Essays in Behavioral and Experimental Economics

Inaugural-Dissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften

durch

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

vorgelegt von

Sebastian Schaube

aus Berlin

Bonn, 2019

Dekan:	Prof. Dr. Daniel Zimmer, LL.M.
Erstreferent:	Prof. Dr. Sebastian Kube
Zweitreferent:	Prof. Dr. Armin Falk
Tag der mündlichen Prüfung:	19. Dezember 2018

Acknowledgments

This thesis would not exist without the support and encouragement I received from many people. Most importantly, Sebastian Kube fostered my interest in behavioral economics. I was lucky to have him as my supervisor and as a co-author. His unique perspective and uncompromising feedback on research ideas, experimental designs, writing papers, and teaching have been invaluable. I also want to thank Armin Falk for his feedback and support, and Matthias Kräkel who always has an open ear and door and whose advice I greatly appreciate.

Lukas Kießling and Jonas Radbruch were the perfect co-authors and friends for brutal discussions, for spending countless hours in school gyms, and for making research joyful again. For advice in various matters of life, I thank my office-mates Jonas Sobott and Jana Wilrodt. Special thanks go to Daniel Becker, Tobit Gamp, and Benjamin Hartung for their friendship and for all the fun we had over the last years. I benefited greatly from the administrative and generous financial support of the Bonn Graduate School of Economics and the Institute for Applied Microeconomics.

I am thankful to my family for encouraging my curiosity in the social sciences and for their lifelong support in general. Last, I cannot stress enough how crucial Eva was for everything included in this thesis. She always reminds me of the truly important things in life and helps me to get back on track whenever I feel lost.

Contents

List of Figures

List of Tables

1

Introduction[⋆]

Few economic decisions are ever made in a social vacuum. Many of us look constantly to others for guidance or for comparison. Our actions might benefit or hurt others, so we use said actions strategically to reward or to punish. We are appalled if our sense of distributional justice gets wronged and refuse to cooperate with those who offended it. Over the past decades the emergence of behavioral economics has put these social aspects of decision making into sharp focus. Behavioral economics has advanced our collective understanding of social preferences and comparisons by incorporating insights from neighboring disciplines such as psychology or political science. A key ingredient for this evolution lies in the controlled examination of human behavior – frequently using laboratory or field experiments – to uncover its underlying mechanisms. The causal identification of these hidden motives in turn gave rise to new theoretical models that include reciprocity or preferences for status and fair outcomes. This thesis consists of five essays that each contribute to different facets of the literature on the behavioral impact of social comparisons and social preferences. All of them employ laboratory or field experiments and combine them with theoretical frameworks that incorporate different kinds of social motives.

In Chapter 2 (joint work with Sebastian Kube, Hannah Schildberg-Hörisch and Elina Khachatryan), I study how social preferences can hamper the adoption of efficient legislation. Institutions and their endogenous adaption are increasingly thought to facilitate cooperation and to mitigate the free-rider problem inherent in the provision of public goods. In this paper, we test within a unified framework how the process of institution formation is affected by three key aspects of natural environments: i) heterogeneity among players in the benefits of cooperation, ii) (a)symmetry in players' institutional obligations, and iii) potential trade-offs between efficiency and equality in payoff allocations. We observe social preferences to be limiting the scope for institution formation. Inequality-

[⋆] I would like to thank Holger Gerhardt for outstanding TeXnical assistance and numerous helpful comments.

averse players frequently object to institutions that fail to address differences in players' benefits from cooperation – even if rejecting the institution causes monetary losses to all players.

Building on these results, Chapter 3 (joint work with Sebastian Kube) investigates the interplay between cooperation and redistribution if individuals profit heterogenously from public goods. We analyze such situations, both from a theoretical and empirical perspective, to explore to what extent formal redistribution can alleviate the implementation problem. We find that the answer to this question depends on whether redistribution and governance i) form two distinct institutions to be decided upon separately or ii) are "bundled" into a single institution. Implementation and cooperation rates are higher in the latter case, where parties decide over the joint implementation of the combined institution. By contrast, coordination problems arise if redistribution is available separately from, rather than being an integral component of, the governing institution; resulting in significantly lower cooperation rates compared to the bundled case.

In Chapter 4 I study how social preferences can be leveraged to obtain valuable information about employees. Peer evaluations are frequently used if an employee's performance is hard or even impossible to observe by a principal. If, however, the evaluating peer is in direct competition with the evaluated peer for bonuses or promotions, incentives for truthful reporting might be reduced. I explore to what extent team-incentives – in contrast to fixed wages – can encourage truthful evaluations. I use a laboratory experiment that combines a real effort task and a rank-order tournament, where the prizes are distributed on basis of the mutual evaluations of the tournament participants. I find that the outcome depends on whether the participants i) can evaluate their peers individually or ii) have to rank them. Team-incentives have significant positive impact on the sincerity of peer evaluations only in the latter case. Without team-incentives and forced rankings the evaluation behavior carries little to no meaning.

In Chapter 5 and Chapter 6 (joint work with Lukas Kießling and Jonas Radbruch) I investigate whom individuals choose as peer and how these self-selected peers impact individual performance. While the influence of peers on our consumption behavior, general well-being, and individual performance on the job or in school is widely accepted, we know relatively little about how these peers arise in the first place. Frequently they are not randomly selected, but might be carefully chosen. We conduct a field experiment in physical education classes at secondary schools. Students participate in a running task twice: first, the students run alone, then with a peer. Before the second run, we elicit preferences for peers. We experimentally vary the matching in the second run and form pairs either randomly or based on elicited preferences. We find that students on average prefer peers of higher ability, but these preferences vary with their personality traits. Higher competitiveness, lower extraversion, and an internal locus of control are associated with preferences for superior peers. Taking social network information into account, we find homophily in agreeableness and attitudes for social comparisons even conditional on friendship ties. These self-selected peers improve individual performance by .14-.15 SD relative to randomly assigned peers. While self-selection leads to more social ties and lower performance differences within pairs, this altered peer composition does not explain performance improvements. Rather, we provide evidence that self-selection has a direct effect on performance and provide several markers that the social interaction has changed. Regarding the design of peer assignment mechanisms, our results also highlight the importance of accounting for the multidimensionality of peer preferences. In summary, this thesis documents effects on individual behavior that are not predicted by standard economic theory, but underscore the relevance of social preferences for our understanding of economic decision making and behavior, and the design of efficient institutions.

2

Institution Formation and Cooperation with Heterogeneous Agents *

2.1 Introduction

> "[...] a set of rules used in one physical environment may have vastly different consequences if used in a different physical environment."
>
> *(Ostrom, 1990, p.22)*[1]

Cooperation problems are ubiquitous in many areas in economics, ranging from teamwork or hold-up problems in managerial economics, over community governance or property rights security in development economics, natural resource management or climate protection in environmental economics, trade obstacles or treaty formation in international economics, to tax compliance and the provision of public goods in public economics. Each example certainly has its own distinctive issues, but when it comes to mitigating the underlying cooperation problems, there is usually a common approach: the modification of individuals' incentive-compatibility constraints, such that "free-riding" is no longer the dominant strategy (e.g., Shavell and Polinsky, 2000). These modifications (implicitly or explicitly) impose restrictions on individuals' choice sets, which raises the question whether they will be implemented in the first place (e.g., Gürerk et al., 2006; Tyran and Feld, 2006; Kosfeld et al., 2009; Bierbrauer and Hellwig, 2011; Markussen et al., 2014). In the present paper, we will shed light on this central question – asking in particular to which extent i) the heterogeneity

* Financial support from the German Research Foundation (DfG) through SFB-TR 15 is gratefully acknowledged.

[1] Reported, inter alia, in Decker et al. (2003).

of the involved players and ii) the (a)symmetry of the restrictions affects their implementation.

Consider the following example that we use throughout the paper, namely the provision of a public good. If members of a society are perfectly identical and all benefit equally from overcoming this social dilemma, one might expect them to mutually agree on establishing an institution that eliminates the social dilemma.[2] However, controversies might arise when members are heterogeneous and have different stakes in overcoming the social dilemma. In particular when equality considerations are taken into account, the exact content of the institution is key to successful implementation. Symmetric institutions, in which all members have the same obligations, might be rejected in favor of asymmetric institutions with member-specific obligations – even if this implies monetary losses for all members.

To causally identify how institution formation is affected by selected aspects of natural environments, we conducted a series of laboratory experiments. The basic underlying game, a public-good game, is a prominent workhorse for studying cooperation problems. Each player receives an endowment and has to decide on its allocation between private consumption and contributions to a public good. Provision of the public good creates benefits for all group members and is socially efficient in terms of the sum of monetary payoffs.[3] However, the individual marginal return from the public good is below the marginal return from private consumption, such that free-riding incentives exist which jeopardize public good provision. To offer players the opportunity to endogenously mitigate the cooperation problem, we add an additional stage that is played prior to the public good game. At this first stage, players decide on implementing an institution using unanimity voting. If all players in the group vote in favor of the institution, they are committed to certain efficiency-enhancing contribution levels in the subsequent public good game.[4] If at least one player votes against the implementation of the institution, the regular public good game is played and each player can freely decide how much to contribute in the second stage.

Players in our setup thus start in the absence of institutions and subsequently decide on the implementation of a joint institution to foster cooperation. In such

[2] Of course, expected benefits must exceed the costs of implementing the institution. Throughout the paper, we take this for granted by assuming the institution to be costless – notwithstanding that the case of positive costs would be interesting to study (e.g., Kamei et al., 2015).

[3] Throughout the paper, efficiency refers to monetary payoffs.

[4] One could also think of the institution as consisting of two elements: i) It states a certain obligation for each player, i.e., the exact amount that he is required to contribute in the second stage, and ii) it installs a deterrent sanctioning technology, i.e., players' contributions are monitored and a player receives harsh punishment when deviating from the required contribution. For reasons of simplicity, the second component is not an explicit part of the experiment. Instead, it is implicitly modeled by restricting a player's choice set in the second stage to the required contribution (see Kosfeld et al., 2009; Gerber et al., 2013, for similar approaches).

an initial, lawless state of nature that is characterized by sovereign players facing a social dilemma, it seems natural to use unanimity voting for deciding on the implementation of institutions.[5] In fact, unanimous decision-making is the easiest possible, if not the only, voting procedure that players do not have to explicitly agree upon prior to voting. It does not require players to give up sovereignty, since each player can veto any decision. This is different for non-unanimous voting rules, such as majority voting, where players need to forfeit part of their sovereignty and which therefore typically only emerges after a joint history of cooperation.[6]

Since our focus is on how institution formation is affected i) by heterogeneity in players' benefits from cooperation, and ii) by the (a)symmetry of obligations, we vary these factors in a controlled manner while fixing the decision rule to unanimity voting in all treatments. First, in some treatment conditions (*Homogeneous types*), all players are of the same type and, thus, receive the same benefits from the public good, while in other conditions (*Heterogeneous*), there are two types that differ in their marginal benefits. Second, we vary the content of the institution. All players are either obliged to contribute their entire endowment to the public good (*Symmetric Institution*), or obligations differ between the two player types (*Asymmetric*). While the symmetric institution implies efficient public good provision, but inequality in payoffs for heterogeneous players, obligations in the asymmetric case are chosen such that final payoffs are equalized. This setup allows us to clearly identify the roles of inequality aversion and efficiency concerns in the process of institution formation.

We find that inequality considerations can hamper the formation of efficient institutions meant to foster cooperation. With heterogeneous player types, those with low marginal benefits frequently object to the symmetric institution (about 40% reject it). The same is observed for homogeneous player types with asymmetric institutions (about 45% reject it). On the other hand, support is high when the institution implements equal payoff allocations: the asymmetric institution seems perfectly acceptable for heterogeneous player types, as does the symmetric institution for homogeneous types. In both cases, more than 90% of all votes are in favor of the implementation.

With respect to the sum of monetary payoffs, we observe that efficiency is always lower when institution formation failed than when the institution was implemented. The symmetric institution for homogeneous player types performs

[5] The idea of an initial state of nature that is characterized by sovereign agents in a lawless environment goes back to Rousseau (1762) and Hobbes (1651).

[6] Cooperation in past periods may foster trust and reciprocal behavior among players, which may make them willing to forfeit part of their sovereignty. To give just one example, international organizations, most notably the League of Nations as the precursor of what is now the United Nations, used to apply the unanimity voting rule for voting on matters of substance before World War II. It was only during the post-war growth in international coordination through permanent organizations that non-unanimous voting rules were increasingly applied.

best (average efficiency is above 90% of the maximally obtainable sum of payoffs). Compared to this, under heterogeneity both the symmetric and the asymmetric institution lead to lower rates of efficiency, albeit for different reasons. In the former case, average efficiency is lower because the symmetric institution is frequently rejected. In the latter case, heterogeneous player types frequently implement the asymmetric institution, but average efficiency is lower since total obligations and the level of public good provided are lower. The asymmetric institution for homogeneous players performs worst.

The striking differences in average efficiency and implementation rates between treatments underline at least three important issues. First, our results stress that inequality aversion can have a strong impact on the process of institution formation. In most of the existing studies on institution formation, introducing social preferences to the theoretical models usually leads to stronger support for the institution; be it because more players want to be part of a coalition than is predicted under standard preferences (e.g., Kosfeld et al., 2009; McEvoy et al., 2015), or because the institution to be implemented allows them to reduce free-riders' payoffs (e.g., Markussen et al., 2014). By contrast, in those cases where inequality-aversion makes a difference in our setup, inequality-averse players are predicted to be less inclined to support the formation of the institution – a phenomenon that has not been discussed so far in the corresponding literature. As can be seen in our data, this easily leads to situations where players forego monetary payoffs by objecting to efficient institutions; in particular given the requirement of unanimous decisions.

However, and this is the second point we would like to stress, the use of unanimity voting for implementing institutions must not always be detrimental to efficiency. On the contrary, it can even help to foster cooperation.[7] Already Wicksell (1964) discusses that institutions based on unanimity or consensus voting can be ideally suited to overcome the canonical problem of free-riding. Unanimity makes individual activism implicitly conditional on the activism of all other parties involved. This mitigates the dilemma of institution formation: those who agree on implementing an institution do not face the subsequent risk of free-riding by non-supporting, and thus non-participating, players (see also Maggi and Morelli, 2006). Consequently, there is no drawback in supporting institutions that are based on unanimous decisions; either all players participate and the institution is formed, or the institution is not created at all. This can be clearly seen when comparing our data to related studies that implic-

[7] Apart from this, there is also another desirable feature of unanimity. It is easy to agree on a principle of unanimity, since every party has veto power and freedom of choice is thus granted (at least ex ante, before an institution is implemented). Moreover, recent evidence implicitly suggests that many people value unanimous decisions, and that they have a strong preference for involving all players in the decision-making process (see Decker et al., 2003; Sutter et al., 2010; Linardi and McConnell, 2011, and the references therein).

itly allow players to “opt out” of institutions (Kosfeld et al., 2009; Gerber et al., 2013). While efficient and equitable institutions are frequently not implemented in those other studies, we observe that such institutions receive strong support and are implemented most of the times when unanimity is required.

Of course, this is not to say that unanimity will lead to stronger cooperation all the time. The unanimity voting rule grants de facto veto rights to every party involved. Therefore, it is crucial that the institution to be voted on addresses idiosyncratic interests amongst the involved parties. We see this in our study, since homogeneous players frequently reject asymmetric institutions, and heterogeneous players regularly reject symmetric institutions. Support for the latter is also found in lab experiments by Banks et al. (1988), and Kesternich et al. (2014), as well as in the survey evidence reported in Reuben and Riedl (2013). The importance of fixing appropriate institutional obligations beforehand is also reflected in the literature that studies homogeneous players’ acceptance thresholds on minimum contribution requirements in public good games (Birnberg et al., 1970; Dannenberg et al., 2014; Rauchdobler et al., 2010). Taken together, the evidence strongly suggests that prior to the ultimate voting about the implementation of an institution, great care has to be taken ex ante in designing the institution.

The institution at hand is implicitly built around a centralized authority with a deterrent sanctioning technology, but also other institutional mechanisms could be implemented to foster cooperation (e.g., Apesteguia et al., 2013; Falkinger et al., 2000; Andreoni and Gee, 2012). One could even think about implementing decentralized sanctioning regimes. Of course, the seminal papers by Ostrom et al. (1992), Gächter and Fehr (2000), and Fehr and Gächter (2002) started with the basic idea of mutual monitoring and punishment among the members of a group; focusing in particular on the question whether certain behavioral norms can emerge, even in the absence of formal institutions with a centralized structure. Still, there are some studies where players do vote over the implementation of decentralized sanction regimes (Putterman et al., 2011; Markussen et al., 2014; Kamei et al., 2015). Those studies exclusively focus on majority voting and homogeneous agents. It might be interesting to reconsider their results in our setup with heterogeneous players, or to see how behavior would change when using unanimity voting procedures.

Finally, in particular the results of our two benchmark treatments, where homogenous or heterogeneous players face only a regular public goods game and are not given the option to implement an institution, directly add to a broad strand of literature on “asymmetric” public goods. This literature has different approaches to studying the impact of asymmetries on cooperation; most commonly by varying the ratio of costs to benefits between players of the same group (see, e.g., the seminal paper by Fisher et al. (1995), and the recent work by McGinty and Milam (2013)), or by introducing inequalities in players’ endowments, be it explicitly, (e.g., Cherry et al., 2005), or implicitly by using different action sets for different agents, (e.g., Khadjavi et al., 2014). Just like in our data, where contributions are lower with heterogeneous agents but the difference falls short of being significant, the existing empirical evidence is mixed: with some studies reporting lower contributions in the presence of an asymmetry, some reporting higher contributions, and others finding no effect at all on cooperation levels (see, for example, Anderson et al. (2008), p.1014f, for a detailed review of the findings). Maybe most notable for our context are the findings from Riedel and Schildberg-Hörisch (2013). In their setup, asymmetry is implemented by exogenously imposing requirements on minimum contribution levels that differ between the two (otherwise symmetric) players that form a group. These obligations are non-binding and are backed up by non-deterrent sanctions only. Still, the authors observe that non-binding obligations shape contribution behavior, but in contrast to our findings, the effect is only temporary and vanishes over time.

The outline of the paper is as follows. Section 2.2 describes the experiment design. In Section 2.3, behavioral predictions for subjects' behavior will be derived, using both standard and social preferences (inequality aversion). Section 2.4 presents and discusses the empirical results. Section 2.5 concludes.

2.2 Experiment

In natural environments, the complexity of the process of institution formation makes it particularly difficult to draw causal conclusions about the conditions under which institutions come into being. As a starting point, we therefore use the controlled environment of laboratory experiments to study central aspects of the endogenous formation of institutions. In this section, we present the design of our experiment and describe the implemented procedures.

2.2.1 Experimental Design

Our design builds on a standard public goods game (VCM game), a frequently used workhorse to study elements of social dilemmas in the lab (e.g., Isaac and Walker, 1988). Each player has a private endowment $E = 20$. Players simultaneously decide on the amount c_i that they contribute to a public good, with $0 \leq c_i \leq E, i = 1, ..., n$. The benefits from the public good are enjoyed by all players, independent of their individual contribution c_i. In some treatments, players are heterogeneous, i.e., not all players benefit from the public good to the same extent. To model heterogeneity, we allow the marginal per capita return (MPCR)

γ_i from the public good to vary across players.[8] Given the contributions of all players $(c_1, ..., c_n)$, player i's material payoff π_i is thus given by

$$\pi_i = E - c_i + \gamma_i \textstyle\sum_{i=1}^{n} c_i.$$

In all treatments, parameters for γ_i are chosen such that players face a social dilemma. Efficiency, defined as the sum of payoffs of all players, is maximized if all players contribute their entire endowment. Yet, from an individual perspective, each player's material payoff is maximized by not contributing to the public good, regardless of the other players' contributions. Formally, this implies $\sum_{i=1}^{n} \gamma_i > 1$ and $\gamma_i < 1\ \forall i$.

We form groups of three players ($n = 3$). Between treatments, we vary two components. First, we vary the composition of players' types γ_i. In some treatments (HOM), players are homogeneous, i.e., all players are of the same type and thus receive the same benefits from the public good ($\gamma_i = 2/3$). In other treatments (HET), players are heterogeneous: two players have a high return from the public good ($\gamma_i = 3/4$) and one player has a low return ($\gamma_i = 1/2$).[9] The different marginal per capita returns are chosen as to keep total efficiency gains constant between treatments ($2 \cdot 3/4 + 1 \cdot 1/2 = 3 \cdot 2/3$).

Second, we vary availability and content of the institution. In the benchmark treatments (VCM), there is no institution formation stage and players play a regular public goods game. In the main treatments, there is an institution formation stage first, followed by a contribution stage. In the institution formation stage, a single institution is available and can be implemented via unanimity voting, i.e., the institution is implemented if and only if all players vote in favor of adopting the institution. If the institution is rejected, the regular public goods game without any restrictions on contributions is played. The institution states each player's obligation $\bar{c}_i$, the amount that each player has to contribute to the public good in the second stage if the institution has been implemented. Voting and the implementation of the institution are costless.[10]

The main treatments vary in the type of institution that is available. In general, treatments are designed to reflect a tradeoff between efficiency and equality of payoffs. In treatments with the symmetric institution (SYM), all players are obliged to contribute their entire endowment to the public good if the institution has been implemented. The symmetric institution maximizes the sum of

[8] To give just two among many possible examples, nation states differ in their benefit from climate protection or researchers at different stages of their career benefit from joint publications to a different extent.

[9] We choose a single player with a lower return because this setup is sufficient to illustrate the potential weakness of unanimity voting, i.e., already a single player can prevent successful institution formation by vetoing.

[10] These are simplifying assumptions. Qualitatively, the theoretical predictions do not change as long as the gains in individual material payoffs due to implementing the institution outweigh the individual implementation costs.

Table 2.1. Treatments

	VCM	SYM	ASYM
HOM	$\gamma = {}^2/_3$ no obligations	$\gamma = {}^2/_3$ $\bar{c} = 20$ $\Pi = 40$	$\gamma = {}^2/_3$ $c(\bar{c} = 20) = 20$, $c(\bar{c} = 8) = 8$ $\Pi(\bar{c} = 20) = 32$, $\Pi(\bar{c} = 8) = 44$
HET	$\gamma_h = {}^3/_4$, $\gamma_l = {}^1/_2$ no obligations	$\gamma_h = {}^3/_4$, $\gamma_l = {}^1/_2$ $\bar{c} = 20$ $\Pi_h = 45$, $\Pi_l = 30$	$\gamma_h = {}^3/_4$, $\gamma_l = {}^1/_2$ $\bar{c_h} = 20$, $\bar{c_l} = 8$ $\Pi_h = \Pi_l = 36$

payoffs of all players and, thus, induces the efficient outcome. In treatments with the asymmetric institution (ASYM), one player is required to contribute 8 units, while the two others are obliged to contribute all 20 units to the public good. In treatments with heterogeneous players, the obligation is 20 for the high types, and 8 for the low types. Obligations are chosen such that the asymmetric institution implies equal payoffs for both types of players (36 each), which comes at an efficiency cost. In contrast, with heterogeneous players, the symmetric institution implies inequality in final payoffs (45 for the high types and 30 for the low type). If the asymmetric institution is combined with homogeneous players, one randomly chosen player has to contribute 8 units, while the other two players are obliged to contribute 20 units. The design results in the 2 × 3-treatment matrix shown in Table 2.1.

2.2.2 Procedures

The computerized experiments (using z-Tree; Fischbacher (2007)) were run at the BonnEconLab of the University of Bonn, Germany in 2012. Student subjects were recruited randomly from all majors (using Orsee; Greiner (2015)) and were randomly assigned to one of the six treatments (between-subject design). For each treatment, we ran two sessions with 24 subjects each. In each session, subjects first received written instructions (see Appendix 2.C). To create common knowledge, instructions were read out aloud to the subjects. Afterwards, subjects answered a set of control questions and could pose clarifying questions to ensure understanding of the game's structure and payoffs. Subjects then played the game repeatedly for 20 periods. Interaction took place within the same group of three subjects (partner matching protocol), it was anonymous and decisions were taken in private at the computer. After each voting stage, subjects received feedback on the voting result and the voting behavior of the other two subjects in their matching group. After each contribution stage, subjects were informed about their own payoff and the payoffs and contributions of the other two subjects in their group. After all 20 periods, subjects answered a questionnaire covering socio-demographic characteristics. Each session lasted about 80 minutes. Accumulated earnings were converted at a rate of 40 tokens = 1

Table 2.2. Behavioral Predictions Based on Standard Preferences

		VCM	SYM	ASYM
HOM	voting	-	implement institution	implement institution
	contribution	$c = 0$	$c = 20$	$c(\bar{c} = 20) = 20$, $c(\bar{c} = 8) = 8$
HET	voting	-	implement institution	implement institution
	contribution	$c_h = c_l = 0$	$c_h = c_l = 20$	$c_h = 20$, $c_l = 8$

Euro. Total earnings per subject ranged between 10 Euro and 22.5 Euro, with an average of about 16.4 Euro.

Altogether, we had 282 subjects, and observations on 5640 individual decisions. Given the allocation of subjects to the six treatments, repeated interaction in 20 periods and matching groups of 3, we have 16 independent observations per treatment.[11] 39% of our subjects are male, their age ranges from 16 to 42, with an average age of 22 years.

2.3 Behavioral Predictions

For each treatment, we characterize players' equilibrium behavior under two alternative assumptions concerning the shape of the utility function. First, we assume that each player's utility function coincides with the monetary payoff of the game, π_i, i.e., that players have standard preferences. Second, we assume that (at least some) players have social preferences. From the large set of available approaches to social preferences, e.g., including, among many others, Bolton and Ockenfels (2000) or Charness and Rabin (2002), we use the specification from Fehr and Schmidt (1999): in addition to valuing own monetary payoff, a player suffers from inequality in monetary payoffs, i.e., from others being worse or better off than himself. In our treatment with heterogeneous benefits from the public good, players might vote against implementing an institution that obliges all players to contribute equally to the public good in order to avoid inequality in payoffs. In the remainder of this section, we will provide an intuition for the behavioral predictions for each treatment under the two alternative assumptions on the shape of players' utility functions using the parameters of our design. More general proofs are provided in Appendix 2.A.

Table 2.2 summarizes the behavioral predictions for players with standard preferences. In basic VCM games, they are predicted not to contribute to the public good at all. Whenever $\gamma_i < 1$, contributing does not pay off from an individual perspective. Condition $\gamma_i < 1$ is met for all players in treatments HOM-VCM ($\gamma = 2/3$) and HET-VCM ($\gamma_l = 1/2$ and $\gamma_h = 3/4$).

[11] Exceptions are treatments HET-VCM and HOM-ASYM, for which we have 15 independent observations since some subjects did not show up.

In all two-stage treatments, predictions are derived using backward induction. Let U^{INST} denote utility when the institution has been implemented, with INST=SYM for the symmetric and INST=ASYM for the asymmetric institution. In the contribution stage, players will compare the utility they receive with the respective institution being in place, U^{INST}, to the utility of the VCM game that is played if the institution has not received unanimous support in the voting stage, U^{VCM}. Unanimity voting ensures that, whenever $U^{INST} \geq U^{VCM}$, it is a best response to the voting behavior of the other players to vote in favor of the institution. If all other players also vote in favor of implementing the institution, the institution will be implemented and the player's preferred outcome is achieved. If, in contrast, at least one other player votes against implementing the institution, the institution will not be implemented and the VCM game will be played. However, the approving player is still equally well off as if he had voted against implementing the institution. Whenever $U^{INST} < U^{VCM}$, a player will vote against installing the institution. In our design, $U^{INST} > U^{VCM} = E = 20$ for all player types in treatments HOM-SYM, HOM-ASYM, HET-SYM and HET-ASYM (see payoffs in Table 2.1). Consequently, for all treatments, players with standard preferences are predicted to vote in favor of the respective institution. The institution will be implemented and players will contribute according to their individual obligation. To summarize, if players have standard preferences, unanimity voting on the formation of institutions is predicted to help to overcome the social dilemma of public good provision. This result holds irrespective of whether players are homogeneous or heterogeneous and whether a symmetric or an asymmetric institution is voted on.

Table 2.3 displays the behavioral predictions for players with social preferences in terms of inequality aversion (Fehr and Schmidt, 1999). If players have social preferences, there are multiple equilibria in treatment HOM-VCM.[12] The intuition is as follows: If all players are sufficiently averse to advantageous inequality (β sufficiently high)[13], they will exactly match the contribution level $c \in [0, E]$ of the other players to equalize payoffs. If players are not or only mildly averse to advantageous inequality (β low), the only equilibrium that remains is the one with zero contributions of all players. In treatment HET-VCM, the basic mechanism driving the existence of equilibria with positive contributions is the same. If all players are sufficiently averse towards earning more than others, they contribute positive amounts as soon as the other players contribute positive amounts to prevent an unequal payoff distribution. However, to achieve equal payoffs for all three players, the low type contributes less than the two high types.

[12] The proof is provided in Fehr and Schmidt (1999).

[13] In the model of Fehr and Schmidt (1999), the parameter β captures the intensity of aversion to advantageous inequality, while the parameter α measures the degree of aversion to disadvantageous inequality.

Table 2.3. Behavioral Predictions Based on Fehr-Schmidt Preferences

		VCM	SYM	ASYM
HOM	voting	-	implement institution	type $\bar{c} = 20$ rejects if α high, type $\bar{c} = 8$ rejects if β high and $c > 12$ for all players in the VCM
	contribution	(c,c,c), $c \in [0, 20]$ if $\beta_i > 1/3 \forall i$; (0,0,0) otherwise	$c = 20$	if reject: as in HOM-VCM otherwise: $c(\bar{c} = 20) = 20$, $c(\bar{c} = 8) = 8$
HET	voting	-	low type rejects if α_l high	implement institution
	contribution	$(c_h, c_h, c_l = 2/5 c_h)$, $c_h \in [0, 20]$ if $\beta_h > 2/7$ and $\beta_l > 2/5$; (0,0,0) otherwise	if reject: as in HET-VCM otherwise: $c_h = c_l = 20$	$c_h = 20$, $c_l = 8$

In treatments HOM-SYM and HET-ASYM, assuming social instead of standard preferences does not change the predictions. In both cases, the proposed institution guarantees equality of payoffs while simultaneously maximizing utility of players who are sufficiently averse to unequal payoffs. Hence again, all players are predicted to vote in favor of the respective institution, it will be implemented, and players will contribute according to their obligation. In treatments HET-SYM and HOM-ASYM, however, predictions based on standard preferences and social preferences differ. In both treatments, players with standard preferences always support the formation of the institution as it offers a higher monetary payoff than the VCM and they do not suffer from unequal payoffs that arise from implementing the institution. In contrast, in treatment HET-SYM, low type players with social preferences who suffer sufficiently from being worse off than the high types (α sufficiently high), object to institution formation. They prefer a lower monetary payoff, but equal payoffs across players in the VCM, to a higher monetary payoff, but disutility from inequality due to the symmetric institution being in place. Consequently, low type players drive rejections of the proposed symmetric institution. Similarly, in treatment HOM-ASYM, all players potentially have a motive for voting against the asymmetric institution that introduces inequality in payoffs: Players with an obligation of 8 tokens, if they are sufficiently averse to advantageous inequality, and players with an obligation of 20 tokens if they are sufficiently averse to disadvantageous inequality.[14]

2.4 Results

This section is structured along five sets of results concerning differences in voting and contribution behavior across treatments. All results are qualitatively in line with the behavioral predictions presented in Section 2.3 and Appendix 2.A, when assuming that at least some players are inequality averse to an extent that induces their behavior to deviate from the predictions based on standard preferences. We report detailed predictions in Appendix 2.A.4 and the corresponding results in the text below.

[14] Preferences for efficiency, see, e.g., Charness and Rabin (2002), are an alternative explanation for rejecting an asymmetric institution. Efficiency seekers should reject institutions that do not induce full contributions in order to contribute more than they were obliged to with the institution being in place (expecting others to contribute more after rejection provides a reason for selfish agents to reject institutions, too). Charness-Rabin preferences predict that rejections of institutions are possible in all treatments. However, all predicted rejections require fairly large amounts of contributions in the subsequently played VCM to be justified. Yet, our data (see section 2.4.4) do not indicate higher levels of contributions after an institution is rejected than under the institution, suggesting that efficiency seeking is not a predominant motive for rejections. Behavioral predictions based on Charness-Rabin preferences (with no reciprocity, see Appendix I in Charness and Rabin (2002)) are available from the authors upon request.

First, we will briefly present results in treatments HOM-VCM and HET-VCM that provide baseline scenarios for comparing whether unanimity voting on institutions increases efficiency. We proceed by discussing under which circumstances unanimity voting on symmetric or asymmetric institutions helps to increase public good provision. We thereby focus on treatment comparisons in which changes in behavior can be attributed to a single change in setup. That means, we either compare treatments with different institutions, while keeping constant the composition of player types (HOM or HET) or we compare treatments with a different composition of player types, while keeping constant the nature of the institution to be voted on (SYM or ASYM).

Table 2.4 and Table 2.5 contain first descriptive results. Table 2.4 displays contributions averaged over all periods by treatment. Table 2.5 shows the share of affirmative votes and implementation rates averaged over all periods by treatment. Moreover, Figures 2.1 and 2.2 display the treatment-specific development of contributions and share of affirmative votes over time.

Table 2.4. Average Contributions by Treatment

	VCM	SYM	ASYM
		HOM	
overall	10.72	18.18	11.44
	(7.83)	(5.27)	(8.35)
types $\bar{c} = 20$	–	–	12.12
	–	–	(8.79)
types $\bar{c} = 8$	–	–	10.09
	–	–	(7.23)
		HET	
overall	8.05	14.21	13.85
	(6.58)	(7.79)	(7.20)
high types	9.42	14.77	17.18
	(7.07)	(7.42)	(6.36)
low types	5.33	13.08	7.21
	(4.33)	(8.38)	(2.90)

Notes: Contributions are in tokens. Standard deviations are presented in parentheses.

2.4.1 Baseline Treatments: Homogeneous versus Heterogeneous Players in the VCM

On average, contributions in the standard VCM tend to be lower with heterogeneous than with homogeneous agents. Subjects contribute 10.7 out of 20 units

Table 2.5. Share of Affirmative Votes and Implementation Rate by Treatment

	Hom		Het	
	Sym	Asym	Sym	Asym
	Affirmative votes			
overall	.95	.54	.84	.91
types $\bar{c} = 20$	–	.48	.96	.90
types $\bar{c} = 8$	–	.68	.60	.94
	Implementation Rates			
	.87	.27	.56	.77

Notes: The share of affirmative votes as well as the implementation rate is significantly higher in treatment HOM-SYM than in HET-SYM (MWU, $p = 0.01$ for both), in HOM-SYM than in HOM-ASYM (MWU, $p < 0.01$ for both), lower in HET-SYM than in HET-ASYM (MWU, $p = 0.16$ for the share of affirmative votes and $p = 0.08$ for the implementation rate), and significantly higher in HET-ASYM than in HOM-ASYM (MWU, $p < 0.01$ for both). The share of low types' affirmative votes in HET-ASYM is significantly higher than in HOM-ASYM (MWU, $p < 0.01$). The share of high types' affirmative votes in HET-ASYM is significantly lower than in HOM-ASYM (MWU, $p = 0.04$).

in treatment HOM-VCM and 8.1 units in treatment HET-VCM (Mann-Whitney ranksum test (MWU), $p = 0.11$).[15]

Result 1:

> Average contributions in treatment HET-VCM are slightly, but not significantly lower than in treatment HOM-VCM.

Moreover, we observe that average contributions of low and high types differ in HET-VCM: while low type players contribute only 5.3 units, high type players contribute 9.4 units on average. As a consequence, average payoffs for the two player types are similar, 26.8 and 28.7 units, respectively. Players of both types seem to intuitively strive for equal payoffs.

2.4.2 Unanimity Voting on the Symmetric Institution: Homogeneous versus Heterogeneous Players

We first consider the voting behavior of homogeneous players who are confronted with the decision whether to install the symmetric institution that obliges each player to contribute the efficient amount, 20 units. Overall, 95.2% of votes (914 out of 960 votes) are in favor of implementing the symmetric institution. As a result, in 86.6% of all cases, all three players of a group unanimously agree to implement the symmetric institution and it is indeed implemented.

[15] Throughout the paper, we report two-sided p-values. Each matching group's average contribution is one independent observation.

