a wage that exceeds their reservation wage more than females.

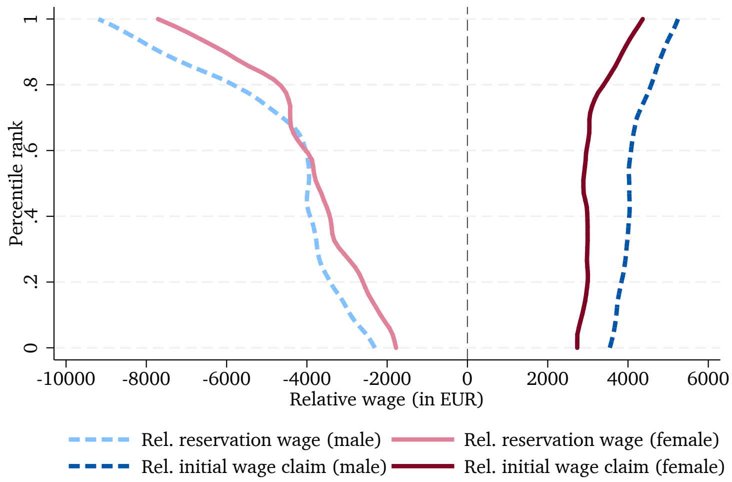
When expressing initial wage claims, expected wages and reservation wages in terms of ranks on the expected wage distribution of males, we observe that the previous finding persists: gender differences in initial wage claims are larger than differences in expected and reservation wages. The difference in ranks between initial wage claims and reservation wages is thus higher for males (16 ranks) than for females (13 ranks) (see panel B of Table 4.2). Figure 4.4a presents the distribution of negotiation patterns. About one third of female students in our sample leave very

19. For recent evidence on the importance of male-female differences in reservation wages for the gender gap, see Caliendo, Lee, and Mahlstedt (2017).

20. The close association between initial wage claims, reservation wages and expected wages is further confirmed by the results displayed in Table 4.A.2. It indicates that the difference between expected wages and initial wage claims remains constant along the expected wage distribution (coefficient close to 1). Nonetheless, the difference between reservation and expected wages increases along the distribution (coefficient < 1). This implies that at the top individuals expect a negotiation result that exceeds their reservation wage relatively more (this can also be seen graphically in Figure 4.3b).



(a) Initial wage claims, exp. and res. wages

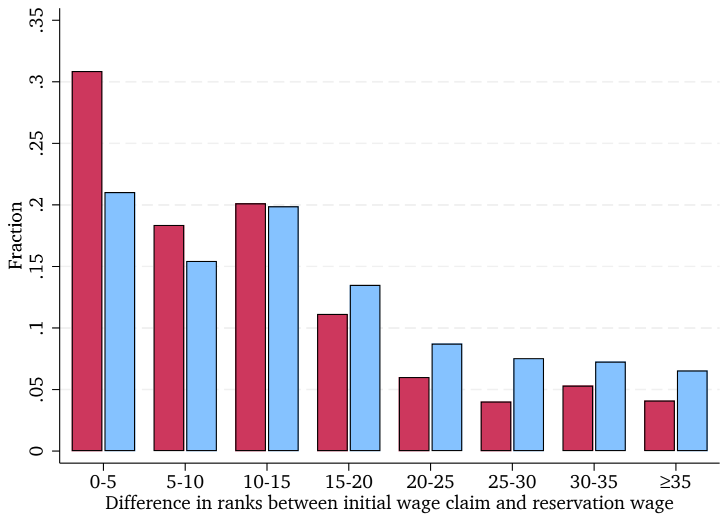


(b) Relative reservation wages and wage claims

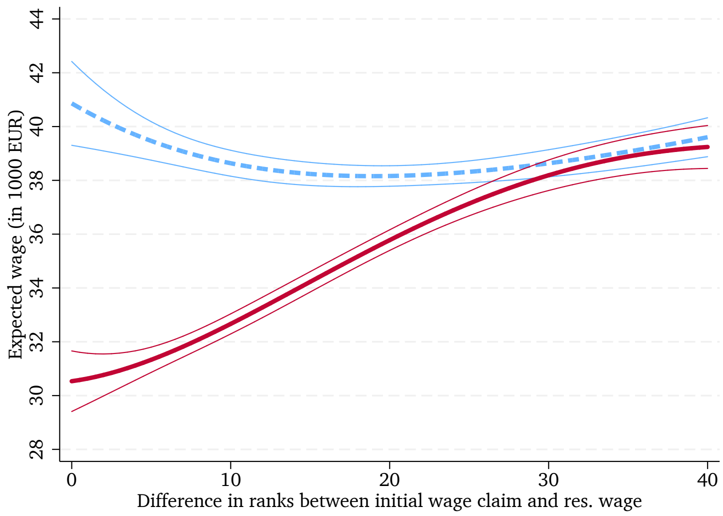
Figure 4.3. Initial wage claims, expected and reservation wages

Notes: Figure 4.3a presents reservation wages (w_R, light), expected wages (w_{exp}, medium) and initial wage claims (w_I, dark) ordered according to their percentile rank in the expected wage distribution of female (red, solid) and male (blue, dashed) students in our sample. Figure 4.3b presents the same distributions net of expected wages. All wages are winsorized at the 1% and 99% level.

little scope for negotiations, as there are only five ranks or fewer between their initial wage claims and their respective reservation wages. By contrast, males tend to enter negotiations with much bolder wage claims, with the majority planning to claim a wage that lies fifteen ranks or more above their reservation wage.



(a) Distribution of negotiation patterns



(b) Association of negotiation styles and expected wages

Figure 4.4. Negotiation styles by gender

Notes: Figure 4.4a presents the negotiation styles NS_i defined by the difference in ranks measured on the male (log) expected wage distribution between the initial wage claims and reservation wages for both female (red, left bars) and male (blue, right bars) students in our sample. Figure 4.4b presents the association between these negotiation styles NS_i and expected starting wages, $w_{i,st}$, from regressions of the type $w_{i,st} = \beta_0 + \beta_1 NS_i + \beta_2 NS_i^2 + \beta_3 NS_i^3 + \beta_4 NS_i^4 + \epsilon_i$ estimated separately for male and female students including 95% confidence intervals. All wages are winsorized at the 1% and 99% level.

These differences in negotiation styles prompt the question whether bolder negotiation styles pay off. While our data do not permit establishing causality, they allow us to investigate the relationship between negotiation styles and expected wages. Figure 4.4b uncovers a striking pattern: while males expected wages are nearly unaffected by their negotiation patterns, negotiations have large perceived returns for female students. A larger scope for negotiations with a higher initial wage claim increases the wage females expect to earn after graduation. Increasing boldness in negotiation styles by one standard deviation (approx. 12 ranks) for a female at the mean is associated with an 3,453 EUR increase in her expected wage, while a corresponding increase for males only amounts to an increase of 171 EUR.

The results in this section provide a novel view on negotiation styles as a driver of gender differences in labor market outcomes. While previous research suggests that females are less likely to initiate negotiations (Babcock and Laschever, 2009; Bowles, Babcock, and Lai, 2007; Leibbrandt and List, 2015), we provide field evidence suggesting that females ask for less in wage negotiations, thereby complementing evidence from laboratory experiments (Rigdon, 2012). In addition, we show that negotiating pays off for female students, who expect large returns for higher initial wage claims. This finding might be consistent with the notion that women "know when to ask" (Exley, Niederle, and Vesterlund, forthcoming). By contrast, there is no such effect for males.

4.4.3 Decomposing the Gender Gap in Wage Expectations

Previous sections have documented the extent to which males and females differ in their prospective child-rearing and negotiation patterns. In this section, we use Oaxaca-Blinder decompositions to approximate the extent to which these and other factors contribute to the gender gap in wage expectations. Thus, for example, sorting into specific academic majors has been shown to hold particular importance for expected and actual wage gaps (Francesconi and Parey, 2018; Zafar, 2013), as is sorting into different occupations and industries (Goldin, 2014; Wiswall and Zafar, 2018b). Nonetheless, sorting into occupations and industries might not only reflect preferences, but might also be driven by individual perceptions about discrimination or class ceilings (Blau and Kahn, 2017), ability, perceived relative ability, personality or economic preferences (Cortes and Pan, 2018; Fouarge, Kriechel, and Dohmen, 2014), all of which may also have a direct effect on expected wages. We thus subsume all potential drivers of the gender wage gap by forming three groups: (A) sorting into majors, occupations, industries as well as perceived/actual ability and personality, (B) labor supply and family planning, and (C) negotiation styles.

To obtain relative shares of these factors, we compute the share of the gap that is attributable to sorting (comprised of sorting into majors, occupations, and industries as well as perceived relative ability, personality and economic preferences), family planning, and negotiation styles based on a twofold Oaxaca-Blinder decomposition using regression coefficients from a pooled regression model.[21] The results of this

21. We use pooled coefficients to obtain an estimate about the importance of differences in characteristics rather than their (perceived) prices. Differences in coefficients enter the unexplained differ-

model suggest that each of the above factors matter for expected wages and that the estimated relationship mimics results of models with actual wages as dependent variable (see Tables 4.A.3 and 4.A.4). Thus, for example, majors in medical sciences, law, economics/business, and STEM each yield a large and significant premium over a major in humanities. Similarly, conscientiousness and extraversion yield a wage premium, while agreeableness is associated with lower wages (for a comparison using actual wages, see Heineck and Anger, 2010). Finally, working hours are positively associated with expected wages as is boldness in wage negotiations.

Table 4.3 and Figure 4.5a present the results from an Oaxaca-Blinder decomposition of the gender gap in wage expectations for both starting wages as well as expected wages earned over the life cycle (see Figure 4.5b). Consistent with previous research (Arcidiacono, Hotz, and Kang, 2012; Wiswall and Zafar, 2015; Zafar, 2013), we find that a sizable share of the gender gap in wage expectations relates to differential sorting into majors, occupations, and industries, with occupations as the finest category being most important. By contrast, our vast battery of perceived/actual ability, personality and economic preference measures explains only 3% of the male-female difference in expected starting wages.[22] However, this share rises to 10% once we decompose expected lifetime wages. We interpret this as suggestive evidence of anticipated employer learning (see, e.g., Altonji and Pierret, 2001), i.e., the idea that employers are unable to fully price a graduate's non-cognitive characteristics at the beginning of the career, but only with increasing experience. The notion that majors explain a smaller share of the gap in lifetime wages relatively to starting wages is also consistent with this idea.

Compared to sorting, labor supply and family planning together make up for a somewhat smaller share of around 12%, where most of the variance is explained by anticipated working hours rather than child-related career breaks. In fact, we observe hardly any expected child penalty after we control for occupations and industries, indicating that women may opt for somewhat more family-friendly occupations (with flatter wage trajectories as described in Section 4.3.3), but then do not experience a relative decline in expected wages due to family planning and child-related career breaks (see Kuziemko et al., 2018, for related evidence). Finally, negotiation styles explain 14% of the gender gap and this is true on average even within occupation categories and after controlling for measures of perceived and actual ability. Moreover, the importance of negotiation styles remains similar at 9%

ence, usually attributed to discrimination. Note that this yields a lower bound of the estimated effect of wage negotiations, given our estimates displayed in Figure 4.4b. There are no differences in the pricing of child-related labor force interruptions.

22. Overconfidence, measured by perceived and actual ability, thus proves much less important in our data than suggested by some of the previous evidence on elite students (see, e.g., Reuben, Wiswall, and Zafar, 2017).

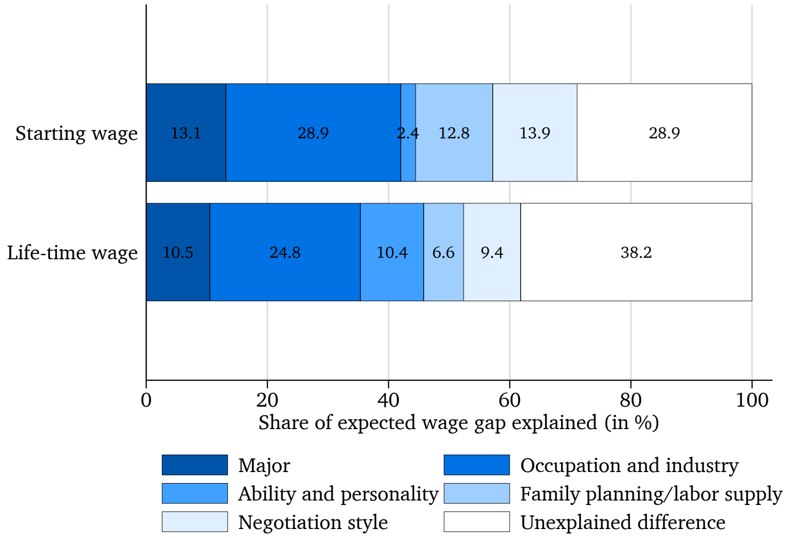
over the life cycle, indicating that negotiation strategies set individuals on different initial wage trajectories with important ramification throughout their entire career.

Table 4.3. Oaxaca-Blinder decomposition of the gender gap in wage expectations

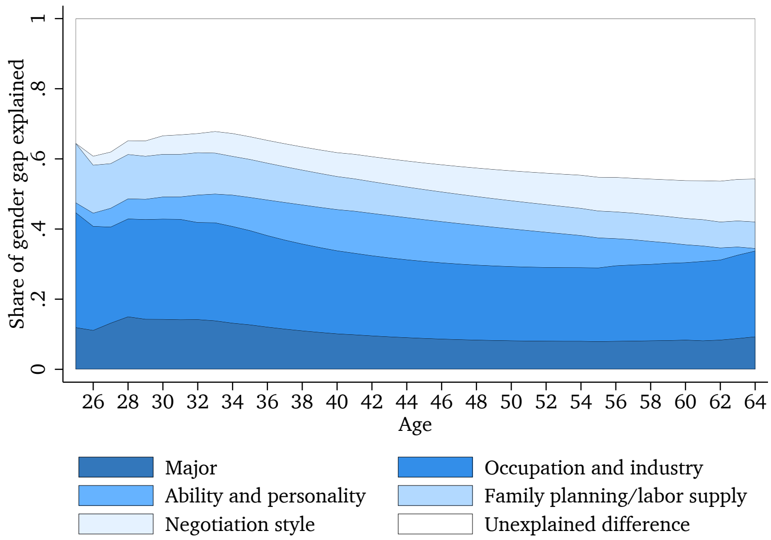
	log(Expected starting wage)				**log(Expected lifetime wage)**			
	(1) with occ. sorting		(2) without occ. sorting		(3) with occ. sorting		(4) without occ. sorting	
Unadjusted difference	0.181	100.000	0.181	100.000	0.230	100.000	0.230	100.000
	(0.010)		(0.010)		(0.011)		(0.011)	
Explained difference	0.129	71.104	0.101	55.752	0.142	61.812	0.119	51.740
	(0.009)		(0.008)		(0.009)		(0.008)	
Composition effects attributable to								
A. Sorting								
Major	0.024	13.143	0.044	24.535	0.024	10.504	0.046	19.944
	(0.005)		(0.004)		(0.004)		(0.004)	
Occupation	0.029	16.165			0.041	17.653		
	(0.006)				(0.006)			
Industry	0.023	12.708			0.017	7.195		
	(0.004)				(0.004)			
Perc./actual ability	0.004	2.432	0.005	2.899	0.024	10.443	0.027	11.786
& personality	(0.005)		(0.005)		(0.005)		(0.005)	
B. Labor supply/family planning								
Hours worked	0.018	9.783	0.018	9.714	0.017	7.299	0.019	8.234
	(0.003)		(0.003)		(0.002)		(0.003)	
Children	0.005	2.969	0.007	4.065	-0.002	-0.723	0.003	1.388
	(0.005)		(0.005)		(0.004)		(0.004)	
C. Negotiation styles								
Boldness	0.025	13.904	0.026	14.539	0.022	9.440	0.024	10.388
	(0.002)		(0.002)		(0.002)		(0.002)	
Observations	10788		10788		9146		9146	

Notes: This table decomposes the differences in log expected starting or lifetime wages into components attributable to (A) sorting into majors, occupations, and industries as well as perceived ability, personality and economic preferences (perceived ability on the job and in university, IQ, Big Five personality traits, altruism, impatience, positive and negative reciprocity, risk aversion and trust), (B) labor supply and family planning (expected hours per week, expected number of children, months at home with children, indicator for early parenthood), and (C) negotiation styles (as defined in Section 4.2.4) using Oaxaca-Blinder decompositions. For each decomposition, we also present the share of the difference that is attributable to the respective component and present results with and without controls for sorting into occupation and industries. Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level.