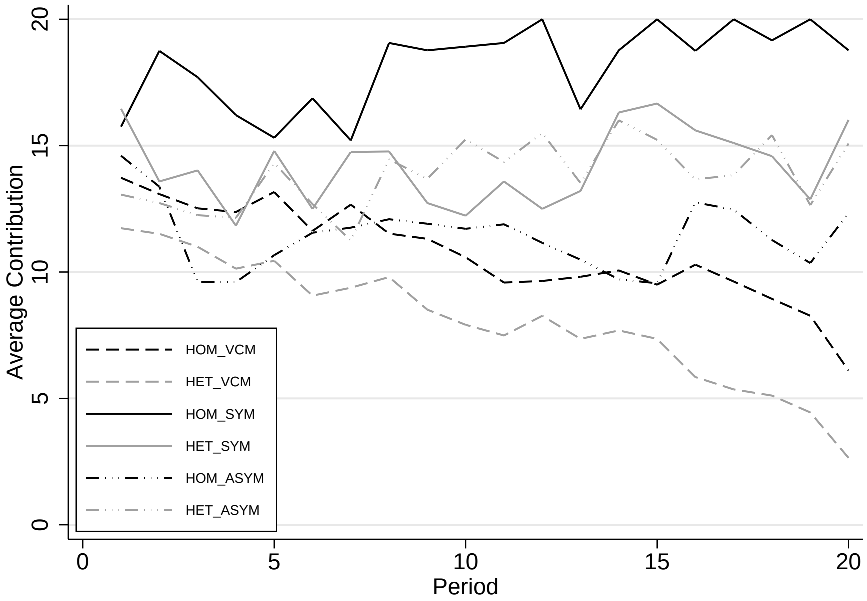


Figure 2.1. Development of Average Contributions over Time

Notes: In treatments $HOM - VCM$ and $HET - VCM$, average contributions decrease over time ($HOM - VCM$: Spearman's Rho $r = -0.27, p < 0.01$ and $HET - VCM$: $r = -0.47, p < 0.01$). In treatments $HOM - SYM$ and $HET - ASYM$, average contributions increase over time ($HOM - SYM$: $r = +0.27$, $p < 0.01$, $HET - ASYM$: $r = +0.21$, $p < 0.01$). In treatments $HET - SYM$ and $HOM - ASYM$, time trends in contributions are not significant ($HET - SYM$: $r = +0.08$, $p = 0.16$, $HOM - ASYM$: $r = -0.03$, $p = 0.64$).

Result 2:

> In treatment HOM-SYM, average contributions are significantly higher than in treatment HOM-VCM.

On average, subjects contribute 18.2 units in treatment HOM-SYM instead of 10.7 units in treatment HOM-VCM (MWU, $p < 0.01$). After some periods of initial learning efficiency is close to 100% (see also Figure 2.1). To summarize, in our setup with homogeneous players, unanimity voting on the symmetric institution increases efficiency substantially.

Does unanimity voting on the efficient institution also yield high support if players are heterogeneous, i.e., if the efficient institution introduces unequal payoffs? Again, we start by analyzing behavior in the voting stage. In treatment HET-SYM, the overall share of affirmative votes is lower than in treatment HOM-SYM, 83.9% instead of 95.2% (MWU, $p = 0.01$). Heterogeneous players object the implementation of the efficient symmetric institution more often than homogeneous players. The difference in affirmative votes between treatment HOM-SYM and HET-SYM persists over time (see Figure 2.2). Similarly, the overall

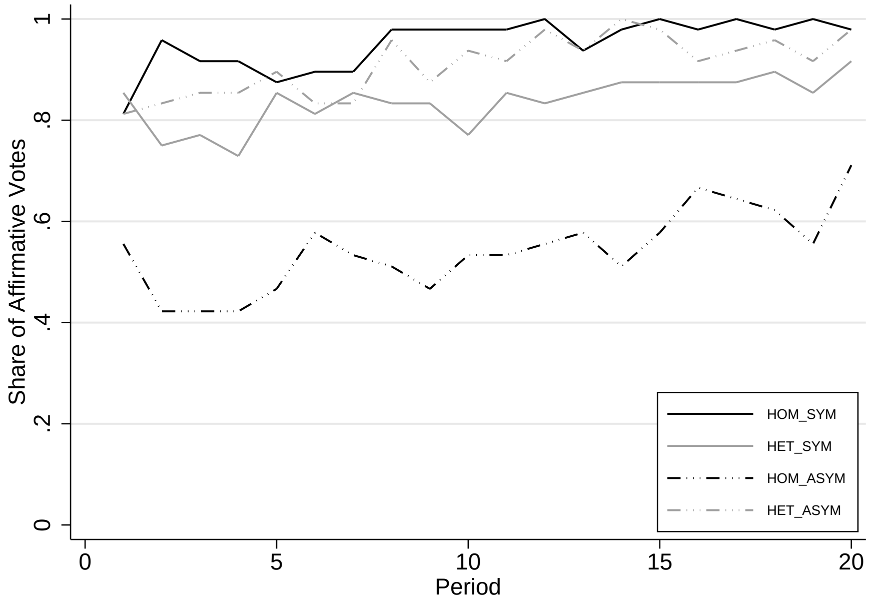


Figure 2.2. Share of Affirmative Votes over Time

Notes: In all four two-stage treatments, the share of affirmative votes increases over time ($HOM-SYM$: Spearman's Rho $r = +0.29$, $HET-SYM$: $r = +0.18$, $HOM-ASYM$: $r = +0.18$, $HET-ASYM$: $r = +0.26$, all $p < 0.01$).

implementation rate in treatment HET-SYM is 56.3%, substantially lower than in treatment HOM-SYM, 86.6% (MWU, $p = 0.01$). Rejections of the institution are largely due to the voting behavior of low types. In our data, 95.9% of high types vote in favor of implementing the institution in treatment HET-SYM, but only 59.7% of low types do.

As a consequence of the lower implementation rate, average contributions are significantly lower in treatment HET-SYM than HOM-SYM: 14.2 instead of 18.2 (MWU, $p = 0.01$). However, average contributions in treatment HET-SYM are significantly higher than in the VCM with heterogeneous players (MWU, $p < 0.01$). Result 3 summarizes our results for treatment HET-SYM.

Result 3:

a) In treatment HET-SYM, average contributions are significantly higher than in treatment HET-VCM.

b) In treatment HET-SYM, both implementation rate and average contributions are significantly lower than in treatment HOM-SYM.

Overall, if players are heterogeneous rather than homogeneous in their marginal returns from the public good, unanimity voting on the efficient institution does not always result in its successful implementation. Still, compared to the standard public good game in which no institution is available, unanimity voting on the efficient institution increases efficiency substantially – even if players are heterogeneous.[16]

2.4.3 Unanimity Voting on the Asymmetric Institution: Homogeneous versus Heterogeneous Players

A potential remedy to the frequent rejections of the symmetric institution by low type players is to design an asymmetric institution that ensures the maximum possible payoffs among the set of all equitable payoff allocations. Obviously, under the asymmetric institution, the low type players' obligation must be lower than under the symmetric institution. As a drawback, the implementation of the asymmetric institution results in a lower level of public good provision than the implementation of the symmetric institution.

Concerning results in treatment HET-ASYM, 91.0% of players vote in favor of implementing the asymmetric institution which results in 77.2% successful implementations. Thus, with heterogeneous players, the asymmetric institution that guarantees equal payoffs for both player types is more than 20% points more likely to be implemented than the symmetric one that induces the efficient outcome, but unequal payoffs across player types (MWU, $p = 0.08$ for the implementation rate and $p = 0.16$ for the share of affirmative votes). The higher implementation rate is due to the substantially higher likelihood of low types to vote in favor of the asymmetric institution than the symmetric one: 94.1% instead of 59.7% (MWU, $p < 0.01$). With 89.5%, the high types' share of affirmative votes for the asymmetric institution is only slightly lower than the 95.9% affirmative votes for the symmetric institution (MWU, $p = 0.04$).

While implementation rates differ markedly for treatment HET-SYM and HET-ASYM, average contributions do not: 13.9 units in HET-ASYM compared to 14.2 units in HET-SYM (MWU, $p = 0.97$).[17] There are two opposing effects

[16] Although the focus of our paper is on distributive fairness, subjects' behavior could be driven by procedural fairness concerns, too. While economists have started studying the latter approach only lately (the first economic experiments are reported in Bolton et al. (2005); see also Krawczyk (2011), for a theoretical model), the idea of procedural fairness has been prominent in psychology for some decades already (see, for example, Tyler and Lind (2000) for a summary). Barrett-Howard and Tyler (1986) report that procedural fairness is equal in importance to distributive fairness for subjects who are confronted with allocation decisions. This could provide a potential explanation for the 84% approval rate in HET-SYM, namely if subjects think that equal obligations are procedurally fair in general and/or randomly assigning heterogeneity in the MPCRs to subjects is procedurally fair. Interestingly, if this line of reasoning indeed applies it seems to be done in a self-serving manner, because the support for the institution is much stronger among high types than among low types. We thank an anonymous referee for pointing this out.

[17] In line with the theoretical predictions, we observe that in HET-ASYM low and high types are equally well off on average. High types' average payoff is 34.0, low types' average payoff is 33.6. In contrast, in treatment HET-SYM, average payoffs of low types are substantially lower

that cancel each other out: while the higher implementation rate in HET-ASYM increases contributions, implementing the asymmetric institution instead of the symmetric one reduces contributions of the low types from 20 to 8 units. Compared to the benchmark VCM game with heterogeneous players, average contribution levels are significantly higher in treatment HET-ASYM than in treatment HET-VCM (MWU, $p < 0.01$). We summarize results for treatment HET-ASYM below.

Result 4:

a) In treatment HET-ASYM, the implementation rate is higher than in treatment HET-SYM. In contrast, average contributions in treatments HET-ASYM and HET-SYM are very similar.

b) In treatment HET-ASYM, average contributions are significantly higher than in treatment HET-VCM.

Overall, designing institutions that address players' demand for equal benefits from institution formation seems to be very successful in raising the implementation rate. In many contexts, a higher rate of institution formation could be considered beneficial per se, e.g., due to raising reliability of public good provision or by potentially triggering future institutionalized cooperation. However, increasing the implementation rate by voting on an asymmetric institution will always come at the cost of institutionalizing less than efficient levels of public good provision.

To rule out that the high implementation rate in HET-ASYM is due the asymmetry in contributions per se, we now turn to treatment HOM-ASYM. Here, we can explore how the asymmetric institution performs if players are homogeneous, i.e., when it introduces binding rules concerning contributions to potentially increase efficiency, but those rules induce unequal payoffs across players.

Proposing an asymmetric institution to homogeneous players receives relatively low levels of support. The average share of affirmative votes ranges between 40% and 70% over time, resulting in an average implementation rate of only 26.7%. For players with an obligation of 8 units, the share of affirmative votes is 67.7%, while it is 20 percentage points lower for those with an obligation of 20. This might be because both types of players possibly have a motive to vote against the institution, namely aversion to advantageous inequality (for players with an obligation of 8 units) and aversion to disadvantageous inequality (for players with an obligation of 20 units). The share of affirmative votes as well as the implementation rate is significantly higher in treatment HET-ASYM than in HOM-ASYM (MWU, $p < 0.01$ for both).

than those of high types: 28.2 instead of 37.2. Again, this finding is in line with the theoretical predictions.

We have already shown that, with homogeneous players, proposing a symmetric institution helps to overcome the social dilemma of public good provision. This is not the case with an asymmetric institution. The average contributions in treatment HOM-ASYM are not significantly different from average contributions in treatment HOM-VCM (MWU, $p = 0.75$) and significantly lower than in treatment HOM-SYM (MWU, $p < 0.01$).

Finally, the asymmetric institution performs worse for homogeneous than for heterogeneous players, i.e., when it introduces inequality instead of addressing it. With homogeneous players, both the share of affirmative votes and the average contributions are lower (MWU, $p < 0.01$ for affirmative votes and $p = 0.04$ for contributions). This strongly suggests that the success of the asymmetric institution for heterogeneous agents is indeed due to addressing payoff inequalities between agents. Below, we summarize results for treatment HOM-ASYM.

Result 5:

a) Average contributions in treatment HOM-ASYM and HOM-VCM do not differ significantly.

b) In treatment HOM-ASYM, implementation rate and average contributions are significantly lower than in treatment HOM-SYM.

c) In treatment HOM-ASYM, implementation rate and average contributions are significantly lower than in treatment HET-ASYM.

2.4.4 Contributions by Institution Formation Status

So far, we have analyzed average contributions in a given treatment, averaging over cases of successful institution formation and those of failure to form an institution. We have not studied yet how failure to implement the proposed institution affects contribution levels. If motives for objecting to institution formation differ across treatments, contribution levels in case of failed institution formation could also differ across treatments. For example, inequality aversion could be a plausible motive for voting against institution formation in treatments HET-SYM and HOM-ASYM in which institutions induce unequal payoffs. In treatments HOM-ASYM and HET-ASYM, a preference for efficient levels of public good provision could drive rejections (compare footnote 14). Rejections of the institution are harder to rationalize in treatment HOM-SYM because implementation of the institution results in maximal and equal payoffs. Consequently, rejections could be due to, e.g., mistakes or pleasure from exerting (destructive) power. These motives could induce negative reciprocity, resulting in contribution levels well below the corresponding VCM. In contrast, efficiency seekers could reject an asymmetric institution aiming at contribution levels that exceed institutional obligations. Players who reject an institution due to inequality aversion have motives to contribute as in the baseline VCM whose equilibria ensure equality of payoffs across players.

While we did not elicit subjects' individual beliefs about the preferences of players which rejected the institution, presenting results on average contributions in case of failed institution formation is still informative. Table 2.6 and Figure 2.B.1 in the Appendix show that, in the relatively rare case of institution failure (13%), average contribution levels in treatment HOM-SYM are substantially below those of the corresponding VCM (6.4 instead of 10.7 units). In treatments HET-SYM, HET-ASYM, and HOM-ASYM, average contributions are much closer to those of the corresponding baseline VCM. Taken together, our results do not point at a large, "hidden cost" of failed institution formation, namely substantially and frequently reduced contributions in case of failed institution implementation (except for treatment HOM-SYM).

Table 2.6. Average contributions after failed institution formation

	VCM	SYM	ASYM
HOM	10.72 (7.83)	6.43 (6.90)	9.78 (8.56)
HET	8.05 (6.58)	6.76 (6.33)	6.59 (7.12)

Notes: Contributions are in tokens. Standard deviations are presented in parentheses.

2.5 Conclusion

The paper at hand studied the process of institution formation in social dilemmas, in particular the role of heterogeneity among players i) in their benefits from cooperation and ii) in their institutional obligations. We found that the potential tension between efficiency and equality in payoffs, originating from these heterogeneities, strongly affected implementation rates of institutions. With heterogeneous players, aggregate implementation rates were significantly lower for institutions featuring equal rather than unequal obligations; and vice versa for homogeneous players – even though failed implementation usually implied severe cutbacks in monetary payoffs. Both with homogeneous and heterogeneous players, failed implementations arose primarily, but not exclusively, from rejections by the disadvantaged players that profited to a lesser degree from the implemented institution. Consequently, institutions which tailored obligations to players' specific heterogeneities were able to gather higher degrees of support. In fact, if benefits from institution formation were evenly distributed across players, we observed strikingly higher implementation and cooperation rates than what has typically been found in related studies that only require non-unanimous support for institutions to be implemented for all members (e.g., Kosfeld et al., 2009).

A potential reason for the latter finding is that, in contrast to other decision rules, unanimity voting entails a very strong notion of conditional cooperation. The veto right inherent in unanimity voting makes each player's cooperation decision contingent on the decision of all other players involved. Consequently, the supporting players do not face the risk of being exploited by non-supporting players.[18] On a similar note, no player will ever be governed by an institution that he did not support himself. Both, the notion of conditional cooperation and the retained sovereignty, make unanimity voting an attractive rule to settle on in the first place.

On the other hand, these advantages come at the cost of an increased likelihood of rejecting efficient institutions as well as potentially low levels of cooperation after a rejection has occurred. Already with three players, we saw that these problems exist. With larger groups, one might expect successful institution formation to be even more difficult, in particular if benefits from institution formation are not equally distributed across players. Moreover, our data suggest that voting against the institution is sometimes connected with the implicit costs of making subsequent cooperation more difficult. One might even imagine that rejecting players become the target of retaliation in other, seemingly unrelated, domains. Both threats might be bigger in large groups, simply because there are more players who might potentially opt against the institution and/or who might retaliate rejections. Yet, for groups deciding on the implementation of an institution that takes care of players' idiosyncrasies, these threats might instead strengthen the power of an unanimity rule. Furthermore, under institutions that lead to inequalities in payoffs, payoff differences might be less salient in large groups because they are harder to recognize – in particular if players do not compare themselves with everyone else in a large population, but rather choose a small reference group consisting of similar others. It would therefore be interesting to check in future studies whether the positive or negative effects dominate when group size is increased.

Follow-up studies might also investigate if aggregate behavioral patterns are affected by changes in other parameters of our design, like the marginal per capita return from cooperation or the exact content of the institution. We observed in our data on heterogeneous agents that, overall, the symmetric and asymmetric institution lead to similar average cooperation rates. This was due to two opposing effects that cancel each other out: while the higher implemen-

[18] Parts of this reasoning rely on the strong enforcement mechanism underlying our institutions. It would be interesting to study setups that allow to discriminate between i) support and ii) adherence to an institution, and how these factors are affected by the voting mechanism in place. We thank an anonymous referee for pointing this out.

tation rate for the asymmetric institution generally increases cooperation, total obligations (and thus cooperation rates) are lower than when the efficient symmetric institution is implemented. Although this qualitative finding is not at the heart of our paper, it is still intriguing. Given the quantitative behavioral effects that we observe, one could imagine that average outcomes between symmetric and asymmetric institutions start diverging as the most efficient payoff-equalizing mechanism becomes more inferior to the efficient mechanism.

Along similar lines, natural next steps for future extensions also include more complex institutional arrangements. For example, redistribution might allay disadvantaged member's doubts about the implementation of efficient institutions for heterogeneous agents. The implementation of institutions with hierarchical structures, from simple leader-follower arrangements to multi-layered structures, yield the potential to increase implementation rates and cooperation, too (e.g., Gächter et al., 2010; Hamman et al., 2011; Falk and Kosfeld, 2012). Complementing these variations, one could also shed more light on the performance of different voting rules for implementing given institutions (e.g., Young, 1995; Gillet et al., 2009; Austen-Smith and Feddersen, 2006). More generally, allowing for richer environments with competing institutions and voting rules opens up the possibility to learn even more about the type of institutions that *endogenously* arise within a group. Of course, in contrast to our approach, self-selection would make proper causal interpretation more difficult. Still, it would be a nice complement to the current research agenda: understanding what kind of institutions are created by groups, which voting rules are adopted for implementing these institutions, and how these institutions perform under a variety of circumstances.

References

Anderson, Lisa R., Jennifer M. Mellor, and Jeffrey Milyo (2008): "Inequality and public good provision: An experimental analysis." *Journal of Socio-Economics*, 37, 1010–1028. [10]

Andreoni, James and Laura K. Gee (2012): "Gun for hire: delegated enforcement and peer punishment in public goods provision." *Journal of Public Economics*, 96 (11), 1036–1046. [9]

Apesteguia, Jose, Patricia Funk, and Nagore Iriberri (2013): "Promoting rule compliance in daily-life: Evidence from a randomized field experiment in the public libraries of Barcelona." *European Economic Review*, 64, 266–284. [9]

Austen-Smith, David and Timothy J. Feddersen (2006): "Deliberation, preference uncertainty, and voting rules." *American Political Science Review*, 100 (2), 209–217. [26]

Banks, Jeffrey S., Charles R. Plott, and David P. Porter (1988): "An experimental analysis of unanimity in public goods provision mechanisms." *Review of Economic Studies*, 55 (2), 301–322. [9]

Barrett-Howard, Edith and Tom Tyler (1986): "Procedural justice as a criterion in allocation decisions." *Journal of Personality and Social Psychology*, 50 (2), 296–304. [21]

Bierbrauer, Felix J. and Martin F. Hellwig (2011): "Mechanism Design and Voting for Public-Good Provision." *MPI Collective Goods Preprint* (2011/31). [5]

Birnberg, Jacob G., Louis R. Pondy, and C. Lee Davis (1970): "Effect of three voting rules on resource allocation decisions." *Management Science*, 16 (6), 356–372. [9]

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels (2005): "Fair procedures: Evidence from games involving lotteries." *Economic Journal*, 115 (506), 1054–1076. [21]

Bolton, Gary E. and Axel Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition." *American Economic Review*, 90 (1), 166–193. [13]

Charness, Gary and Matthew Rabin (2002): "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*, 117 (3), 817–869. [13, 16]

Cherry, Todd L., Stephan Kroll, and Jason F. Shogren (2005): "The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab." *Journal of Economic Behavior & Organization*, 57 (3), 357–365. [10]

Dannenberg, Astrid, Andreas Lange, and Bodo Sturm (2014): "Participation and Commitment in Voluntary Coalitions to Provide Public Goods." *Economica*, 81, 257–275. [9]

Decker, Torsten, Andreas Stiehler, and Martin Strobel (2003): "A comparison of punishment rules in repeated public good games an experimental study." *Journal of Conflict Resolution*, 47 (6), 751–772. [5, 8]

Falk, Armin and Michael Kosfeld (2012): "It's all about connections: Evidence on network formation." *Review of Network Economics*, 11 (3), Article 2. [26]

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebmer (2000): "A simple mechanism for the efficient provision of public goods: Experimental evidence." *American Economic Review*, 90 (1), 247–264. [9]

Fehr, Ernst and Simon Gächter (2002): "Altruistic punishment in humans." *Nature*, 415, 137–140. [9]

Fehr, Ernst and Klaus M. Schmidt (1999): "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114 (3), 817–868. [13, 14, 30, 31]

Fischbacher, Urs (2007): "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10 (2), 171–178. [12]

Fisher, Joseph, R. Mark Isaac, Jeffrey W. Schatzberg, and James M. Walker (1995): "Heterogenous demand for public goods: Behavior in the voluntary contributions mechanism." *Public Choice*, 85 (3-4), 249–266. [9]

Gächter, Simon and Ernst Fehr (2000): "Cooperation and punishment in public goods experiments." *American Economic Review*, 90 (4), 980–994. [9]

Gächter, Simon, Daniele Nosenzo, Elke Renner, and Martin Sefton (2010): "Sequential vs. simultaneous contributions to public goods: Experimental evidence." *Journal of Public Economics*, 94 (7-8), 515–522. [26]

Gerber, Anke, Jakob Neitzel, and Philipp C. Wichardt (2013): "Minimum participation rules for the provision of public goods." *European Economic Review*, 64, 209–222. [6, 9, 34]

Gillet, Joris, Arthur Schram, and Joep Sonnemans (2009): "The tragedy of the commons revisited: The importance of group decision-making." *Journal of Public Economics*, 93 (5-6), 785–797. [26]

Greiner, Ben (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association*, 1 (1), 114–125. [12]

Gürerk, Ozgür, Bernd Irlenbusch, and Bettina Rockenbach (2006): "The competitive advantage of sanctioning institutions." *Science*, 312 (5770), 108–11. [5]

Hamman, John R., Roberto Weber, and Jonathan Woon (2011): "An experimental investigation of electoral delegation and the provision of public goods." *American Journal of Political Science*, 55 (4), 738–752. [26]

Hobbes, Thomas (1651): *Leviathan*. Meiner Verlag (1996), Hamburg. [7]

Isaac, R. Mark and James M. Walker (1988): "Group size effects in public goods provision: The voluntary contributions mechanism." *Quarterly Journal of Economics*, 103 (1), 179–199. [10]

Kamei, Kenju, Louis Putterman, and Jean-Robert Tyran (2015): "State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods." *Experimental Economics*, 18, 38–65. [6, 9]

Kesternich, Martin, Andreas Lange, and Bodo Sturm (2014): "The impact of burden sharing rules on the voluntary provision of public goods." *Journal of Economic Behavior & Organization*, 105, 107–123. [9]

Khadjavi, Menusch, Andreas Lange, and Andreas Niklisch (2014): "The social value of transparency and accountability: Experimental evidence from asymmetric public good games." *WiSo HH Working Paper Series 12 2014*. [10]

Kosfeld, Michael, Akira Okada, and Arno Riedl (2009): "Institution Formation in Public Goods Games." *American Economic Review*, 99 (4), 1335–1355. [5, 6, 8, 9, 25]

Krawczyk, Michal W. (2011): "A model of procedural and distributive fairness." *Theory and Decision*, 70 (1), 111–128. [21]

Linardi, Sera and Margaret A. McConnell (2011): "No excuses for good behavior: Volunteering and the social environment." *Journal of Public Economics*, 95 (5), 445–454. [8]

Maggi, Giovanni and Massimo Morelli (2006): "Self-enforcing Voting in International Organizations." *American Economic Review*, 96 (4), 1137–1158. [8]

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran (2014): "Self-organization for collective action: An experimental study of voting on sanction regimes." *Review of Economic Studies*, 81 (1), 301–324. [5, 8, 9]

McEvoy, David M., Todd L. Cherry, and John Stranlund (2015): "Endogenous Minimum Participation in International Environmental Agreements: An Experimental Analysis." *Environmental and Resource Economics*, 62 (4), 729–744. [8]

McGinty, Matthew and Garrett Milam (2013): "Public goods provision by asymmetric agents: Experimental evidence." *Social Choice and Welfare*, 40 (4), 1159–1177. [9]

Ostrom, Elinor, James Walker, and Roy Gardner (1992): "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review*, 86 (2), 404–417. [9]

Putterman, Louis, Jean-Robert Tyran, and Kenju Kamei (2011): "Public goods and voting on formal sanction schemes." *Journal of Public Economics*, 95 (9), 1213–1222. [9]

Rauchdobler, Julian, Rupert Sausgruber, and Jean-Robert Tyran (2010): "Voting on Thresholds for Public Goods: Experimental Evidence." *FinanzArchiv: Public Finance Analysis*, 66 (1), 34–64. [9]

Reuben, Ernesto and Arno Riedl (2013): "Enforcement of contribution norms in public good games with heterogeneous populations." *Games and Economic Behavior*, 77 (1), 122–137. [9]

Riedel, Nadine and Hannah Schildberg-Hörisch (2013): "Asymmetric obligations." *Journal of Economic Psychology*, 35, 67–80. [10]

Rousseau, Jean Jaques (1762): *Der Gesellschaftsvertrag*. Röder-Taschenbuch (1988), Köln. [7]

Shavell, Steven and A. Mitchell Polinsky (2000): "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature*, 38 (1), 45–76. [5]

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]

Tyler, Tom and E. Allen Lind (2000): "Procedural justice." In *Handbook of Justice Research in Law*. Ed. by J. Sanders and V.L. Hamilton. New York: Kluwer, 63–91. [21]

Tyran, Jean-Robert and Lars P. Feld (2006): "Achieving Compliance when Legal Sanctions are Non-deterrent." *Scandinavian Journal of Economics*, 108 (1), 135–156. [5]

Wicksell, Knut (1964): "A New Principle of Just Taxation." In *Classics in the Theory of Public Finance*. Ed. by Richard A. Musgrave and Alan T. Peacock. London: Macmillan. [8]

Young, Peyton (1995): "Optimal voting rules." *Journal of Economic Perspectives*, 9 (1), 51–64. [26]

Appendix 2.A Model and Theoretical Predictions

2.A.1 Model

We study the following two-stage game in which players have perfect information on other players' preferences:

Voting stage: First, all players simultaneously and independently vote either in favor of or against adopting an institution. The institution specifies a contribution level that each player is obliged to contribute to the public good and introduces sanctions for deviant contribution levels. Sanctions are sufficiently severe to ensure that the prescribed contribution levels are indeed implemented.

Contribution stage: Second, all players simultaneously and independently choose their contribution level to the public good. If the institution has been implemented, players will contribute the amount specified by the institution. If the institution has not been implemented, there is no sanctioning mechanism and players play a standard public goods game (VCM).

In the contribution stage, players know how other players in their group voted in the voting stage. In the following, we demonstrate that rejecting the institution can increase utility in some treatments, while not in others. A multitude of equilibria exists. In order to keep the subsequent analysis tractable and short, when analyzing equilibria in which at least one player rejects, we focus on those equilibria in which rejecting the institution strictly increases the rejecting player's utility.[19] For each treatment, we will first characterize equilibria if play-

[19] There also exist equilibria, in which players reject the institution although the resulting utilities are lower than in the state of successful institution formation: As soon as one player rejects, the decision of the other players does not affect institution formation under unanimity voting. Consequently, further equilibria exist in which at least two players reject the institution.

ers' utility functions coincide with the monetary payoff of the game, π_i, i.e., if players have standard preferences. We will then proceed by analyzing equilibria of the game if (some) players have social preferences, i.e., suffer from inequality in monetary payoffs (compare, among others, Fehr and Schmidt (1999)). Fehr and Schmidt (1999) assume that players compare their own monetary payoff with the monetary payoff of all other players. They introduce the following utility function:

$$U_i = \pi_i - \alpha_i \tfrac{1}{n-1} \textstyle\sum_{j=1}^{n} max\{\pi_j - \pi_i; 0\} - \beta_i \tfrac{1}{n-1} \textstyle\sum_{j=1}^{n} max\{\pi_i - \pi_j; 0\}$$

The first term represents the monetary payoff obtained in the game. The second term captures utility losses due to being worse off than other players. α_i measures the degree of individual envy. The last term denotes utility losses that players receive from being better off than other players. β_i is typically interpreted as a measure for the degree of compassion. Additionally, two important properties are assumed. First, $\alpha_i \geq \beta_i$ or, in words, envy is at least as strong as compassion. Second, $\beta_i < 1$, which prevents agents from "burning their own money" to achieve a more equal outcome. In our setup with heterogeneous benefits from the public good, players might vote against implementing an institution that obliges all players to contribute equally to the public good in order to avoid inequality. Hence, we consider the model of Fehr and Schmidt (1999) as a natural choice to derive predictions for our setup.

In the following, subscript l (h) stands for low (high) type, i.e., players with a low (high) MPCR. Given the focus of this paper, (only) the analysis of the treatments featuring heterogeneous players with social preferences focuses on the case of three players: one low type with low MPCR γ_l, two high types with high MPCR γ_h, and $\Delta\gamma = \gamma_h - \gamma_l < 1/2$. Additionally, the propositions presented in the text further specify results by setting $\gamma_l = 1/2$ and $\gamma_h = 3/4$, the parameters we have used in the experiment implementation.

2.A.2 Treatments $HOM - VCM$ and $HET - VCM$

The standard Voluntary Contribution Mechanism (VCM) is a one-stage game without voting on implementing an institution and without any sanctioning mechanism for low contributions.

Proposition 1. *If players are money-maximizers, they contribute $c_i = 0, \forall i$ in treatments $HOM - VCM$ and $HET - VCM$.*

Whenever $\frac{\partial \pi_i}{\partial c_i} = -1 + \gamma_i < 0$, the marginal individual cost of contributing to the public good exceeds the marginal individual benefit. Consequently, in any standard VCM game, a money-maximizing player will not contribute to the public good for all $\gamma_i < 1$. Condition $\gamma_i < 1$ is met by definition of the public goods game for all players in treatments $HOM - VCM$ ($\gamma = 2/3$) and $HET - VCM$ ($\gamma_l = 1/2$ and $\gamma_h = 3/4$).

Proposition 2. *Let us assume that players have social preferences.*
In treatment $HOM-VCM$, if $\gamma_i+\beta_i<1$ for at least one player, there is a unique equilibrium in which all players contribute $c_i=0$. If all players have $\gamma_i+\beta_i>1$, other equilibria with $c_i>0$ exist, in which all players contribute $c_i=c_j\in[0,E],\forall j\neq i$.
In treatment $HET-VCM$, if $\beta_h>2/7$ for both high types and $\beta_l>2/5$ for the low type, equilibria with positive contributions exist in which $c_h\in[0,E]$ and $c_l=2/5c_h$. All players earn equal payoffs. Otherwise, there exists a unique equilibrium in which all players contribute $c_i=0$.

The proof of $HOM-VCM$ is provided in Fehr and Schmidt (1999). The intuition is as follows: If players are sufficiently averse to advantageous inequality (β sufficiently high), they are willing to exactly match the contribution levels of the other players to equalize payoffs. Using the parameters of our experiment, the proposition boils down to the result that equilibria with positive contribution levels only exist if $\beta>1/3$ for all players. In treatment $HET-VCM$, the basic mechanism that drives the existence of equilibria with positive contributions is the same as in the VCM with homogeneous players. If players are sufficiently averse towards earning more than others, they contribute positive amounts to prevent an unequal payoff distribution as soon as other players contribute a positive amount. To achieve an equal payoff distribution, the low type contributes less than the high types. In the following, we provide a formal analysis of the behavior of players with social preferences in treatment $HET-VCM$.

We start by analyzing the **behavior of the low type**.
Case 1: $\pi_l\leq\pi_1$ and $\pi_l\leq\pi_2$
Let us assume that the low type contributes such that his monetary payoff is not larger than the payoff of both high types (that are labeled by indices 1 and 2). Then, the utility function of the low type that is relevant for the marginal analysis is denoted by: $U_l=E-c_l+\gamma_l(c_l+c_1+c_2)-\frac{\alpha_l}{2}(c_l-c_1+\Delta\gamma(c_l+c_1+c_2))-\frac{\alpha_l}{2}(c_l-c_2+\Delta\gamma(c_l+c_1+c_2))$. The derivative with respect to c_l is given by $\frac{\partial U_l}{\partial c_l}=-1+\gamma_l-\alpha_l(1+\Delta\gamma)$ and will always be negative as $\gamma_l<1$ and $\Delta\gamma\geq0$. Hence the low type will never increase his contribution, but at least decrease his contribution until $\pi_l=\pi_1\leq\pi_2$ or $\pi_l=\pi_2\leq\pi_1$. In sum, the low type will never contribute such that his payoff will be lower than the payoffs of both high types.
Case 2: $\pi_1<\pi_l<\pi_2$ or $\pi_2<\pi_l<\pi_1$
If the low type's payoff is larger than the payoff of one high type, but still smaller than the other high type's payoff, the derivative of the utility function is given by $\frac{\partial U_l}{\partial c_l}=-1+\gamma_l-1/2(\alpha_l-\beta_l)(1+\Delta\gamma)$. This derivative is strictly negative, as disadvantageous inequality is assumed to affect utility at least as strong as advantageous inequality ($\alpha_i\geq\beta_i$), $\gamma_l<1$, and $\Delta\gamma\geq0$. The low type will decrease his contribution until his payoff equals the payoff of the better off high type. Intu-

itively, by reducing his contribution the low type will increase his own monetary payoff and simultaneously decrease disutility from disadvantageous inequality at a faster rate than increasing disutility from advantageous inequality.

Case 3: $\pi_l > \pi_1$ and $\pi_l > \pi_2$

Let us now assume that the payoff of the low type is strictly larger than the payoffs of both high types. The utility function that is relevant for the marginal analysis is now denoted by $U_l = E - c_l + \gamma_l(c_l + c_1 + c_2) - \frac{\beta_l}{2}(c_1 - c_l - \Delta\gamma(c_l + c_1 + c_2)) - \frac{\beta_l}{2}(c_2 - c_l - \Delta\gamma(c_l + c_1 + c_2))$. Thus, $\frac{\partial U_l}{\partial c_l} = -1 + \gamma_l + \beta_l(1 + \Delta\gamma)$, which is positive if $\beta_l > \frac{1-\gamma_l}{1+\Delta\gamma}$. If this condition is fulfilled, the low type will contribute in such a way that his payoff will equal the payoff of the high type with the lower contribution to the public good. If $\beta_l < \frac{1-\gamma_l}{1+\Delta\gamma}$, the low type does not contribute to the public good at all since he does not suffer sufficiently from advantageous inequality.

The next section analyzes **behavior of** one **high type**, player 1, given the actions of the other high type, player 2, and the low type *l*. Without loss of generality, we will only analyze the decisions of high type 1 who is representative for behavior of both high types.

Case 1: $\pi_1 < \pi_l$ and $\pi_1 < \pi_2$

If player 1 obtains the lowest monetary payoff, $U_1 = E - c_1 + \gamma_h(c_l + c_1 + c_2) - \frac{\alpha_1}{2}(c_1 - c_l - \Delta\gamma(c_l + c_1 + c_2)) - \frac{\alpha_l}{2}(c_1 - c_2)$. The derivative $\frac{\partial U_1}{\partial c_1} = -1 + \gamma_h - \alpha_1(1 - \frac{\Delta\gamma}{2})$ is always negative as $\gamma_h < 1$ and $\Delta\gamma \leq 1/2$. Thus, player 1 will never increase his contribution, but, in contrast, decrease it until his payoff at least equals the payoff of one other player. By reducing his contribution, player 1 can increase his monetary payoff and simultaneously decrease inequality.

Case 2: $\pi_l \leq \pi_1 < \pi_2$

Player 1 is worse off than the other high type, but weakly better off than the low type. As the analysis of the low type's behavior has shown, this case can never arise in equilibrium.

Case 3: $\pi_2 < \pi_1 < \pi_l$

Player 1 is better off than the other high type, but worse off than the low type. The utility function that is relevant for the marginal analysis is given by $U_1 = E - c_1 + \gamma_h(c_l + c_1 + c_2) - \frac{\alpha_1}{2}(c_1 - c_l - \Delta\gamma(c_l + c_1 + c_2)) - \frac{\beta_1}{2}(c_2 - c_1)$. Setting the derivative $\frac{\partial U_1}{\partial c_1} = -1 + \gamma_h - \frac{\alpha_1}{2}(1 - \Delta\gamma) + \frac{\beta_1}{2}$ larger than zero, results in the condition $\beta_1 > 2(1 - \gamma_h) + \alpha_1(1 - \Delta\gamma)$. If this condition is met, player 1 will match the contribution of the other high type no matter what the low type does. The low type may either choose his contribution to equalize payoffs of all three players or not contribute to the public good at all. In the following, equilibria that result in unequal payoffs will be called asymmetric.