We conduct several additional analyses and robustness checks. First, we notice that the above Oaxaca-Blinder decomposition explains a substantial portion, but not all of the difference in male-female expected starting (lifetime) wages. Given the breadth of available measures on individual characteristics in our data, unmeasured differences in personal characteristics are unlikely to account for the remaining difference. Instead, differences in regional contexts may prove important (see, e.g., Kuchler and Zafar, forthcoming; Malmendier and Nagel, 2011; Malmendier and



(a) Decomposition of starting and lifetime wages



(b) Decomposition over the life-cycle

Figure 4.5. Decomposition of expected wages

Notes: Figure 4.5a illustrates the decomposition of expected starting and lifetime wages presented in Table 4.3. Figure 4.5b presents this decomposition for all ages over the life cycle. Categories are aggregated such that labor supply/children corresponds to the sum of hours worked and children, negotiation style/personality corresponds to negotiation style, perceived ability/discrimination as well as personality.

Nagel, 2015), since individuals stem from very different regional labor markets with very different actual gender wage gaps (see Figure 4.A.2). Nonetheless, our findings displayed in Tables 4.A.5 and Figure 4.A.3 indicate that regional differences in gender wage gaps are largely unrelated to actual expected wages. Second, along the same lines, we investigate the importance of having experienced different degrees of female wage discrimination in previous student jobs. Here, again we find that the wage earned in previous student jobs does not explain the wage differences as shown in Table 4.A.6. Third, we replicate Table 4.3 for students who do not aim to enter the public sector as for them negotiation styles might be more important than for prospective civil servants. As Table 4.A.7 documents, we do not find substantial differences when focusing on this subsample. Finally, in Appendix Table 4.A.8 and 4.A.9, we also present unconditional quantile decompositions corresponding to the decompositions in Table 4.3 at different points along the distribution. The results of these decompositions are similar to the Oaxaca-Blinder decompositions at the mean, with one exception: the importance of negotiation styles decreases along the distribution, while personality traits become more important in explaining the gap.

4.5 Conclusion

This study provides first large-scale evidence on the gender gap in wage expectations. Already prior to labor market entry, women expect much lower wages than men and this gender gap in expected wages is significant and large across all subgroups. Moreover, it prevails along the entire distribution, and increases over the prospective life cycle. In terms of relative magnitudes, females would need to work on average around four hours more per week in the same occupation and industry, or major for instance in medical sciences rather than humanities to catch up with the starting wages of their male peers. Similarly, in expectation, it would take them about nine years more of accumulated work experience to make up for the gender penalty.

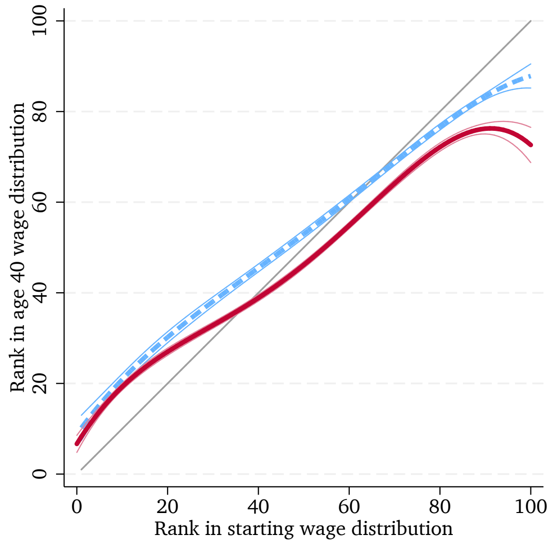
The overall pattern of results confirms previous findings on the importance of sorting into certain majors, industries or occupations, and a female preference for jobs with flatter wage schedules (Blau and Ferber, 1991; Blau and Kahn, 2017; Brunello, Lucifora, and Winter-Ebmer, 2004; Zafar, 2013). Yet, except for a wage penalty of having children early, women seem to underestimate the extent and importance of child-related career breaks. We also document a striking relationship between expected wages, initial wage claims and reservation wages, and use this information to construct a measure of negotiation styles, which reveals that women plan to enter wage negotiations with more modest wage claims relative to their reservation wage. A decomposition of starting and lifetime wages into components related to sorting, perceived/actual ability as well as personality, child-rearing responsibilities and negotiation styles unveils that after sorting is accounted for, working hours matter but child-related career breaks are largely unimportant. What does matter, however, is boldness in initial wage negotiations, with important consequences for expected starting and lifetime wages.

The above findings have implications for our understanding of wage-setting processes, expectation formation, and economic modeling. In particular, the documented systematic and accurate gender differences in wage expectations and their strong relation with wage claims and reservation wages suggest that expected wages drive actual wage differences and persistent gender wage gaps. At the same time, the expectation formation process for wages is non-adaptive, given that relative wage expectations are not affected by contextual labor market variables. Instead, expected wages seem a prospective, preference-related component in wage setting, which might thus be more easily malleable than, e.g., expectations about aggregate economic relationships that are indeed shaped by experiences (Fuster, Laibson, and Mendel, 2010; Kuchler and Zafar, forthcoming; Malmendier and Nagel, 2011, 2015). Given their accuracy and forward-looking nature, relative expected wage disparities likely matter for financial decision-making, household bargaining, as well as education and labor market choices. In this respect, our results also inform the economic modeling of such decisions and associated learning processes (see, e.g., Breen and Garcia-Penalosa, 2002; Reuben, Wiswall, and Zafar, 2017; Wiswall and Zafar, 2018b; Xia, 2016).

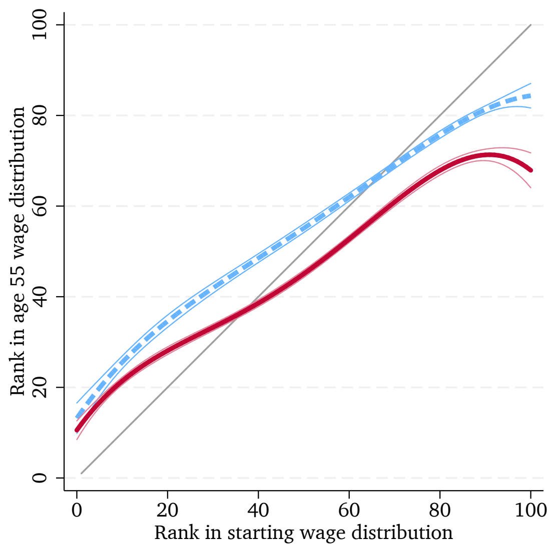
The findings presented in this paper also provide an explanation for several empirical patterns. First, our results suggest that women are aware of the career cost of having children early, which may explain the observational tendency to delay child birth among highly-educated women (Bratti, 2015). However, aside from considerations of timing, women underestimate the child-related dampening in their wage trajectories, with potential implications for household bargaining and the distribution of child-rearing tasks. Thus, women may stay home at a higher rate not only because they expect lower labor market returns than their spouses, but also because they underestimate the wage loss associated with raising children (Kuziemko et al., 2018). Second, it seems as if reluctant negotiation behavior leads to lower reference points and lower subsequent wage expectations. While we cannot make strong causal statements given the nature of our data, our evidence strongly supports the idea that initial negotiation styles matter for starting wages and differences in starting wages lead to different wage trajectories. Hence, these findings may explain why wage gaps are larger among university students entering labor markets in which unionized wage setting is rare and where employer-employee negotiations hold particular importance in the wage-setting process (Blau and Kahn, 2017).

Our results also deliver insights regarding the effective implementation of policies aimed at leveling the playing field between genders. First, our findings suggest that negotiation trainings – rather than encouraging more negotiations per se (Exley, Niederle, and Vesterlund, forthcoming) – might be an effective measure to improve female labor market outcomes and reduce the gender wage gap (Ashraf, Bau, Low, and McGinn, 2018). In fact, such measures seem to be more effective than policies that encourage women to enter male-dominated fields, for which the gender gap in expectations tends to be somewhat higher. They may also be more effective than exposure to low actual gender gaps (e.g., by enforcing equal pay in student jobs), which we find to be unrelated to differences in wage expectations. Second, the above evidence suggests that information treatments on child-related wage penalties might help women to gain a more realistic view of the career costs of raising a family and they might also lead women to bargain for a more equal distribution of child-rearing responsibilities within households. In future research, it would thus be informative to ascertain how our measure of negotiation styles elicited before labor market entry translates into realized wages, and whether randomly-assigned information treatments about negotiation styles or child-related labor market penalties can reduce actual wage gaps to the same extent as suggested in this paper.

Appendix 4.A Additional Figures and Tables



(a) Age 40



(b) Age 55

Figure 4.A.1. Marginal effects of increases in starting wage ranks on later earnings

Notes: This figure presents the associations between the an individual's rank in the starting wage distribution (R_{st}) and the rank in the distribution of ranks later in life (R_a, $a = 40, 55$) including 95% confidence intervals. Marginal effects are from regressions of the type $R_{a,i} = \beta_0 + \beta_1 R_{st,i} + \beta_2 R_{st,i}^2 + \beta_3 R_{st,i}^3 + \beta_4 R_{st,i}^4 + \epsilon_i$ ($a = 40, 55$) estimated separately for female (red, solid) and male (blue, dashed) students.