With the parameters chosen in our experiment the condition for asymmetric equilibria is reduced to $\beta_1 > \frac{1}{2} + \frac{3}{4}\alpha_1$. This condition can never be satisfied. Consider the limiting case of $\alpha_1 \geq \beta_1$: $\alpha_1 = \beta_1$. This results in $\frac{\beta_1}{4} > \frac{1}{2}$, which cannot

hold as $\beta_i < 1$ is another assumption of the Fehr-Schmidt model.
If $\beta_1 < 2(1-\gamma_h) + \alpha_1(1-\Delta\gamma)$, player 1 will at least reduce his contribution until $\pi_2 < \pi_1 = \pi_l$.
Case 4: $\pi_1 > \pi_l$ and $\pi_1 > \pi_2$
If the payoff of player 1 is larger than the payoffs of the two other players, his utility function is now denoted by: $U_1 = E - c_1 + \gamma_h(c_l + c_1 + c_2) - \frac{\beta_1}{2}(c_l - c_1 + \Delta\gamma(c_l + c_1 + c_2)) - \frac{\beta_1}{2}(c_2 - c_1)$. The derivative $\frac{\partial U_1}{\partial c_1} = -1 + \gamma_h + \beta_1(1 - \frac{\Delta\gamma}{2})$ turns positive for $\beta_1 > \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$. This implies that for sufficiently large values of β_1, player 1 will increase his contribution to the public good until at least one other player obtains the same payoff as he does. Intuitively, a player 1 who is sufficiently averse to advantageous inequality will contribute in order to reduce inequality towards both other players. If $\beta_1 < \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$, player 1 will not contribute at all.

In the following, we summarize the resulting equilibria:
In treatment $HET-VCM$, if $\beta_h < \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$ for at least one high type player, there exists a unique equilibrium in which all players contribute $c_i = 0$.
If $\beta_h > \frac{1-\gamma_h}{1-\frac{1}{2}\Delta\gamma}$ for both high types and $\beta_l > \frac{1-\gamma_l}{1+\Delta\gamma}$ for the low type, equilibria with positive contributions exist with $c_h \in [0, E]$ and $c_l = c_h \frac{1-2(\gamma_h-\gamma_l)}{1+(\gamma_h-\gamma_l)}$ (symmetric equilibria).
If $\beta_l < \frac{1-\gamma_l}{1+\Delta\gamma}$ for the low type and $\beta_h > 2(1-\gamma_h) + \alpha_h(1-\Delta\gamma)$ for both high types, another class of equilibria with $c_i \geq 0$ exists, in which both high types contribute the same amount $c_h \in [0, E]$ and the low type contributes $c_l = 0$ (asymmetric equilibria). Proposition 2 summarizes the results using the parametrization of our experiment.

2.A.3 Two-stage treatments with voting stage and contribution stage

In all two-stage treatments, players are assumed to apply backward induction. Let U^{INST} denote utility when the institution has received unanimous support and has been implemented, with $INST = SYM$ for the symmetric and $INST = ASYM$ for the asymmetric institution. In the contribution stage, players will compare the utility they receive with the respective institution being in place, U^{INST}, to U^{VCM}, the utility of the VCM that is played if the institution has not received unanimous support in the voting stage. Whenever $U^{INST} \geq U^{VCM}$, a player will vote in favor of implementing the institution. With unanimity voting, if all other players also vote in favor of implementing the proposed institution, the institution will be implemented and the player's preferred outcome is achieved. If, in contrast, at least one other player votes against implementing the institution, the institution will not be implemented and the VCM will be played. However, the approving player is still equally well off as if he had voted against implementing the institution. Thus, unanimity voting ensures that it is always a best response to the voting behavior of the other players to vote in favor of the institution if $U^{INST} \geq U^{VCM}$. A player will never be hurt from voting for his preferred outcome no matter how the other players vote. Whenever $U^{INST} < U^{VCM}$, a player will vote against installing the institution.

2.A.3.1 Treatment $HOM-SYM$

In treatment $HOM-SYM$, homogeneous players vote on implementing the symmetric institution.

Proposition 3. *The following statements hold both for money-maximizing players and for players with social preferences. In treatment $HOM-SYM$, all players vote in favor of implementing the institution. The symmetric institution is always implemented and all players contribute according to the institutional rules, i.e., $c_i = E \forall i$.*

If players are homogeneous ($\gamma_i = \gamma$) and **money-maximizing**, they compare $U^{SYM} = \gamma n E$ to $U^{VCM} = E$ to decide on voting in favor of or against the symmetric institution that requires each player to contribute the efficient contribution level E. $\gamma n E > E$ if $\gamma > 1/n$, a condition that is always met by definition in a VCM game with homogeneous players. Consequently, with unanimity voting, all players will vote in favor of the symmetric institution.

In treatment $HOM-SYM$, assuming **social preferences** instead of pure money-maximizing does not change predictions. With homogeneous players, the symmetric institution guarantees equality of payoffs while simultaneously maximizing them. Hence again, all players are predicted to vote in favor of the symmetric institution. Formally, $U^{SYM} = \gamma n E \geq U^{VCM} = E - \hat{c} + \gamma n \hat{c}$, where $\hat{c}$ denotes contributions in equilibrium with $\hat{c} \in [0, E]$. As U^{VCM} is strictly increasing in $\hat{c}$ due to $n\gamma > 1$, the utility after successful implementation of the symmetric institution is always at least as large as the utility from the VCM. This analysis is equivalent to the one done by Gerber et al. (2013) for the 4 player case.

2.A.3.2 Treatment $HET-SYM$

In treatment $HET-SYM$, heterogeneous players vote on implementing the symmetric institution.

Proposition 4. *In treatment $HET-SYM$, money-maximizing players vote in favor of implementing the symmetric institution. The symmetric institution is always implemented and all players contribute according to the institutional rules, i.e., $c_i = E$, $i \in \{h, l\}$.*

If players are heterogeneous and **money-maximizing**, they compare $U^{SYM} = \gamma_i n E$ with $\gamma_i \in \gamma_l, \gamma_h$ to $U^{VCM} = E$ to decide on voting in favor of or against the symmetric institution that requires each player to contribute the efficient contribution level E. $U^{SYM} > U^{VCM}$ whenever $\gamma_i > 1/n$. Given the parametrization of our experiment ($\gamma_l = 1/2$, $\gamma_h = 3/4$, $n = 3$), this condition is met for both high and low type players. Consequently, all players vote in favor of the symmetric institution.

In treatment $HET - SYM$, predictions based on standard preferences and **social preferences** differ markedly. Players with standard preferences always support the formation of the symmetric institution as it offers a higher monetary payoff than the VCM and they do not suffer from inequality that arises from symmetric contributions of players with different MPCRs. In contrast, low type players with social preferences who suffer sufficiently from being worse off than the high types if the symmetric institution is implemented object to institution formation. They prefer a possibly lower payoff, but equal payoffs across players in the VCM to a higher monetary payoff, but disutility from inequality with the symmetric institution being in place.

Proposition 5. *High type players with social preferences will always vote in favor of installing the symmetric institution. In contrast, low type players with social preferences will reject the installation of the symmetric institution if they are sufficiently averse to disadvantageous inequality, more precisely, if $\alpha_l > \frac{2}{3} - \frac{4}{75}\hat{c_h}$, where $\hat{c_h}$ is the equilibrium contribution of high types in the VCM. If players have social preferences, the symmetric institution will not always be implemented.*

The proof of proposition 5 is provided below. We first analyze the **behavior of the low type**. If the symmetric institution is implemented, the low type's utility is $U_l^{SYM} = 3E(\gamma_l - \alpha_l \Delta\gamma)$. As has been shown in the previous analysis of treatment $HET - VCM$, if players have social preferences and heterogeneous MPCRs the VCM has both symmetric and asymmetric equilibria. Hence, U_l^{VCM} depends on the kind of equilibrium that is played in the VCM. If a symmetric equilibrium is played, the utility of the low type is $U_{l,sym}^{VCM} = E + c_h(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$, where c_h denotes the contribution level of the high types in the VCM. The low type will reject the symmetric institution, if $U_{l,sym}^{VCM} > U_l^{SYM}$, i.e., if $\alpha_l > \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$. If an asymmetric equilibrium is played in the VCM, utility in the VCM is $U_{l,asym}^{VCM} = E + \gamma_l 2c_h - \beta_l(1-2\Delta\gamma)$. The critical threshold for rejecting the symmetric institution is given by $\alpha_l > \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(2\gamma_l - \beta_l(1-2\Delta\gamma))$.

Next, we will analyze the voting **behavior of the high types**. Again, we must distinguish between symmetric and asymmetric equilibria being played in the VCM. If a symmetric equilibrium is played in the VCM, all players' payoffs are equal: $U_{h,sym}^{VCM} = U_{l,sym}^{VCM} = E + c_h(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$. If $U_{h,sym}^{VCM} > U_h^{SYM}$, the symmetric institution will be rejected. Setting $U_{h,sym}^{VCM} > U_h^{SYM} = 3E(\gamma_h - \frac{\beta_1}{2}\Delta\gamma)$, leads to the condition $\beta_h > \frac{2}{3}\frac{3\gamma_h-1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$. This can be interpreted as follows:

if the high types' sensitivity towards advantageous inequality and the contributions in the VCM are large enough, high types will vote against the symmetric institution to achieve an outcome with equal payoffs, rather than potentially higher, but unequal payoffs. If an asymmetric equilibrium is played in the VCM, the utility of the high types is given by $U_{h,asym}^{VCM} = E + c_h(2\gamma_h - 1 - \frac{\alpha_h}{2}(1 - 2\Delta\gamma))$. Rearranging $U_{h,asym}^{VCM} > U_h^{SYM}$ leads to $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{c_h}{E}(2\gamma_h - 1 - \frac{\alpha_1}{2}(1 - 2\Delta\gamma))$.

Let us summarize behavior of players with social preferences in treatment $HET - SYM$: If symmetric equilibria are played in the VCM, the low type votes against implementing the symmetric institution if $\alpha_l > \frac{3\gamma_l - 1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$. The high types vote against implementing the symmetric institution if $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h + \gamma_l - 1}{1 + \Delta\gamma})$. If an asymmetric equilibrium is played in the VCM, the low type rejects the institution if $\alpha_l > \frac{3\gamma_l - 1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(2\gamma_l - \beta_l(1 - 2\Delta\gamma))$, while the high types reject it for $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{c_h}{E}(2\gamma_h - 1 - \frac{\alpha_h}{2}(1 - 2\Delta\gamma))$. If the institution is not implemented, contribution levels are identical to those in the treatment $HET - VCM$. If the symmetric institution is implemented, all players contribute $c_i = E,\ i \in \{h, l\}$.

Using the parametrization of the experiment simplifies results drastically. Equilibria with asymmetric payoffs cannot arise. Low types will reject the symmetric institution if $\alpha_l > \frac{2}{3} - \frac{4}{75}c_h$. High types will reject the institution for $\beta_h > \frac{10}{3} - \frac{8}{75}c_h$. Since $\beta < 1$ by assumption of the Fehr-Schmidt model and $c_h \in [0, 20]$, this condition is never met and high types will never reject the institution.

2.A.3.3 Treatment $HET - ASYM$

In treatment $HET - ASYM$, heterogeneous agents vote on implementing the asymmetric institution.

Proposition 6. *The following statements hold both for money-maximizing players and for players with social preferences. In treatment $HET - ASYM$, all players vote in favor of implementing the asymmetric institution. The institution is always implemented and all players contribute according to the institutional rules, i.e., high types contribute $c_h = E$ and low types contribute c_l.*

If the asymmetric institution has been implemented the utility of **money-maximizing** low types is denoted by $U_l^{ASYM} = E - c_l + \gamma_l(n_1 E + n_2 c_l)$, the utility of money-maximizing high types by $U_h^{ASYM} = \gamma_h(n_1 E + n_2 c_l)$. Contribution levels in the asymmetric institution are designed to equalize payoffs across heterogeneous player types, i.e., the contribution level of the low types, c_l, is determined by $U_l^{ASYM} = U_h^{ASYM}$. Solving for c_l and restricting contributions to be non-negative results in $c_l = \max\{E\frac{1 - n_1\Delta\gamma}{1 + n_2\Delta\gamma}; 0\}$. Whenever $U_l^{ASYM} = U_h^{ASYM} >$

$U^{VCM} = E$, players vote in favor of the asymmetric institution. Inserting c_l and rearranging $U_l^{ASYM} = U_h^{ASYM} > E$ leads to $\gamma_h n_1 + \gamma_l n_2 > 1$, which is the necessary condition for public good provision to be efficient, a condition that is met by definition of the public goods game.

Both high and low type players with **social preferences** will vote in favor of implementing the asymmetric institution. Implementing the asymmetric institution guarantees both player types the highest attainable payoff among all equilibrium payoffs of the VCM and does not induce payoff inequalities. This intuitive line of reasoning summarizes the part of the formal analysis provided below that is relevant for the parameters used in the experiment. In sum, in treatment $HET - ASYM$, we will show that the low type with social preferences will only reject the asymmetric institution if an asymmetric equilibrium is played in the VCM. Otherwise, the utility level with the asymmetric institution in place represents the highest attainable equilibrium utility level and the implementation of the asymmetric institution will always be supported.

If a symmetric equilibrium is played in the VCM, the **low type** will compare the utility obtained under the asymmetric institution, $U_l^{ASYM} = U_h^{ASYM} = \gamma_h E(\frac{3}{1+\Delta\gamma})$, to the utility level in a symmetric equilibrium of the VCM, $U_{l,sym}^{VCM} = E + c_h(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$. Simplifying $U_l^{ASYM} \geq U_{l,sym}^{VCM}$ results in the condition $E \geq c_h$ that is always met. Consequently, the low type will support the implementation of the asymmetric institution. If an asymmetric equilibrium is played in the VCM, the low type will compare U_l^{ASYM} to $U_{l,asym}^{VCM} = E + \gamma_l 2c_h - \beta_l c_h(1 - 2\Delta\gamma)$. Setting $U_{l,asym}^{VCM} > U_l^{ASYM}$ and rearranging results in $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{(1+\Delta\gamma)(2\gamma_l-\beta_l)(1-2\Delta\gamma)}$, i.e., the low type will only reject the asymmetric institution if the amount contributed to the public good in the asymmetric equilibrium of the VCM is relatively large compared to the total endowment and if the conditions for the existence of an asymmetric equilibrium in the VCM are met, i.e., if $\beta_l < \frac{1-\gamma_l}{1+\Delta\gamma}$ for the low type and $\beta_h > 2(1-\gamma_h) + \alpha_h(1-\Delta\gamma)$ for both high types.

High types will always support the implementation of the asymmetric institution. First, the asymmetric institution guarantees them the highest attainable utility level in the VCM among all possible symmetric equilibria of the VCM. Second, also if an asymmetric equilibrium is played in the VCM, the high types' utility obtained under the asymmetric institution must at least be as large as the utility in every possible asymmetric equilibrium of the VCM, since the low type additionally contributes a non-negative amount to the public good and equal payoffs are ensured. Technically, let us compare the high types' utility from the asymmetric equilibrium in the VCM, $U_{h,asym}^{VCM} = E + c_h(2\gamma_h - 1 - \frac{\alpha_1}{2}(1-2\Delta\gamma))$ to the utility from the asymmetric institution $U_h^{ASYM} = E\gamma_h(\frac{3}{1+\Delta\gamma}) = E\gamma_h(2 + \frac{1-2\Delta\gamma}{1+\Delta\gamma})$. Setting $U_{h,asym}^{VCM} > U_h^{ASYM}$ results in $c_h(2\gamma_h - 1 - \frac{\alpha_1}{2}(1-2\Delta\gamma)) > E(\gamma_h(2 + \frac{1-2\Delta\gamma}{1+\Delta\gamma}) - 1)$. Since $E \geq$

c_h, $-\frac{\alpha_1}{2}(1-2\Delta\gamma) > \gamma_h(\frac{1-2\Delta\gamma}{1+\Delta\gamma})$ which can be simplified to $-\frac{\alpha_1}{2} > \frac{\gamma_h}{1+\Delta\gamma}$ must hold for $U^{VCM}_{h,asym} > U^{ASYM}_{h}$ to be true. However, $-\frac{\alpha_1}{2} > \frac{\gamma_h}{1+\Delta\gamma}$ can never be true, since the left side of the inequality is negative, while the right one is positive. Consequently, high types will always vote in favor of implementing the asymmetric institution.

For the parameters used in the laboratory experiment asymmetric equilibria in the VCM do not exist. The only source of rejecting the asymmetric institution is eliminated and all players are predicted to vote in favor of implementing the asymmetric institution.

2.A.3.4 Treatment $HOM-ASYM$

In treatment $HOM-ASYM$, homogeneous players vote on implementing the asymmetric institution.

Proposition 7. *For money-maximizing players, it is a weakly dominant strategy to vote in favor of implementing the asymmetric institution in treatment $HOM-ASYM$. The asymmetric institution is always implemented and all players contribute according to the institutional rules.*

The asymmetric institution obliges n_1 players to contribute their whole initial endowment E, while the other n_2 players are obliged to contribute only $\bar{c} < E$ with $n_1 + n_2 = n$. **Money-maximizing players** who are obliged to only contribute $\bar{c}$ will vote in favor of the asymmetric institution because it will increase their earnings: $U^{ASYM}_{\bar{c}} = E - \bar{c} + \gamma(n_1 E + n_2 \bar{c}) > E$, the payoff in the VCM, since $\bar{c} < E$ and $\gamma(n_1 + n_2) > 1$. However, for players who are obliged to contribute E, the formation of the asymmetric institution does not pay off when the share of players with low contributions gets too large or these players' contribution level $\bar{c}$ gets too small. Their payoff from the asymmetric institution is denoted by $U^{ASYM}_{E} = \gamma(n_1 E + n_2 \bar{c})$. Only if $\gamma(n_1 E + n_2 \bar{c}) > E$, it is a weakly dominant strategy for all players to support the installment of the asymmetric institution. For the parametrization of our experiment ($\gamma = 2/3$, $n_1 = 2$, $n_2 = 1$, $E = 20$, and $\bar{c} = 8$), this is indeed the case.

Players with **social preferences** will reject the asymmetric institution if the inequality introduced by the asymmetric institution outweighs its monetary gains. The utility of the two players who contribute fully is denoted by $U^{ASYM}_{E} = \gamma(n_1 E + n_2 \bar{c}) - \frac{\alpha_i}{2}(E - \bar{c})$, the utility of the player who contributes $\bar{c}$ is given by $U^{ASYM}_{\bar{c}} = E - \bar{c} + \gamma(n_1 E + n_2 \bar{c}) - \beta_i(E - \bar{c})$. For three players with social preferences, $U^{VCM} = E - \hat{c} + \gamma(n_1 + n_2)\hat{c}$, where $\hat{c}$ denotes the equilibrium contribution to the public good in the VCM. Comparing U^{ASYM}_{E} to U^{VCM} shows that the players contributing fully will reject the asymmetric institution if $\alpha_E > 2\frac{(\hat{c}-E)+\gamma(n_1 E+n_2\bar{c}-(n_1+n_2)\hat{c})}{E-\bar{c}}$, while the player contributing $\bar{c}$ will reject it if $\beta_{\bar{c}} > \frac{(\hat{c}-\bar{c})+\gamma(n_1 E+n_2\bar{c}-(n_1+n_2)\hat{c})}{E-\bar{c}}$. Inserting the experimental parameters, the two conditions simplify to $\alpha_E > 2 - \frac{\hat{c}}{6}$ and $\beta_{\bar{c}} > 2 - \frac{\hat{c}}{12}$. These inequalities also show that the asymmetric institution is more attractive if equilibrium contributions in the VCM $\hat{c}$ are low.

Proposition 8. *Players with social preferences who are obliged to contribute fully will vote against the asymmetric institution if $\alpha_{\bar{c}} > 2 - \frac{\hat{c}}{6}$, while players who are obliged to contribute $\bar{c}$ will vote against the asymmetric institution if $\beta_E > 2 - \frac{\hat{c}}{12}$. Hence, if homogeneous players have social preferences, the asymmetric institution will not always be implemented.*

2.A.4 Behavioral predictions

The results section is structured along five sets of predictions and corresponding results concerning differences in voting and contribution behavior across treatments. The predictions are listed below. They build on the theoretical predictions for the different treatments. We focus on predictions that are based on treatment comparisons in which changes in behavior can be attributed to a single change in setup (either a change in the institution or a change in the composition of player types). Moreover, the predictions are based on three assumptions. First, we assume that at least some players are inequality averse to an extent that induces their behavior to deviate from the predictions based on standard preferences. Second, when comparing two-stage treatments to the corresponding baseline VCMs, we assume that whenever an institution is rejected in the voting stage of a two-stage treatment, subjects play the equilibrium in the VCM of the contribution stage that the same group of subjects would play in the baseline VCM. Finally, for each of the two treatments HOM-VCM and HET-VCM, all possible equilibria can be ranked according to efficiency on a continuous scale from 0 to 1. When comparing the two baseline VCMs in treatments HOM-VCM and HET-VCM, we assume that the same group of subjects would play the equilibrium of the same efficiency rank in treatment HOM-VCM and HET-VCM, e.g., a given group of subjects that chooses the most efficient equilibrium in treatment HOM-VCM, would also choose the most efficient equilibrium in treatment HET-VCM. This results in the following predictions concerning treatment comparisons:

Prediction 1:

> Average contributions in treatment HET-VCM are lower than in treatment HOM-VCM.

Prediction 2:

> In treatment HOM-SYM, average contributions are (weakly) higher than in treatment HOM-VCM.

Prediction 3:

a) In treatment HET-SYM, average contributions are (weakly) higher than in treatment HET-VCM.

b) In treatment HET-SYM, both implementation rate and average contributions are lower than in treatment HOM-SYM.

Prediction 4:

a) In treatment HET-ASYM, the implementation rate is higher than in treatment HET-SYM. There is no unambiguous prediction whether average contributions are higher in treatment HET-ASYM or in treatment HET-SYM.

b) In treatment HET-ASYM, average contributions are (weakly) higher than in treatment HET-VCM.

Prediction 5:

a) There is no unambiguous prediction whether average contributions are higher in treatment HOM-ASYM or in treatment HOM-VCM.

b) In treatment HOM-ASYM, the implementation rate is lower and average contributions are (weakly) lower than in treatment HOM-SYM.

c) In treatment HOM-ASYM, the implementation rate is lower than in treatment HET-ASYM. There is no unambiguous prediction whether average contributions are higher in treatment HOM-ASYM or in treatment HET-ASYM.

Appendix 2.B Contributions Without Institutions

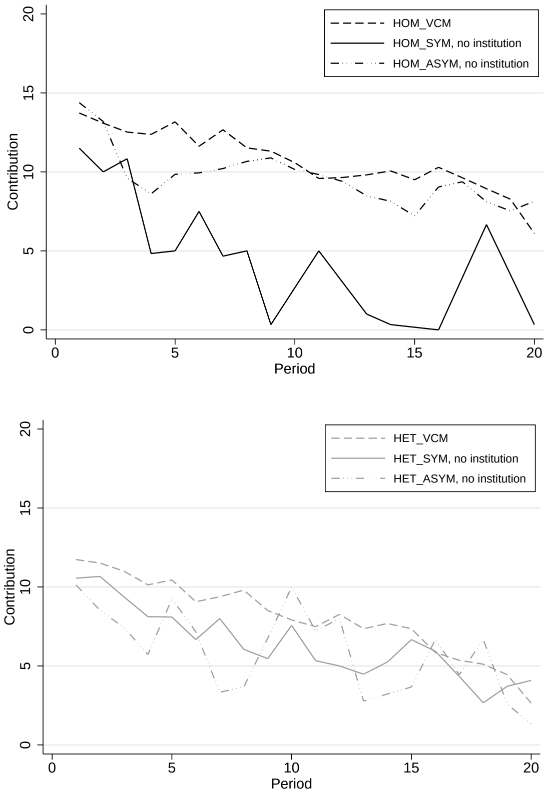


Figure 2.B.1. Contributions in Case of No Institution over Time

Appendix 2.C Instructions

The instructions below are translations of the German instructions for the experiment. The instruction are for treatment $HET-ASYM$. Instructions for the other treatments were as similar as possible except for the necessary adjustments concerning the composition of types (in treatments with homogeneous players), the level of obligations (in treatments with the symmetric institution), and the omittance of the first stage in the baseline VCM treatments.

General instructions for the participants

You are now participating in an economic experiment. If you read the following explanations carefully, you will be able to earn a considerable amount of money – depending on your decisions and those of the other participants. Thus it is very important to read these instructions carefully and to understand them.

During the experiment, it is absolutely prohibited to communicate with the other participants. If you have any questions, please ask us: please raise your hand and we will come to your seat. If you violate this rule, you will be dismissed from the experiment and forfeit all payments.

How much money you will receive after the experiment depends on your decisions and those of the other participants. During the experiment, payoffs will be calculated in Taler instead of Euro. Your total income will be calculated in Taler first. The total amount of Taler that you have accumulated during the experiment will be converted into Euro and paid to you in cash at the end of the experiment. The exchange rate from Taler to Euro is as follows:

$$40 \text{ Taler} = 1 \text{ Euro}$$

The experiment consists of exactly one part. This part is divided into **20 periods**. At the beginning of the experiment you are randomly assigned to a group of three. Thus, there are two other participants in your group. In each group of three, there are **two participants of type A** and **one participant of type B** (the difference between type A and type B will be explained in detail shortly). Whether you are of type A or of type B is determined randomly. **In all periods your type remains the same, just as the types of the other participants in your group remain the same**. You will be interacting with the same two participants in all periods. Neither during, nor after the experiment will you receive any information about the identities of the other participants in your group.

Detailed Information about the Course of each Period

Each period is divided into three stages:

1. In the **second stage** you have to decide on how many Taler you contribute to a project and how many Taler you keep for yourself.

2. In the **first stage** you can decide if you want to commit yourself and the other participants in your group to certain contributions to the project in stage 2. Only if **all** participants decide in stage 1 to commit all participants in your group to certain contributions to the project, the contributions will actually be fixed. If not all participants decide to fix the contributions, then you and the other participants in your group will be able to choose any contribution level in the second stage.

3. In the **third stage** you get to know the contributions of all participants in your group to the project in stage 2 and the payoffs of all participants in your group in this period.

At the beginning of each period every participant receives **20 Taler**. In each period you have to decide on how to use these 20 Taler. You can contribute Taler to a **project** or put them on a **private account**. Every Taler that you don't contribute to the project is automatically put on your private account.

Income from your private account:
For each Taler you put on your private account, you earn exactly one Taler. For example, if you put 20 Taler on your private account (thus contributing zero Taler to the project), you would earn 20 Taler from your private account. If, e.g., you would put 2 Taler on your private account (thus contributing 18 Taler to the project), your income from the private account would be 2 Taler. Nobody but you receives Taler from your private account.

Income from the project:
For each Taler that you or another participant in your group contributes to the project, you (and each other participant in your group) earn a certain number of Taler. Each participant's income from the project depends on his or her type and is determined as follows:

$$\text{\underline{\textit{Type A}}\textit{'s income from the project}} = \tfrac{3}{4} * \textit{sum of all contributions to the project}$$

$$\underline{\textit{Type B's}}\ \textit{income from the project} = \tfrac{1}{2} * \textit{sum of all contributions to the project}$$

Example 1: The sum of contributions from all participants to the project is 12 Taler (e.g., if you and the two other participants contribute 4 Taler each, or if one of the three participants contributes 12 Taler and the two other participants contribute 0 Taler). Then the two participants in your group who are of type A each receive an income of $\frac{3}{4}$ * 12 = 9 Taler from the project, and the participant in your group who is of type B receives an income of $\frac{1}{2}$ * 12 = 6 from the project.

Example 2: The sum of contributions from all participants to the project is 36 Taler. Then the two participants in your group who are of type A each receive an income of $\frac{3}{4}$ * 36 = 27 Taler from the project, and the participant in your group who is of type B receives an income of $\frac{1}{2}$ * 36 = 18 from the project.

Income at the end of a period:
Your income at the end of a period is the sum of your income from your private account and your income from the project:

Type A:
Income from the private account (20 – contribution to the project)
*+ Income from the project ($\frac{3}{4}$ * sum of contributions to the project)*
= Income at the end of the period

Type B:
Income from the private account (20 – contribution to the project)
*+ Income from the project ($\frac{1}{2}$ * sum of contributions to the project)*
= Income at the end of the period

Let us illustrate how your income at the end of a period is calculated using two examples:

Example 1: Assume that you are of type A and contribute 16 Taler to the project, just as the other two participants. The sum of contributions is then 16 + 16 + 16 = 48 Taler. Your income in this example would be:

4 Taler from the private account + $\frac{3}{4}$ * 48 Taler from the project = 4 + 36 = 40 Taler

Example 2: Assume that you are of type A and contribute 0 Taler to the project, while the other two participants contribute 16 Taler each. The sum of contributions is then 16 + 16 + 0 = 32 Taler. Thus, your income would be:
20 Taler from the private account + $\frac{3}{4}$ * 32 Taler from the project = 20 + 24 = 44 Taler

The first stage

In the **first stage** you can decide whether you want to commit yourself and the other participants in your group to a certain contribution to the project in the second stage. All participants decide simultaneously. Only if **all** participants in your group decide to commit themselves and the other participants to certain contributions, are the contributions in stage 1 actually fixed. In this case contributions will be fixed as follows:

***Type A:** Contribution of 20 Taler to the project*

***Type B:** Contribution of 8 Taler to the project*

If **not all** participants decide to fix the contributions, you and the other participants in your group can freely contribute any number of your 20 Taler to the project in the second stage.

The second stage
At the beginning of the second stage you get to know how each participant in your group decided in the first stage.
If in the first stage all participants decided to fix the contributions in the second stage, then in the second stage you have to contribute the corresponding amount. Thus, if you are of type A you have to enter a contribution of 20 Taler and if you are of type B you have to enter a contribution of 8 Taler. Other inputs are not possible and will automatically be adjusted by the computer program.
In this case the period income of the participants of type A is $\frac{3}{4} * 48 = 36$ Taler each and the period income of the participant of type B is $12 + \frac{1}{2} * 48 = 36$ Taler.
If in the first stage not all participants decided to fix the contributions in the second stage, then in the second stage all participants can freely choose any integer contribution between 0 and 20 to the project (0, 1, 2, . . . , 19, 20).
In this case your period income is computed as indicated above:

Type A: 20 – your contribution to the project + $\frac{3}{4}$ * (sum of all contributions to the project in your group)
Type B: 20 – your contribution to the project + $\frac{1}{2}$ * (sum of all contributions to the project in your group)

The third stage
In the third stage you get to know the contributions to the project by all participants in your group, as well as their period income. Furthermore, you will again see how each participant in your group decided in the first stage.
Then the current period ends and the next period begins with the same participants. Your type and the types of the other participants remain the same. All participants can then again decide in the first stage whether they want to fix contributions in the second stage. Again, the second stage follows and finally the third stage.

Conclusion of the experiment and payment

The experiment ends after 20 periods. Subsequently, we will ask you to answer a few general questions on the computer. Your answers to these questions have no influence on how much money you earn in the experiment. When all participants have filled out the questionnaire, payments will be made. Your total income from the 20 periods will be converted into Euro and paid to you in cash.

Do you have any questions? If so, please raise your hand.

3

Cooperation and Redistribution: Does “bundling” foster institution formation? *

3.1 Introduction

When it comes to cooperation problems, the formation of a “governing institution” provides a promising means to overcome the problem.[1] However, potential tensions arise if the returns from cooperation are not evenly distributed among participating parties (e.g., Reuben and Riedl, 2013; Kube et al., 2015). In particular if neither side payments nor renegotiation are possible, parties profiting less from successful cooperation might be inclined to reject institutions or bills that favor others disproportionally.[2] Such concerns might be alleviated in the presence of formal redistribution rules. Yet, the implementation of these rules could present a challenge in itself, because not all parties benefit to the same extent from redistribution. In this paper, we shed light on this issue by exploring procedures under which parties might potentially agree to implement formal redistribution rules.

To this end, we take a prominent social dilemma as the underlying cooperation problem and compare the effectiveness of two different procedures for deciding over the implementation of a redistribution rule: either a bundled or a separate approach. In the bundling approach, redistribution forms an integral

[*] We thank audiences at the EEA 2015, M-BEES 2015, the ETH Workshop on Social Norms and Institutions 2015 and Norms Actions and Games 2016 for their helpful comments.

[1] Throughout the paper, “forming a governing institution” is meant as parties agreeing on a set of rules (e.g., laws) that are backed up by deterrent sanctions.

[2] Examples for this behavior are legion. The behavior of the republican party during the financial shutdown 2013 can be thought of as prime example: willingly accepting the harsh consequences – also for their own party – in order to prevent the current administration from gaining large credit for their ability to avoid the fiscal cliff.

part of the governing institution and parties vote over the *joint* implementation. If they vote against implementation, neither a governing institution nor a formal redistribution rule will be in place. Contrarily, in the separate approach governance and redistribution are *separately* available and parties have two distinct votes at their disposal. Thus, in addition to the potential outcomes under the bundling approach of having both redistribution and governance in place, or neither of them, it is also possible that only the formal redistribution rule or only the governing institution is implemented. As such, the separate approach is more flexible in that it allows to foster efficiency-enhancing cooperation via the governing institution even if the involved parties dislike redistribution (or, vice versa, to have redistribution rules in place without modifying parties’ choice sets through governance). Yet, this comes at the cost of introducing potential frictions if the rejection of one proposal lowers the acceptance and induces rejection of the other proposal.

Based on a theoretical model with social preferences, we show that behavior under these two approaches might indeed differ. Of course, whether there are environments where the outcomes actually do differ is ultimately an empirical question. We use the controlled environment of laboratory experiments to test for this. As in our model, the lab experiments allow us to vary the implementation approach while keeping constant all other relevant aspects of the environment. The workhorse that we use for our design is a linear public-good game with a voluntary-contribution mechanism, as it is very frequently used in the related strand of literature. Each player decides how to allocate a given monetary endowment between a private good and a public good. Public-good provision is socially efficient, but for each player the individual marginal return rates from the private and public good are such that there exist free-riding incentives. Within this framework, we test two institutions that might potentially mitigate the inherent cooperation problem, namely a governing institution and a formal redistribution rule. The governing institution (if implemented) commits all players to contribute their entire endowment to the public good. Yet, players are heterogenous in their returns from the public good, which creates a tension between equality and efficiency. This is addressed by the formal redistribution rule. If implemented, it ensures equality in payoffs among all players at all times, simply by redistributing payoffs from players that have a high return or low contributions from the public good to those with a low return or high contributions.

Players decide on the implementation prior to the public-good stage. In line with previous literature (e.g., Kosfeld et al., 2009; Gerber and Wichardt, 2009; Potters et al., 2005) we use the unanimity voting rule for the institution selection process. It ensures that players cannot be governed by institutions against their will.[3] Between treatments, we vary the (combination of) institutions that are available to the players. In our main treatments, there is both a formal redistribution rule and a governing institution. They are bundled in treatment Bun, i.e., players cast a single vote for or against their *joint* implementation. Contrarily, redistribution and governance are available *separately* in treatment Sim and players cast two individual votes, i.e., one per institution. We compare the outcomes between these two approaches, but also relate them to three potential benchmarks: no institution available at all (treatment Vcm), only redistribution (treatment Re), or only governance (treatment Fix).

We find that both institutions are potentially able to increase cooperation rates compared to the baseline Vcm, but cooperation rates and, in particular, implementation rates differ significantly between treatments. If only a single institution is available (Re or Fix), the formal redistribution rule is implemented more frequently than the governing institution (79% versus 56%). Given that any institution that is able to eliminate the social dilemma should receive unanimous support from purely selfish players, the result suggests that behavior (at least in parts) is driven by equality considerations.

Interestingly, this seems to create additional tensions when both redistribution and governance are separately available (Sim). Not only is there a drop in the implementation rate of the governing institution (implemented in 66% of all cases), but also the redistribution rule is implemented in only 43% of all cases. 26% of all situations result in total failure to implement any institution. By contrast, the bundled approach, where governance and redistribution are jointly available and players cast a single vote (Bun), induces the highest implementation rate (87%). Strikingly, we observe in a subsequent, independent decision experiment that the majority of subjects seem to dislike being restricted in their choice set. When being asked which regime they prefer, 60% of people indicate that they themselves would prefer the situation where the two decisions are not interlinked but separately available (Sim instead of Bun).

Taken together, our findings suggest that if returns from cooperation are heterogeneous, redistribution might generally serve as a way to mitigate the cooperation problem. They also stress – and this is the more important and novel contribution of our paper – that the exact details of the implementation approach can make a difference. If parties decide on governance and redistribution separately, lacking commitment to either one might render the process of institution formation difficult. Thus, it might be worthwhile to restrict parties' choice set by taking a bundled approach; i.e., to link the decision on the governing institution with the decision on redistribution.

[3] As such, we consider it to be a natural starting point for our setup. It mimics that there are sovereign players facing a social dilemma that try to overcome this initial, lawless state of nature by forming an institution. Furthermore, the unanimity rule can be understood as a "stress test" since already a single player's veto is sufficient to reject an institution.

This might explain why we observe many instances in natural environments where different decision problems are intertwined although one could easily decide on them separately. For example, the practice to bundle different bills is extremely common in the Congress of the United States.[4] In fact, outside the domains of social dilemmas and redistribution there is a whole strand of literature that examines whether the restriction of players' choice sets is able to align conflicting interests among heterogeneous players. Such mechanisms are designed by Jackson and Sonnenschein (2007) and Hortala-Vallve and Llorente-Saguer (2010). They restrict the number of votes that players can cast on different initiatives in order to implement the ex-ante socially efficient solution. The efficiency of this mechanism is supported empirically by Engelmann and Grimm (2012). Casella and Gelman (2008) include the intensity of preferences by allowing for a limited number of additional votes for every agent. However, both the choice of the voting mechanism and the items on a ballot grant substantial power over the final outcome to the legislator (Romer and Rosenthal, 1978), which might in turn be responsible for a surprisingly low match between voters' preferences and outcomes in referenda (Romer and Rosenthal, 1979). This negative view on the outcome of referenda stands in contrast to Lupia and Matsusaka (2004) and Matsusaka (2010) which find that the availability of referenda leads to a better representation of voters' preferences in passed bills and less pronounced money-power-money relationships; or Torgler (2005), who claims positive spillovers of referenda to other social decisions (e.g., tax moral).

Our paper also ties into the literature on (endogenous) institution formation and social dilemmas. Over the last decade, a broad literature has been established on institutions that are designed specifically to overcome social dilemmas; with some focusing on decentralized institutions (e.g., Ostrom et al., 1992; Gächter and Fehr, 2000), and others focusing on centralized institutions (e.g., Falkinger et al., 2000; Kosfeld et al., 2009; Gerber and Wichardt, 2009). Most of these articles study behavior in settings where the institution is exogenously given, i.e., if it is already in place. Within this subset of the literature it has been shown that redistribution of profits, be it carried out centrally (Falkinger et al., 2000) or by participating subjects (Sausgruber and Tyran, 2007), is able to overcome the social dilemma if players' returns from cooperation are homogeneous. Our paper extends on this by exploring the case of heterogenous returns from cooperation. Moreover, we focus on the endogenous implementation of redistribution. In a similar vain, other articles have studied the process of institution formation, e.g., by explicit voting mechanisms (Ertan et al., 2009; Tyran and

[4] For instance, each year congress is supposed to adopt twelve bills in order to finance the operations of government, which are not necessarily interrelated. However it is also possible to bundle several or even all bills into one "omnibus" spending bill. This option has been found to be much more likely to pass and to receive fewer amendments (Hanson, 2014) than individual bills. This method was just recently used for the fiscal year 2014.

Feld, 2006; Markussen et al., 2013; Putterman et al., 2011; Ones and Putterman, 2007), by "voting by feet" (Gürerk et al., 2009; Ahn et al., 2008) or by allowing participants to delegate their decisions to a central authority (Hamman et al., 2011; Fleiß and Palan, 2013), but did only focus on governing institutions because agents were homogenous (Gerber et al., 2013; Fischer and Nicklisch, 2007). Or they did study the case of heterogenous agents, but did not focus on redistribution (e.g., Kube et al., 2015; Fisher et al., 1995; Reuben and Riedl, 2013). To the best of our knowledge, the only other study that has looked at the endogenous formation of formal redistribution rules in the presence of heterogenous agents is Kesternich et al. (2014). In line with our findings from treatment Re, they also find that redistribution can be an effective means to overcome cooperation problems. Yet – and this is the novel aspect where our paper takes the existing academic discussion on redistribution further – our findings show that the actual impact of a formal redistribution rule depends on the details of the approach that is used for deciding over the implementation of the redistribution rule.

The outline of the paper is as follows. Section 3.2 describes the experimental design. In Section 3.3 the theoretical predictions for subjects' behavior will be derived, using both standard and social preferences (Fehr and Schmidt, 1999). Section 3.4 presents and discusses the empirical results. Section 3.5 concludes.

3.2 Experiment

In natural environments, the complexity of the process of institution formation makes it particularly difficult to draw causal conclusions about the conditions under which institutions come into being. As a starting point, we therefore use the controlled environment of laboratory experiments to study the endogenous formation of institutions. In this section, we present the design of our laboratory experiment and describe the implemented procedures.

3.2.1 Experimental Design

The basic game underlying our experiments is a standard public-goods game (VCM game), as it is frequently used in the literature to study elements of social dilemmas in the lab. In the game, each player has a private endowment E. Players simultaneously decide on the amount c_i that they want to contribute to a public good, with $0 \leq c_i \leq E$, $i = 1, ..., n$. The returns from the public good are enjoyed by all players, independently of their individual contribution c_i. In some treatments, players are heterogeneous, i.e., not all players benefit from the public good to the same extent. To model heterogeneity, we allow the marginal per capita return (MPCR) γ_i from the public good to vary across players. Given the contributions of all players $(c_1, ..., c_n)$, player i's material payoff π_i is thus given by

$$\pi_i \;=\; E - c_i + \gamma_i \sum_{i=1}^{n} c_i. \tag{3.1}$$

In all treatments, parameters for γ_i are chosen such that a social dilemma arises. Efficiency, defined as the maximized sum of payoffs of all players, is reached if all players contribute their entire endowment. Yet, from an individual perspective, each player's material payoff is maximized by not contributing to the public good, given any set of contributions by the other players. Formally, this implies $\sum_{i=1}^{n} \gamma_i > 1$ and $\gamma_i < 1 \quad \forall i$.

Before the game starts, groups consisting of three players are formed. In each period of the game, each player receives an endowment of $E = 20$. Within every group there are two types of players, which vary in their return from the public good γ_i (MPCR). Each group consists of two subjects with a high return from the public good of $\gamma_h = 0.75$ and one subject with a low return of $\gamma_l = 0.5$. Consequently the overall return from the public good is equal to 2. A players' type stays constant throughout the entire game.

Except for the baseline treatment (Vcm), all treatment conditions feature an additional institution formation stage that takes place before players make their actual contribution decision to the public good. In this first stage the players decide on the implementation of one or two institutions that govern the contribution stage. Implementation of the institution is based on the unanimity rule, i.e., an institution is only implemented if all three players agree to implement it. Voting and implementation of the institutions are assumed to be costless.[5]

There are two kinds of institutions employed in the different treatments, a governing institution and a formal redistribution rule. The **governing** institution prescribes the contributions made by the agents to the public good. In order to reach efficiency, obligations are set to the maximum amount for all agents, i.e., their entire endowment E.[6] Thus, if the governing institution is implemented, all contributions are fixed during the ensuing VCM to $c_i = E = 20$. The formal **redistribution** rule focuses on the distribution of the profits from the public good after all contributions are made. If implemented, the institution will ensure equal total payoffs for all parties involved; irrespectively of their individual contribution to the public good. This is done by redistributing payoff from players with high payoffs to those with lower payoffs, which is automatically done by

[5] In general the theoretical predictions are not affected by costs, as long as they do not outweigh the gain provided by the corresponding institution.

[6] This implies that the governing institution not only sets each player's obligations, but also installs a deterrent sanctioning technology to enforce the required contributions. We follow the established approach of previous papers to not explicitly model this part of the institution as to focus on our main effects.

the computer at the end of a period (if the redistribution rule was implemented in that particular period).

The availability of institutions is varied across treatments. In the benchmark case (treatment Vcm) no institution at all is available and only the regular public-good game is played. The second type of treatment employs one of the described institutions separately. In treatment Fix, only the governing institution is available. If the institution is adopted, it forces all players to contribute their entire endowment of 20 tokens. This results in payoffs of 45 tokens for the high-type players with $\gamma_h = 0.75$ and of 30 tokens for the low-type player with $\gamma = 0.5$. In treatment Re, only the redistribution rule is available. If adopted, all players receive the same payoff independently of their contributions and their individual type.[7]

In treatments Sim and Bun, governance and redistribution are both available at the same time, but the approach to decide on them differs between treatments. In treatment Sim, both institutions are separately available and players have to cast two distinct votes (one per institution). An institution that receives unanimous support is installed with the same consequences as above, regardless of the voting outcome on the other institution. Hence, the ensuing VCM can have 4 different states: (1) the contributions are fixed but payoffs are not redistributed, (2) contributions can be chosen by the players, but payoffs are redistributed, (3) fixed contribution and redistributed payoffs or (4) just the regular VCM. This implies that the resulting payoffs in (1) are as in Fix, in (2) as in Re, and in (4) as in Vcm. In (3), the adoption of both institutions leads to a payoff of 40 tokens for each player. By contrast, in treatment Bun both institutions are bundled such that players cast only a single vote for or against the joint implementation of redistribution and governance. Hence players have to support either both of them or neither. Consequently, if all players in a group affirm the bundle with their vote, each of them receives a payoff of 40 tokens.

3.2.2 Procedures

The experiment was computerized by using z-Tree (Fischbacher, 2007) and conducted at the BonnEconLab at the University of Bonn in January and June 2012. Students were recruited from all majors using Orsee (Greiner, 2015). In order to keep the results of the different treatments as comparable as possible, all setup details and parameters were kept constant throughout the experiment.

[7] The following numerical example illustrates this mechanism: The player with $\gamma = 0.5$ contributes 15 tokens and one of the players with $\gamma = 0.75$ contributes 5 tokens to the public good, while the other high-type player contributes 0 tokens. If the redistribution rule is not implemented, the payoffs would be 15, 30 and 35 tokens. If distribution is implemented, it distributes the total return from the public good (in this example 40 tokens) such that every subject receives the same payoff (taking the remaining endowment into account). Here the total payoff is 80 tokens. Consequently, if the redistribution rule is implemented, each subject would receive 26.66 tokens.

We ran two sessions per treatment with 24 participants per session.[8] In total 237 subjects participated. The subjects were randomly allocated into groups of three, resulting in 16 independent observations per treatment (resp. 15 in Vcm). Interaction took place within the same group of three players (partner-matching protocol), but it was anonymous and decisions were taken in private at the computer. The experiment consisted of 20 identical rounds. After each round the subjects were informed about the voting decisions and contributions of the other two players. Written instructions were distributed prior to the experiment and read out aloud. Afterwards subjects had the possibility to ask questions for clarification and had to answer several control questions. Throughout the entire experiment tokens were used as artificial currency, with 40 tokens equalling 1 Euro. The average payment for the subjects ranged between 14.03 Euro in the VCM treatment and 19.22 Euro in treatment Bun. The payments were made privately in cash directly after the experiment. Each session lasted about 90 minutes.

3.3 Behavioral Predictions

For each treatment, we characterize players' equilibrium behavior under two alternative assumptions concerning the shape of the utility function. First, we assume standard risk-neutral agents, i.e., each player's utility function coincides with the monetary payoff of the game, π_i. Second, we generalize this framework to include potential social preferences: in addition to valuing own monetary payoff, players might suffer from inequality in monetary payoffs between themselves and other players (inequality-aversion as in Fehr and Schmidt (1999)). In the remainder of this section, we will provide an intuition for the behavioral predictions for each treatment under the two alternative assumptions on the shape of players' utility functions using the parameters of our design. More general proofs are provided in Appendix A.

Table 3.1 summarizes the behavioral predictions for players with standard preferences. In any regular VCM game, players with standard preferences are predicted not to contribute to the public good at all. Whenever $\gamma_i < 1$, contributing does not pay off from an individual perspective.

In every treatment featuring an institution the players are assumed to apply backward induction. If the institution has not been implemented in the voting stage, the players on the contribution stage are back in the regular VCM game analyzed above. They are predicted not to contribute to the public good. Therefore, each player's monetary payoff will be equal to the initial endowment of 20 to-

[8] Note that one session in treatment Vcm consists of only 21 participants because some subjects did not show up for the experiment. Moreover, note that the data that we use for additional illustrative purposes from treatments Vcm and Fix are also used in Kube et al. (2015) (all being conducted using the same subject pool, experimental procedures, lab, and instructor).

Table 3.1. Behavioral Predictions Based on Standard Preferences

	VCM	FIX	RE	SIM	BUN
voting	-	implement institution	implement institution	implement 1 of the 2 institutions	implement both institutions
contribution	$c = 0$	$c = 20$	$c = 20$	$c = 20$	$c = 20$

kens. The player will be supporting an institution whenever the utility obtained under the institutional regime is larger than without ($U(INST) \geq U(VCM)$).

In the case of the governing institution, subjects are obliged to contribute their entire endowment. The resulting payoffs would be 45 tokens for the player with a high return and 30 tokens for the player with a low return. Compared to the outcome under the regular VCM, the institution increases the monetary payoffs. Hence in treatment FIX, the institution is unanimously supported in equilibrium.

In treatment RE, the formal redistribution rule (if implemented) changes the payoff function of the ensuing VCM to

$$\pi_i = \frac{1}{3}(60 + \sum_{j=1}^{3} c_j). \tag{3.2}$$

Consequently, under redistribution every player's payoff is increasing in her own contribution. In equilibrium, this institution will incentivize every player to contribute the entire endowment of 20 tokens to the public good in order to maximize the payoff. Thus, in equilibrium the payoff with institution is 40 tokens for every agent and the players will support it in the voting stage.

The same is true for treatment BUN. The combination of governing institution and formal redistribution rule will be unanimously supported, which results in payoffs of 40 tokens for every participating player.

The case of the separate approach in treatment SIM is more complex. As we just saw, each institution on its own will increase each players earnings. Consequently, at least one institution will always be used. Yet, both institutions at the same time will never be used. If all agents vote for the governing institution and fix contributions, the players with the high return from the public good will reject redistribution in order to avoid being stuck with 40 tokens instead of 45. If, however, the low type is rejecting the initiative to fix contributions, the best response of the high types is to support the redistribution. Consequently, if both institutions are potentially available and can be implemented separately (SIM), either the contributions will be fixed (governance in place) or the payoffs will be redistributed – but never both. Assuming successful coordination on either redistribution or governance, standard theory still predicts that the efficient outcome will be reached.

Table 3.2. Behavioral Predictions Based on Social Preferences

	Vcm	Fix	Re
voting	-	low type rejects if α_l high	implement institution
contribution	$(c_h, c_h, c_l = 2/5c_h)$, $c_h \in [0, 20]$ if $\beta_h > 2/7$ and $\beta_l > 2/5$; (0,0,0) otherwise	if reject: as in VCM otherwise: $c_h = c_l = 20$	$c = 20$
	Sim		Bun
voting	low type rejects Fix if α_l high high type supports Re if β_h high		implement both institutions
contribution	$c = 20$		$c = 20$

Table 3.2 displays the behavioral predictions for players with social preferences. If players have social preferences, there are multiple equilibria in the standard public good game with homogenous agents. The intuition is as follows: If all players are sufficiently averse to advantageous inequality (β sufficiently high)[9], they are all willing to exactly match any possible contribution level $c \in [0, E]$ of the other players to equalize payoffs. If players are not or only mildly averse to advantageous inequality (β low), the only equilibrium remains the one with zero contributions of all players. The same basic mechanism is also driving the existence of equilibria with positive contributions in the regular VCM with our heterogeneous players. If players are sufficiently averse towards earning more than others, they contribute positive amounts as soon as the other players contribute positive amounts, in a manner such that unequal payoff distributions are prevented (i.e., to achieve equal payoffs for all three players, the low type contributes less than the two high types).

In treatments Re and Bun, the inclusion of inequity aversion into the utility function does not change the behavioral predictions at all, because the formal redistribution rule takes care of inequity considerations. Both institutions ensure equality in monetary payoffs and they induce efficiency, either by forcing (Bun) or by incentivizing (Re) the players to contribute their entire endowment. Hence the payoff will be strictly larger than in the regular VCM, as no equilibrium with full contributions by all players exists in treatment Vcm (recall that the low type contributes less in the VCM to achieve equality in payoffs).

The predictions change in treatments Fix and Sim if players are inequality averse. Here the potential inequality that is created by fixing the contributions at the maximum without redistribution is the potential source of rejection. The low-type players that would receive only 30 tokens, whilst the others receive 45, would reject if their measure for disadvantageous disutility is large (α). Low type players might even prefer the outcome of a regular VCM without any con-

[9] In the model of Fehr and Schmidt (1999), the parameter β captures the intensity of aversion to advantageous inequality, while the parameter α measures the degree of aversion to disadvantageous inequality.

tributions over this unequal split. Consequently no institution at all might be installed in treatment Fɪx. In treatment Sɪᴍ, the additional availability of the redistribution rule could be able to mitigate this problem, but the actual prediction depends on the degree of inequality aversion. If the players are sufficiently inequality averse, they are able to overcome the social dilemma by implementing only the formal redistribution rule.[10] If players are less inequality averse, only the governing institution will be implemented. If high types are highly adverse to advantageous disutility they also prefer an equal split over an unequal one and both institutions will be implemented. In all cases, the maximum amount is contributed to the public good ensuring efficiency.

3.4 Results

The results will be presented along the lines of five central results. At large these results are in line with the behavioral predictions that were presented in the previous section, given the assumption that some players exhibit social preferences. When comparing the different treatments we will focus on differences in the contribution rates to the public good as well as the implementation rates of the respective institution(s). Additionally, we will compare the voting behavior of the different types within each treatment. All descriptive statistics are summarized in Tables 3.3 and 3.4. Figure 3.1 plots contributions over time to the public good across all treatments. Figure 3.2 displays the share of groups with an institution over time. The voting behavior of the high and low types in treatment Sɪᴍ are presented in Figure 3.3. The corresponding figures for the remaining treatments can be found in Appendix 3.B.

In order to determine the effect of a multitude of different institutions we need to understand the players' behavior under the presence of each individual institution first. Consequently, we start by describing the results of the treatments that feature either the governing institution or the formal redistribution rule. The treatments that feature both institutions at the same time follow suit. The baseline Vᴄᴍ will only serve as a baseline measure of contribution levels if agents are heterogenous in their returns from the public good.

Contribution behavior in treatment Vᴄᴍ confirms the results of Fisher et al. (1995). Players with a high return contribute 9.4 tokens on average, while low types average only 5.3 tokens. Consequently, payoffs for both types are rather similar, and there is sufficient scope for improving cooperation by implementing a governing institution, a formal redistribution rule or both.

[10] Consequently the interest of the low type would then be able to dictate the entire outcome of the game, as the claim to vote against the governing institution is entirely credible.

3.4.1 Voting on a single institution

We first consider the contribution behavior in treatment Fix. The players contribute 14.2 tokens on average. This level is significantly higher than in treatment Vcm (Mann-Whitney ranksum test (MWU) $p < 0.001$).[11] The governing institution is installed in 56% of all possible instances and drives this increase. If the governing institution is not implemented, the players contribute 8.1 (high type) and 4.2 (low type) on average, which is similar to the corresponding contribution levels in Vcm. The higher contribution rates are thus a result of successful institution formation, and not an effect of institution availability per se. In line with our prediction, we observe differences in players' voting behavior: While 95.9% of all high types support the governing institution, only 59.7% of the low types do so. The difference in support for the institution across type potentially hints at inequality consideration as a major reason for the rejection.

Result 1:

> Successful implementation of the governing institution in treatment Fix results in average contributions that are significantly higher than in treatment Vcm.

The equity concerns that cause rejections in treatment Fix are eliminated in treatment Re. This results in the institution being installed more frequently (79%) (MWU $p = 0.053$). The voting behavior across types is similar. 90% of all high types and 97% of all low types vote in favor of the institution. This indicates that the elimination of equity concerns allowed the institution to garner support across both types of players. Despite the higher support, the players contribute 15.6 tokens, which is only slightly more than in treatment Fix (MWU $p = 0.40$). The contrast of increased institution formation and almost constant contributions can be explained by the contribution behavior under redistribution. Even if the formal redistribution rule is implemented and thus payoffs are redistributed, players do not contribute their entire endowment but "only" 18.1 tokens on average. In summary, the treatments featuring the single institution confirm the intuition presented before.

Result 2:

> a) Average contribution rates in treatment Re are not significantly larger than in treatment Fix.
>
> b) The formal redistribution rule in Re receives a larger support and is implemented more frequently than the governing institution in Fix.

[11] Throughout the paper, all statistical significances are computed based on two-tailed tests on the group level.

Table 3.3. Average Contributions by Treatment

Type	Vcm	Fix	Re	Bun	Sim
Without Institution					
High	9.42	8.05	7.27	11.05	8.22
	(7.07)	(6.75)	(7.24)	(7.93)	(7.54)
Low	5.33	4.17	5.26	2.88	3.67
	(4.33)	(4.41)	(6.07)	(5.55)	(5.69)
With Institution					
High	-	20	18.18	20	19.4
	-	-	(4.4)	-	(3.04)
Low	-	20	18.0	20	19.18
	-	-	(5.04)	-	(3.34)
Combined					
High	-	14.77	15.72	18.8	16.5
		(7.42)	(7.09)	(4.21)	(6.75)
Low	-	13.07	15.43	17.7	15.16
	-	(8.38)	(7.14)	(6.18)	(7.93)
Aggregate					
	8.05	14.21	15.63	18.43	16.05
	(6.57)	(7.79)	(7.11)	(5.0)	(7.19)

Notes: Contributions are in tokens. Standard deviations are presented in parentheses.

Table 3.4. Share of Affirmative Votes and Implementation rates by treatment

Institution	Type	Fix	Re	Bun	Sim
		Affirmative Votes			
Fixed contributions					
	High	.96	-	.93	.94
	Low	.6	-	.98	.73
Redistribution					
	High	-	.9	.93	.68
	Low	-	.97	.98	.88
		Implementation Rates			
Fixed contributions		.56	-	-	.31
Redistribution		-	.79	-	.08
Both		-	-	.87	.35

Notes: For treatment Sim the implementation rates denote the shares of the specific institutional regimes that were installed.

3.4.2 Voting over both institutions

Above, we saw that the governing institution in Fix guarantees efficiency by design (if it is implemented), but the lacking equality reduces its implementa-

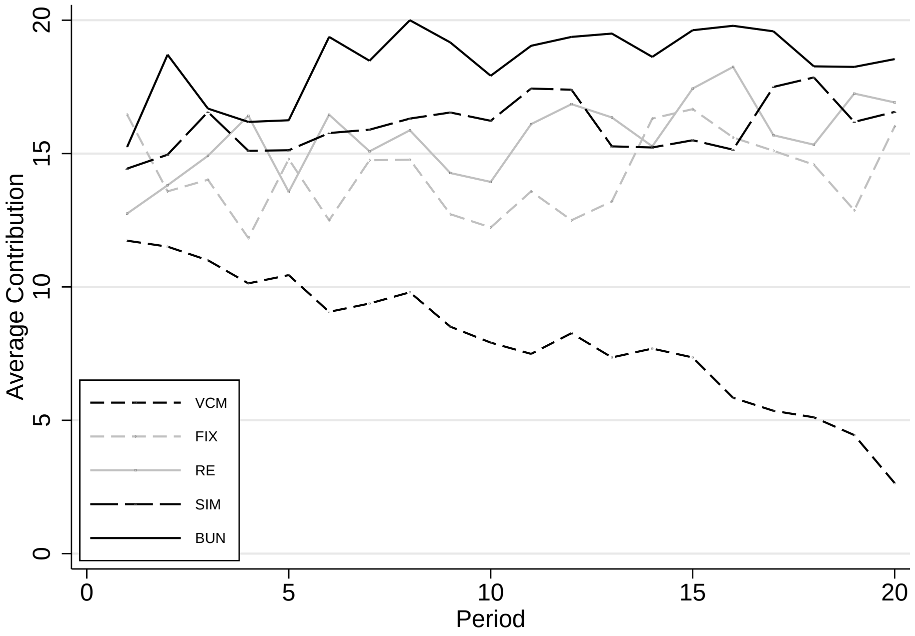


Figure 3.1. Development of Average Contributions over Time

Notes: In all treatments (except for Vcm) contributions exhibit an increasing time-trend. The trend is significant for Re (Spearman's rho $r = 0.18$, $p < 0.01$) and Bun ($r = 0.13$, $p = 0.02$). In treatments Fix and Sim they fall short of being significant (Fix: r=0.08, p=0.16; Sim: $r = 0.09$, $p = 0.12$). For Vcm, the trend is significantly negative ($r = 0.47$, $p < 0.01$).

tion rate. In Re, the forced equity garners high support among all types, but fails to ensure efficiency. In essence, the singular institutions are only able to address a singular matter of the efficiency-equity tradeoff. In the next step, we analyze whether the joint availability of both institutions will be able to handle this tradeoff successfully.

3.4.2.1 Treatment Sim

First, we look at the behavior in treatment Sim, where both institutions are available and can be implemented separately. Here the voting process can have one of four outcomes: (1) if both institutions are supported by all three players, both will be installed and efficiency and equity will be ensured simultaneously. (2) if only the governing institution receives unanimous support, the ensuing VCM is governed by the same institution as in treatment Fix; (3) if only redistribution

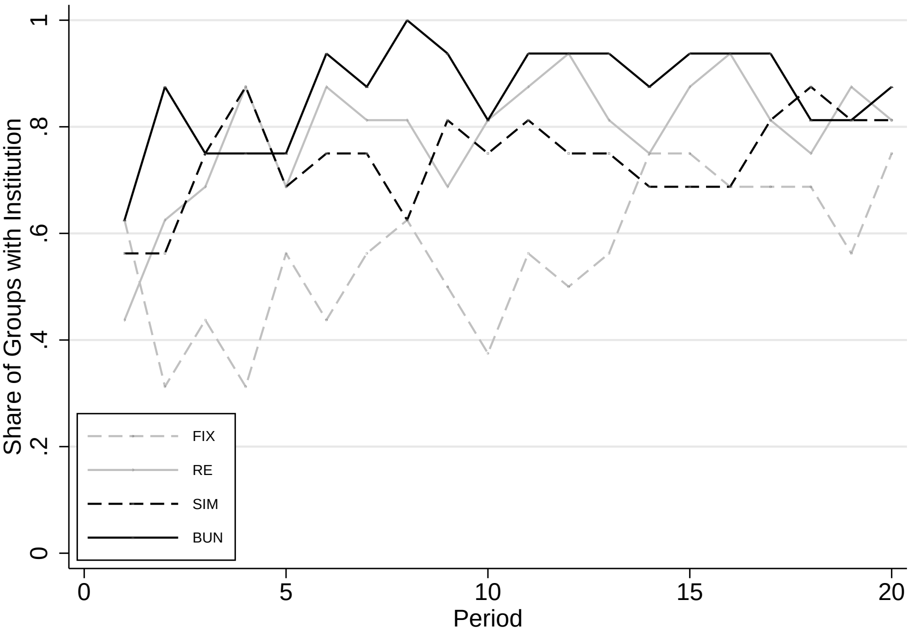
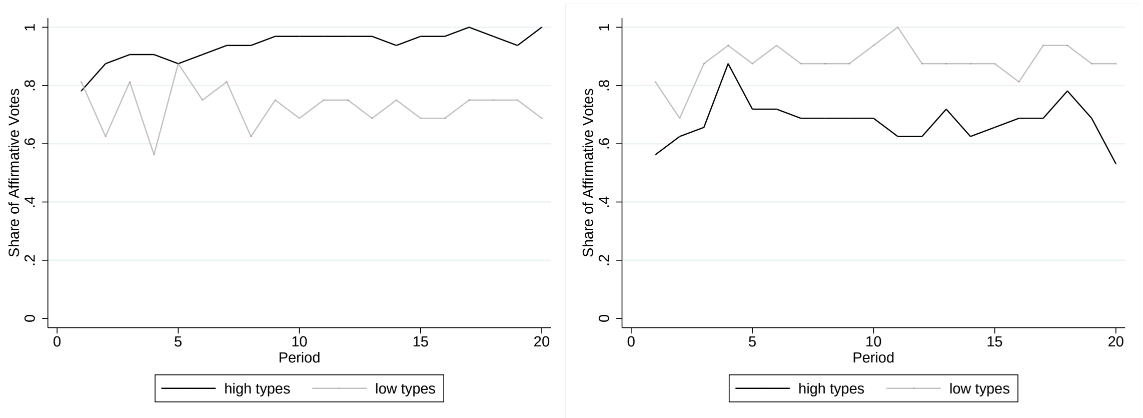


Figure 3.2. Share of Groups with Institution over Time

Notes: In the treatment with single institutions the share of successful formations increases significantly over time (Fix: $r = 0.654$, $p < 0.01$; Re: $r = 0.45$, $p = 0.05$). The trend is similar but weaker in other treatments (Sim: $r = 0.403$, $p = 0.058$; Bun: $r = 0.34$, $p = 0.14$).

is adopted, the institution of treatment Re is present; (4) if both institutions are rejected, the regular VCM is played.

When faced with the two institutions simultaneously, the players vote such that at least one institution is installed in 74% of all cases. The share of groups with an institution is thus higher than in treatment Fix, but slightly lower than in treatment Re. Only the governing institution is implemented in 31%, while solely redistribution is established in only 8% of all cases and both institutions are installed in 35% of all cases. Especially the high prevalence of both institutions being installed hints either at a high degree of inequality aversion by the high-type subjects or at the fear of a reciprocal exchange of votes among the different types (rejection of one institution inducing rejection of the other institution). Turning to voting behavior, we see that each type of player votes more frequently for the institution that offers the larger benefits to himself. High types vote in favor of the governing institution in 94% of all instances, but they support redistribution in only 68% of all cases. For low types the picture is just the opposite: 73% vote for the governing institution with fixed contributions



(a) Governing Institution

(b) Redistribution

Figure 3.3. Voting behavior on the two institutions by subject type in treatment Sim

Notes: Left displays voting on the governing institution, right displays voting on redistribution. Support of high types for the governing institution increases significantly over time ($r = 0.227$, $p < 0.01$).

and 88% vote for redistribution.[12] Overall, the implementation rates are similar to the other treatments, and the contribution level of 16 tokens is only slightly higher (vs. Fix MWU $p = 0.17$; vs. Re MWU $p = 0.87$). If an institution is implemented, the contributions reach 19.2 tokens, which is very close to the efficient level of 20.

Result 3: If governance and redistribution are separately available,

a) the governing institution receives more support from high type players,

b) the redistribution rule receives more support from low type players,

c) and average contributions are not significantly higher compared to treatments with only a single institution.

Overall, we find that the simultaneous presence of both institutions is not able to increase contributions significantly above the level that was already achieved by a single institution. The inverse voting patterns of high and low types highlight their conflict to settle on a common institutional setup. We will now check whether this conflict is eliminated by the bundling procedure in treatment Bun.

[12] If examined using a Wilcoxon matched-pairs signed-ranks test that uses the difference in average voting behavior of single group as independent observation, the differences in voting behavior across types are significant at the 5 % level ($p = 0.012$ for voting on governance, $p = 0.014$ for voting on redistribution).

3.4.2.2 Treatment Bun

Recall that the only difference between treatments Bun and Sim is the more restrictive choice set in Bun, because the decision on the formal redistribution rule is linked to the decision on the governing institution. Thus, subjects either have to support both institutions or none. In principle, both institutions could also be implemented in Sim. Yet, our data show that bundling is better at eliminating the tradeoff between equality and efficiency. The bundled institutions are installed in 87% of all instances. The share of successful institution formation is larger than in treatment Sim, even if the difference falls short of significance (MWU p=0.109). The high implementation rate is the result of nearly identical voting behavior by both types. High types vote in support in 93% of all cases, compared to 98% of low types. Already after a few periods, the support is nearly unanimous across both types.[13] The high rates of support come along with high contribution rates. Average contribution reaches 18.4 tokens, which is significantly more than in treatment Sim (MWU p=0.048). The average contribution rate being close to the efficient maximum indicates that the bundling approach is able to overcome the conflicting interest between heterogenous types (that was still present in treatment Sim, as seen by the frequent rejections). By bundling the governing institution and the redistribution rule, i.e., by combining equity and efficiency, the negative effects of heterogeneity and social preferences on institution formation are eliminated.

Result 4: The details of the implementation approach make a difference:

a) Support and implementation rates for the institutions are higher in treatment Bun than in treatment Sim.

b) Contribution rates are significantly higher in treatment Bun than in treatment Sim.

3.4.3 Reciprocal exchanges of votes among types

The support of the governing institution by the low types increases between treatments Fix and Sim. This difference could hint at some kind of trust in the support of equal outcomes by the high types. This might have two causes. Either the low-type players expect their counterparts to be inequality averse or they expect reciprocity for their own support of the fixed contribution.[14] Both situations would lead to support for the formal redistribution rule. As a corollary from this observation, the question is raised whether the remaining difference

[13] The voting behavior of both types over time is displayed in Appendix 3.B.

[14] As the voting decision is made simultaneously, this kind of reciprocity would need to be belief driven. The beliefs on the action of the other subjects' actions might be wrong in the first place and are corrected in the rounds afterwards.

between types is a result of the low types’ expectations that the high types will reject the redistribution. If this hypothesis is true, one should expect a decrease in the support for the governing institution if the formal redistribution rule has been rejected previously. In line with this, we find that the institution formation process hinges critically on the results of the previous period and its voting decisions (see Tables 3.B.1 and 3.B.2 in Appendix 3.B for further details on the determinants of individual voting behavior).

The observed trend of continued institution formation has, of course, direct implications on contributions and exemplifies the repeated nature of the game. The specific outcomes of a given group determine their future voting behavior to a large extent. This results in large variations in profits over time. Contributions spread apart as they adhere to different trajectories. Successful institution formation in previous rounds increases the subjects’ probability to vote for this particular institution again. Note the special role of treatment Sim. We observe that low types tend to reject the governing institution more frequently when it was the only institution that was implemented previously, i.e., if the redistribution rule was rejected in the previous period. This indicates that the low types are willing to forego efficiency in order to secure them a more equal payoff, similar to what we observe in treatment Fix. This behavior points at a struggle of “choosing” the preferred equilibrium or retaliation effects. The analog behavior by high types cannot be observed. The likely reason for that is that the payoff for high types does not change under redistribution as long as all player contribute fully. Moreover, the establishment of the governing institution only impacts the voting decision on the governing institution but not on the redistribution. This supports the hypothesis that the low types react upon the voting behavior of high types. They change their pattern if an undesired outcome was chosen, but do not change their voting decision on their preferred institution.

3.4.4 Preferring the separate (Sim) over the bundled (Bun) approach

As explained previously, we introduced the integration of redistribution into the governing institution to determine whether this will lead to higher implementation rates and payoffs. However, it remains unclear how subjects evaluate the two frameworks. Do subjects anticipate the higher payoffs that are associated with the bundling approach? Which framework would they prefer if they would be able to select between the two? To answer these questions, we conducted a simple decision experiment with 50 additional subjects in a follow-up study. The subjects were from the same pool as in the main experiment, but did not participate in the main study. They received the original instructions, were told that other subjects had previously participated in this study, and were then asked two questions.

First, they had to estimate which framework yielded the higher average payoffs, Sim or Bun. A correct estimate was rewarded with 3 Euro. A majority of subjects (68%) predicted correctly that the restricted voting option in Bun would result in higher payoffs for the participants. Second, they were asked which regime they themselves would prefer if they had to choose between the two treatments. Strikingly, 60% of subjects indicated that they themselves would prefer Sim over Bun, that is, they stated a preference for the regime where redistribution and the governing institution are not interlinked.

Given this data, the potential of a bundled approach indeed seems to be an open question ex ante, since for many subjects it was not trivial to determine which treatment condition resulted in higher average payoffs. Moreover, even if they correctly anticipated that payoffs would be higher in Bun, a substantial share of these subjects (and 28% of all subjects) still prefers the unbundled scenario Sim, where they can vote separately between implementing redistribution and implementing a governing institution. This implies that they are willing to forego higher monetary payoffs for a richer menu of potential institutional arrangements – suggesting that people might intrinsically value the freedom to choose and not being restricted in their choice set. In view of the outcomes from our main experiment, the follow-up study highlights a tradeoff between optimal design from an efficiency point of view and the individuals' preferences for freedom of choice.

3.5 Conclusion

This paper has investigated the effect of bundling institutions on the voting decisions of heterogeneous agents in the context of a public good game. The different institutions allowed subjects to overcome the free-rider problem of the social dilemma by either fixing contributions at a maximal level (governance), redistributing payoffs (redistribution), or by doing both at the same time. Players were allowed to vote on different institutions. In order to ensure that the support of each interest group is required, voting was governed by the unanimity voting rule. In each scenario, it was rational for self-centred agents to support the establishment of at least one of the available institutions. In order to create a conflict of interest among subjects, different marginal returns to the public good were introduced. If one considered inequity averse subjects, a rejection of the fixed contributions was to be expected. The predictions were partially supported, as a larger tendency of the players with a high MPCR to vote for governance and against redistribution was observed. The behavior of the low types exhibited the opposite tendency. This led to frequent rejections of either or even both institutions if they were offered simultaneously. We addressed this problem by offering the bundled institution, which limited subjects' decision on both institutions (or none at all).