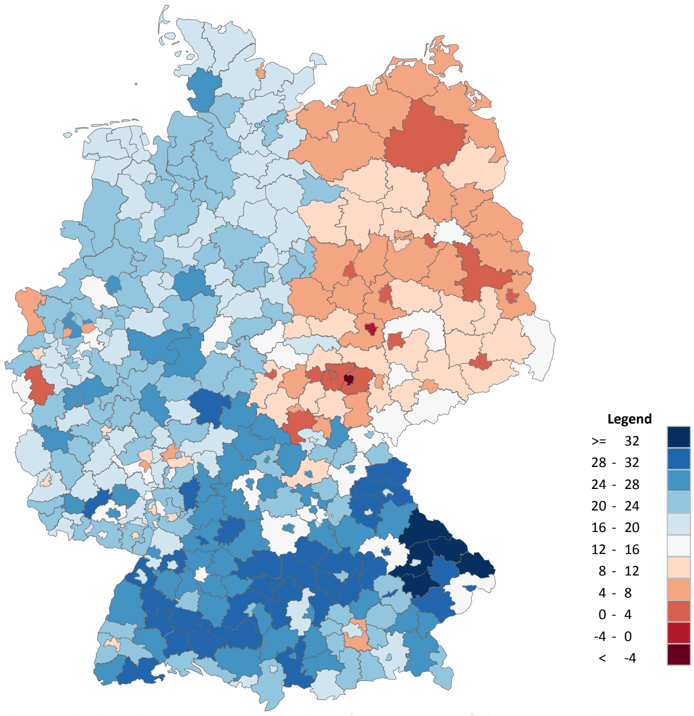
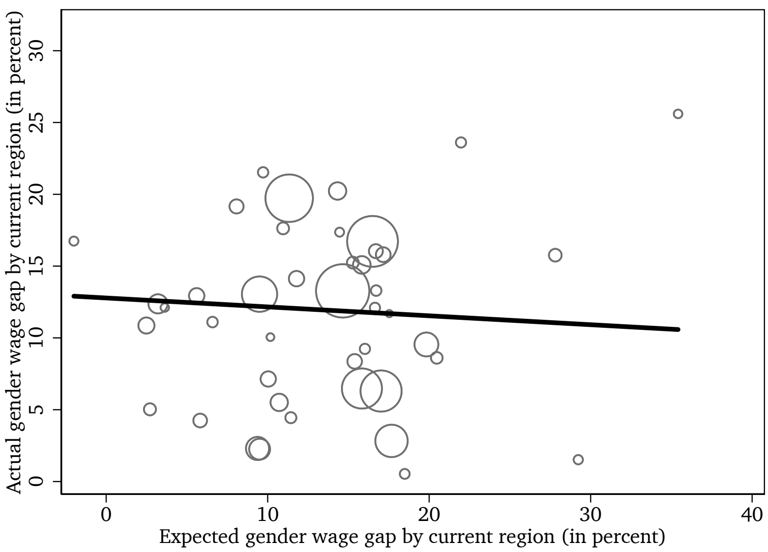


Figure 4.A.2. Regional differences in actual gender wage gaps

Notes: This figure displays percentage differences in actual wages for the year 2012 across regions (Kreise) in Germany using data from the German statistical office.

(a) Region of origin



(b) Current region

Figure 4.A.3. Regional differences in gender wage gaps

Notes: These figures display the relationship between the expected and actual gender gap by region (Kreise) in Germany using either the region of origin (Figure 4.A.3a; slope of -0.035, standard error: 0.037) or the current region (Figure 4.A.3b; slope of -0.062, standard error: 0.152).

Table 4.A.1. Level and rank gaps by major

	Quantiles				
	10th	25th	50th	75th	90th
A. Level gap					
Baseline	-0.236	-0.221	-0.238	-0.138	-0.108
	(0.012)	(0.003)	(0.009)	(0.005)	(0.011)
Control for majors	-0.178	-0.148	-0.129	-0.137	-0.121
	(0.012)	(0.010)	(0.006)	(0.009)	(0.012)
Separately by major					
Med./Health Sciences	-0.135	-0.149	-0.071	-0.183	-0.179
	(0.058)	(0.036)	(0.031)	(0.025)	(0.028)
STEM	-0.219	-0.232	-0.134	-0.145	-0.114
	(0.019)	(0.022)	(0.008)	(0.013)	(0.015)
Law	-0.116	-0.131	-0.187	-0.220	-0.140
	(0.085)	(0.049)	(0.036)	(0.057)	(0.081)
Econ./Business	-0.128	-0.115	-0.109	-0.108	-0.092
	(0.028)	(0.017)	(0.007)	(0.012)	(0.020)
Hum./Soc. Sciences	-0.165	-0.124	-0.131	-0.078	-0.106
	(0.032)	(0.019)	(0.017)	(0.020)	(0.036)
B. Rank gap					
Baseline	-5.2	-12.6	-20.6	-19.1	-8.5
	(0.3)	(0.5)	(0.7)	(0.8)	(0.8)
Control for majors	-4.0	-8.1	-12.4	-13.7	-7.0
	(0.4)	(0.6)	(0.7)	(0.9)	(1.0)
Separately by major					
Med./Health Sciences	-3.0	-5.9	-9.5	-20.2	-10.1
	(1.2)	(1.8)	(3.1)	(3.3)	(2.5)
STEM	-7.9	-15.2	-17.7	-14.3	-6.3
	(1.0)	(1.2)	(1.3)	(1.2)	(1.0)
Law	-2.7	-12.4	-21.0	-14.0	-1.7
	(2.4)	(3.9)	(4.8)	(3.6)	(1.1)
Econ./Business	-7.3	-10.7	-12.7	-11.7	-7.5
	(1.1)	(1.2)	(1.2)	(1.4)	(1.3)
Hum./Soc. Sciences	-1.3	-2.6	-7.1	-11.0	-10.6
	(0.3)	(0.5)	(1.0)	(1.9)	(3.7)

Notes: Each cell of this table reports the female coefficient, which characterizes the gender differences for different quantiles and sample specification. Panel A uses log expected wages as an outcome and thus reports level gaps, while panel B uses percentile ranks of expected wages measured in the expected wage distribution of males and therefore reports rank gaps as outlined in section 4.2.4. Log gross annual wages are winsorized at the 1% and 99% level.

Table 4.A.2. Comparison of initial wage claims, reservation and expected wages

	log(Initial claim)		**log(Reserv. wages)**	
	(1)	(2)	(3)	(4)
A. Complete sample				
Log average expected wage (starting)	0.954***	0.903***	1.061***	1.012***
	(0.016)	(0.018)	(0.021)	(0.023)
Gender, major, occupation, industry, labor supply	No	Yes	No	Yes
R^2 (adj.)	.44	.44	.41	.42
Observations	15346	15346	15346	15346
p-value: Coefficient=1	0.00	0.00	0.00	0.60
	log(Initial claim)		**log(Reserv. wages)**	
	(1)	(2)	(3)	(4)
	Females	Males	Females	Males
B. Subsamples by gender				
Log average expected wage (starting)	0.884***	0.931***	1.006***	1.014***
	(0.025)	(0.023)	(0.032)	(0.032)
Gender, major, occupation, industry, labor supply	Yes	Yes	Yes	Yes
R^2 (adj.)	.4	.49	.39	.45
Observations	8720	6626	8720	6626
p-value: Coefficient=1	0.00	0.00	0.85	0.65