The experiment demonstrated that the bundling of institutions is able to foster coordination and cooperation among subjects which have conflicting interests. The bundled institution was installed about 17% more often than at least one institution in the treatment in which both institutions could be voted upon separately. Contributions were significantly higher in the Bun treatment as well. The bundling of institution seems to work as a commitment device for subjects. The data support the hypothesis that institutions are often rejected if the other subjects are not expected to cooperate by supporting both institutions themselves. Additionally, our results highlight the importance of path dependency: Successful institution formation in past periods was shown to predict institution formation in the present period.

Even though the lab experiment has to abstract from features that might potentially affect behavior in natural environments, our experimental observation supports evidence from political events as described by Burkhart and Manow (2006). Large "packages" of bills in parliament often are a bundle of several different initiatives. Commitment to support such a bundle of bills can be credible. Similarly "issue bundling" has been proposed as a means of increasing support for the reduction of green house gas emissions if they are combined with individual interests such as morbidity, mortality or stress reduction (Koehn, 2008). In the context of referenda, a bundling of bills would allow politics to "sell the good with the bad" and thereby reach support for both initiatives. Naturally, it is not clear whether the bundling of different initiatives is indeed welfare enhancing. The overall effects depend very much on the specific environment and initiatives at stake. A related negative example can be found in the literature on industrial organization: The bundling of initiatives is equal to the bundling of products which can be used to extract additional surplus from consumers.

Given the strength and significance of our findings, we believe that bundling should be able to overcome the phenomenon of mutual blockades that is inherent in many political interactions. It might be very interesting to see in future studies whether this indeed holds true for other institutional arrangements and voting mechanisms; like the use of majority voting or the use of different institutions (e.g., partial redistribution of payoffs, enforcement costs, suboptimal obligations, delegation, etc.). Future treatments might also further enrich the situation by including entitlement considerations, too. Subjects frequently mentioned in our questionnaire that they perceived redistribution to be "fair", since low types were not viewed to be responsible for their ex-ante disadvantage of lower marginal returns from the public good. Thus, following up on the discussion in Cappelen et al. (2013), it might be interesting to see how (the perception about) such feelings of entitlement might interact with the support of institutional arrangements, in particular if they include formal redistribution rules.

References

Ahn, Toh-Kyeong, R. Isaac, and Timothy C. Salmon (2008): "Endogenous group formation." *Journal of Public Economic Theory*, 10 (2), 171–194. [51]

Burkhart, Simone and Philip Manow (2006): "Veto Antizipation - Gesetzgebung im deutschen Bikameralismus." *Max-Planck-Institut für Gesellschaftsforschung*, MPIfG Discussion Paper 06/3. [66]

Cappelen, Alexander W., James Konow, Erik Ø. Sørensen, and Bertil Tungodden (2013): "Just luck: An experimental study of risk-taking and fairness." *American Economic Review*, 103 (4), 1398–1413. [66]

Casella, Alessandra and Andrew Gelman (2008): "A simple scheme to improve the efficiency of referenda." *Journal of Public Economics*, 92 (10), 2240–2261. [50]

Engelmann, Dirk and Veronika Grimm (2012): "Mechanisms for Efficient Voting with Private Information about Preferences." *Economic Journal*, 122, 1010–1041. [50]

Ertan, Arhan, Talbot Page, and Louis Putterman (2009): "Who to punish? Individual decisions and majority rule in mitigating the free rider problem." *European Economic Review*, 53 (5), 495–511. [50]

Falkinger, Josef, Ernst Fehr, Simon Gächter, and Rudolf Winter-Ebmer (2000): "A simple mechanism for the efficient provision of public goods: Experimental evidence." *American Economic Review*, 90 (1), 247–264. [50]

Fehr, Ernst and Klaus M. Schmidt (1999): "A theory of fairness, competition, and cooperation." *Quarterly Journal of Economics*, 114 (3), 817–868. [51, 54, 56, 73]

Fischbacher, Urs (2007): "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10 (2), 171–178. [53]

Fischer, Sven and Andreas Nicklisch (2007): "Ex interim voting: an experimental study of referendums for public-good provision." *Journal of Institutional and Theoretical Economics*, 163, 56–74. [51]

Fisher, Joseph, R. Mark Isaac, Jeffrey W. Schatzberg, and James M. Walker (1995): "Heterogenous demand for public goods: Behavior in the voluntary contributions mechanism." *Public Choice*, 85 (3-4), 249–266. [51, 57]

Fleiß, Jürgen and Stefan Palan (2013): "Of coordinators and dictators: A public goods experiment." *Games*, 4 (4), 584–607. [51]

Gächter, Simon and Ernst Fehr (2000): "Cooperation and punishment in public goods experiments." *American Economic Review*, 90 (4), 980–994. [50]

Gerber, Anke, Jakob Neitzel, and Philipp C. Wichardt (2013): "Minimum participation rules for the provision of public goods." *European Economic Review*, 64, 209–222. [51, 70]

Gerber, Anke and Philipp C. Wichardt (2009): "Providing public goods in the absence of strong institutions." *Journal of Public Economics*, 93 (3-4), 429–439. [48, 50]

Greiner, Ben (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association*, 1 (1), 114–125. [53]

Gürerk, Özgür, Bernd Irlenbusch, and Bettina Rockenbach (2009): "Voting with feet: community choice in social dilemmas." *IZA Discussion Papers* (4643). [51]

Hamman, John R., Roberto A. Weber, and Jonathan Woon (2011): "An experimental investigation of electoral delegation and the provision of public goods." *American Journal of Political Science*, 55 (4), 738–752. [51]

Hanson, Peter C. (2014): "Abandoning the Regular Order Majority Party Influence on Appropriations in the US Senate." *Political Research Quarterly*, 67 (3), 519–532. [50]

Hortala-Vallve, Rafael and Aniol Llorente-Saguer (2010): "A simple mechanism for resolving conflict." *Games and Economic Behavior*, 70 (2), 375–391. [50]

Jackson, Matthew O. and Hugo F. Sonnenschein (2007): "Overcoming Incentive Constraints by Linking Decisions." *Econometrica*, 75 (1), 241–257. [50]

Kesternich, Martin, Andreas Lange, and Bodo Sturm (2014): "The impact of burden sharing rules on the voluntary provision of public goods." *Journal of Economic Behavior & Organization*, 105, 107–123. [51]

Koehn, Peter H. (2008): "Underneath Kyoto: emerging subnational government initiatives and incipient issue-bundling opportunities in China and the United States." *Global Environmental Politics*, 8 (1), 53–77. [66]

Kosfeld, Michael, Akira Okada, and Arno Riedl (2009): "Institution Formation in Public Goods Games." *American Economic Review*, 99 (4), 1335–1355. [48, 50, 69]

Kube, Sebastian, Sebastian Schaube, Hannah Schildberg-Hörisch, and Elina Khachatryan (2015): "Institution Formation and Cooperation with Heterogeneous Agents." *European Economic Review*, 78, 248–268. [47, 51, 54]

Lupia, Arthur and John G. Matsusaka (2004): "Direct democracy: new approaches to old questions." *Annual Review of Political Science*, 7, 463–482. [50]

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran (2013): "Self-organization for collective action: An experimental study of voting on sanction regimes." *Review of Economic Studies*, 81 (1), 301–324. [51]

Matsusaka, John G. (2010): "Popular control of public policy: A quantitative approach." *Quarterly Journal of Political Science*, 5 (2), 133–167. [50]

Ones, Umut and Louis Putterman (2007): "The ecology of collective action: A public goods and sanctions experiment with controlled group formation." *Journal of Economic Behavior & Organization*, 62 (4), 495–521. [51]

Ostrom, Elinor, James Walker, and Roy Gardner (1992): "Covenants With and Without a Sword: Self-Governance is Possible." *American Political Science Review*, 86 (2), 404. [50]

Potters, Jan, Martin Sefton, and Lise Vesterlund (2005): "After you—endogenous sequencing in voluntary contribution games." *Journal of Public Economics*, 89 (8), 1399–1419. [48]

Putterman, Louis, Jean-Robert Tyran, and Kenju Kamei (2011): "Public goods and voting on formal sanction schemes." *Journal of Public Economics*, 95 (9), 1213–1222. [51]

Reuben, Ernesto and Arno Riedl (2013): "Enforcement of contribution norms in public good games with heterogeneous populations." *Games and Economic Behavior*, 77 (1), 122–137. [47, 51]

Romer, Thomas and Howard Rosenthal (1978): "Political resource allocation, controlled agendas, and the status quo." *Public Choice*, 33 (4), 27–43. [50]

Romer, Thomas and Howard Rosenthal (1979): "Bureaucrats versus voters: On the political economy of resource allocation by direct democracy." *Quarterly Journal of Economics*, 563–587. [50]

Sausgruber, Rupert and Jean-Robert Tyran (2007): "Pure redistribution and the provision of public goods." *Economics Letters*, 95 (3), 334–338. [50]

Torgler, Benno (2005): "Tax morale and direct democracy." *European Journal of Political Economy*, 21 (2), 525–531. [50]

Tyran, Jean-Robert and Lars P. Feld (2006): "Achieving Compliance when Legal Sanctions are Non-deterrent." *Scandinavian Journal of Economics*, 108 (1), 135–156. [50]

Appendix 3.A Theoretical Framework

In this section the following public good game with n players will be analyzed formally. Each player possesses an identical private endowment E. The player can decide individually on a contribution $c_i \in 0, E$ which she contributes to the public good. Given all players contributions $(c_i, ..., c_n)$ the material payoff of player i is denoted by

$$\pi_i = E - c_i + \gamma_i \sum_{j=1}^{n} c_j.$$

In order to create a social dilemma $0 < \gamma_i < 1$ and $\sum_{i=1}^{n} \gamma_i > 1$ is assumed. The first condition ensures that a self-centered player never profits from contributing to the public good, while the second assumption creates the effect that a contribution equal to the endowment by all players would be socially efficient. The individual return to the public good (MPCR) induces heterogeneity among the players. Throughout the complete analysis heterogeneity will always refer to this difference in MPCRs and not to any other difference in individual preferences. The mechanism described in chapter two is a two stage n player coordination game. During the first stage the players can vote on the establishment of an institutional regime. The treatments as introduced before differ in the kind of institution that is available. The second stage of the game is the contribution stage with each player choosing the personal contribution to the public good simultaneously. In order to keep the analysis comprehensible perfect information about all characteristics of the players is assumed. From this can be inferred that the players on the voting stage posses complete information about the outcome on the later stage of the game. The method to solve the game will always be that of a subgame perfect Nash equilibrium in pure strategies. We treat the experiment as an one shot game with the behavior within a single period being predicted. The addition of additional rounds does not change the predictions as the number of rounds within the experiment is exogenously given and fixed. Hence, if all players apply backward induction, the number of rounds has no influence on the decisions compared to the one shot game. Due to the unanimity voting rule a large number of SPNE arise. If one player rejects an institution it will never be established. Hence the other players will always be indifferent between voting in favor or rejecting the institution. This problem has already been discussed intensively in the literature. Kosfeld et al. (2009) for instance, focus on stagewise strict equilibria. By definition a Nash equilbrium is stagewise strict if on every stage game each player's strategy is a unique best response towards the other players' strategies. Unfortunately this refinement is too strict for most situations considered here. If a rejection of an institution is justified for a player, the other players' decision does not influence the voting outcome. Thus, no stagewise strict equilibrium can exist if rejections are reasonable. Hence we will deploy a slightly different refinement. Gerber et al. (2013) introduce a more relaxed version of strictness in that sense, that they consider only these Nash equilibria, for which in every stage game exists at least one player whose equilibrium strategy is an unique best response to the other players' equilibrium strategies. In the following analysis we will concentrate on this class of "semi-strict" equilibria exclusively.

3.A.1 Standard Predictions

Proposition 1 (voting behavior and contributions of heterogeneous players)**.** *In treatment VCM, all players contribute $c_i = 0$. In treatments FIX, RE and BUN it is a subgame perfect Nash equilibrium for all players to vote in favor of implementing the proposed institutions. The institutions are always implemented. In case of the treatments FIX and BUN all players contribute according to the institutional rules, i.e., $c_i = E$, $i \in h, l$. In the treatment RE the players will contribute their complete endowment, $c_i = E, \forall i$ voluntarily. In the treatment SIM exists a subgame perfect Nash equilibrium in which the low types support both institutions. The high type players' unique best response is support for the fixed contributions only.*

The proof of proposition 1 will be structured the following way: First the baseline scenario will be analyzed. The results of the VCM treatment will then be used in order to compare the payoffs under the different institutions using backward induction.

Heterogeneous VCM. In treatment VCM both the two high types and the low type are predicted not to contribute to the public good at all: $\frac{\partial \pi_i}{\partial c_i} = -1 + \gamma_i < 0$ for both types of players because $\gamma_i < 1$ by definition of the VCM game. This behavior will be used in all following treatments as result from the VCM. All other potential behavior will be ignored.

Fixed Contributions. In the two-stage game defined in treatment FIX, players will apply backward induction. If the institution has not been implemented in the voting stage, the players on the contribution stage are back in the VCM game analyzed above. They are predicted not to contribute to the public good. Therefore, each player's monetary payoff will be equal to the initial endowment E. If players have unanimously agreed on implementing the institution in the voting stage, all players are obliged to contribute their whole initial endowment E and will earn $\gamma_i nE$. $\gamma_i nE > E$ whenever $\gamma_i > \frac{1}{n}$. This condition has to hold for all players of type $\gamma_i = \gamma_l$. Otherwise these players obtain a payoff smaller than their initial endowment under the symmetric institution and will consequently reject it. The values used for γ in our experiment fulfill this condition ($\gamma_l = \frac{1}{2} > \frac{1}{3}$) With unanimity voting it is consequently a unique best response for all players to vote in favor of the symmetric institution that requires each player to contribute the efficient contribution level E. Intuitively this is clear from the beginning as the players will support the institution if they profit from it.

Pure Redistribution. In order to determine the behavior of the players in the treatment in which only the redistribution rule is available, the behavior in the subsequent VCM with the redistribution rule must be analyzed and contrasted with the predictions for the standard VCM without any institution. The redistribution rule changes the payoff from the public good entirely. As assumed the redistribution takes place in such way that every player receives the same monetary payoff. Hence for the decision of the player only the sum of all payoffs is important. As efficiency of the public good was assumed in that sense, that $\sum_{i=1}^{n} \gamma_i > 1$, the total payoff from public good under full contributions, is larger than the endowments. Technically this can be seen by the fact that the payoff from the public good is now denoted by $\pi_i = \frac{nE+(\sum_{j=1}^{n} \gamma_j - 1)\sum_{j=1}^{n} c_j}{n}$. Whenever $\frac{\partial \pi_i}{\partial c_i} = \frac{\sum_{j=1}^{n} \gamma_j - 1}{n} > 0$ the costs of the contribution are smaller than the accumulated benefit. Hence under the given redistribution rule a self-centred, money maximizing player will always contribute the complete endowment E. This is the condition for the efficiency of the public good. Thus all players will always contribute their complete endowment E. Their contributions lead to a payoff of $\pi_i = \sum_{j=1}^{n} \gamma_j E$. Naturally this payoff will larger than the endowment E whenever $\sum_{j=1}^{n} \gamma_j > 1$ – the condition for the efficiency of the public good. Thus every player will always earn a higher payoff with the institution in place and consequently support it on the voting stage. The redistribution rule eliminates the free-rider problem by distributing the benefits of the public good equally among all players. Hence the players profit from their own contributions the same way they profit from other players' contributions. Costless redistribution is predicted eliminate the social dilemma. The result holds costly redistribution mechanisms as well. As long as the marginal costs for the redistribution are lower than the marginal gain achieved by the public good, contribution can be expected.

Bundled Institutions. As the players will apply backward induction, they will compare their payoff from the bundled institution with the payoff from the normal VCM. As players are predicted not to contribute at all in the VCM, the payoffs which need to be compared are again the endowment E in the case of the VCM and $\sum_{j=1}^{n} \gamma E$ under the institutional regime. Hence it is once more the mutual unique best response to support the institution. Thereby the predicted results for the bundled institution are identical to the treatment in which only the institution for redistribution is available.

Simultaneous Availability. The analysis of the decision in the presence of two individually selectable institutions is not as easy as the analysis for a single institution, because the players have to reach two decisions. Hence each player has now four possible bundles of actions in the first stage. In order to create a simultaneous move game with only one stage, it is again assumed that backward induction is applied. Hence, E is used as payoff if no institution at all is established, and $\sum\limits_{j=1}^{n} \gamma_i E$ if only the redistribution rule or both institutions are implemented. If one assumes that the players of the same type will always act identical their payoffs can be presented in the 2x2 matrix below.[15] Here the possible actions have the following form: The first part denotes the vote towards governance (fixed contributions) and the second part the vote towards the redistribution.

		high type			
		yes, yes	yes, no	no, yes	no, no
	yes, yes	$\sum\gamma_i E, \sum\gamma_i E$	$\gamma_h nE, \gamma_l nE$	$\sum\gamma_i E, \sum\gamma_i E$	E, E
low type	yes, no	$\gamma_h nE, \gamma_l nE$	$\gamma_h nE, \gamma_l nE$	E, E	E, E
	no, yes	$\sum\gamma_i E, \sum\gamma_i E$	E, E	$\sum\gamma_i E, \sum\gamma_i E$	E, E
	no, no	E, E	E, E	E, E	E, E

As $\sum\limits_{j=1}^{n} \gamma_i \geq \gamma_l n$, it is obvious that the low type’s voting decision “yes, yes” offers a payoff at least as large as all other voting options. The unique best response of the high type would then be to play “yes, no”, which would result in the acceptance of the governing institution and the rejection of the redistribution. However, this concentration on these Nash equilibria excludes other equilibria. There are some more pure strategy equilibria to find. In that sense the strategies “no, no” played by at least two players forms a Nash equilibrium as well, like in all other treatments. The last two equilibria arise if at least one of the low type players chooses “no, yes”. Then the best response by the high types is to play either “yes, yes” or “no yes”, which both result naturally in the same payoff. Nevertheless no strategy is a unique best response as always the decisions “yes, yes” and “no, yes” if played by all players induce the same outcome. It can be concluded that in any Nash equilibrium in pure strategies at most one, but never both of the institutions will be accepted.

Under standard predictions in all institutional treatments efficiency is expected independently of the institution available. This is either created by forced contributions or voluntary full contributions in the treatment Re. Hence the voting procedure should have no influence on the contributions. In all treatments

[15] Their payoffs will always be identical independently of their actions. If one drops this simplifying assumption some additional equilibria arise.However, these lead only to more rejections of the institutions (e.g., one high type and one low type reject both institutions).

at least one institution should be adapted at all times if the players behave according to the standard preferences.

3.A.2 Fehr-Schmidt (1999) Preferences

In the next part behavioral predictions will be determined not using the standard model of rational self-centered players. Instead the model of inequity aversion developed by Fehr and Schmidt (1999) will be used. This model assumes that players compare their outcome with the outcome of all other players. In order to model this departure from the standard model they introduce the following utility function:

$$U_i = \pi_i - \alpha_i \tfrac{1}{n-1} \sum_{j=1}^{n} max\{\pi_j - \pi_i; 0\} - \beta_i \tfrac{1}{n-1} \sum_{j=1}^{n} max\{\pi_i - \pi_j; 0\}$$

The first term represents the monetary payoff obtained from the game. The second term captures disadvantageous utility derived from being worse off than other players. α_i is hereby the individual envy parameter. The last term denotes losses the player receives from being better off than the other players. β_i is typically interpreted as a measure for compassion. Additionally two important properties are assumed. The first is $\alpha_i \geq \beta_i$, which indicates that envy is at least as strong as compassion and secondly $\beta_i < 1$, which prevents potential players from "burning" money to achieve a larger degree of equality. In the described setup with heterogeneous returns to the public good players might vote against the establishment of an institution with symmetric obligations in order to prevent inequality. Hence we consider the model of Fehr and Schmidt (1999) as a natural choice to derive predictions for this setup. In order to keep the following part as comprehensible as possible, the analysis of the treatments featuring heterogeneous players with social preferences will be restricted to the case of three players. For the case of heterogeneous players one player with $\gamma_i = \gamma_l$, two players with $\gamma_i = \gamma_h$ and with $\gamma_h - \gamma_l = \Delta\gamma \leq \frac{1}{2}$ will be described during the analysis. A generalization of this results to n players would create a larger amount of asymmetric equilibria and thereby crowd the analysis unnecessarily. In the following the subscript l will used to mark all decisions, characteristics and consequences for the player with the low level of MPCR (low type). Similar the other two players will be marked using the numbers 1 and 2 (high types). Without any restriction we will only analyze the decisions of the high type player 1. Let further $\bar{c}$ denote that level of contribution by the low type that induces equal payoffs between the low type and at least one other high type, given both high types players level of contribution. c_h will always denote the equilibrium level contribution of the high types. The formal analysis of the experiment using Fehr and Schmidt (1999) preferences will be structured as follows: First the behavior in the standard VCM with heterogeneous players is determined.

Afterwards the possible payoffs will be compared to the payoffs created by the institutional regimes, in order to derive the predictions for the treatments Fix, Re, Bun and Sim. First the decision of the low type in the heterogeneous treatments will be analyzed, the decision of the high types follows afterwards.

3.A.2.1 Heterogeneous-VCM

The behavioral predictions for the heterogenous VCM are derived in greater detail in the Appendix 2.A.2 of Chapter 2. Hence, only the resulting equilibrium behavior is highlighted here. If one uses the calibration of the experiment the predicted behavior of the low type player can be summarized as follows:

1. If $\beta_l \geq \frac{2}{5}$ and both high types contribute a positive amount, the low type will contribute until the payoff of herself and the high type contributing less will be equalized.
2. Otherwise the low type will never contribute a positive amount to the public good.

The predicted behavior of the high type players using the calibration of the experiment can be summarized as follows:

1. If $\beta_{1/2} \geq \frac{2}{7}$ and both other players contribute a positive amount to the public good the high type will contribute in such a way that the payoff of herself and the monetary payoff of the player with the second highest payoff will be equalized.
2. Otherwise the high type will never contribute a positive amount to the public good.
3. Both high types will always contribute the same amount to the public good.

These findings result in the following possible equilibria for the treatment Vcm:

1. If $\beta_{1/2} \geq \frac{2}{7}$ and $\beta_l \geq \frac{2}{5}$, every level of contribution with $c_1 = c_2 \in [0; E]$ and $c_l = c_1 \frac{2}{5}$ is an equilibrium.
2. If $\Delta\gamma = \frac{1}{2}$, the contribution of the low type resulting in equal payoffs will be zero. Thus the low type will never contribute in any equilibrium. Hence if $\beta_{1/2} \geq \frac{2}{7}$, but $\beta_l < \frac{2}{5}$ it is still possible that both high types will contribute $c_1 = c_2 \in \{0; E\}$
3. Otherwise the only remaining equilibrium is characterized by $c_1 = c_2 = c_l = 0$

The other possible, asymmetric equilibria have been eliminated due to the chosen parametrization of the experiment.

3.A.2.2 Governance (Fixed Contributions)

The behavioral predictions for the institutions that fix the contributions to the public good are derived in greater detail in Appendix 2.A.3 of Chapter 2. Hence, only the resulting equilibrium behavior is highlighted here. Using the parametrization of the experiment will simplify the results drastically as it has be shown previously that equilibria with asymmetric payoffs cannot arise. Hence for treatment Fix the predicted behavior of players with Fehr-Schmidt preferences can be summarized as follows:

1. The players with $\gamma_l = 0.5$ will reject the establishment of the proposed contribution rule for $\alpha_l > \frac{2}{3} - \frac{4}{75}c_h$.

2. The players with $\gamma_h = 0.75$ will never reject the establishment of the contribution rule.

3.A.2.3 Pure Redistribution

Proposition 2 (Behavior in treatments Re and Bun)**.** *The high types will always prefer the establishment of the institution over the VCM in treatment Re. Low type players will only reject the redistribution if an asymmetric equilibrium is played in the subsequent VCM. The low type players will then reject whenever* $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{2\gamma_l-\beta_l(1-\Delta\gamma)}$*. Otherwise the low type will always support the institution. With redistribution all players will always contribute their complete endowment. (i.e* $c_i = E; \forall i$*) The voting behavior in treatment Bun is identical to the voting behavior of all players in treatment Re.*

The predictions for the VCM with redistribution are mainly the same as under standard preferences. Since the redistribution rule prevents the existence of inequality between the players the utility function remains the same as above and thereby the predictions do not change. All players will contribute completely and receive $U_i^r = (2\gamma_h + \gamma_l)E$. Nonetheless, differences might arise if the players compare this outcome with their payoffs from the normal VCM, in which under the existence of social preferences other equilibria are possible. If the players compare this payoff with the one from the VCM, it is obvious that if a symmetric equilibrium is played in the VCM all players will support the redistribution. In the VCM all players obtain the same payoff, but efficiency is not reached, as the low type does not contribute her complete endowment. As both efficiency and equity is reached by redistribution, this payoff must consequently be larger than the payoff from the VCM. Technically this can be seen by the fact that the highest possible payoff from the VCM is denoted by $U_l^g = U_h^g = \gamma_h E(\frac{3}{1+\Delta\gamma})$. In order to induce a rejection this must be a larger than $U_i^r = (2\gamma_h + \gamma_l)E$, the payoff under redistribution with full contribution. Comparing the two results leads to $2\gamma_h + \gamma_l > 1$ as condition for the support of the institution. This is the condition for the efficiency of the public good and will always be fulfilled. Hence if only symmetric equilibria are possible, the institution will always be supported by all players. In a second step the behavior of the low type in the presence of asymmetric equilibria in a potential VCM will now be discussed. As shown previously the low type’s payoff from an asymmetric equilibrium is denoted by $U_l^g = E + c_h(2\gamma_l - \beta l(1-\Delta\gamma))$. A comparison of the payoffs establishes the result, that for $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{2\gamma_l-\beta_l(1-\Delta\gamma)}$ the low type will reject the establishment of the redistribution as the material payoff will be larger in the VCM. This condition is fulfilled only if the high types contribute a fraction of their endowment large enough to the public good. The high type will never reject the redistribution. This is obvious, as the highest payoff from a asymmetric equilibrium is smaller than the highest payoff from the symmetric equilibrium, since the low types contributes a positive amount. As shown above, the redistribution rule implies a higher payoff for every player than the symmetric equilibrium of the standard VCM. Thus the payoff from the asymmetric equilibrium must be smaller than the payoff under the institutional regime as well. Again the chosen parameters exclude the existence of asymmetric equilibria. This simplifies the result such that the support of the institution guarantees all players a payoff, which is at least as large as the maximum payoff from the VCM.

3.A.2.4 Bundled Institutions

The predictions for the bundled institutions are identical to the to the analysis done for the treatment Re. This can be seen easily. As has been shown previously under the existence of the redistributional regime, all players will contribute their complete endowment and thereby induce the same result as the bundled institutions, which enforce full contribution. Hence the voting decisions on the bundled institutions are predicted to be identical to the decisions on the redistribution rule.

3.A.2.5 Simultaneous Availability

Proposition 3 (Behavior in treatment Sim). *If no asymmetric equilibria are played the low type will prefer the implementation of both institutions as long as* $\alpha_l < \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$. *The best response for the high types will be the rejection of the redistribution if* $\beta_{1/2} < \frac{2}{3}$, *and the support of both institutions if* $\beta_{1/2} \geq \frac{2}{3}$. *If* $\alpha_l > \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$ *and only symmetric equilibria are played, the low type will support only the redistribution. The best response of the high types is then to support this one as well. If asymmetric equilibria would be played during the VCM, the low type will reject both institutions if* $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{2\gamma_l-\beta_l(1-\Delta\gamma)}$. *Otherwise the behavior remains the same as in the case of symmetric equilibria,* *with the threshold for the rejection of the governing institution by the low type now being* $\beta_l > \frac{2}{3}\frac{3\gamma_h-1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$.

The treatment in which the two institutions are available independently is the most complex treatment to analyze. The results from all the treatments above will be combined. The following table shows the outcomes for each decision combination under the simplification that the two high players are summarized here as one player. This is done solely for reasons of visualization. Nonetheless the two high type players might vote differently due to differences in their specific values of α and β. When using the following representation, one must keep in mind that each institution will only be established if it is supported by all three players. The cells display to which other treatment the voting results correspond. Then the already established knowledge about the treatments will be used to determine subgame perfect Nash Equilibria in pure strategies.

		high type			
		yes, yes	yes, no	no, yes	no, no
low type	yes, yes	Bun	Fix	Re	Vcm
	yes, no	Fix	Fix	Vcm	Vcm
	no, yes	Re	Vcm	Re	Vcm
	no, no	Vcm	Vcm	Vcm	Vcm

In order to determine possible equilibria, the proof for proposition 5 will be divided into two parts. Like before in the first part the behavior of the players will be analyzed under the assumption that a symmetric equilibrium is played during the VCM. The second part considers the behavior under possible asymmetric equilibria. In the next step a preference relation between the different treatments for the two types of players types will be established in order to solve the three player game above. Abover it has been shown that in the presence of symmetric equilibria the low type will always prefer the identical treatments Re and Bun over the treatment Vcm. Obviously, they will also prefer these two over the treatment Fix as they offer a higher monetary payoff with less inequality among the players. Furthermore, we know that the Vcm offers a higher utility than Fix if $\alpha_l > \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$ if a symmetric equilibrium is being played. This results in two possible preference ordering of the treatments:

1. $RE \sim BUN \succ FIX \succsim VCM$, if $\alpha_l \leq \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$ (type 1)
2. $RE \sim BUN \succ VCM \succ FIX$, else (type 2)

If one takes now a closer look at the table presented above, it becomes obvious that the support of both institution is weakly preferred over all other voting decisions by type 1, given the behavior in the ensuing VCM. Conversely type 2 weakly prefers the voting decision “no, yes” over all other options and supports thereby only redistribution. The next step will determine the high types’ possible best answers to the proposed strategies. For this a preference ordering similar to the one presented above needs to be established. As shown before, the high type will always prefer the establishment of the identical outcome of the treatments Bun and Re over the Vcm. High type players will actually prefer the case of full contribution with redistribution (Bun) to the case without redistribution (Fix) if their sensitivity towards advantageous utility is too strong. In order to derive this threshold the payoff from the two different scenarios will be compared. The payoff without redistribution is $U_1^f = 3\gamma_h E - \frac{\beta_1}{2} 3E\Delta\gamma$ and with redistribution $U_1^r = (2\gamma_h + \gamma_l)E$. Thus the high type will prefer the redistribution institution whenever $\beta_{1/2} > \frac{2}{3}$. Additionally it has been shown, that the high type will prefer the Vcm over Fix if $\beta_h > \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$, in case of a symmetric equilibrium. Moreover, it is known, that the treatments Re and Bun are equivalent and preferred over the outcome of treatment Vcm. Thus we derive the following three possible preference relations for the high type.

1. $FIX \succ RE \sim BUN \succ VCM$, if $\beta < \frac{2}{3}$ (type 1)
2. $RE \sim BUN \succsim FIX \succ VCM$, if $\beta \geq \frac{2}{3}$ and $\beta_h < \frac{2}{3}\frac{3\gamma_h - 1}{\Delta\gamma} - \frac{2c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$ (type 2)
3. $RE \sim BUN \succ VCM \succsim FIX$, else (type 3)

The high type player of type 1 will react to a support of both institutions by the low type with a rejection of the redistribution institution. Under these circumstances Fix would be played. Players of type 2 and 3 will react to a support of both institutions by either supporting institutions as well or only the redistributing one. This behavior results in the treatments Re or Bun being played and in the same payoffs for all players. If, however, the low type is of type 2 and will always support only the redistribution institution the high type can decide between the treatments Vcm and Re being played. Here all high types are predicted to support the institution Re. Re is hence always established in equilibrium if the low type is of type 2. That results in the following equilibria:[16]

1. If the low type is of type 1 and both high types are of type 1, only the governing institution is established. The results are the same as for standard predictions.
2. If the low type is of type 1 and one of the high types is of type 1, while the other is of type 2 or 3, only the redistribution is supported.

[16] These are not the only equilibria possible. Nevertheless, they are the only equilibria in which only voting decisions are selected whose outcome are weakly preferred over the outcomes of all other voting decisions, given the behavior in the VCM game. Thus we concentrate on these during the analysis. The description of all equilibria existing in pure strategies would crowd the analysis further. An example of an omitted Nash equilibrium is the rejection of all institutions by all players.

3. If the low type is of type 1 and both high types are of type 2 or 3, either both institutions or redistribution is established.

4. If the low type is of type 2, governance is always rejected and redistribution is always established.

The next section will analyze the behavior if an asymmetric equilibrium would be played during the VCM. The different equilibria change only the payoff of the VCM. In this case it has been shown previously that the low type will prefer the treatment VCM over the institutions in the treatments RE and BUN whenever $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{2\gamma_l-\beta_l(1-\Delta\gamma)}$ and over the governing institution whenever $\alpha_l > \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(2\gamma_l - \beta_l(1-2\Delta\gamma))$. This implies again that the contributions by the high types must be large enough warrant a rejection of the institutions. As before the redistribution and the bundled institutions are preferred over governance only. In this case three preference orderings are possible.

1. $RE \sim BUN \succ FIX \succ VCM$, if $\alpha_l < \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$ (type 1)

2. $RE \sim BUN \gtrsim VCM \gtrsim FIX$, if $\alpha_l \geq \frac{3\gamma_l-1}{3\Delta\gamma} - \frac{c_h}{3E\Delta\gamma}(\frac{2\gamma_h+\gamma_l-1}{1+\Delta\gamma})$ and $\frac{c_h}{E} \leq \frac{2\gamma_h+\gamma_l-1}{2\gamma_l-\beta_l(1-\Delta\gamma)}$ (type 2)

3. $VCM \succ RE \sim BUN \succ FIX$, if $\frac{c_h}{E} > \frac{2\gamma_h+\gamma_l-1}{2\gamma_l-\beta_l(1-\Delta\gamma)}$ (type 3)

The preference relation of the first two types is the same as in the case of symmetric equilibria. Hence their behavior must be the same as well. The only difference is type 3. For this type it is always the best response to reject both institutions in order to force that the unregulated VCM is played during the second stage of the game. Hence the decision of the high types does not matter in this case at all. Thus only the behavior in case of support of governance and redistribution (type 1) and the redistribution rule (type 2) must be analyzed. Fundamentally the behavior of the of the high types remains unchanged. They still prefer the bundled institutions over the singular governing institution whenever $\beta \geq \frac{2}{3}$ and prefer them always over the VCM. However the cutoff value for the high types to prefer the VCM over the governing institution is now different. As derived in section 2.A.3.2, it is now denoted by $\beta_1 > \frac{2}{3}\frac{3\gamma_h-1}{\Delta\gamma} - \frac{c_h}{E}(2\gamma_h - 1 - \frac{\alpha_1}{2}(1-2\Delta\gamma))$. Thus the following equilibria are possible under asymmetric equilibria in the VCM and if only voting decisions are made that are weakly preferred over all other voting options, given the behavior of all players in the VCM:

1. If the low type is of type 1 and both high types are of type 1, only the governing institution is established. The results are the same as for standard predictions.

2. If the low type is of type 1 and one of the high types is of type 1, while the other is of type 2 or 3, only the redistribution rule is supported.

3. If the low type is of type 1 and both high types are of type 2 or 3, either governance and redistribution or only redistribution are established.

4. If the low type is of type 2, governance is always rejected and the redistribution rule is always established.

5. If the low type is of type 3, no institution is established.

The results are much less crowded if the parametrization of the experiment is used. As before the asymmetric equilibria are excluded. Then it is known that the high type will always prefer the governing institution over the Vcm. Thus it can be established, that the low type will support both institutions or only the redistribution rule. If $\alpha_l > \frac{2}{3} - \frac{4}{75}c_h$, the fixed contribution of the governing institution will be rejected. High types will support the redistribution rule only if the governing institution are already rejected by the low type or if their sensitivity towards disadvantageous disutility is too large ($\beta > \frac{2}{3}$). Thus three possible outcomes remain. If the high types care about advantageous disutility, both institutions should be established. If the low type is susceptible to disadvantageous disutility, only the redistribution rule is established. If inequity aversion is rather small for all participants, the outcome of the standard predictions will be realized.

3.A.2.6 Behavioral Predictions

In the result section we will test explicit behavioral predictions on the players' voting and contribution behavior. For these predictions we assume that at least some players' true preference are different from standard preferences. Additionally we assume that the players will play the VCM in case of institution failure in the same manner they would play the baseline in treatment Vcm.

Prediction 1:

a) In treatment Fix the contributions are weakly higher than in treatment Vcm.

b) In treatment Fix the institution is not installed in all cases.

c) Rejections of the institution are caused by the voting behavior of low types.

Prediction 2:

a) In treatment Re the average contribution and the implementation rates are higher than in treatment Fix and treatment Vcm.

b) In treatment Re no difference in voting behavior exists.

Prediction 3:

a) In treatment Sim the average contribution and the implementation rates are (weakly) higher than in treatment Fix.

b) In treatment Sim the average contribution and the implementation rates are not higher than in treatment Re.

c) High types will support governing institution more frequently than low types.

d) Low types will support the redistribution rule more frequently than high types.

Prediction 4

a) In treatment Bun the average contribution and the implementation rates are weakly higher than in treatment Sim.

b) In treatment Bun no difference in voting behavior exists.

Appendix 3.B Determinants of Voting Behavior

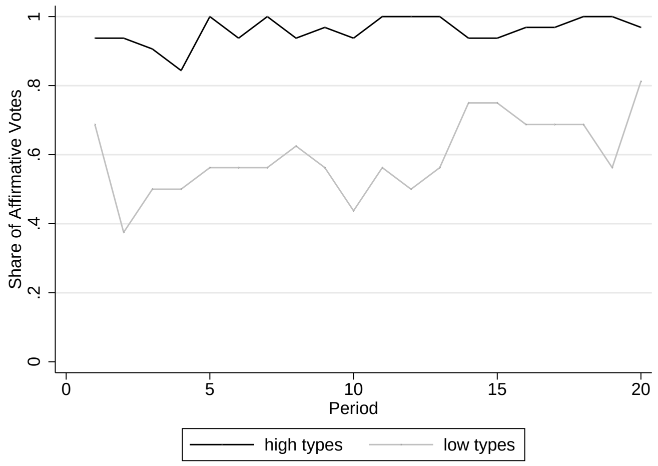


Figure 3.B.1. Voting Behavior by Type in Treatment Fix

Notes: For both types the affirmative votes exhibit an increasing time trend. The trend is significant both for high types (Spearman’s rho $r = 0.14$, $p = 0.013$) and low types (Spearman’s rho $r = 0.13$, $p = 0.019$).

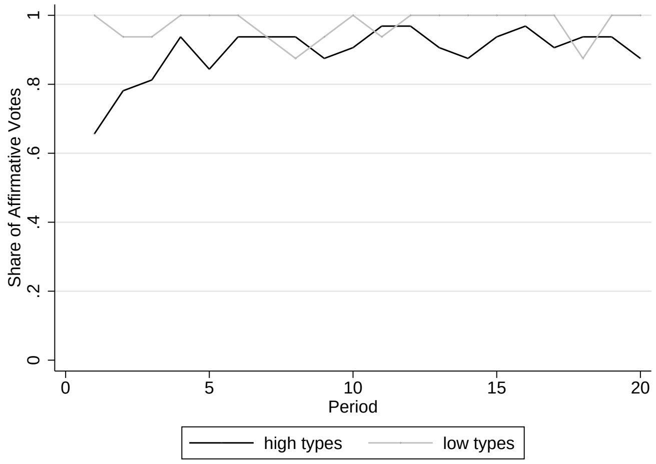


Figure 3.B.2. Voting Behavior by Type in treatment Re

Notes: For the high types the affirmative votes exhibit an increasing time trend. The trend is significant for high types (Spearman's rho $r = 0.17$, $p < 0.01$). For the low types there is almost no time-trend (Spearman's rho $r = 0.03$, $p = 0.54$).

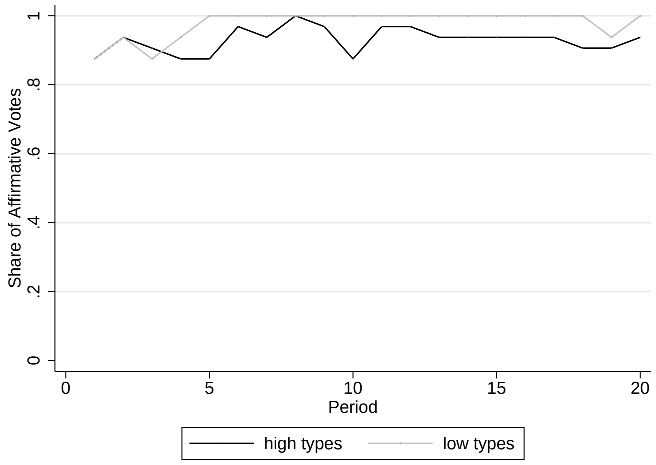


Figure 3.B.3. Voting Behavior by Type in treatment Bun

Notes: For the low types the affirmative votes exhibit an increasing time trend. The trend is significant for low types (Spearman's rho $r = 0.15$, $p < 0.01$). For the high types there is almost no time-trend (Spearman's rho $r = 0.07$, $p = 0.22$).

Table 3.B.1. Determinants of individual voting behavior on fixing contribution

	vote on fixing contributions				
	(1)	(2)	(3)	(4)	(5)
Fix	0.072 (0.19)	0.004 (0.01)	0.297 (0.87)		
Bun	0.964** (2.49)	0.860** (2.47)	0.589 (1.64)		
hightype		1.111*** (4.02)	1.225*** (4.41)	-0.101 (-0.34)	1.541*** (3.25)
own vote fix prev. period		0.455*** (3.92)		0.629*** (2.71)	0.194 (0.71)
outcome fix prev. period			0.239*** (3.80)	0.670*** (2.94)	1.057*** (2.96)
low*fix only prev. period				-2.663*** (-7.14)	
only fix prev. period				0.108 (0.33)	-1.325*** (-4.51)
only re prev. period					0.150 (0.53)
other controls	Yes	Yes	Yes	Yes	Yes
N	2880	2736	2736	912	912

Notes: This table reports the results of probit panel regressions. We use subjects per period as unit of analysis. Regression (4) and (5) include only treatment Sim. *, **, and *** denote significance at the 10, 5, and 1 percent level. textitz statistics in parentheses and observations clustered on matching-group-level.

Table 3.B.2. Determinants of individual voting behavior on redistribution

	vote on redistribution				
	(1)	(2)	(3)	(4)	(5)
Re	0.929***	0.729***	0.834***		
	(3.21)	(3.10)	(2.97)		
Bun	1.406***	1.171***	1.306***		
	(4.56)	(4.49)	(4.26)		
hightype		-0.787***	-0.967***	-1.214***	-1.225***
		(-3.47)	(-3.82)	(-3.65)	(-3.72)
own vote re		0.896***		0.995***	0.966***
prev. period		(7.71)		(5.78)	(5.49)
outcome re			0.714***		
prev. period			(6.68)		
high*re only				-0.044	
prev. period				(-0.09)	
only fix					-0.113
prev. period					(-0.74)
only re				-0.525	-0.584**
prev. period				(-1.15)	(-2.53)
other controls	Yes	Yes	Yes	Yes	Yes
N	2880	2736	2736	912	912

Notes: This table reports the results of probit panel regressions. We use subjects per period as unit of analysis. Regression (4) and (5) include only treatment Sim. *, **, and *** denote significance at the 10, 5, and 1 percent level. textitz statistics in parentheses and observations clustered on matching-group-level.

Appendix 3.C Instructions

The instructions below are translations of the German instructions for the experiment. Differences between the treatments are marked by T Vcm:“...”, T Fix:“...”, T Re:“...”, T Bun:“...”, T Sim:“...”

General instructions for the participants

You are now participating in an economic experiment. If you read the following explanations carefully, you will be able to earn a considerable amount of money – depending on your decisions and those of the other participants. Thus it is important to read these instructions very carefully.

During the experiment, it is absolutely prohibited to communicate with the other participants. Should you have any questions, please ask us. If you violate this rule, you will be dismissed from the experiment and forfeit all payments. How much money you will receive after the experiment depends on your decisions and those of the other participants. The experimental payoffs will be calculated in Taler. The total amount of Taler that you have accumulated during the experiment will be converted into Euro and paid to you in cash at the end of the experiment. The exchange rate from Taler to Euro is as follows:

$$40 \text{ Taler} = 1 \text{ Euro}$$

The experiment consists of exactly one part. This part is divided into **20 periods**. At the beginning of the experiment you are randomly assigned to a group of three. Thus, there are two other participants in your group. In each group of three, there are **two participants of type A** and **one participant of type B** (the exact difference between type A and type B will be explained shortly). Whether you are of type A or of type B is determined randomly. **In all periods your type remains the same, just as the types of the other participants in your group remain the same**. You will be interacting with the same two participants in all periods. Neither during, nor after the experiment will you receive any information about the identities of the other participants in your group.

Detailed Information about the Course of each Period

Each period is divided into three stages:

1. In the **second stage** you have to decide on how many Taler you contribute to a project and how many Taler you keep for yourself.

2. T Fix, Bun, Sim: In the **first stage** you can decide if you want to commit yourself and the other participants in your group to certain contributions to the project in stage 2 (T Bun: and redistribute the incomes such that all participants receive the same payoff at the end of the period.) Only if **all** participants decide in stage 1 to commit all participants in your group to certain contributions to the project, will the contributions actually be fixed (T Bun: and the income will be redistributed). If not all participants decide to fix the contributions, then you and the other participants in your group will each be able to contribute any number of your 20 Taler to the project in the second stage (T Bun: and the incomes will be not redistributed).
   T Re, Sim: (T Sim: Afterwards) In the **first stage** you can decide, whether you want to distribute your income and the income of the other participants, such all participants receive the same payoff at the end of the period, independently of their contribution in stage 2 . Only if all participants decide in stage 1 to redistribute the incomes, the incomes will actually be redistributed at the end of the period.

3. In the **third stage** you get to know the contributions of all participants in your group to the project in stage 2 and the payoffs of all participants in your group in this period.

At the beginning of each period every participant receives **20 Taler**. In each period you have to decide on how to use these 20 Taler. You can contribute Taler to a **project** or put them on a **private account**. Every Taler that you don't contribute to the project is automatically put on your private account.

Income from your private account:
For each Taler you put on your private account, you earn exactly one Taler. For example, if you put 20 Taler on your private account (thus contributing zero Taler to the project), you would earn 20 Taler from your private account. If, e.g., you would put 2 Taler on your private account (thus contributing 18 Taler to the project), your income from the private account would be 2 Taler. Nobody but you receives Taler from your private account.

Income from the project:
For each Taler that you or another participant in your group contributes to the project, you (and each other participant in your group) earn a certain number of Taler. Each participant’s income from the project depends on his or her type and is determined as follows:

$$\textit{Type A's income from the project} = \frac{3}{4} * \textit{sum of all contributions to the project}$$

$$\textit{Type B's income from the project} = \frac{1}{2} * \textit{sum of all contributions to the project}$$

Example 1: The sum of contributions from all participants to the project is 12 Taler (e.g., if you and the two other participants contribute 4 Taler each, or if one of the three participants contributes 12 Taler and the two other participants contribute 0 Taler). Then the two participants of type A each receive an income of $\frac{3}{4} * 12 = 9$ Taler from the project, and the participant of type B receives an income of $\frac{1}{2} * 12 = 6$ from the public good.

Example 2: The sum of contributions from all participants to the project is 36 Taler. Then the two participants of type A each receive an income of $\frac{3}{4} * 36 = 27$ Taler from the project, and the participant of type B receives an income of $\frac{1}{2} * 36 = 18$ from the project.

Income at the end of a period:
T Bun: If contributions have not been fixed and income is not redistributed, T Re: If not all participants decided to adopt the redistribution, Your income at the end of a period is the sum of your income from your private account and your income from the project:

Type A:
Income from the private account (20 – contribution to the project)
*+ Income from the project ($\frac{3}{4}$ * sum of contributions to the project)*
= Income at the end of the period

Type B:
Income from the private account (20 – contribution to the project)
*+ Income from the project ($\frac{1}{2}$ * sum of contributions to the project)*
= Income at the end of the period

Let us illustrate how your income at the end of a period is calculated using two examples:

Example 1: Assume that you are of type A and contribute 16 Taler to the project, just as the other two participants. The sum of contributions is then 16 + 16 + 16 = 48 Taler. Your income in this example would be:

4 Taler from the private account + $\frac{3}{4}$ * 48 Taler from the project = 4 + 36 = 40 Taler

Example 2: Assume that you are of type A and contribute 0 Taler to the project, while the other two participants contribute 16 Taler each. The sum of contributions is then 16 + 16 + 0 = 32 Taler. Thus, your income in this example would be:

20 Taler from the private account + $\frac{3}{4}$ * 32 Taler from the project = 20 + 24 = 44 Taler

T Re, Sim: If all participants supported the redistribution, the incomes calculated above will added up and be distributed evenly among the participants:

Total income from all private accounts *+ Total income from the project ($\frac{3}{4}$ * sum of contributions to the project)*
= Total income of all participants
÷ 3 participants
= Income at the end of the period

Let us illustrate this calculation using one example, as well:
Example: Assume that you are of type B and contribute 0 Taler to the project. The other participants contributed 12 Taler each. Your income in this example would be

20 Taler from the private account + $\frac{1}{2}$ * 24 Taler from the project = 20 + 12 = 32 Taler

The income of the other participants would be

8 Taler from the private account + $\frac{3}{4}$ * 24 Taler from the project = 8 + 18 = 26 Taler

If the redistribution was adopted every participants receives one third of the total income of 26 + 26 + 32 = 84 Taler. The income would be 28 Taler.

T Fix, Re Bun Sim: **The first Stage**
T Fix, Bun Sim: In the **first stage** you can decide whether you want to commit yourself and the other participants in your group to a certain contribution to the project in the second stage. (T Bun: and redistribute the income such that all participants receive the same payoff.) All participants decide simultaneously. Only if **all** participants in your group decide to commit themselves and the other participants to certain contributions, are the contributions actually fixed. In this case contributions would be fixed as follows:

Type A: *Contribution of 20 Taler to the project*

Type B: *Contribution of 20 Taler to the project*

T Bun: The income of all participants is then given by:

Income type A (20 - 20 + $\frac{3}{4}$*60=45 Taler)
+ Income type A (20 - 20 + $\frac{3}{4}$**60=45 Taler)*
+ Income type B (20 - 20 + $\frac{1}{2}$**60=30 Taler)*
= Total income of all participants (120 Taler)
÷ 3 participants
= Income at the end of the period (40 Taler)

T Fix, Bun, Sim: If **not all** participants decide to fix the contributions, you and the other participants in your group can freely contribute any number of your 20 Taler to the project in the second stage.

T Re, Sim: In the **first stage** you can decide (T Sim: independently from your previous decision) whether you want to commit yourself and the other participants in your group to redistribute all income such that everybody in your group receives the same payoff at the end of the period. All participants decide simultaneously. Only if **all** participants in your group decide to commit themselves and the other participants to the redistribution, the incomes are actually redistributed. Only then the incomes will be redistributed such that every participant receives the same payoff.

T Re, Bun, Sim: If **not all** participants decide to redistribute the incomes, your income is given by your income from your private account and from the project.

The second (T Vcm: **first**) **stage**

T Fix, Re, Bun, Sim: At the beginning of the second stage you get to know how each participant in your group decided in the first stage.

T Fix, Bun, Sim: If in the first stage all participants decided to fix the contributions in the second stage, then in the second stage you have to contribute the corresponding amount. Thus, if you are of type A you have to enter a contribution of 20 Taler and if you are of type B you have to enter a contribution of 8 Taler. Other inputs are not possible and will automatically be adjusted by the computer program.

In this case the period incomes of the participants of type A is $\frac{3}{4} * 60 = 45$ Taler and the period income of the participant of type B is $\frac{1}{2} * 60 = 30$ Taler.

T Bun: This is redistributed automatically such that every participant receives 40 Taler.

T SIM: If the participants committed themselves to redistribute the incomes during the first stage, the incomes will be redistributed at the end of the period such that every participant receives 40 Taler.
T FIX, BUN, SIM: If in the first stage not all participants decided to fix the contributions in the second stage (T BUN: and redistribute the incomes), then in the second stage all participants can freely choose any integer number between 0 and 20 to contribute to the project (0, 1, 2, ..., 19, 20).
T VCM, RE: In the second (T VCM first)stage all participants can freely choose any integer number between 0 and 20 to contribute to the project (0, 1, 2, ..., 19, 20).
T SIM, RE: If the participants did not commit to redistribute the incomes at the end of the period,
The period incomes of the participants are computed as indicated above:
Type A: 20 – contribution to the project + $\frac{3}{4}$ * (sum of all contributions to the project)
Type B: 20 – contribution to the project + $\frac{1}{2}$ * (sum of all contributions to the project)
T SIM: If the participants committed themselves to redistribute the incomes, the total income calculated above is divided by three and every participant receives the same payoff.

The third (T VCM: **second**) **stage**
In the third (T VCM: second) stage you get to know the contributions to the project by all participants in your group, as well as their period income. T FIX, RE, BUN, SIM: Furthermore, you will again see how each participant in your group decided in the first stage.
Then the current period ends and the next period begins with the same participants. Your type and the types of the other participants remain the same. T FIX, RE, BUN, SIM: All participants can then again decide in the first stage whether they want to fix contributions (T RE: redistribute incomes) in the second stage. Again, the second stage follows and finally the third stage.
T VCM: All participants can then again decide upon their contributions to the project. Again, the second stage follows.

Conclusion of the experiment and payment

The experiment ends after 10 periods. Subsequently, we will ask you to answer a few general questions on the computer. Your answers to these questions have no influence on how much money you will earn in the experiment. When all participants have filled out the questionnaire, payments will be made. Your total income from the 10 periods will be converted into Euro and paid to you in cash.

Do you have any questions? If so, please raise your hand.

4

Peer Evaluation and Compensation Schemes in a Real Effort Experiment*

4.1 Introduction

Team based work has seen a large rise over the last 60 years. While allowing tremendous gains from close cooperation and specialization, a single worker's production in such teams might be hard to observe for employers and supervisors. This poses a serious challenge for incentive based contracts, as they no longer can be tailored directly to an individual's performance. In many situations, however, the teammates can be assumed to have a good knowledge about the individual contributions to a project (May and Gueldenzoph, 2006): researchers know what their co-authors contributed to an article; employees know who prepared the slides for the CEO. In order to leverage this knowledge firms employ payment schemes that depend on the evaluation by team-members (peer evaluation). Roughly a third of all Fortune 500 companies uses them for incentive purposes (see Bohl, 1996; Johnson, 2004).

While this additional information is undoubtedly valuable for principals, extracting them is not necessarily straightforward. Especially the monetary consequences of the underlying reporting mechanism might affect the reporting behavior. For example, if the evaluations are not subject to any further constraints overly positively evaluations might arise due to a likability or in-group bias as reported by Golman and Bhatia (2012). A different negative effect arises if colleagues compete for promotions or under fixed bonus schemes (Huang et

* I thank Sebastian Kube for his guidance as well as audiences at Bonn, Trier and ESA Europe 2016 for their helpful comments. Laura Ehrmantraut provided excellent research assistance. Financial support by the BGSE is gratefully acknowledged.

al., 2017). These circumstances might put team-members in direct competition with each other. Thus they could be inclined to evaluate certain competitors overly negative in order to increase their own chances of receiving a high payoff. While these arguments put the meaning of performance appraisals in doubt, there exists only scant empirical evidence on the causal determinants of evaluation behavior in different work environments. In this paper I shed light on how incentives, reciprocity and the reporting mechanism itself affect the peer evaluation behavior if workers compete for bonuses.

To study the impact of team-incentives and the reporting mechanism on evaluation behavior I conducted a laboratory experiment. Subjects perform a real effort task in groups. Afterwards they have to evaluate the performance of their fellow teammates. The resulting ranking determines bonus payments for the subjects. Across treatments the evaluation scheme is varied. Subjects either evaluate their teammates independently or are forced to rank them. In addition, between treatments I vary whether teammates profit from each others work; on top of a bonus payment subjects receive either a flat wage or are paid on cumulative team performance. Hence, this design enables me to causally identify how team-incentives and reciprocity affect the willingness to evaluate others truthfully.

In order to analyze the potential implications of peer evaluation in a tournament setting I derive behavioral predictions based on both standard theory and action-based reciprocity. I consider the situation of teams, featuring three workers. The workers play a two-stage game, consisting of an real effort task and an ensuing evaluation stage. During the evaluation stage the workers compete for prizes through a tournament, which are distributed based on the result of the workers' evaluations. Effort levels are assumed to be common knowledge among workers. If no team-incentives are employed subjects are always predicted to evaluate their co-workers in the worst way possible. Consequently resulting evaluations and rankings should bear no similarity with the true performance of the workers. This changes as soon as incentives based on team output (here defined as the sum of all workers' outputs) are introduced. Standard theory still predicts that rankings should not reflect a worker's true performance. If subjects are assumed to be reciprocal, however, the theoretical framework predicts subjects to evaluate each other truthfully once the differences in output among them become sufficiently large. The differences in output that induce truthful rankings are predicted to be lower under the forced ranking treatment. Profound consequences of this behavior are expected on effort choice. If evaluation behavior is indeed truthfully, an incentive for higher effort choices is created. The model is consequently able to endogenize costs of lying in a team-setting by making them conditional on reciprocal feelings.

Nevertheless, whether reciprocity and team-incentives are indeed able to induce workers to evaluate each other truthfully remains ultimately an empirical question. Analyzing the results of the laboratory experiment I find that, if subjects do not profit from their teammates work, their evaluation behavior is very close to the selfish benchmark. They do not evaluate each other positively (under unconstrained evaluations) or seem to randomize the ranking (under constrained evaluations). The results are similar, if team-incentives are introduced under unconstrained evaluation. Subjects do not respond to good performances of teammates by rewarding them with positive evaluations. This behavior, however, changes once they are asked to rank their teammates. In this treatment the share of truthful evaluations increases significantly by 15 percentage points. This results in more than half of all subjects receiving the bonus in line with their actual output rank, in contrast to only a third of subjects receiving it without team-incentives. In both treatments, subjects react to performance differences between their teammates. If one of their teammates performed much better than the other, it becomes significantly more likely that they will be ranked in line with their actual performance. The effect is even stronger under team-incentives. This implies that it is indeed not the existence of the team-incentives per se which induces more truthful evaluations; team-incentives rather change how the participants react to individual performance differences. Interestingly, across all treatments the highest performing individuals tend to evaluate other participants more truthfully. Under unconstrained evaluations this might be a purely technical effect, as here evaluating nobody positively can be part of a true ranking. However, no such technical reasons apply for the other treatments. In general these high performing individuals might be intrinsically motivated to perform and contribute to a public good, potentially signifying preferences for fairness and cooperation in general. In line with this, those participants that do not work at all in a given period tend to evaluate less truthfully as well.

Taken together, these results suggest that evaluations by teammates can be used as a basis for compensation schemes – even if wages are partially determined in a tournament based on said evaluations. However, they indicate that the specific details of the evaluation mechanism, as well as the overall compensation scheme, can have an impact: If employees do not profit from each other's work, or are not forced to rank their co-workers, the resulting evaluations carry little to no meaning. Only if these two aspects are combined, the workers tend to evaluate in line with the true performance and their informational advantage can be leveraged. Given that higher performing individuals seem to be more honest at evaluating (as in Davison et al., 2014) supervisors could try to identify these workers from past periods and might want to weigh their evaluations stronger.

Even if no formal way of performance appraisals by peers is introduced, the results presented here still have implications that should be considered. Employees will always have the opportunity to report on their co-workers. Naturally these "office politics" can hardly be regulated. Combined with the results from above, this implies that these unofficial ways might rarely be used to tout a teammate's good work, but rather to report shirking behavior. Nonetheless, team-incentives might be useful in these situations to increase the likelihood of truthful peer reports, as in Carpenter et al. (forthcoming). While this might help to identify underperforming employees it does not incentivize them to stand out in a positive way.

The results are also in line with evidence that negative reciprocity (e.g., Offerman, 2002; Abbink et al., 2000) is stronger than positive reciprocity, as an unwillingness to reward lower performing co-workers seems to be the driver of the increased meaning of evaluations. While larger performance differences between teammates are associated with more truthful evaluations, subjects seem to punish performance below their own level. This implies that the workers are especially unwilling to reward a teammate at the expense of higher performing ones.

More generally, the results presented are also of interest for the literature on public goods with rewards or punishment. Frequently, individuals are observed to reward high contributions – even at their own costs (Fehr and Gächter, 2000). The treatments that feature a team payment constitute a public good game as well; the only difference being that the contributions are based on the performance in the real effort task. In contrast to the previous literature voluntary rewards are rarely observed. This indicates that the results from standard VCM with rewards might be more difficult to generalize to settings that feature nonmonetary contributions or larger rewards.

The remainder of the paper is structured as follows. Section 4.2 describes the experimental design. Section 4.3 introduces a theoretical framework and presents predictions for the subjects' behavior, using both standard and social preferences. Section 4.4 presents and discusses the results of the laboratory experiment. Section 4.5 concludes.

4.1.1 Related Literature

The results in this paper relate to several recent strands of literature on tournaments, performance evaluation and wage setting (for a broad overview see Dechenaux et al., 2015).

The complexity of modern day work relations makes it harder to correctly asses individual workers' performances and to determine the according wages. A natural way for principals is to use performance appraisals from supervisors. While these were found to have a positive impact on effort (Berger et al., 2012; Engellandt and Riphahn, 2011) they also tend to be biased by social ties (Breuer et al., 2013). All these studies focus on the evaluation decision of a principal towards his employees. This might not always be possible, as in some situations even direct supervisors are not able to observe individual effort.

Given the high prevalence of close work ties in modern work environment a strand of literature tries to leverage the presumably superior information of teammates about each other's performance. For instance, co-workers might be the only one's who witness an individual trying to avoid unpleasant tasks (Fedor et al., 1999). Towry (2003) finds that teams are able to successfully govern themselves once their identity is strong enough; given a strong team-identity individual payments can be agreed upon endogenously without the help of a supervisor. Even when agents do not have any kind of social preferences, peer evaluations can help to extract additional information: Kim (2011) demonstrates theoretically that peer evaluations can be employed optimally to gain information on whom to promote, while Marx and Squintani (2009) show that they can be used to implement first-best efforts in a team-production setting.

Although theoretical models have stressed the advantages of peer evaluations, some empirical findings question the meaning of resulting evaluations. Huang et al. (2017) find that workers indeed alter their evaluations to harm highly qualified colleagues and support lesser qualified ones. Team-incentives were found to increase the propensity to lie in peer evaluations (Conrads et al., 2013), increase in-group favoritism (Hammermann et al., 2012) and increase effort through peer pressure (Mohnen et al., 2008).

In a laboratory experiment, Carpenter et al. (2010) find that subjects do evaluate each other overly negative once these evaluations determine the outcome of a tournament. The lack of positive evaluations under fixed wages basically replicates this finding. The closest paper to mine is Carpenter et al. (forthcoming): The authors look at the interplay between team-incentives and reporting teammates for shirking. Subjects can report their teammates if they deem their work unsatisfactory. The principal in turn has then the opportunity to alter their wages. In contrast to my paper, reporting is never expected to influence a subject's payoff directly, but a principal has the opportunity to punish upon a negative report. In addition the desire to report relies solely on negative reciprocity. The authors find that workers tend to report their teammates for shirking only under team-incentives. My results support the notion that team-incentives can help to encourage truthful evaluations of co-workers. However, they also indicate that this crucially depends on the fact that the evaluating individuals do not sacrifice their own income when doing so.

In my paper, the teammates are confronted with a potentially competitive tournament situation. This might affect effort provision, as well as their behavior towards teammates, as competitive mindsets were found to increase effort choices and prevent collusion if subjects were informed about the competitors behavior (Maas et al., 2011). On the flip side competitive situations are known to increase sabotage (Schwieren and Weichselbaumer, 2010). While no explicit form of sabotage[1] is present in this paper, negative evaluations towards a high performing teammate might be seen that way. Gürtler et al. (2013) find that sabotage is frequently directed at exceptionally strong subjects. This is in line with my finding that additional output by high performing subjects does not always increase their chances to be evaluated positively, but might even be harmful in the absence of team-incentives.

4.2 Experiment

To empirically test the effect of team-incentives on performance appraisals by teammates I conducted a series of laboratory experiments. The laboratory setting allowed me to implement team-incentives as well as to vary the evaluation scheme. Even more crucial, clear and objective performance levels can be recorded and are observed by evaluating teammates. Lastly, all interaction was anonymous and participants could not track each other across time periods. In contrast to a field setting, this has two major advantages: No personal bias can affect evaluations and positive or negative evaluations cannot be exchanged across different time periods through collusion or vendettas thus allowing for a clear interpretation of the treatment intervention.

4.2.1 Experimental Design

In each period of the experiment, the subjects were randomized into groups of three workers. The period itself consists of two stages: During the first stage the participants work on a real effort task. In the second stage they are asked to evaluate other participants on their performance during the previous stage. In each period of the experiment the subjects work on the slider task (as introduced in Gill and Prowse, 2012). They see 48 bars on their computer screen. Their task is to center a slider on each bar using only the mouse of the computer. The goal is to center as many sliders as possible during a two minute stretch. This task yields outcomes that can easily be observed and evaluated by other participants. Moreover, the outcome is mainly determined by the willingness to exert effort during each period and cannot be completed by guessing or knowing.

After the effort stage subjects were informed about their own performance (the number of correctly positioned sliders) and about the performance of the other two workers within their group. The participants then had to evaluate the performance of these two group members. The resulting evaluations were used to pay out an individual bonus that was a part of their payment.

Across treatments two evaluation procedures (FREE and FORCED) were employed. In both procedures the participants were asked to award points for

[1] Sabotage in organizations has been studied extensively, both theoretically as well as empirically (e.g., Gürtler, 2008; Gürtler and Münster, 2010; Kräkel, 2005).

"good" work to their team-members. In condition Free the subjects could evaluate their group-members individually. They had the option to evaluate each of the two other subjects either positively by assigning them a point or negatively by assigning no point. That means they could assign points to both, one or none of them. By contrast, in condition Forced the subjects *had* to hand out exactly one positive and one negative evaluation to the other two group-members (i.e., they had to assign exactly one point to one other group-member).[2] After all participants finished their evaluations, the number of points that each of them received were counted. The individual with the most points gained a fixed bonus payment that was not impacted by the number of points received. The subject with the second most points received also an, albeit smaller, bonus payment. The subject with the least number of bonus points received no bonus at all. That means that the three subjects that form a group compete in a setting, where the resulting payments are determined by their own evaluations.

This bonus B and a base wage w determined the subjects' payoffs. The structure of the base wage was varied across treatments. In treatments Fix the base wage was fixed and independent of any performance. Each subject received a flat wage of $w = 80$ points per period. The other treatments (Team) featured a variable base wage. Subjects are paid for the cumulative number of correctly positioned sliders by all three subjects within their group. For each slider correctly positioned by a group-member every subject within the group received a point. This results in two possible wage functions:

$$\pi = 80 + B$$

and

$$\pi = \sum_{i=1}^{3} Output_i + B.$$

As detailed above, the second part of the payment – the individual bonus B – is determined via the peer evaluation. The subject with the highest number of points received a bonus of $B^1 = 80$, the subject with the second most points received $B^2 = 40$, and the subject with the least points was paid $B^3 = 0$. If there was a tie between two or three subjects the corresponding boni were equally split amongst them. That means if two subjects received a point each, whilst the third group-member received no point at all, the two subjects with a point shared B^1 and B^2 and received 60 tokens each. Combining the two evaluation schemes and the two wage regimes results in a 2 × 2 treatment matrix, featuring two different evaluation stages and two different base wages. A summary of the treatments is shown in the Table 4.1 below.

[2] In this sense the forced ranking scheme mirrors the classical Borda count. This evaluation mechanism can be found in Eberlein and Walkowitz (2008).

Table 4.1. Summary of Treatments

	fixed wage $w = 80$	team payment $w = \sum Output$
free evaluation	Fix-Free	Team-Free
forced ranking	Fix-Forced	Team-Forced

4.2.2 Procedures

The experiment was computerized using z-Tree (Fischbacher, 2007) and conducted at the BonnEconLab at the University of Bonn in 2014. Students were recruited randomly from all majors using Orsee (Greiner, 2015). Two sessions were conducted for each treatment featuring 24 participants each. In total 192 subjects participated. The subjects were randomly divided into two matching groups of 12 subjects each. The experiment consisted of 10 identical periods. For each period new groups of three workers were formed (stranger matching protocol). The three workers were always drawn from the same group of 12 participants. Subjects were informed about this procedure. This was done to prevent the subjects from carrying positive or negative reciprocal feelings over into following periods. During two practice periods prior to the experiment the subjects had the opportunity to familiarize themselves with the slider task.

Each session lasted about 90 minutes. All interaction between the participants was anonymous and decisions were taken in private at the computer. After each period the subjects were informed about the evaluation decision of their team-members, as well as the results and the payoffs for the period. Written instructions were distributed prior to the experiment and read out aloud. Afterwards subjects had the possibility to ask questions for clarification and had to answer control questions to ensure their understanding. Throughout the entire experiment tokens were used as artificial currency, with 75 tokens equaling 1 Euro. The average payment for the subjects was 14.24 Euro, ranging from 12.48 in the team-paid treatments to 16 Euro in the treatments with fixed payment. The payments were made in cash in private directly after the experiment.

4.3 Theoretical Framework

In this section, I present a simple model of action based reciprocity to motivate how team-incentives can endogenously create sincere evaluations.[3] Following

[3] Reciprocity is not the only non-standard preference that might explain truthful evaluation in a one-shot experiment. Aversion to lying might be another possible motive. Nonetheless this motive would be the same across all treatments and would not predict any treatment differences. A similar argument can be made for other social preferences such as inequity aversion. If solely

the experimental design, I will analyze how an individual reacts to observed effort (or output) choices. Workers might want to reward their teammates for high effort with positive evaluations or punish them with bad ones. Hence, I consider a model that features action-based-reciprocity as a natural choice to derive behavioral predictions for the evaluation behavior. The form of reciprocity considered here will thus be similar to the models of Cox et al. (2007) or Englmaier and Leider (2012).[4] Reciprocity between teammates should only be present if workers actually profited from each other's effort in the first place. As a consequence this version of reciprocity will predict sincere evaluations only under the presence of team payments. If wages are fixed, reciprocal individuals will act the same as selfish individuals. Assume that each individual incurs effort costs $c(e_i)$, with $c' > 0, c'' > 0$, for performing the task. π_i denotes that individual's monetary payoff. This consists of the bonus B that the individual receives based on the evaluation as well as either the fixed payment w or the team-payment $\sum_{i=1}^{3} e_i$. I assume the output to be a deterministic function of effort, thus random events do not play a role.[5] In the following I assume that an individual's own level of effort serves as a reference point for the evaluation of teammates.[6] This means that individuals react to efforts that are larger than their own with positive feelings and to lower efforts with negative feelings. I assume the following utility function:

$$U_i = \pi_i - c(e_i) + \phi \sum_{j \neq i} (c(e_j) - c(e_i))(\pi_j - \pi_j^{fair}) \tag{4.1}$$

This means that individuals profit from higher payoffs for their co-workers, if these co-workers exerted more effort than they did. The opposite holds true for teammates that exerted lower effort. The marginal utility from higher or lower payoffs increases in the effort distance. This implies that more extreme effort choices carry more weight. In short, if the positive reciprocity is strong enough, the worker might be willing to reward a high-performing teammate. In

differences in effort costs would induce favorable evaluation for those that worked more, the presence of team wage should not matter.

[4] Throughout this paper I focus on action-based reciprocity, i.e., reciprocity arises in response to previous kind or unkind actions. Nonetheless belief based reciprocity could play a role in all treatments as well. That would imply that subjects would reciprocate to positive evaluations that they believe they will receive during that period. This form of reciprocity, however, would not predict any treatment differences.

[5] If output was at least partially random, it would not be fully attributable to an individual. Given that lack of responsibility, reciprocal feelings are harder to justify between teammates.

[6] This setting of a reference point is not restrictive. Any other reference point for the evaluation of performances will lead to qualitatively similar results. Higher reference points simply require higher performances by teammates in order to induce positive reciprocal feelings. Consequently, higher reference points simply result in stronger requirements for truthful evaluations but do not rule them out.

general this willingness increases in the effort difference as well as the workers' individual propensity to reciprocate (ϕ_i).

In the following, I shortly describe the equilibrium evaluation behavior in response to observed effort choices across treatments. First, for individuals without reciprocal preferences, then for individuals with. Extensive derivations of these predictions are provided in Appendix 4.A. As the paper focuses on the evaluation behavior in response to effort differences, I will only present prediction on the workers' evaluation behavior and not on the effort decision itself. Appendix 4.A provides a detailed analysis of the effort choice in response to sincere evaluations.[7]

If an individual does not posses any reciprocal attitudes, this person will seek to maximize their own monetary payoff. Hence, neither previous effort levels of teammates nor the payment scheme in the first stage are predicted to have any influence on the workers' evaluation decisions. Consequently the standard model predicts no behavioral differences between treatments that feature group wages or fixed wages. Effort differences among the workers are not expected to impact the decision as well. A worker with standard preferences will always evaluate as to maximize her own monetary payoff. Thus, positive evaluations in the unrestricted condition are never predicted to occur. A positive evaluation will always (weakly) decrease the bonus payment of the individual that handed it out. If no positive evaluations are used, this results in all workers being ranked equally. In expectation the results are not different for the Forced treatments. Again, effort differences are not predicted to impact evaluation decisions. Hence, in equilibrium the subjects would randomize the positive vote among their teammates, as their own payoff is never impacted by their evaluation decision.[8] Consequently, in expectation all workers will be ranked equally. In general, under standard preferences the resulting ranking will never be affected by the individual performances.