Notes: This table presents the relation of expected starting wages to initial wage claims and reservation wages. In panel (a), we present results for the whole sample, while we replicate columns (2) and (4) of panel (a) for each gender separately. Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 4.A.3. Determinants of the gender gap in starting wage expectations

	log(expected starting wage)				
	(1)	(2)	(3)	(4)	(5)
Female	-0.184***	-0.105***	-0.087***	-0.063***	-0.052***
	(0.009)	(0.010)	(0.012)	(0.013)	(0.012)
A. Sorting					
Medical/health sciences		0.117***	0.150***	0.123***	0.107***
		(0.029)	(0.035)	(0.034)	(0.033)
STEM		0.126***	0.140***	0.137***	0.114***
		(0.020)	(0.021)	(0.020)	(0.020)
Law		0.189***	0.139**	0.084*	0.079
		(0.044)	(0.055)	(0.050)	(0.050)
Economics/business		0.174***	0.178***	0.156***	0.133***
		(0.017)	(0.019)	(0.019)	(0.018)
Civil servant		-0.040***	-0.031**	-0.018	-0.020
		(0.013)	(0.015)	(0.015)	(0.014)
Agreeableness			-0.009	-0.007	-0.006
			(0.007)	(0.006)	(0.006)
Conscientiousness			0.021***	0.018***	0.019***
			(0.007)	(0.007)	(0.007)
Emotional Stability			-0.001	0.000	-0.003
			(0.007)	(0.006)	(0.006)
Extraversion			0.028***	0.021***	0.020***
			(0.006)	(0.006)	(0.006)
Openness			0.003	-0.002	-0.005
			(0.006)	(0.005)	(0.005)
B. Labor supply/family planning					
Exp. working hours per week				0.016***	0.015***
				(0.002)	(0.002)
Exp. number of chidren				0.013**	0.011*
				(0.007)	(0.007)
Exp. months at home				-0.000	-0.000
				(0.000)	(0.000)
Exp. children before age 30				-0.006	-0.008
				(0.010)	(0.010)
C. Negotiation Style					
Boldness					0.007***
					(0.000)
Occupation and industry	No	Yes	Yes	Yes	Yes
Subjective ability/perc. discrimination	No	No	Yes	Yes	Yes
IQ and economic preferences	No	No	Yes	Yes	Yes
R^2 (adj.)	.025	.087	.11	.16	.18
Observations	15346	15346	10788	10788	10788

Notes: This table presents regressions of log expected starting wages on varying sets of controls: variables that relate to (A) sorting based on majors (with humanities as the omitted baseline major category), occupations, industries and standardized measures of personality, (B) labor supply and family planning, and (C) negotiation styles. Column (5) corresponds to the specification underlying the decomposition in Table 4.3. Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 4.A.4. Determinants of the gender gap in lifetime wage expectations

	log(expected lifetime wage)				
	(1)	(2)	(3)	(4)	(5)
Female	-0.239***	-0.153***	-0.113***	-0.097***	-0.088***
	(0.009)	(0.009)	(0.011)	(0.012)	(0.012)
A. Sorting					
Medical/health sciences		0.163***	0.203***	0.199***	0.182***
		(0.020)	(0.024)	(0.024)	(0.023)
STEM		0.137***	0.146***	0.147***	0.127***
		(0.017)	(0.018)	(0.018)	(0.017)
Law		0.221***	0.166***	0.133***	0.131***
		(0.045)	(0.055)	(0.051)	(0.051)
Economics/business		0.207***	0.198***	0.185***	0.165***
		(0.016)	(0.018)	(0.017)	(0.017)
Civil servant		-0.083***	-0.074***	-0.065***	-0.066***
		(0.012)	(0.014)	(0.014)	(0.014)
Agreeableness			-0.021***	-0.022***	-0.020***
			(0.006)	(0.006)	(0.006)
Conscientiousness			0.016***	0.015***	0.015***
			(0.006)	(0.006)	(0.006)
Emotional Stability			0.003	0.003	0.000
			(0.006)	(0.005)	(0.005)
Extraversion			0.034***	0.030***	0.029***
			(0.006)	(0.005)	(0.005)
Openness			0.022***	0.020***	0.017***
			(0.005)	(0.005)	(0.005)
B. Labor supply/family planning					
Exp. working hours per week				0.010***	0.010***
				(0.001)	(0.001)
Exp. number of chidren				0.021***	0.019***
				(0.006)	(0.006)
Exp. months at home				-0.000	-0.000
				(0.000)	(0.000)
Exp. children before age 30				0.038***	0.036***
				(0.010)	(0.010)
C. Negotiation Style					
Boldness					0.006***
					(0.000)
Occupation and industry	No	Yes	Yes	Yes	Yes
Subjective ability/perc. discrimination	No	No	Yes	Yes	Yes
IQ and economic preferences	No	No	Yes	Yes	Yes
R^2 (adj.)	.052	.19	.23	.26	.28
Observations	12734	12734	9146	9146	9146

Notes: This table presents regressions of log expected starting wages on varying sets of controls: variables that relate to (A) sorting based on majors (with humanities as the omitted baseline major category), occupations, industries and standardized measures of personality, (B) labor supply and family planning, and (C) negotiation styles. Column (5) corresponds to the specification underlying the decomposition in Table 4.3. Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level. *, **, and *** denote significance at the 10, 5, and 1 percent level.

Table 4.A.5. Association of actual gender gaps with expected gender gaps

	log(expected starting wage)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.190***	-0.152***	-0.154***	-0.147***	-0.177***	-0.073***
	(0.010)	(0.013)	(0.012)	(0.013)	(0.016)	(0.016)
Avg. wage in county of origin (in 1,000 EUR)		0.010***		0.004	0.002	0.000
		(0.002)		(0.003)	(0.003)	(0.003)
Avg. wage in current county (in 1,000 EUR)			0.012***	0.009***	0.002	0.003
			(0.002)	(0.003)	(0.004)	(0.003)
State fixed effects	No	No	No	No	Yes	Yes
Major, occupation, industry, labor supply	No	No	No	No	No	Yes
R^2 (adj.)	.027	.029	.029	.03	.033	.16
Observations	11759	11759	11759	11759	11759	11759

Notes: This table presents regressions of log expected starting wages on a female indicator and measures of actual regional wage levels for a respondent's own gender. The sample is restricted to those with valid information on their county of origin (i.e., where students received their high school diploma) and current county (i.e., where they are currently living). Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level.

Table 4.A.6. Oaxaca-Blinder decomposition of the gender gap in wage expectations including past wages in student jobs

	log(Expected starting wage)				log(Expected lifetime wage)			
	(1) with occ. sorting		(2) without occ. sorting		(3) with occ. sorting		(4) without occ. sorting	
Unadjusted difference	0.181	100.000	0.181	100.000	0.230	100.000	0.230	100.000
	(0.010)		(0.010)		(0.011)		(0.011)	
Explained difference	0.129	71.295	0.102	56.106	0.142	61.832	0.119	51.784
	(0.009)		(0.008)		(0.009)		(0.008)	
Composition effects attributable to								
A. Sorting								
Major	0.024	13.507	0.045	24.921	0.024	10.535	0.046	19.989
	(0.005)		(0.004)		(0.004)		(0.004)	
Occupation	0.029	16.052			0.041	17.646		
	(0.006)				(0.006)			
Industry	0.023	12.673			0.017	7.192		
	(0.004)				(0.004)			
Perc./actual ability & personality	0.004	2.227	0.005	2.659	0.024	10.421	0.027	11.751
	(0.005)		(0.005)		(0.005)		(0.005)	
B. Labor supply/family planning								
Hours worked	0.018	9.764	0.018	9.709	0.017	7.302	0.019	8.239
	(0.003)		(0.003)		(0.002)		(0.003)	
Children	0.005	2.694	0.007	3.766	-0.002	-0.747	0.003	1.354
	(0.005)		(0.005)		(0.004)		(0.004)	
C. Negotiation styles								
Boldness	0.025	13.838	0.026	14.477	0.022	9.433	0.024	10.379
	(0.002)		(0.002)		(0.002)		(0.002)	
D. Student jobs								
Wage in student jobs	0.001	0.538	0.001	0.574	0.000	0.050	0.000	0.072
	(0.000)		(0.001)		(0.000)		(0.000)	
Observations	10788		10788		9146		9146	

Notes: This table decomposes the differences in log expected starting or lifetime wages into components attributable to (A) sorting into majors, occupations, and industries as well as perceived ability, personality and economic preferences (perceived ability on the job and in university, IQ, Big Five personality traits, altruism, impatience, positive and negative reciprocity, risk aversion and trust), (B) labor supply and family planning (expected hours per week, expected number of children, months at home with children, indicator for early parenthood), (C) negotiation styles (as defined in section 4.2.4), and (D) past wages in student jobs using Oaxaca-Blinder decompositions. For each decomposition, we also present the share of the difference that is attributable to the respective component and present results with and without controls for sorting into occupation and industries. Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level.