The evaluation behavior is predicted to differ for subjects with reciprocal preferences (ϕ_i is sufficiently large). In the treatments that feature no constraints on the evaluation scheme subjects do not have to hand out positive evaluations. A positive evaluation – depending on the teammates' evaluation behavior – might or might not impact a subject's own monetary payoff. If their own wage is not affected by the evaluation, they will always be (weakly) better

[7] Given that the evaluation behavior is truthful, I argue that no equilibrium in pure strategies for the effort choice exists. Rather, effort is selected from a certain interval.

[8] If an individual does not receive any positive evaluation from her teammates, she will be ranked last no matter her decision. If she receives two positive evaluations, she will be ranked first, independently of her action. If she receives a single positive evaluation, she can hand her positive evaluation to that teammate that has already received one. Then she will be ranked second. Otherwise she hands it to the teammate, that has not yet received a positive evaluation. Then all team-members received one positive evaluation and will be ranked equally. Given the parameters of the experiment this results in the same payoff.

off by reporting truthfully. An example of this would be two workers awarding a point to each other but none to the third co-worker. No matter the third worker's action, she will always receive the smallest bonus. Her evaluation, however, can impact the bonuses of the two other subjects. Rewarding the better performing worker is always the best reply in this situation: The downside of reducing the higher performing subject's payoff will always be larger than the upside of improving the payoff of the lower performing one. If the subject's own monetary payoff is affected by a potential evaluation decision, this decision is not as clear cut as before. A positive evaluation will always lower the own wage, while it increases the wage of the positively evaluated teammate. This is only desirable for the subject, if the utility from the money forfeited is smaller than the gains from awarding the teammate the higher bonus. This second utility from awarding the "deserved" bonus gets larger with increasing effort differences. This means that if her own monetary payoff is affected by positive evaluations, a subject will hand them out only to those co-workers that performed sufficiently better than herself.

If the evaluation scheme is constrained, the evaluation decisions become interrelated. Awarding a point to one co-worker simultaneously serves as a punishment for the other one, as her payoff will (weakly) decrease. As the worker suffers from rewarding co-workers with lower effort levels, the forced evaluation scheme increases incentives to reward workers with higher effort levels. Hence, awarding the point to the co-worker which exerted the lesser effort will decrease the reciprocity utility the worker receives. The subject will always be at least indifferent between the two options as long as her own monetary payoff remains unaffected. Given the parametrization used in the experiment this will always be the case.[9] Hence, a reciprocal worker is always predicted to evaluate truthfully. In case of a tie, the worker will randomize.

In summary, reciprocal preferences predict that more truthful evaluations arise if the teammates profit from each other's effort. Moreover, the share of truthful evaluations is expected to be higher under the constrained evaluation scheme and to be higher for larger differences in the effort level among the teammates.

Prediction 1:

> Bonus payments reflect performance differences more in treatments with team-incentives than in treatment without.

[9] In the Appendix 4.A I demonstrate, that only one worker can affect her own ranking by switching from truthful to untruthful evaluations. This deviation results in all subjects being ranked equally. As the bonus is linear in the rank in this experiment, this subject is not able to increase her monetary payoff this way.

Prediction 2:

Bonus payments reflect performance differences more in treatments Team-Forced than in treatment Team-Free.

4.4 Results

The experimental design allows me to study the impact of team-incentives on peer evaluations. The evaluation scheme was varied across treatments such that it either allowed subjects to evaluate their teammates independently (Free-Treatments) or forced them to rank their teammates (Forced-Treatments). While the subsequent analysis focuses on the evaluation behavior, I start by describing the performance levels across the different treatments. If subjects would have refused to work at all in the treatments without team-incentives, this would make the subsequent analysis meaningless. In a next step the evaluation behavior of the subjects as well as the resulting rankings will be analyzed. Here, I will mainly compare evaluation behavior across treatments that feature the same evaluation scheme, as evaluation options differ fundamentally across schemes.

4.4.1 Performance levels

Across all treatments the participants position on average 17 sliders per period correctly. Reciprocal preferences predict that workers react to output differences under the team-wage conditions but not under the fixed-wage conditions. If participants would not work at all if wages are fixed, identifying the impact of team-incentives on evaluation behavior would be impossible. While the subjects indeed reposition less sliders in the treatments with fixed wages, almost all participants still work on the task. Only 8.2% of all observations in the treatments with fixed wages feature participants that did not work on the sliders at all. In treatments with fixed wages the participants position on average 15.9 sliders per period. This increases to 18 sliders per period in treatments that feature the team payment. This means the average output is 12% lower in the treatments where wages are fixed compared to those with team-incentives. The detailed outcomes by treatment are summarized in Table 4.2. Looking at the dispersion of performances I find that the standard deviation is higher in the two treatments that feature the fixed wages.[10] The lower part of Table 4.2 which breaks the output level down by the subject's rank within her team reveals that the

[10] By design the incentives to produce any output are different across the two payment schemes. Only in the treatments without team-incentives participants ever decided not to work at all. I check whether the output is different beyond this unwillingness to work at all in Table 4.B.1 in Appendix 4.B. This table presents the results of a corresponding regression analysis. If one accounts for the subjects that do not work at all, only the output levels for subjects in treatments Fix-Forced and Team-Free differ at the 10%-level.

fixed wage primarily affected the performance of the less productive subjects within a group (i.e., the bottom of the performance distribution). In contrast, the outputs of the highest performing subject in a team are very similar across treatments and never statistically different from each other.[11]

Table 4.2. Summary Statistics for Sliders Solved

	Fix-Free	Team-Free	Fix-Forced	Team-Forced	Total
Aggregate	15.86	18.37	15.91	17.68	16.96
	(7.22)	(4.26)	(6.80)	(4 .85)	(6.01)
Observations	480	480	480	480	1920
By Outputrank					
Output Rank = 1	21.12	21.66	20.71	21.33	21.20
	(4.68)	(3.49)	(4.04)	(4.50)	(4.21)
Observations	171	174	176	177	698
Output Rank = 2	14.82	18.12	15.70	17.23	16.44
	(6.41)	(3.06)	(5.21)	(2.91)	(4.84)
Observations	173	162	161	160	(656)
Output Rank = 3	10.59	14.67	10.25	13.66	12.32
	(6.35)	(2.94)	(6.65)	(3.40)	(5.44)
Observations	136	144	143	143	566

Notes: Standard deviations are presented in parentheses. Note that in some groups ties between the participants occurred. If their output was the highest within the group both were counted as rank=1, if the tied occurred not for the first rank, both were counted as rank=2.

In general, the performance levels of the subjects vary greatly within their respective teams, independent of the wage scheme. Nonetheless, the average output level still differs across treatments. Consequently, the subsequent section will take this into account by including the output levels and their differences within a team into the subsequent analysis.

4.4.2 Outcome of the Evaluation Stage

In this section I will compare the results of the evaluation stage across treatments. First, I will analyze how many subjects received a bonus that was in line with the actual output rank within their group, i.e., are ranked correctly.[12] Standard theory predicts no differences between treatments: In each treatment a third of all subjects is predicted to receive the wage in line with their output

[11] Even if all observations in a treatment are viewed as independent, the highest output in a group does not differ significantly across groups (MWU p>0.10 for all possible treatment comparison).

[12] Note, that implies also that whenever all subjects receive identical evaluation and are ranked equally the second highest performing subject receives the correct bonus.

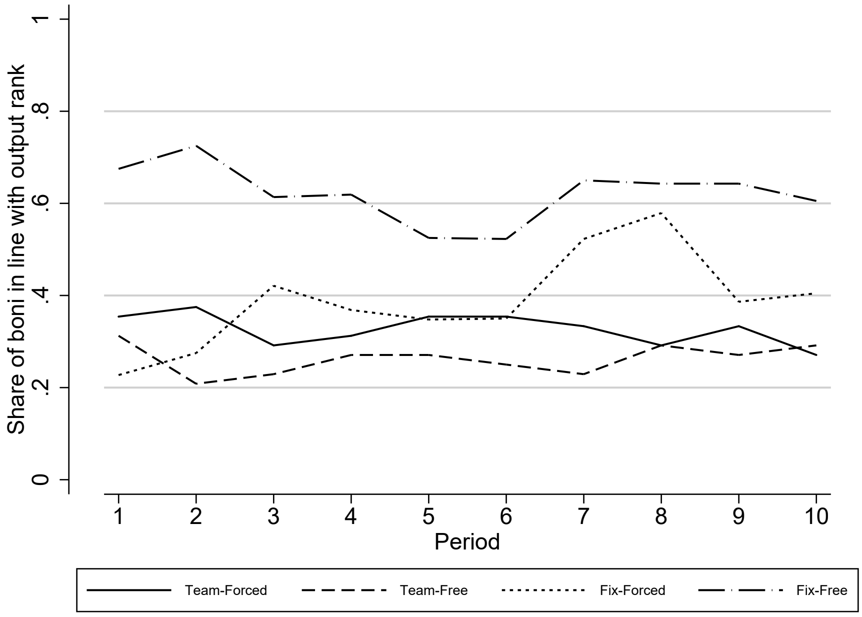


Figure 4.1. Share of subjects with "correct" bonus over time

Notes: The figure presents the share of subjects that received the bonus associated with their output rank across periods.

rank. If reciprocal preferences are assumed, the share is predicted to be higher in treatments with team-incentives.

In the next step I will investigate how the evaluation decisions that lead to these outcomes differ across treatments. As highlighted in section 4.3 the two evaluation schemes introduce different motives to evaluate the other workers sincerely. While the unconstrained evaluation scheme is predicted to rely only on positive reciprocity for positive evaluations through rewards, under constrained evaluation positive and negative reciprocity are predicted to play a role. A positive evaluation (reward) for one co-worker always implies a negative (punishment) for the other. As the set of evaluation option differs fundamentally, I will focus on the direct comparison of the two treatments that feature the constrained evaluation first, and then on the ones with the unconstrained evaluation scheme. In the Forced treatments only awarding a point to the better performing teammate can be regarded as sincere evaluation. In the Free treatments, however, other evaluations can also lead to a correct ranking (e.g.: Not awarding points at all by the highest performing subject, or awarding points to both other teammates by the lowest performing subject). Infrequently two teammates had the same output. In the treatments with the constrained evaluation scheme, neither the remaining teammates evaluations nor the resulting rankings could reflect that tie. Hence, these observations were excluded from the analysis. Figure 4.1 displays the share of subjects that are correctly ranked in each treatment across time. Table 4.3 shows the corresponding averages for each treatment.

Table 4.3. Share of Subjects with "correct" Bonus by Treatment and Rank

	Fix-Free	Team-Free	Fix-Forced	Team-Forced
Aggregate	32.71	26.25	38.65	62.14
Observations	480	480	414	412
By Outputrank				
Output Rank = 1	26.90	21.84	26.06	54.61
Observations	171	174	142	141
Output Rank = 2	50.29	40.74	50.39	72.66
Observations	171	162	129	128
Output Rank = 3	17.65	15.28	40.56	60.14
Observations	136	144	143	143

Notes: Note that in some groups ties between the participants occurred. If their output was the highest within the group they were counted as rank=1; if the tie occurred not for the first rank, they were counted as rank=2. Their received bonus was counted as correct only if it reflected that tie, i.e., in a case of two subjects both having the highest output, this means they share the two largest boni. This was only possible in the FREE-treatments. In the FORCED-treatments these observations were excluded.

Across all periods the share of correctly ranked subjects is highest in treatment TEAM-FORCED. The results of the other three treatments are relatively similar. On aggregate 62% of subjects receive the bonus that is in line with their output rank within their group. This share is significantly lower in treatment FIX-FORCED, where only 38% of all subjects are ranked correctly (MWU $p < 0.01$)[13]. In both treatments with unconstrained evaluation these shares are slightly, but not significantly lower than in FIX-FORCED (MWU $p > 0.17$, for all possible comparisons). 32% of subjects are ranked correctly in FIX-FREE and only 26% in TEAM-FREE – both values are even below the theoretically predicted share for selfish subjects. Notably, the treatments display different patterns across periods: In FIX-FORCED the share of correctly ranked subjects increases significantly over time (Spearman's $\rho = 0.83$, $p < 0.01$), whereas the other three treatments display stable or slightly decreasing time trends ($p > 0.05$). The corresponding regressions that utilize the entire data are displayed in Table 4.4. Column (1)

[13] While the shares quoted encompass all periods, the statistical tests on the treatment comparisons are based on group averages in the first period only. Here the voting behavior in each single group can be seen as independent from all other groups. As this approach disregards the overwhelming part of the observed behavior, the later parts of the analysis will focus on regression analysis that clusters on the matching group level and controls for possible time-trends.

Table 4.4. Impact of treatments on ranking outcomes

	Correctly ranked			
	(1)	(2)	(3)	(4)
Team-Forced	0.95***	0.99***	0.98***	1.08***
	(0.20)	(0.21)	(0.21)	(0.28)
Fix-Forced	0.35***	0.39***	0.36**	0.35
	(0.12)	(0.14)	(0.14)	(0.23)
Fix-Free	0.19	0.21	0.17	0.32
	(0.16)	(0.17)	(0.19)	(0.30)
Output Rank=2		0.58***	0.62***	-0.45 ***
		(0.13)	(0.14)	(0.13)
Output Rank=3		0.02	0.03	-0.07
		(0.10)	(0.10)	(0.14)
abs. Output size		-0.00	-0.00	-0.01
		(0.01)	(0.01)	(0.01)
Output=0		-0.40**	-0.52***	-0.17
		(0.18)	(0.19)	(0.22)
$\lvert \Delta Output \rvert$ to teammate			0.01	
with higher output			(0.01)	
$\lvert \Delta Output \rvert$ to teammate			0.01	
with lower output			(0.01)	
Period dummies	Yes	Yes	Yes	Yes
N	1786	1786	1786	1114
$PseudoR^2$	.06	.09	.09	.09

Notes: This table presents probit regressions using the resulting rank as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

shows again that the forced evaluation scheme leads to boni being more often in line with the subjects' output ranks; this effect is higher under team-incentives.

Result 1:

a) Significantly more subjects receive a bonus in line with their output in treatment Team-Forced than in all other treatments.

b) Team-incentives do not lead to more subjects being ranked under the free evaluation scheme.

Table 4.3 highlights a strong heterogeneity by the output rank of the subject. Across all treatments subjects that performed second-best in their group received the correct bonus most often – almost twice as often as the other subjects. Hence, output ranks are included in column (2) to (4) of Table 4.4. Column (4) excludes ties in the ranking outcome and demonstrates that they are the sole reason for this effect.[14] The theoretical framework predicted that subjects might be ranked in line with their output rank if the output differences to their teammates are larger. Thus column (3) includes the output differences to the two teammates – separately for the one with the higher and the lower output of the two. However, these output differences do not have any significant effect on the evaluation outcome.[15]

In the next step I will present the sources of the treatment differences by analyzing the individual evaluation behavior and its determinants. As the bonus payments are a direct result of the subjects' evaluation behavior, this broadly mirrors these results. As mentioned above, when interpreting the share of sincere evaluations one has to separate between the two evaluation schemes. Under the Forced scheme, subjects have only two options to evaluate their teammates – either they evaluate one or the other positively. This implies that even a randomizing individual will evaluate half of the time truthfully. The rational benchmark for the Free-treatments is lower (33%). If nobody evaluates anyone else positively, this can be considered a sincere evaluation for the best performing subject. The subsequent analysis will be separated between the two evaluation schemes.

4.4.2.1 Constrained Evaluation Scheme

Under fixed wages (Fix-Forced) the share of truthful evaluations is close to the selfish benchmark: Evaluations are correct in 59.3% of all cases. This share rises significantly to 74.5% under team-incentives (MWU $p < 0.01$). Table 4.5 displays the share of sincere evaluations by treatments and output rank within the group. The share of correct evaluation differs markedly between output ranks. In both treatments, subjects that had the highest output in their group evaluate their teammates sincerely with the highest frequency. There is no such gap in evaluation behavior between the subjects that ranked second or third in their respective groups. In treatment Fix-Forced they evaluate sincerely in about half of all cases, this rises to more than 70% with team-incentives. The best perform-

[14] Tables 4.C.2 and 4.D.4 in the Appendix repeat this analysis for the two evaluation schemes separately. Under constrained evaluation all ranks are equally likely to receive their correct bonus. Under unconstrained evaluation second-ranked subject receives it significantly less often if ties are excluded. A correct bonus for the second best performing subject in the absence of ties requires that three positive and correct evaluation are awarded within a group. As presented only very few positive evaluation were handed out, making this event very unlikely. In addition, Table 4.C.2 in Appendix 4.C presents a further analyses that includes the cumulative absolute output differences to the two teammates for treatments with constrained evaluation scheme.

[15] As the treatments alter the incentives to produce any output, the treatments might also change who is ranked first or second within a given group. Hence these controls are at least partially endogenous and I do not claim a causal relationship between a person's output rank and the probability to receive the correct bonus.

Table 4.5. Share of True Evaluations by Treatment and Rank

	Fix-Free	Team-Free	Fix-Forced	Team-Forced
Aggregate	48.33	42.08	59.32	74.49
Observations	480	480	444	443
By Outputrank				
Output Rank = 1	88.89	81.61	70.00	80.75
Observations	171	174	160	141
Output Rank = 2	32.95	17.90	53.80	70.70
Observations	171	162	158	157
Output Rank = 3	16.91	21.53	52.38	71.20
Observations	136	144	126	125

Notes: Note that in some groups ties between the participants occurred. If their output was the highest within the group they were counted as rank=1, if the tied occurred not for the first rank, they were counted as rank=2. In the FORCED-treatments evaluations could not reflect ties between the teammates. These observations were excluded.

ing subjects also evaluate sincerely in 70% of all cases without team-incentives. In treatment TEAM-FORCED this share rises even further to over 80%.

Result 2: Under constrained evaluation

a) Evaluations are significantly more likely to be truthful under team-incentives.

b) Particularly the best performing subjects evaluate truthfully.

The regression analysis presented in Table 4.6 confirms that there is indeed a significant impact of team-incentives (column (1)), as well as the individual rank on the sincerity of the evaluation (column (2)). Dummy variables for each period are included to flexibly control for potential time trends. Based on the theoretical framework a subject is predicted to evaluate more truthfully the higher the output differences between the two teammates are. Hence, this output difference is included in column (3), in addition to a subject's own output.[16] In line with the theoretical prediction higher output differences between the two teammates make true evaluations more likely. Column (4) interacts the output differences and treatment. While higher output differences make truthful evaluation more likely in both treatments their impact is greater in treatment TEAM-FORCED; the difference between the estimated coefficients is weakly significant.

[16] Table 4.C.3 in Appendix 4.C presents alternate specifications that include additionally the output difference between the evaluator and the two teammates individually. Results on the effect of the treatment remain similar, but they indicate that a lower performance of the worse performing teammate increases the likelihood of sincere evaluations.

Table 4.6. Determinants of Evaluation Behavior Forced treatments

	True Evaluation			
	(1)	(2)	(3)	(4)
Team pay	0.43** (0.17)	0.42** (0.18)	0.46** (0.19)	0.34* (0.20)
Output Rank=2		-0.41*** (0.11)	-0.48*** (0.12)	-0.50*** (0.12)
Output Rank=3		-0.42** (0.17)	-0.40** (0.16)	-0.42*** (0.16)
abs. Output size		-0.01 (0.02)	-0.01 (0.02)	-0.01 (0.02)
Output=0		-0.32 (0.43)	-0.36 (0.43)	-0.37 (0.44)
$\|\Delta Output\|$ between teammates			0.02** (0.01)	
Fix pay × $\|\Delta Output\|$ between teammates				0.01** (0.01)
Team pay × $\|\Delta Output\|$ between teammates				0.03*** (0.01)
Period dummies	Yes	Yes	Yes	Yes
N	887	887	887	887
$PseudoR^2$	.03	.04	.04	.05
p-value: Team vs. Fix				.098

Notes: This table presents probit regressions using the truthfullness of the evaluation behavior as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

The treatment indicator is only weakly significant once this interaction is accounted for. This reveals that the higher truthfulness of evaluations is not only an effect of the team-incentives per se, but that they change how the subjects respond to output differences.

Result 3: Under constrained evaluation

a) Larger differences between the teammates' outputs are associated with more sincere evaluations.

b) This effect is slightly stronger with team-incentives than without.

4.4.2.2 Unconstrained Evaluation Scheme

If subjects can allocate positive and negative evaluations freely, selfish individuals are predicted to never evaluate positively. This behavior can never be seen as sincere evaluation except for the best performing subject. This would mean that only the best performing subjects would ever evaluate truthfully.

In treatment Fix-Free 48.3% of all subjects evaluate their teammates correctly. The share of sincere evaluations rises to 88% if a subject actually evaluated at least one teammate positively. However, positive evaluations are rather scarce; only 0.27 positive evaluations are handed out per period and subject. Positive evaluations are mainly used to reward teammates that produced a high output. If a subject created the highest output, she received on average 0.49 positive evaluations, this drops to 0.26 for the next best teammate and to 0.09 positive evaluations for the lowest performing teammate. At 42.1% the share of truthful evaluations is also low in treatment Team-Free. The differences between the two treatments Team-Free and Fix-Free is not significant at any conventional level (MWU $p > 0.2$). If there is at least one positive evaluation handed out, these evaluations are truthfully in 80.9% of all cases. However, at only 0.3 positive evaluations per period, they are similarly rare as in Fix-Free. Their distribution differs only slightly: The most productive group member received on average 0.44, the second highest 0.35 and the lowest output 0.14 positive evaluations respectively.

While outcomes of the evaluation process did not vary substantially across periods, the positive evaluations that are used exhibit a strong time trend. 47.3% of all points are already awarded in the first three periods and very rarely afterwards.[17]

Table 4.7 displays the results of the corresponding regression analysis. Column (1) shows that team-incentives do not seem to have a significant impact on the share of truthful evaluations.

Result 4:

> If evaluation options are not constrained, evaluations are not more sincere with team-incentives than without.

Column (2) additionally includes their own output level as well as output ranks. The inclusion of these additional mediators result even in the treatment indicator becoming significantly negative. Similar to the treatments with constrained evaluation there are strong differences in the truthfulness of evaluation behavior across the relative placements within the group. As can be seen in Table 4.5 most correct evaluations come from subjects that performed the best; they are truthful in 88.9% (Fix-Free) and 81.6% (Team-Free) of all cases. However, this has clear mechanical reason. For the subjects with the highest output in a group awarding no positive evaluation is still sincere. Sincere evaluations are less frequent for the other subjects, with slightly more truthful evaluations occurring in treatment Fix-Free. Combined with a nearly identical number of

[17] As a result the share of true evaluation decreases in both treatments slightly across time (Spearman's ρ =-0.66, p=0.04 for Fix-Free and ρ=-0.59, p=0.07 for Team-Free). This means that the aggregate numbers discussed above will even overstate the equilibrium-level meaning of the evaluations. This decreasing share is presented in Figure 4.D.2 in Appendix 4.D.

Table 4.7. Average treatment effects for Free evaluation

	(a) True Evaluation			
	(1)	(2)	(3)	(4)
Team pay	-0.14 (0.11)	-0.37** (0.17)	-0.41** (0.19)	-0.73*** (0.21)
Output Rank=2		-2.40*** (0.17)	-2.41*** (0.17)	-2.40*** (0.16)
Output Rank=3		-2.32*** (0.23)	-2.26*** (0.23)	-2.27*** (0.23)
abs. Output size		0.03* (0.02)	0.03 (0.02)	0.02 (0.02)
Output=0		-0.13 (0.21)	0.04 (0.22)	0.13 (0.24)
$\|\Delta Output\|$ to teammate with higher output			-0.03* (0.02)	
$\|\Delta Output\|$ to teammate with lower output			0.00 (0.01)	
$\|\Delta Output\|$ to teammate with higher output × Fix pay				-0.04** (0.02)
$\|\Delta Output\|$ to teammate with higher output × Team pay				-0.00 (0.03)
$\|\Delta Output\|$ to teammate with lower output × Fix pay				-0.00 (0.02)
$\|\Delta Output\|$ to teammate with lower output × Team pay				0.03*** (0.01)
Period dummies	Yes	Yes	Yes	Yes
N	960	960	960	960
$PseudoR^2$	.0091	.41	.41	.41
p-value: Team vs. Fix High				.36
p-value: Team vs. Fix Low				.03

Notes: This table presents probit regressions sing the truthfullness of the evaluation behavior as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

points awarded across treatments, this implies that in Team-Free positive evaluations are given out to the worse performing of the two teammates more often than in Fix-Free (25.3% vs. 16.6%).[18]

[18] Within the theoretical framework rewarding of worse performing subjects is never predicted. However during the questionnaire that followed the experiment some subjects stated that they sent points to low performing subjects in order to motivate them for upcoming periods – even in the presence of stranger matching.

If a teammate produced a much higher output than the subject, she might be willing to reward that teammate with a positive evaluation. Column (3) includes these differences in output to the two teammates. The subjects evaluate each teammate individually. Hence, I include differences in output to them separately, i.e., one difference in output for the higher performing teammate, and one difference in output for the lower performing teammate. Column (4) separates these effects by treatment. While column (3) seems to indicate that larger output differences are not associated with more truthful evaluations, column (4) reveals an interaction with the presence of team-incentives. Without them subjects tend to evaluate less truthfully for larger output differences. The opposite holds for the lower performing subjects under team-incentives: Subjects evaluate more often truthfully if the output difference gets larger, e.g., the teammate performs worse. This means that subjects rarely reward their teammates for high output.[19] In summary, these findings show that team-incentives are not able to encourage subjects to reward their teammates with positive evaluations for high output. If anything, evaluation behavior is more truthful under fixed wages. This implies that either no reciprocal feeling were induced by the task among the teammates or that positive reciprocity was too weak to encourage the subjects to reward their teammates.

In general, the results imply that a restriction of the evaluation options can have beneficial effects for the meaningfulness of evaluation behavior. Still evaluations tend to be truthful only if they are coupled with some kind of team-incentive. If not, evaluation behavior occurs to be almost random. Consequently only treatment Team-Forced resulted in a high share of correctly ranked subjects in equilibrium. The share of correct ranks in the other treatments is close to the prediction for workers without reciprocity. Combining these observations with the results on the unconstrained evaluation scheme implies that only the combination of positive and negative reciprocity is able to impact the reporting behavior significantly. This supports the view that there is a high degree of unwillingness to reward participants for their low output. These results are in line with the behavior observed for negative reciprocity by Carpenter et al. (forthcoming).

[19] In general, the impact of the output differences might differ depending on their sign. If subjects were reciprocal, only suitably large positive output difference should lead to truthful evaluations, large negative output differences should not matter at all. Table 4.D.5 in Appendix 4.D separates the output differences by sign. In contrast to the theoretical predictions a better performance of the higher performing teammates decreased the likelihood of a truthful evaluation significantly.

4.5 Conclusion

In many workplace-situations individual work efforts are hard to observe. Asking well informed co-workers is a prominent way to determine individual payments. In this paper I demonstrated how the existence of team incentives, as well as the evaluation mechanism, affect peer evaluations. Evaluations were used to determine bonus payments for workers. Across treatments, subjects had to evaluate each other either independently or were forced to rank their teammates. In addition to the bonus the subjects received a base wage. This wage was either fixed or depended on the cumulative performance of all teammates. In all treatments money-maximizing subjects' evaluations are never predicted to be related to the actual performance levels: If not forced, they should not evaluate anyone positively; if forced to rank their teammates, the ranking should be random. In the presence of reciprocal subjects, this might change under team-incentives. If the performance difference was large enough, the theoretical framework predicted that reciprocal subjects should report a truthful ranking or use positive evaluations. These predictions were only partially supported by the data. Team-incentives indeed increased the amount of truthful evaluations, but only if subjects were forced to rank their teammates; otherwise very few positive evaluations were handed out. In line with the predictions, higher performance differences between the teammates encouraged subjects to rank them correctly. Even though this effect could be observed in both treatments, it was stronger under team incentives. Hence, it highlights the importance of inducing reciprocal feelings among teammates for truthful evaluations.

While the quantitative results might be specific to this experiment and the laboratory setting, the general influence of the evaluation scheme is of broader interest. The results suggest that peer evaluation should be more informative under some form of profit sharing. Additionally, workers seem reluctant to reward co-workers for good performances if their own wage might suffer from doing so. However, they respond to negative performances. Thus, principals might be well advised to minimize this spillover on own wages, and – even more importantly – making bad as well as good evaluations mandatory in order to punish low-performers as well as to incentivize and reward high-performers. Although informal evaluations, such as peer reporting or whistle-blowing, might be able to inform managers about bad behavior, handing out rewards to the better performing individuals becomes far more attainable in a more formalized evaluation process that forces workers to rank each other.

In general, behavioral biases or social preference frequently affect optimal contracts or wage setting. In this situation reciprocity makes individual incentives possible in the first place. Even if the principal was not present in this experiment, the design presented here can be readily extended to include a supervisor. Further research could possibly analyze payment decisions under constrained and unconstrained evaluation schemes or give principals leeway over whose evaluations to follow. Another starting point for further research arises from the observation that subjects evaluated teammates positively to motivate them for future periods. Whereas this behavior might not drastically affect a subject's future performance much given the one-shot nature of interaction, positive or negative evaluations might carry a different meaning if teammates interact repeatedly. Similar to repeated public good games (see, e.g., Sutter et al., 2010), workers might select rewards and punishment schemes to ensure sustained cooperation.

In this paper I highlight that peer evaluations can serve as a tool to distribute bonus payments to high performing employees only under some strong constraints. Consequently, further research on how to leverage co-worker insights on individual work behavior to design appropriate incentives is needed.

References

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner (2000): "The moonlighting game: An experimental study on reciprocity and retribution." *Journal of Economic Behavior & Organization*, 42 (2), 265–277. [96]

Berger, Johannes, Christine Harbring, and Dirk Sliwka (2012): "Performance Appraisals and the Impact of Forced Distribution–An Experimental Investigation." *Management Science*, 59 (1), 54–68. [96]

Bohl, Don (1996): "Minisurvey: 360-Degree Appraisals Yield Superior Results, Survey Stows." *Compensation & Benefits Review*, 28 (5), 16–19. [93]

Breuer, Kathrin, Petra Nieken, and Dirk Sliwka (2013): "Social ties and subjective performance evaluations: an empirical investigation." *Review of Managerial Science*, 7 (2), 141–157. [96]

Carpenter, Jeffrey, Peter Hans Matthews, and John Schirm (2010): "Tournaments and Office Politics: Evidence from a Real Effort Experiment." *American Economic Review*, 100 (1), 504–517. [97]

Carpenter, Jeffrey, Andrea Robbett, and Prottoy Akbar (forthcoming): "Profit Sharing and Peer Reporting." *Management Science*. [96, 97, 114]

Conrads, Julian, Bernd Irlenbusch, Rainer Michael Rilke, and Gari Walkowitz (2013): "Lying and team incentives." *Journal of Economic Psychology*, 34, 1–7. [97]

Cox, James, Daniel Friedman, and Steven Gjerstad (2007): "A tractable model of reciprocity and fairness." *Games and Economic Behavior*, 59 (1), 17–45. [101, 119]

Davison, Kristl, Vipanchi Mishra, Mark Bing, and Dwight Frink (2014): "How individual performance affects variability of peer evaluations in classroom teams: A distributive justice perspective." *Journal of Management Education*, 38 (1), 43–85. [95]

Dechenaux, Emmanuel, Dan Kovenock, and Roman Sheremeta (2015): "A survey of experimental research on contests, all-pay auctions and tournaments." *Experimental Economics*, 18 (4), 609–669. [96]

Eberlein, Marion and Gari Walkowitz (2008): "Positive and negative team identity in a promotion game." [99]

Engellandt, Axel and Regina Riphahn (2011): “Evidence on incentive effects of subjective performance evaluations.” *Industrial & Labor Relations Review*, 64 (2), 241–257. [96]

Englmaier, Florian and Stephen Leider (2012): “Contractual and organizational structure with reciprocal agents.” *American Economic Journal: Microeconomics* (2), 146–183. [101, 119]

Falk, Armin and Urs Fischbacher (2006): “A theory of reciprocity.” *Games and Economic Behavior*, 54 (2), 293–315. [119]

Fedor, Donald, Kenneth Bettenhausen, and Walter Davis (1999): “Peer reviews: employees’dual roles as raters and recipients.” *Group & Organization Management*, 24 (1), 92–120. [97]

Fehr, Ernst and Simon Gächter (2000): “Cooperation and punishment in public goods experiments.” *American Economic Review*, 90 (4), 980–994. [96]

Fischbacher, Urs (2007): “z-Tree: Zurich toolbox for ready-made economic experiments.” *Experimental Economics*, 10 (2), 171–178. [100]

Gill, David and Victoria Prowse (2012): “A Structural Analysis of Disappointment Aversion in a Real Effort Competition.” *American Economic Review*, 102 (1), 469–503. [98]

Golman, Russell and Sudeep Bhatia (2012): “Performance evaluation inflation and compression.” *Accounting, Organizations and Society*, 37 (8), 534–543. [93]

Greiner, Ben (2015): “Subject pool recruitment procedures: organizing experiments with ORSEE.” *Journal of the Economic Science Association*, 1 (1), 114–125. [100]

Gürtler, Oliver (2008): “On sabotage in collective tournaments.” *Journal of Mathematical Economics*, 44 (3-4), 383–393. [98]

Gürtler, Oliver and Johannes Münster (2010): “Sabotage in dynamic tournaments.” *Journal of Mathematical Economics*, 46 (2), 179–190. [98]

Gürtler, Oliver, Johannes Münster, and Petra Nieken (2013): “Information policy in tournaments with sabotage.” *Scandinavian Journal of Economics*, 115 (3), 932–966. [98]

Hammermann, Andrea, Alwine Mohnen, and Petra Nieken (2012): “Whom to choose as a team mate? A lab experiment about in-group favouritism.” *IZA Discussion Paper Series*. [97]

Huang, Yifei, Matthew Shum, Xi Wu, and Jason Zezhong Xiao (2017): “Strategic Manipulation in Peer Performance Evaluation.” [93, 97]

Johnson, Lauren Keller (2004): “The ratings game: Retooling 360s for better performance.” *Harvard Management Update*, 9 (1), 1–4. [93]

Kim, Jin-Hyuk (2011): “Peer Performance Evaluation: Information Aggregation Approach.” *Journal of Economics & Management Strategy*, 20 (2), 565–587. [97]

Kräkel, Matthias (2005): “Helping and sabotaging in tournaments.” *International Game Theory Review*, 7 (2), 211–228. [98]

Maas, Victor, Marcel Van Rinsum, and Kristy Towry (2011): “In search of informed discretion: An experimental investigation of fairness and trust reciprocity.” *Accounting Review*, 87 (2), 617–644. [97]

Marx, Leslie and Francesco Squintani (2009): “Individual accountability in teams.” *Journal of Economic Behavior & Organization*, 72 (1), 260–273. [97]

May, Gary and Lisa Gueldenzoph (2006): “The effect of social style on peer evaluation ratings in project teams.” *The Journal of Business Communication (1973)*, 43 (1), 4–20. [93]

Mohnen, Alwine, Kathrin Pokorny, and Dirk Sliwka (2008): "Transparency, inequity aversion, and the dynamics of peer pressure in teams: Theory and evidence." *Journal of Labor Economics*, 26 (4), 693–720. [97]

Offerman, Theo (2002): "Hurting hurts more than helping helps." *European Economic Review*, 46 (8), 1423–1437. [96]

Schwieren, Christiane and Doris Weichselbaumer (2010): "Does competition enhance performance or cheating? A laboratory experiment." *Journal of Economic Psychology*, 31 (3), 241–253. [97]

Sutter, Matthias, Stefan Haigner, and Martin Kocher (2010): "Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations." *Review of Economic Studies*, 77 (4), 1540–1566. [116]

Towry, Kristy (2003): "Control in a Teamwork Environment—The Impact of Social Ties on the Effectiveness of Mutual Monitoring Contracts." *Accounting Review*, 78 (4), 1069–1095. [97]

Appendix 4.A Behavioral Predictions

4.A.1 Theoretical Framework

In general each subject will receive her monetary payoff π_i and has to pay effort costs $c(e_i)$

$$U_i \;=\; \pi_i - c(e_i). \tag{4.A.1}$$

π_i consist of two parts: the wage (either fixed or based on team-performance) and the respective bonus. For the effort costs I assume that $c' > 0$ and $c'' > 0$. Additionally these costs are assumed to be homogeneous across workers.[20] In order to include reciprocal preferences, I extend the utility function from above:

$$U_i \;=\; \pi_i - c(e_i) + \phi \sum_{j=1}^{2} (c(e_j) - c(e_i))(\pi_j - \pi_j^{fair}) \tag{4.A.2}$$

The third part accounts for reciprocal preferences. ϕ denotes the reciprocity parameter. For simplicity I assume here, that the reciprocity parameter is identical for both positive and negative reciprocity. The reciprocity is induced by the differences in effort choices on the first stage only under the team payment

[20] Even though this might not apply in general, as be people might posses different subject costs for doing certain tasks (e.g., solving brain-teasers or math-questions). This is not a problem for the subsequent derivations. The workers need to believe only, that their cost functions are identical to the ones of their fellow workers, s.t. they treat their own cost function as a reference point for all other workers. This is perfectly reasonable in the context of a laboratory experiment where workers do not have additional information about each other.

regime.[21] If the baseline wage is fixed, the employees will never profit from any additional effort of their fellow workers and thus never establish any reciprocal feelings.[22] Individuals are assumed to have a positive attitude towards employees that exerted more effort that they did and a negative towards those that exerted less. In that sense their own effort choice serves as reference point for evaluating the others performance. Thus they tend to receive a positive marginal utility from an increase of these co-worker's monetary payoff. The co-worker's payoff is evaluated against her "fair"-payoff . The "fair"-payoff corresponds the material payoff if an employee is ranked in accordance with her effort rank.[23] The multiplicative structure ensures that the reciprocal feelings regarding others are larger if these workers chose more extreme effort levels, i.e., an employee will behave much more hostile towards a worker that chose an effort level drastically below hers than towards one that chose an effort level only slightly below. The utility function is similar to the one used by Englmaier and Leider (2012). It can also be identical to the utility function from Falk and Fischbacher (2006) if one abstracts from the underlying belief structure and assumes every action to be intentional.