Table 4.A.7. Oaxaca-Blinder decomposition of the gender gap in wage expectations for students who want to enter the private sector

	log(Expected starting wage)				log(Expected lifetime wage)			
	(1) with occ. sorting		(2) without occ. sorting		(3) with occ. sorting		(4) without occ. sorting	
Unadjusted difference	0.186	100.000	0.186	100.000	0.252	100.000	0.252	100.000
	(0.012)		(0.012)		(0.012)		(0.012)	
Explained difference	0.143	77.001	0.114	61.436	0.158	62.583	0.130	51.349
	(0.011)		(0.010)		(0.010)		(0.010)	
Composition effects attributable to								
A. Sorting								
Major	0.026	14.139	0.051	27.425	0.022	8.816	0.048	19.212
	(0.006)		(0.004)		(0.005)		(0.004)	
Occupation	0.034	18.483			0.046	18.169		
	(0.007)				(0.007)			
Industry	0.023	12.096			0.020	7.966		
	(0.005)				(0.005)			
Perc./actual ability	0.009	5.082	0.011	5.663	0.029	11.639	0.034	13.414
& personality	(0.006)		(0.006)		(0.006)		(0.006)	
B. Labor supply/family planning								
Hours worked	0.019	10.271	0.019	10.042	0.016	6.508	0.018	7.312
	(0.004)		(0.004)		(0.003)		(0.003)	
Children	0.008	4.515	0.010	5.206	0.003	1.071	0.005	2.123
	(0.005)		(0.005)		(0.004)		(0.005)	
C. Negotiation styles								
Boldness	0.023	12.416	0.024	13.099	0.021	8.414	0.023	9.289
	(0.002)		(0.003)		(0.002)		(0.003)	
Observations	8340		8340		7079		7079	

Notes: This table decomposes the differences in log expected starting or lifetime wages into components attributable to (A) sorting into majors, occupations, and industries as well as perceived ability, personality and economic preferences (perceived ability on the job and in university, IQ, Big Five personality traits, altruism, impatience, positive and negative reciprocity, risk aversion and trust), (B) labor supply and family planning (expected hours per week, expected number of children, months at home with children, indicator for early parenthood), and (C) negotiation styles (as defined in section 4.2.4) for individuals who want to enter the public sector (i.e., excluding those who aim for the public sector) using Oaxaca-Blinder decompositions. For each decomposition, we also present the share of the difference that is attributable to the respective component and present results with and without controls for sorting into occupation and industries. Robust standard errors in parentheses. Log gross annual wages are winsorized at the 1% and 99% level.

Table 4.A.8. Quantile decomposition

	Quantiles										OB	
	10th		25th		50th		75th		90th		Mean	
Unadjusted difference	0.225	100.000	0.208	100.000	0.225	100.000	0.122	100.000	0.086	100.000	0.181	100.000
	(0.016)		(0.010)		(0.008)		(0.009)		(0.014)		(0.010)	
Difference explained	0.183	81.465	0.124	59.396	0.119	52.743	0.085	69.452	0.058	67.731	0.129	71.104
	(0.013)		(0.008)		(0.007)		(0.007)		(0.011)		(0.009)	
Composition effects attributable to												
A. Sorting												
Major	0.023	10.302	0.021	10.208	0.026	11.564	0.017	13.810	0.011	12.933	0.024	13.143
	(0.007)		(0.005)		(0.005)		(0.005)		(0.006)		(0.005)	
Occupation	0.046	20.285	0.028	13.556	0.028	12.279	0.018	14.513	0.017	19.616	0.029	16.165
	(0.008)		(0.005)		(0.005)		(0.005)		(0.008)		(0.006)	
Industry	0.032	14.308	0.026	12.611	0.024	10.833	0.022	17.894	0.012	13.607	0.023	12.708
	(0.006)		(0.004)		(0.004)		(0.004)		(0.006)		(0.004)	
Perc./actual ability & personality	0.004	1.951	0.007	3.357	0.013	5.617	0.016	13.492	0.029	33.793	0.004	2.432
	(0.008)		(0.004)		(0.004)		(0.004)		(0.007)		(0.005)	
B. Labor supply/family planning												
Hours worked	0.014	6.307	0.008	3.824	0.006	2.639	0.005	4.291	0.007	8.044	0.018	9.783
	(0.003)		(0.002)		(0.001)		(0.001)		(0.002)		(0.003)	
Children	0.009	4.209	-0.001	-0.675	0.002	0.987	0.006	4.975	0.005	5.563	0.005	2.969
	(0.006)		(0.004)		(0.003)		(0.003)		(0.005)		(0.005)	
C. Negotiation styles												
Boldness	0.054	24.102	0.034	16.517	0.020	8.824	0.001	0.477	-0.022	-25.825	0.025	13.904
	(0.004)		(0.003)		(0.002)		(0.001)		(0.003)		(0.002)	

Notes: Quantile decomposition (using unconditional quantile regressions based on Firpo, Fortin, and Lemieux, 2009) of the gender gap in expected starting wages using the same variables as in Table 4.3. The final column presents results from an Oaxaca-Blinder decomposition at the mean for reference. Log gross annual wages are winsorized at the 1% and 99% level.

Table 4.A.9. Quantile decomposition without sorting

	Quantiles										OB	
	10th		25th		50th		75th		90th		Mean	
Unadjusted difference	0.225	100.000	0.208	100.000	0.225	100.000	0.122	100.000	0.086	100.000	0.181	100.000
	(0.016)		(0.010)		(0.008)		(0.009)		(0.014)		(0.010)	
Difference explained	0.147	65.167	0.098	47.336	0.091	40.473	0.065	53.260	0.045	52.244	0.101	55.752
	(0.012)		(0.007)		(0.006)		(0.006)		(0.010)		(0.008)	
Composition effects attributable to												
A. Sorting												
Major	0.060	26.634	0.047	22.460	0.046	20.542	0.031	25.838	0.020	23.207	0.044	24.535
	(0.005)		(0.003)		(0.003)		(0.003)		(0.004)		(0.004)	
Perc./actual ability & personality	0.005	2.433	0.008	4.034	0.014	6.259	0.017	14.147	0.030	34.730	0.005	2.899
	(0.008)		(0.005)		(0.004)		(0.004)		(0.007)		(0.005)	
B. Labor supply/family planning												
Hours worked	0.014	6.238	0.008	3.817	0.006	2.688	0.006	4.748	0.008	9.024	0.018	9.714
	(0.003)		(0.002)		(0.001)		(0.001)		(0.002)		(0.003)	
Children	0.011	5.091	0.000	0.006	0.004	1.758	0.009	7.072	0.008	8.969	0.007	4.065
	(0.006)		(0.004)		(0.003)		(0.003)		(0.005)		(0.005)	
C. Negotiation styles												
Boldness	0.056	24.770	0.035	17.020	0.021	9.226	0.002	1.455	-0.020	-23.685	0.026	14.539
	(0.004)		(0.003)		(0.002)		(0.001)		(0.003)		(0.002)	

Notes: Quantile decomposition (using unconditional quantile regressions based on Firpo, Fortin, and Lemieux, 2009) of the gender gap in expected starting wages using the same variables as in Table 4.3 without controls for sorting into occupations and industries. The last column presents results from an Oaxaca-Blinder decomposition at the mean for reference. Log gross annual wages are winsorized at the 1% and 99% level.

Appendix 4.B Expected Wage Gaps by Major and Occupation

The gender gap in wage expectations prevails within majors. To determine the respective gaps, we aggregate all majors into five categories (Medicine and health sciences, STEM, Law, Economics and business studies, humanities and social sciences) and present expected overall wages in Table 4.B.1. While there exists substantial heterogeneity in levels across majors female students expect to earn less than their male counterparts within each of the respective study fields. This holds both for starting wages and over the life cycle. However, the gender gap is slightly lower in fields that are traditionally chosen by females than in male-dominated subjects. Thus females on average expect to earn only 84% of the average male starting wage in legal studies, as compared to 93% in humanities. At the age of 55, the respective shares decrease to 72–80%.

Table 4.B.1. Descriptive statistics of gross annual expected wages by major

	Med./Health Sci.				**STEM**			
	Males	Females	Ratio	N	Males	Females	Ratio	N
Starting	38860	34282	0.88	1313	40620	35472	0.87	5234
Age 40	59589	49800	0.84	1313	58214	47314	0.81	5234
Age 55	70977	56474	0.80	1313	69692	52657	0.76	5234
	Law				**Econ./Business**			
	Males	Females	Ratio	N	Males	Females	Ratio	N
Starting	48511	40670	0.84	676	40352	36345	0.90	3427
Age 40	76524	60519	0.79	676	66612	52688	0.79	3427
Age 55	96180	69487	0.72	676	82717	60698	0.73	3427
	Human./Soc. Sci.				**All subjects**			
	Males	Females	Ratio	N	Males	Females	Ratio	N
Starting	31808	29480	0.93	4698	39076	33434	0.86	15348
Age 40	44822	38009	0.85	4698	58301	45765	0.78	15348
Age 55	53151	41489	0.78	4698	70518	51291	0.73	15348

Notes: This table shows average expected starting wages as well as expected wages at the age of 40 and 55 for males and females for majors aggregated into five categories. All wages are winsorized at the 1% and 99% level.

Additionally, Table 4.B.2 presents the gender gap in wage expectations for different occupations. Goldin (2014) suggests that occupations for which earnings are a nonlinear/convex in working hours have larger gender gaps than those with fairly flat/linear relationships. Indeed, we observe the gender gap in wage expectations for occupations with nonlinear hours-earnings profiles (e.g. lawyers) to be larger than for, e.g., teachers, who tend to have very flat hours-earnings profiles.[23] Along these same lines the gap tends to be smallest for authors and journalists, who might even have decreasing hours/earnings profiles due to decreasing marginal productivity. Students thus correctly anticipate that flatter hours-earnings profiles are associated with lower earning gaps.

Table 4.B.2. Gender gap in wage expectations by occupations

	Gender gap by occupation				
	Journalists & authors	Teachers	Engineering professionals	Medical doctors	Lawyers
Gap in EUR	-1423	-1792	-3578	-6630	-9824
Gap in log-points	-0.071	-0.130	-0.123	-0.122	-0.225
Gap in ranks	-5.6	-9.7	-12.6	-13.0	-14.1
Observations	729	1141	1470	464	433

Notes: This table presents the gender gap in wage expectations measured in Euro, log-points and ranks for different occupations. Each coefficient corresponds stems from a regression of expected wages, log expected wages or ranks in the male expected wage distribution on an indicator for females. All wages are winsorized at the 1% and 99% level.

23. The table does not include results for pharmacists, as we cannot distinguish individuals planning to work in pharmacies from those planning to work in the pharmaceutical industry.

Appendix 4.C A Regression-based Comparison of Expected and Actual Wages

We formally compare expected and actual wages by pooling them in a single regression on an indicator for being an actual graduate. Table 4.C.1 reveals that in terms of raw wages, graduates earn 11.2 percentage points lower wages when compared to the expected wages of students.[24] Nonetheless, once we control for gender, sorting patterns and hours worked the difference vanishes. In fact, this difference is entirely driven by differences in hours worked as some graduates start working part-time after finishing their studies and thus earn lower wages than graduates in full-time jobs. Similar to what has been found in the literature (e.g., Webbink and Hartog, 2004; Wiswall and Zafar, 2018a), the wage expectations of students elicited in our survey thus tracks the distribution of realized earnings very well once we account for hours worked. This suggests that the gender gap in expected wages likely translates into differences in realized wages.

Table 4.C.1. Comparison of expected and actual log wages

	log wages (pooled)	
	(1)	(2)
Actual graduate	-0.112	0.016
	(0.022)	(0.025)
Gender, major, occupation, industry, labor supply	No	Yes
R^2 (adj.)	.0022	.13
Observations	16501	16400

Notes: Robust standard errors in parentheses. Sample pools over log gross annual wages of both current students using expected wages and actual graduates with realized wages. All wages are winsorized at the 1% and 99% level.

24. Note that we do not observe a difference in mean actual and expected wages but in log wages, given that taking the logarithm gives more weight on the lower end of the wage distribution. As can be seen from Figure 4.4, this is where the differences between actual and expected wages are more pronounced.

References

Albrecht, James W., Per-Anders Edin, Marianne Sundström, and Susan B. Vroman. (1999). “Career interruptions and subsequent earnings: A reexamination using Swedish data.” *Journal of Human Resources* 34 (2): 294–311. [178]

Albrecht, James, Anders Björklund, and Susan Vroman. (2003). “Is There a Glass Ceiling in Sweden?” *Journal of Labor Economics* 21 (1): 145–177. [171]

Altonji, Joseph G., and Charles R. Pierret. (2001). “Employer learning and statistical discrimination.” *Quarterly Journal of Economics* 116 (1): 313–350. [186]

Angelov, Nikolay, Per Johansson, and Erica Lindahl. (2016). “Parenthood and the gender gap in pay.” *Journal of Labor Economics* 34 (3): 545–579. [178]

Antonczyk, Dirk, Bernd Fitzenberger, and Katrin Sommerfeld. (2010). “Rising wage inequality, the decline of collective bargaining, and the gender wage gap.” *Labour Economics* 17 (5): 835–847. [171]

Arcidiacono, Peter, V. Joseph Hotz, and Songman Kang. (2012). “Modeling college major choices using elicited measures of expectations and counterfactuals.” *Journal of Econometrics* 166 (1): 3–16. [161, 163, 186]

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn. (2018). “Negotiating a Better Future: How Interpersonal Skills Facilitate Inter-Generational Investment.” Working Paper. [190]

Attanasio, Orazio P., and Katja M. Kaufmann. (2014). “Education choices and returns to schooling: Mothers’ and youths’ subjective expectations and their role by gender.” *Journal of Development Economics* 109: 203–216. [175]

Babcock, Linda, and Sara Laschever. (2009). *Women don’t ask: Negotiation and the gender divide*. Princeton University Press. [180, 185]

Baker, Rachel, Eric Bettinger, Brian Jacob, and Ioana Marinescu. (2018). “The Effect of Labor Market Information on Community College Students’ Major Choice.” *Economics of Education Review* 65: 18–30. [163]

Bayer, Patrick, and Kerwin K. Charles. (2018). “Divergent Paths: A New Perspective on Earnings Differences Between Black and White Men Since 1940.” *Quarterly Journal of Economics* 133 (3): 1459–1501. [166, 171]

Bertrand, Marianne. (2011). “New perspectives on gender.” In *Handbook of Labor Economics*. Edited by O. Ashenfelter and D. Card. Vol. 4, Elsevier. Chapter 17, 1543–1590. [164]

Bertrand, Marianne, Claudia Goldin, and Lawrence F. Katz. (2010). “Dynamics of the gender gap for young professionals in the financial and corporate sectors.” *American Economic Journal: Applied Economics* 2 (3): 228–55. [160, 171, 177]

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. (2017). “Gender stereotypes about intellectual ability emerge early and influence children’s interests.” *Science* 355 (6323): 389–391. [164]

Blau, Francine D., and Marianne A. Ferber. (1991). “Career plans and expectations of young women and men: The earnings gap and labor force participation.” *Journal of Human Resources* 26 (4): 581–607. [159, 178, 189]

Blau, Francine D., and Lawrence M. Kahn. (2017). “The gender wage gap: Extent, trends, and explanations.” *Journal of Economic Literature* 55 (3): 789–865. [159, 160, 185, 189, 190]

Borghans, Lex, James J. Heckman, Bart H.H. Golsteyn, and Huub Meijers. (2009). "Gender differences in risk aversion and ambiguity aversion." *Journal of the European Economic Association* 7 (2-3): 649–658. [164]

Bowles, Hannah R., Linda Babcock, and Lei Lai. (2007). "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask." *Organizational Behavior and Human Decision Processes* 103: 84–103. [185]

Bratti, Massimiliano. (2015). "Fertility postponement and labor market outcomes." *IZA World of Labor*, [190]

Breen, Richard, and Cecilia Garcia-Penalosa. (2002). "Bayesian learning and gender segregation." *Journal of Labor Economics* 20 (4): 899–922. [190]

Brunello, Giorgio, Claudio Lucifora, and Rudolf Winter-Ebmer. (2004). "The wage expectations of European business and economics students." *Journal of Human Resources* 39 (4): 1116–1142. [159, 189]

Bütikofer, Aline, Sissel Jensen, and Kjell G. Salvanes. (2018). "What Explains the Gender Gap among Top Earners." *European Economic Review* 109: 103–123. [160]

Caliendo, Marco, Wang-Sheng Lee, and Robert Mahlstedt. (2017). "The gender wage gap and the role of reservation wages: New evidence for unemployed workers." *Journal of Economic Behavior & Organization* 136: 161–173. [182]

Cortes, Patricia, and Jessica Pan. (2018). "Occupation and gender." *Oxford Handbook of Women and the Economy*, 425. [185]

Croson, Rachel, and Uri Gneezy. (2009). "Gender Differences in Preferences." *Journal of Economic Literature* 47 (2): 448–474. [164]

Daniel, Fernandez-Kranz, Aitor Lacuesta, and Nuria Rodriguez-Planas. (2013). "The motherhood earnings dip: Evidence from administrative records." *Journal of Human Resources* 48 (1): 169–197. [160, 177]

Destatis. (2013). "Geburtentrends und Familiensituation in Deutschland 2012." Federal Statistical Office Germany. [180]

Destatis. (2014). "Auf dem Weg zur Gleichstellung?" Federal Statistical Office Germany. [160]

Destatis. (2017a). "Kinderlosigkeit, Geburten und Familien. Ergebnisse des Mikrozensus 2016." Federal Statistical Office Germany. [180]

Destatis. (2017b). "Verdienste auf einen Blick." Federal Statistical Office Germany. [160, 176]

Dominitz, Jeff, and Charles F. Manski. (1997). "Using expectations data to study subjective income expectations." *Journal of the American Statistical Association* 92 (439): 855–867. [161]

Eurostat. (2018). "Employment rates of recent graduates." http://ec.europa.eu/eurostat/statistics-explained/index.php/Employment_rates_of_recent_graduates. accessed: August 14, 2018. [164]

Exley, Christine L., Muriel Niederle, and Lise Vesterlund. (Forthcoming). "Knowing When to Ask: The Cost of Leaning-in." *Journal of Political Economy*, [185, 190]

Fabian, Gregor, Torsten Rehn, Gesche Brandt, and Kolja Briedis. (2013). *Karriere mit Hochschulabschluss*. HIS: Forum Hochschule. [180]

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. (2018). "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences." Working Paper. [165]

Filippin, Antonio, and Andrea Ichino. (2005). “Gender wage gap in expectations and realizations.” *Labour Economics* 12 (1): 125–145. [175]

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. (2009). “Unconditional quantile regressions.” *Econometrica* 77 (3): 953–973. [202, 203]

Fouarge, Didier, Ben Kriechel, and Thomas Dohmen. (2014). “Occupational sorting of school graduates: The role of economic preferences.” *Journal of Economic Behavior & Organization* 106: 335–351. [185]

Francesconi, Marco, and Matthias Parey. (2018). “Early gender gaps among university graduates.” *European Economic Review* 109: 63–82. [160, 166, 171, 176, 185]

Fuster, Andreas, David Laibson, and Brock Mendel. (2010). “Natural expectations and macroeconomic fluctuations.” *Journal of Economic Perspectives* 24 (4): 67–84. [190]

Goldberg, Lewis R., John A. Johnson, Herbert W. Eber, Robert Hogan, Michael C. Ashton, C. Robert Cloninger, and Harrison G. Gough. (2006). “The international personality item pool and the future of public-domain personality measures.” *Journal of Research in Personality* 40 (1): 84–96. [165]

Goldin, Claudia. (2014). “A grand gender convergence: Its last chapter.” *American Economic Review* 104 (4): 1091–1119. [169, 178, 185, 204]

Goldin, Claudia, and Lawrence F. Katz. (2016). “A most egalitarian profession: pharmacy and the evolution of a family-friendly occupation.” *Journal of Labor Economics* 34 (3): 705–746. [160, 177, 178]

Hall, Robert E., and Alan B. Krueger. (2012). “Evidence on the Incidence of Wage Posting, Wage Bargaining, and On-the-Job Search.” *American Economic Journal: Macroeconomics* 4 (4): 56–67. [164]

Heineck, Guido, and Silke Anger. (2010). “The returns to cognitive abilities and personality traits in Germany.” *Labour Economics* 17 (3): 535–546. [186]

Jensen, Robert. (2010). “The (perceived) returns to education and the demand for schooling.” *Quarterly Journal of Economics* 125 (2): 515–548. [161]

Kaufmann, Katja M. (2014). “Understanding the income gradient in college attendance in Mexico: The role of heterogeneity in expected returns.” *Quantitative Economics* 5 (3): 583–630. [161]

Kirkeboen, Lars J., Edwin Leuven, and Magne Mogstad. (2016). “Field of Study, Earnings, and Self-Selection.” *Quarterly Journal of Economics* 131 (3): 1057–1111. [168]

Kleven, Henrik, Camille Landais, and Jakob Egholt Sogaard. (Forthcoming). “Children and gender inequality: Evidence from Denmark.” *American Economic Journal: Applied Economics*, [160, 177]

Kuchler, Theresa, and Basit Zafar. (Forthcoming). “Personal Experiences and Expectations about Aggregate Outcomes.” *Journal of Finance*, [187, 190]

Kunze, Astrid. (2018). “The Gender Wage Gap in Developed Countries.” *Oxford Handbook of Women and the Economy*, 369. [159]

Kuziemko, Ilyana, Jessica Pan, Jenny Shen, and Ebonya Washington. (2018). “The Mommy Effect: Do Women Anticipate the Employment Effects of Motherhood?” Working Paper. [160, 180, 186, 190]

Leibbrandt, Andreas, and John A. List. (2015). “Do Women Avoid Salary Negotiations? Evidence from a Large-Scale Natural Field Experiment.” *Management Science* 61 (9): 2016–2024. [185]

Malmendier, Ulrike, and Stefan Nagel. (2011). "Depression babies: do macroeconomic experiences affect risk taking?" *Quarterly Journal of Economics* 126 (1): 373–416. [187, 190]

Malmendier, Ulrike, and Stefan Nagel. (2015). "Learning from inflation experiences." *Quarterly Journal of Economics* 131 (1): 53–87. [187, 190]

Manski, Charles F. (2004). "Measuring expectations." *Econometrica* 72 (5): 1329–1376. [161]

OECD. (2015). "OECD family database: Gender pay gaps for full-time workers and earnings differentials by educational attainment." [159]

Piopiunik, Marc, Franziska Kugler, and Ludger Wößmann. (2017). "Einkommenserträge von Bildungsabschlüssen im Lebensverlauf: Aktuelle Berechnungen für Deutschland." *ifo Schnelldienst* 70 (7): 19–30. [163, 168]

Raven, John C., and John H. Court. (1998). *Raven's progressive matrices and vocabulary scales*. Oxford Psychologists Press. [165]

Reuben, Ernesto, Matthew Wiswall, and Basit Zafar. (2017). "Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender." *Economic Journal* 127 (604): 2153–2186. [159, 161, 186, 190]

Rigdon, Mary L. (2012). "An Experimental Investigation of Gender Differences in Wage Negotiations." Working Paper. [185]

Schmitt, David P, Anu Realo, Martin Voracek, and Jüri Allik. (2008). "Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures." *Journal of Personality and Social Psychology* 94 (1): 168. [164]

Schweri, Jürg, and Joop Hartog. (2017). "Do Wage Expectations Influence the Decision to Enroll in Nursing College?" *Journal of Economic Behavior & Organization* 141: 135–150. [175]

Seegers, Philipp, Jan Bergerhoff, Stephan Hartmann, and Anne Knappe. (2016). *Fachkraft 2020: 5. und 6. Erhebung zur wirtschaftlichen und allgemeinen Lebenssituation der Studierenden in Deutschland*. Studitemps, and Maastricht University. [162]

Small, Deborah A., Michele Gelfand, Linda Babcock, and Hilary Gettman. (2007). "Who goes to the bargaining table? The influence of gender and framing on the initiation of negotiation." *Journal of Personality and Social Psychology* 93 (4): 600. [180]

Stinebrickner, Ralph, and Todd R. Stinebrickner. (2014). "A Major in Science? Initial Beliefs and Final Outcomes for College Major and Dropout." *Review of Economic Studies* 81 (1): 426–472. [161]

Stinebrickner, Todd R., Ralph Stinebrickner, and Paul J. Sullivan. (2019). "Job Tasks, College Major, and the Gender Wage Gap." Working Paper. [164]

Webbink, Dinand, and Joop Hartog. (2004). "Can students predict starting salaries? Yes!" *Economics of Education Review* 23 (2): 103–113. [163, 175, 206]

Wiswall, Matthew, and Basit Zafar. (2015). "Determinants of College Major Choice: Identification using an Information Experiment." *Review of Economic Studies* 82 (2): 791–824. [186]

Wiswall, Matthew, and Basit Zafar. (2018a). "Human Capital Investments and Expectations about Career and Family." Working Paper. [175, 206]

Wiswall, Matthew, and Basit Zafar. (2018b). "Preference for the Workplace, Investment in Human Capital, and Gender." *Quarterly Journal of Economics* 133 (1): 457–507. [185, 190]

Xia, Xiaoyu. (2016). "Forming wage expectations through learning: Evidence from college major choices." *Journal of Economic Behavior & Organization* 132: 176–196. [190]

Zafar, Basit. (2011). "Can subjective expectations data be used in choice models? Evidence on cognitive biases." *Journal of Applied Econometrics* 26 (3): 520–544. [161]

Zafar, Basit. (2013). "College major choice and the gender gap." *Journal of Human Resources* 48 (3): 545–595. [185, 186, 189]

Zambre, Vaishali. (2018). "The Gender Gap in Wage Expectations: Do Young Women Trade off Higher Wages for Lower Wage Risk?" Working Paper. [161]