4.A.2 Evaluation behavior

In order to determine the employees' behavior by backward induction the evaluation stage has to be considered first. I consider the case of three workers – called 1, 2 and 3.

Constrained Evaluation. In a first step the constrained option mechanism will be evaluated. Here the workers can choose between voting truly (awarding the point to the teammate with the higher effort level) or not (evaluating the opposite way). If the employees are not reciprocal ($\phi = 0$) their behavior is straight forward. The behavior on the previous stage has no influence on the evaluation behavior as it does not influence the employees utilities. For every strategy combination there is one subject that can guarantee herself either the second highest payoff by evaluating truthfully or the equal split of the entire bonus pool by deviating. This becomes obvious if one considers the following situation: One teammate received two positive evaluations. Then there must be another subject that received a positive evaluation, as the subject with two points had to hand out one, as well. This worker – currently ranked second –

[21] It it is essential for this model that effort and not solely performance can be observed by all employees, otherwise potential output differences could never be tied directly to differences in the effort choice. This particular approach will eliminate the symmetric equilibria in pure strategies that arise in tournaments games.

[22] A higher effort choice under the team payment can be seen as more generous in the sense of Cox et al. (2007), as it enlarges the co-worker's budget set.

[23] This reference point is not necessary for the subsequent predictions. Nonetheless it shortens the subsequent analysis, as under a truthful ranking of all workers the reciprocal utility will be zero for all workers.

evaluated the teammate ranked first positively. However, by shifting her positive evaluation from the co-worker with two positive evaluation to the one currently ranked third, she can ensure that the bonus pool is split evenly among all workers, potentially increasing her own payoff. As long as the bonus scheme is weakly convex, this will always be the case. Naturally, there exists also a mixed equilibrium. Here the every employee distributes the positive evaluation with equal probability among the other two co-workers. Again, the outcome is not informative at all.

Proposition 1 (Existence of true rankings under constraint evaluation options). *The true ranking of the workers can form an equilibrium, whenever the second ranked worker is sufficiently reciprocal, i.e.,* $\phi_2 \geq \frac{\bar{B}-B^2}{(c(e_1)-c(e_2))(\bar{B}-B^1)+(c(e_3)-c(e_2))(\bar{B}-B^3)}$. *Otherwise the ranking will never be tied to the actual performance of the workers during the first stage.*

In a next step I will show that informative equilibria where every worker evaluates truthfully can exist if workers are sufficiently reciprocal. Under these circumstances only the behavior of the worker with the second highest effort during the first stage is important.[24] In order to show that this is sufficient I will look at the option of the two other employees first under the equilibrium assumption that everybody else evaluates truthfully. The employee with the highest effort has no need to manipulate her ranking anyway as she is ranked first by the others, and thus first overall. Her decision determines the ranking of the other two employees. There she is expected to act truthfully. The employee will suffer more from rewarding a worker who exerted less effort. This can be seen by comparing the resulting payoffs: $U_i(true) = \pi_i - c(e_i)$ vs. $U_i(untrue) = \pi_i - c(e_i) + \phi((c(e_2) - c(e_1))(B^3 - B^2) + (c(e_3) - c(e_1))(B^2 - B^3))$. The differing term between these two is clearly negative as the difference in effort costs between the evaluating worker and the last ranked worker is larger than between her and the second ranked worker. Thus the worker suffers more from an increase in the payoff of the last ranked worker, than she does profit from the losses of the second ranked worker. In summary, conditionally on the co-workers acting fair, worker 1 will do so as well. The decision is similar for worker 3. Under fair behavior by the two other workers, she will never receive a point and will always be ranked last. Thus her decision will determine only the ranking of the two other employees. Awarding the point to worker 1 will induce the true ranking. This is predicted as worker 3 profits more from a increase in

[24] Strikingly this does not change if the number of employees is increased but forced ranking system through a Borda count is maintained. In case of a m-worker game, the worker with the nth highest effort will receive $(m-n)(m-n-1)+(n-1)(m-n) = (m-n)(m-2)$ points if everybody evaluates in a fair manner. Thus the difference between the nth and $n+1$th ranked employee will be $(m-2)$ or the highest amount of points to be distributed. Thus only the second ranked worker can bridge the gap to the employee in front of her by awarding ranking her last if everybody else reports truly.

worker 1 payoff as from worker 2. The reciprocal utility is zero under the fair behavior and $\phi((c(e_1)-c(e_3))(B^2-B^1)+(c(e_2)-c(e_3))(B^1-B^2))<0$ otherwise. Thus again the true evaluation is to be expected. As mentioned before worker 2 is the critical worker. Given the true behavior of the other workers, she as well as worker 1 have already been awarded a point. By distributing her point towards worker 1 the true ranking is induced. Otherwise everybody is ranked equally and receives the bonus $\bar{B}=\frac{B^1+B^2+B^3}{3}$. The utility from acting fair is denoted by

$$U_2(true) \;=\; \gamma\sum_{i=1}^{3} f(e_i) + B^2 - c(e_2) \tag{4.A.3}$$

and from acting money-maximizing

$$\begin{aligned} U_2(untrue) \;=&\gamma\sum_{i=1}^{3} f(e_i) + \bar{B} - c(e_2) \\ &+ \phi_2((c(e_1)-c(e_2))(\bar{B}-B^1) + (c(e_3)-c(e_2))(\bar{B}-B^3)). \end{aligned} \tag{4.A.4}$$

Comparing them yields that $U_2(true) \geq U_2(untrue)$, whenever

$$\phi_2 \;\geq\; \frac{\bar{B}-B^2}{(c(e_1)-c(e_2))(B^1-\bar{B})+(c(e_2)-c(e_3))(\bar{B}-B^3)}. \tag{4.A.5}$$

Thus worker 2 will evaluate worker 1 positively, whenever her measure for reciprocity or the differences in efforts are relatively large or the monetary loss endured through fair evaluation is sufficiently small. Notably, not only the effort levels of the worker with the two highest effort levels matter, but a lower effort level of the third worker results in a lower threshold. Increasing effort differences between the workers with the highest and lowest effort levels are predicted to make true evaluation more likely. In general, setting $B^2=\bar{B}$ will always result in the true ranking being the lone equilibrium (as long as $\phi>0$ for all subjects), as the second ranked individual has nothing to gain from evaluating not truthfully, but rather prefers the true ranking over the equal split. Consequently the equal split cannot be an equilibrium. The same is true for the other workers. They prefer the true ranking of their teammates over the other ranking they can induce, but cannot gain in the money dimension by deviating.

Unconstrained Evaluation. Under the more complex free scheme, the workers have four options to evaluate their co-workers. They can distribute a point to both other employees, only one to one of their co-workers or no points at all. If a worker awards no points at all even though at least one co-worker has performed better in the previous stage, this behavior will be called an untruthful evaluation. At the same time it will always maximize a worker's monetary payoff. In fact this strategy is a weakly dominant one. The distribution of a positive amount of points results in the receiving worker being ranked weakly higher as the awarding one.[25] In equilibrium other evaluations might handed out if the employee exhibit the reciprocal preferences introduced above:

Proposition 2 (Existence of true rankings under free evaluation options). *The workers are able to replicate the stable true ranking of the constraint evaluation mechanism, whenever the second ranked worker is sufficiently reciprocal, i.e.,* $\phi_2 \geq \frac{1}{c(e_1)-c(e_2)}$. *A second equilibrium resulting in the true ranking is possible if additionally also the third ranked worker is reciprocal, i.e.,* $\phi_3 \geq \frac{1}{c(e_2)-c(e_3)}$.

Two different evaluation patterns can be considered "true": Awarding points to all employees that exerted more effort or awarding one point to the better performing co-worker. If the workers mix these two behavior from these two equilibria, the resulting ranking will nevertheless coincide with the true ranking in almost all cases.[26] In order to exclude one evaluation possibility from the equilibrium analysis, it can be shown that no worker will reward a certain performance without rewarding a better one.

As stated before standard workers will never award any points. The distribution of points will weakly increase the receiving workers bonus, while simultaneously weakly lowering all other workers' bonuses (including her own). The same logic can be used to exclude actions that reward certain co-workers, without rewarding a higher ranked one. Given the assumed reciprocal utility function the marginal utility change induced by a higher material payoff for the two other workers depends one the difference in effort costs from the first stage. The difference is simply denoted by $\frac{\partial U_i}{\pi_j} - \frac{\partial U_i}{\pi_k} = \phi_i(c(e_j) - c(e_k))$. This is always positive as long as $e_j > e_k$. The marginal utility from increasing the payoff of the higher ranked co-worker is always larger. Thus if it is worthwhile to increase the payoff of the lesser ranked worker, it must be even more utility increasing to increase the payoff of the other one. Simultaneously, workers will profit more from punishing the lesser ranked co-worker, e.g., by not awarding the point to them. In summary this implies that evaluating their co-workers, no worker will ever award a point to the co-worker that exerted the lesser effort without awarding a point to the one with the higher effort level. Consequently one can restrict the subsequent equilibrium analysis to three choices: Awarding one point to

[25] If the receiving worker is already ahead due to the behavior of others, this cannot be altered through the own evaluation. If she is currently ranked equally or worse, awarding any points might lead to draw between the two or to a swap in rankings. As both a draw and a change in ranking will decrease the monetary payoff, under standard preferences the employee will always be (weakly) better off by not awarding any points at all.

[26] The lone exception arises if the worker with the lowest effort awards one point to the worker with the highest effort level and the highest ranked worker awards no points at all.

the worker with the higher effort level, awarding points to both co-workers and awarding no points at all.

In order to keep the analysis in this section short I will focus on two different equilibria inducing the true ranking. The first arises if the workers chose the decisions of the restricted mechanism (i.e., everybody awards a point to the better ranked of the other two workers). Under the equilibrium each worker will award a point to all co-workers that exerted more effort than herself, that means the worker with the highest awards no points at all. The next best worker awards a point towards the first worker 1 and the worker with lowest effort 3 rewards both co-workers.

The first equilibrium is basically the same as the equilibrium presented in the section before. Only worker 2 can behave in such a way to bridge the gap to the next ranked worker. As she would suffer from increasing the payoff of worker 3, she would only deviate from awarding a point to worker 1 by awarding no points at all. This deviation would consequently result in herself and worker 1 being equally ranked, while worker 3 remains last. Thus the free evaluation scheme does not have the drawback, that the unfair evaluation of a higher ranked worker necessarily induced a too positive evaluation of a lesser ranked worker. Consequently, this deviation becomes more attractive. Comparing the utilities while keeping the other decisions fixed reveals, that worker 2 will now deviate whenever

$$\phi_2 < \frac{1}{c(e_1) - c(e_2)}. \tag{4.A.6}$$

Thus the decision whether to deviate will only depend on the difference in effort level between herself and the first ranked worker. If this difference is large enough the second ranked worker will induce the true ranking by awarding a point towards worker 1. Worker 1 and 3 have no incentive to deviate, as their decision will not affect their own ranking. Additionally they prefer the true ranking of co-workers over every other ranking of them.

If worker 2 is not reciprocal enough or the performance differences are too small, there might be other equilibria where some workers still hand out positive evaluations.

As mentioned before a second equilibrium will result in the true ranking of all workers: All workers reward all co-workers that exerted more effort than they did. Thus worker 1 will again have no incentive to deviate, as her action could potentially only decrease her own payoff or change the payoffs of the other workers. However, as shown before the true ranking of the co-workers is always preferred over any other possible ranking of them. Considering worker 2 this equilibrium puts the same requirements on her as the previous one, as her equilibrium decision is the same and a deviation towards awarding no points at all has the same effect. It will not affect the outcome of worker 3, but only the ranking of herself and worker 1. She will support this equilibrium if $\phi_2 \geq \frac{1}{c(e_1)-c(e_2)}$. As worker 3 will be distributing the only point towards worker 2 in equilibrium, she can force a tie between them two by not awarding any points at all or only rewarding worker 1. In both cases deviations are profitable if $\phi_3 < \frac{1}{c(e_2)-c(e_3)}$. Again the decision depends only on the differences in the specific effort costs. If the differences in the selected efforts among all three workers is sufficiently large the true equilibrium should be induced. Notably, this decision is not impacted by the bonus sizes.

Again, if workers 2 and 3 are not reciprocal enough or there is little separation in the performances, a true ranking will not be an equilibrium. However, other equilibria with positive evaluations might still persist.[27]

Comparing the two mechanism yields that the requirements for equilibria with true rankings are stronger in the free evaluation mechanism.[28] If the price structure is flat, the true ranking should always be attainable under the constraint evaluation system, while this is not true under free conditions. This holds additionally true if the prize structure is concave, as under these circumstances the second prize will always be larger than the average prize. This changes naturally if the prize structure becomes convex. Such structures arise if there is only one prize to be distributed, i.e., as in promotion tournaments. Furthermore it is noteworthy that the existence of these equilibria solely depends on the distance between the effort choices but not on the bonus structure. Intuitively, a worker will only acknowledge differences between themselves and their co-workers if they are sufficiently large. Around every effort choice exists an area in which a worker is not willing to separate her choice from other effort choices in that area.

4.A.3 Effort choice

Even though effort choices are not discussed to great detail in this paper, this section sketches the influence of reciprocal attitudes on the previous effort choice for the case of constrained evaluations. In order for the analysis of the evaluation decision to have any bite, it is necessary that it is actually reasonable to expect different effort choices. This section provides an equilibrium analysis under the assumption that a truthful equilibrium will always be played. This is expected under forced evaluation with a linear bonus scheme.

[27] Imagine the following situation: Worker 1 does not hand out any positive evaluations, while worker 2 is not willing to reward worker 1. Then worker 3 might still prefer to give a positive evaluation to worker 1 if the performance differences are large enough, i.e., $\phi_3 \geq 1/(2c_1 - c_2 - c_3)$. In the resulting equilibrium, workers 2 and 3 are ranked equally with worker 1 being in first place.

[28] This is obvious, as each individual has more options to improve only their standing within the free evaluation mechanism.

During the first stage the individuals have to take into account whether an informative equilibrium will be played during the second stage or not. If this happens not to be the case, the employees will exert that amount of effort that maximizes their utility with regard to the team based incentive scheme. Thus effort is chosen s.t. $c'(e_i) = \gamma f'(e_i)$. This effort level will be denoted by e^{min}. The exertion of additional effort is never worthwhile, as it will not increase the likelihood of winning the second stage tournament.[29] This changes however if the employees assume that a truthful equilibrium will be played during the second stage. If one assumes that the employer selects $B^2 = \bar{B}$, the true ranking can always be expected to emerge under the constrained scheme. Under the complex evaluation scheme the true equilibrium will only emerge if and only if the difference in chosen effort levels is sufficiently large. Then the prospect of a potentially higher reward during the second stage serves as incentive to exert additional effort. If a truthful equilibrium is played, the employee with the highest effort will be awarded B^1, the second B^2, the third B^3. This can be interpreted as an all-pay auction with three bidders, three prizes and convex bidding costs. Bidding costs are convex due to the convexity of effort costs. The pure bidding costs are determined by the difference between the effort costs above the efficient level and the additional income generated by the return form the team payment.

If all workers are reciprocal, mutual true evaluation is the only equilibrium in pure strategies under the constrained evaluation scheme. Hence, I will analyze the influence of mutual true evaluations on effort in the first stage. As common in all-pay auctions no equilibrium in pure strategies exists. Otherwise workers would have an incentive to overbid each other up to the point where the bidding costs are equal to the largest bonus payment. However, the other workers are then again not expected to exert any effort, which again renders the exertion of the maximum amount of effort useless. Consequently, I will show that there exists a monotonically increasing atomless distribution function F from which effort choices are randomly selected in equilibrium. By exerting the minimal rational effort e^{min} a worker can always guarantee herself the bonus B^3 with certainty. Her utility is consequently given by

$$E(U(e^{min})) \;=\; \gamma \sum_{j=1}^{2} (2E(f(e_j) + f(e^{min})) - c(e^{min}) + B^3. \tag{4.A.7}$$

For further illustration I will use the standard function $f(e) = e$ and $c(e) = \frac{e^2}{2}$, resulting in $e^{min} = \gamma$ and $E(U(e^{min})) = \frac{\gamma^2}{2} + B^3 + \gamma 2E(e_i)$. In order to allow for

[29] It is important to note this behavior is not maximizing the entire groups welfare. This is due to the public good structure of the team-payment. Every teammate profits from the individual's extra effort. The team-efficient effort level would naturally be $c'(e_i) = n * \gamma f'(e_i)$.

a mixing effort choices the expected payoff of all other effort choices that are played with positive probability must be equal to this amount. Consequently the upper limit of this interval is pinned down by this equation. $c(e^{max})$ ensures the worker the maximum bonus B^1 with surety. This results in the utility being equal to

$$E(U(e^{max})) \;=\; \gamma(2E(e_i) + e^{max}) + B^1 - c(e^{max}). \tag{4.A.8}$$

Setting this equal to the payoff from the lowest effort and using the specified functions results in $e^{max} = \gamma + \sqrt{2(B^1 - B^3)}$. Given the c.d.f. $F(e)$, which is assumed to be chosen s.t. equal utilities across all selections between e^{min} and e^{max} are induced, the utility of certain effort choice e is given by

$$\begin{aligned} E(U(e)) \;=& \gamma(2E(e_i) + e) - c(e) \\ &+ F^2(e)B^1 + 2F(e)(1 - F(e))B^2 + (1 - F(e))^2B^3. \end{aligned} \tag{4.A.9}$$

Equating this function with determined utility leads to the following equilibrium density function

$$F(e) \;=\; \frac{(\gamma - e)^2}{2(B^1 - B^3)}. \tag{4.A.10}$$

This analysis can be extended for settings were the true ranking will only occur if the output differences between the subjects are large enough. This would imply that even a subject with the minimal effort level has a chance to receive a larger bonus (either shared between her and the second highest performing teammate, or shared between all teammates). This increases the expected payoff of that choice. Likewise even very high effort choices cannot longer guarantee the highest possible bonus. This decreases the value of additional effort the subjects exert and compresses the range of output levels. The central argument that eliminates symmetric equilibria in pure strategies is the same.

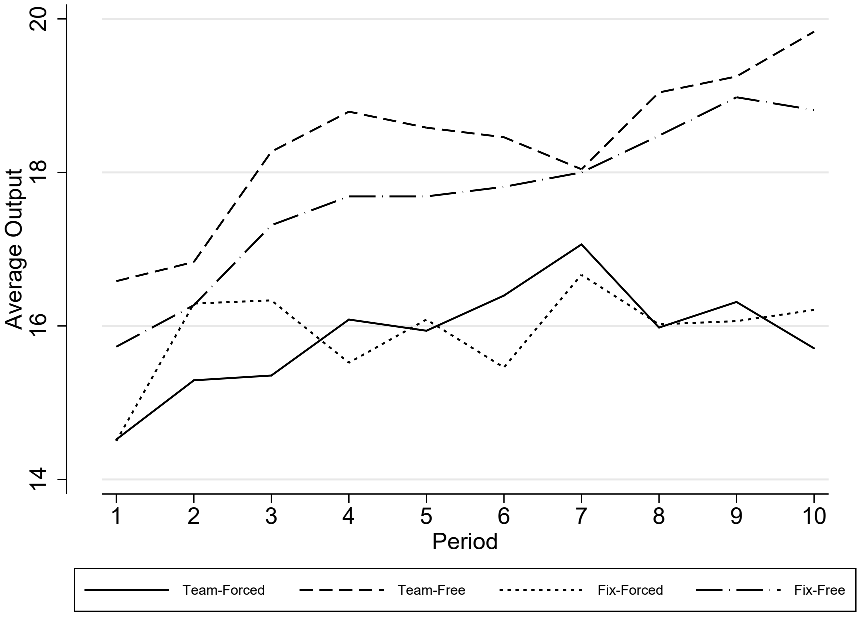


Figure 4.B.1. Sliders solved across periods

Notes: The figure presents average amount of sliders solved across periods.

Appendix 4.B Performance Levels

As discussed in section 4.4 the subjects solve more sliders in the treatments with team-incentives. Notably, these two treatments show significant time trends (Spearman's ρ =0.48, p<0.01 for treatment TEAM-FREE and Spearman's ρ =0.63, p<0.01 for treatment TEAM-FORCED). Similar trends cannot be observed under fixed wages. Table 4.B.1 shows that in the absence of team-incentive the average output is slightly lower in the treatments with fixed wages. The difference, however becomes very small if those observations are excluded, where the subject did not work on the task at all.

Table 4.B.1. Impact of Treatments on Output

	Sliders Correctly Positioned	
	(1) All	(2) Output > 0
Fix-Forced	-1.76*	-0.59
	(0.98)	(0.50)
Team-Free	0.69	0.69
	(0.79)	(0.79)
Fix-Free	-1.81	-0.10
	(1.21)	(0.69)
Period dummies	Yes	Yes
N	1920	1841
R^2	.046	.036

Notes: This table presents regressions using individual output in a single period as dependent variable. Column (2) excludes all observations where a subject did not produce any output at all. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

Appendix 4.C Constrained evaluation

This section reports alternate specifications for the estimation of treatment effects for the treatment with constrained evaluation scheme. The analysis includes different measures of performance between the subjects and their teammates.

Table 4.C.2. Alternate specification for ranking outcomes in Forced treatments

	Correctly ranked				
	(1)	(2)	(3)	(4)	(5)
Team pay	0.60*** (0.19)	0.67*** (0.20)	0.39 (0.26)	0.73*** (0.23)	0.75*** (0.23)
Output Rank=2		0.58*** (0.13)	0.57*** (0.12)		-0.09 (0.09)
Output Rank=3		0.20 (0.16)	0.15 (0.16)		0.18 (0.20)
abs. Output size		-0.01 (0.02)	-0.02 (0.02)		-0.02 (0.02)
Output=0		-0.60*** (0.14)	-0.48** (0.19)		-0.17 (0.26)
cumulative $\lvert\Delta Output\rvert$		0.02* (0.01)			
Fix pay × cumulative $\lvert\Delta Output\rvert$			0.01 (0.01)		
Team pay × cumulative $\lvert\Delta Output\rvert$			0.03** (0.01)		
Period dummies	Yes	Yes	Yes	Yes	Yes
N	826	826	826	651	651
$PseudoR^2$	.05	.08	.084	.067	.079
p-value: Team vs. Fix			.085		

Notes: This table presents probit regressions using resulting ranks as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

Table 4.C.2 presents additional analyses for the determinants of rank outcomes. Again, column (1) replicates the baseline specification from Table 4.4, column (1), however, only for the two treatments with Forced evaluation. In column (2) the cumulative output difference to the other two teammates is included as mediator. Column (3) interacts this with the two treatments. The cumulative output difference is especially large if a subject performed much better or much worse than both of her teammates. This might give both teammates an incentive to evaluate that subject truthfully. As both teammates have to evaluate truthfully to induce the true ranking of a subject, this measure is included here rather than the individual differences. In line with the findings for the evaluation behavior a higher cumulative output difference is associated with a higher likelihood to be ranked correctly. This effect is slightly stronger under team-incentives. Column (4) and (5) exclude those groups were all subjects were ranked equally. Column (4) features no additional controls, while column (5) corresponds to column (4) of Table 4.4. In contrast to the results column (2) Table 4.4, the indicator for the second best performing subject is no longer significant. This indicates that this effect was indeed an artifact of ranking ties resulting in the appropriate bonus for the second ranked subject. The overall treatment effect remains significant and becomes even slightly larger.

Table 4.C.3 extends the analysis in section 4.4. Column (1) corresponds to the specification presented in Table 4.6, column (1). In column (2) the absolute differences in output to the two teammates are included separately. While the effect size is rather similar, there is only a significant effect of a larger output differences to the lower performing teammate. As the impact of output differences might differ depending on their sign, column (3) separates the output differences. In both treatments subjects tend to evaluate more truthfully the the lower the relative performance of the lower performing teammate. Higher output by a better performing subject does not seem to be associated with more truthful evaluations.

Table 4.C.3. Alternate specification for evaluation outcomes in Forced treatments

	True evaluations		
	(1)	(2)	(3)
Team pay	0.43** (0.17)	0.48** (0.19)	0.12 (0.22)
Output Rank=2		-0.34*** (0.08)	-0.16** (0.08)
Output Rank=3		-0.44*** (0.12)	-0.48*** (0.09)
abs. Output size		-0.01 (0.02)	-0.02 (0.01)
Output = 0		-0.59 (0.37)	-0.44* (0.24)
$\lvert \Delta Output \rvert$ to better teammate		0.02 (0.02)	
$\lvert \Delta Output \rvert$ to worse teammate		0.02*** (0.01)	
Fix pay × $\lvert \Delta Output \rvert$ to better teammate			- 0.01 (0.02)
Team pay × $\lvert \Delta Output \rvert$ to better teammate			0.04 (0.03)
Fix pay × $\lvert \Delta Output \rvert$ to worse teammate			0.03*** (0.01)
Team pay × $\lvert \Delta Output \rvert$ to worse teammate			0.04*** (0.01)
Period dummies	Yes	Yes	Yes
N	887	887	887
$PseudoR^2$	.028	.053	.058
p-value: Team vs. Fix High			.18
p-value: Team vs. Fix Low			.15

Notes: This table presents probit regressions using evaluation behavior as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

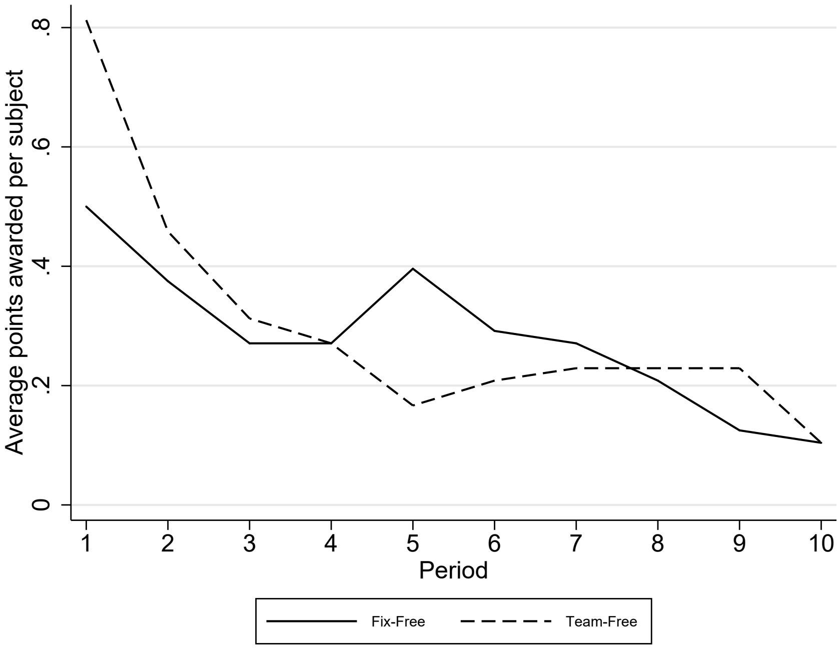


Figure 4.D.2. Positive evaluations over time

Notes: The figure presents the average amount of positive evaluations used per subjects in the two treatments with unconstrained evaluations.

Appendix 4.D Unconstrained evaluation

This section presents additional material on the treatments that feature the unconstrained evaluation scheme.

Figure 4.D.2 displays the average number of positive evaluations subjects used in a given period. Positive evaluations are handed out predominately in the beginning of the experiment. Subjects almost stop using them during the second half; only about a quarter of all subjects hands out a single positive evaluation during that period. Consequently in both treatments the positive evaluations exhibit a negative time-trend: Spearman's ρ =-0.57, p<0.01 for Fix-Free and ρ=-0.57, p<0.07 for Team-Free.

In Table 4.D.4, I check whether the results evaluation outcomes change if one focuses only on the treatments with unconstrained evaluation. Additionally, column (2) excludes those groups where no positive evaluations were handed out all. Here, this is equivalent to all subjects being ranked equally. Column (3) includes additional controls. Results do not change; the treatment has no significant effect on the share of subjects that are ranked correctly. Interestingly, if positive evaluations are actually handed out, they result in actually the highest

Table 4.D.4. Alternate specification for ranking outcomes in FREE treatments

	Correctly Ranked		
	(1)	(2)	(3)
Team pay	-0.19 (0.16)	-0.31 (0.28)	-0.37 (0.31)
Output Rank=2			-0.87*** (0.12)
Output Rank=3			-0.30* (0.16)
abs. Output size			0.02 (0.02)
Output=0			0.18 (0.47)
Period dummies	Yes	Yes	Yes
N	960	493	493
$PseudoR^2$	.0055	0.018	.084

Notes: This table presents probit regressions using resulting ranks as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

performing subjects receiving the correct bonus significantly more often than the others.

In order to highlight the potential impact of output differences for the evaluation behavior the analysis included the absolute output differences between the evaluating subject and each of their teammates. Nonetheless, the impact of the output differences might differ depending on their sign. If subjects were reciprocal, only suitably large positive output difference should lead to truthful evaluations, large negative output differences should not matter at all. While columns (1) and (2) correspond to columns (1) and (3) of Table 4.7, column (3) includes the output differences separated by their sign. In contrast to the theoretical predictions a better performance of the higher performing teammates actually decreased the likelihood of a truthful evaluation significantly.

Table 4.D.5. Alternate specification for evaluation outcomes in FREE treatments

	(a) True Evaluation		
	(1)	(2)	(3)
Team pay	-0.14 (0.11)	-0.41** (0.19)	-0.42** (0.19)
Output Rank=2		-2.41*** (0.17)	-2.16*** (0.28)
Output Rank=3		-2.26*** (0.23)	-2.05*** (0.32)
abs. Output size		0.03 (0.02)	0.02 (0.03)
Output=0		0.04 (0.22)	0.02 (0.23)
$\|\Delta Output\|$ to better teammate		-0.03* (0.02)	
$\|\Delta Output\|$ to worse teammate		0.00 (0.01)	
$\|\Delta Output\| < 0 \times \|\Delta Output\|$ to better teammate			0.04 (0.04)
$\|\Delta Output\| \geq 0 \times \|\Delta Output\|$ to better teammate			-0.04** (0.02)
$\|\Delta Output\| < 0 \times \|\Delta Output\|$ to worse teammate			-0.01 (0.02)
$\|\Delta Output\| \geq 0 \times \|\Delta Output\|$ to worse teammate			0.01 (0.03)
Period dummies	Yes	Yes	Yes
N	960	960	960
$PseudoR^2$	.0091	.41	.41

Notes: This table presents probit regressions using evaluation behavior as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on matching-group-level.

Appendix 4.E Instructions

The instructions below are translations of the German instructions for the experiment. Differences between the treatments are marked by Free:"..." and Team "...".

General instructions for the participants

You are now participating in an economic experiment. If you read the following explanations carefully, you will be able to earn a considerable amount of money – depending on your decisions and those of the other participants. Thus it is important to read these instructions very carefully.
During the experiment, it is absolutely prohibited to communicate with the other participants. Should you have any questions, please ask us. If you violate this rule, you will be dismissed from the experiment and forfeit all payments. How much money you will receive after the experiment depends on your decisions and those of the other participants. The experimental payoffs will be calculated in Taler. The total amount of Taler that you have accumulated during the experiment will be converted into Euro and paid to you in cash at the end of the experiment. The exchange rate from Taler to Euro is as follows:

75 Taler = 1 Euro

The experiment consists of exactly one part. This part is divided into **10 periods**. At the beginning of the experiment you are randomly assigned to a group of three. Besides yourself, there are two other participants in your group. In each period you will interact with different, randomly assigned participants. Neither during the experiment, nor afterwards you will receive any information about the identity of the other participants in your group.

Detailed Information about the Course of each Period

Each period consist of three stages

The 1st Stage

During the 1st stage you will be presented with 48 scales. Each scale goes from 0 to 100. On each scale, there is a slider. You can move this slider with you mouse. You have exactly two minutes to position as many slider as you like on the value 50 on the scale. At the beginning each slider is positioned at the extreme corner of the scale, at zero. Below the scale, the current position of the slider is displayed. You are free to work on the sliders in whichever order you like. Prior to the main part of the experiment, you will have the opportunity to familiarize yourself with the task during two practice periods.

The 2nd Stage

At the beginning of the 2nd stage you will be informed on how many sliders each participant in your group has positioned correctly during the 1st stage.
Now, you will have the opportunity to evaluate the other members of your group. For this you can award points indicating good work. Forced: "You have to award 1 point to exactly one of your teammates." Free: "You can award each of your teammates 1 or 0 bonus-points. That means, you have the option to award both of them 1 point, you can award only one of them 1 point or you award no points at all." Other decisions are not possible. The points awarded by yourself and the other group-members will be used to determine a bonus payment. The teammate with most points will receive 80 Taler, the next teammate will receive 40 Taler and the teammate with the lowest amount of points will receive 0 Taler.

The 3rd Stage

During the 3rd stage you will be informed on the points each group-member received as well as their income during that period. Additionally you will be shown again, how many slider each group-member positioned correctly during the 1st stage.
Afterwards this period ends and a new one begins. The participants within your group are randomly selected only for that single period. During the 1st stage, all participants can reposition sliders, again. The 1st stage is followed by the 2nd stage and finally the 3rd stage.

Detailed Information about the Income in each Period

Your income for each period will always consist of two parts.

FIX: “**Basic income:**

In each period, you and each of your teammates receive a basic income of 80 Taler each. The basic income is completely independent of your decisions in that period and will always be paid out to you.”

TEAM: “**Income from the project:**

You and all other teammates receive 1 Taler each for each slider, that you or one of the participants in your group positioned correctly during the 1st Stage. If you and your teammate positioned a total of 50 sliders correctly, you will earn exactly 50 Taler from the project.”

Income from the bonus payment:

At the end of each period, you and the other members of your group have the opportunity to award a bonus-point to *exactly one of your teammates*. The serve as base for the calculation of the bonus payments. For this the points a participant received are added up. The participant with the most points receives 80 Taler. The participant with the second most points receives 40 Taler. The group-member with the fewest points receives no bonus payment. If more than one teammate received the same amount of points, the corresponding bonus payments will be split among them.

$$\begin{aligned} \textit{Payment for the most bonus-points} &= \textit{80 Taler} \\ \textit{Payment for the 2nd most bonus-points} &= \textit{40 Taler} \\ \textit{Payment for the fewest bonus-points} &= \textit{0 Taler} \end{aligned}$$

Example 1: Participant A received 2 bonus-points, participant B received 0 bonus-points and participant C 1 bonus-point. Then participant A receives a bonus of 80 Taler. Participant B receives 0 Taler and participant C 40 Taler.

FORCED:

“**Example 2:** Participant A and B and C received 1 bonus-point each. Then the participants share the entire sum of bonus-payments. Everyone receives a bonus-payment of $(80 + 40 + 0) / 3 = 120 / 3 = 40$ Taler.”

FREE:

“**Example 2:** Participant A and B received 2 bonus-points, each. Participant C received 1 bonus-point. Then the participants A and B share the largest two bonus-payments. Both of them receive a bonus-payment of $(80 + 40) / 2 = 120 / 2 = 60$ Taler. Participant C receives a bonus-payment of 0 Taler.”

Income at the end of each period:

Your income at the end of each period is equal to the sum of your basic income and the your income from the bonus payments. This means:

	Fix: *"Basic income"* Team: *"Income from the project"*
+	*Income from the bonus payment*
=	*Income at the end of the period*

We will demonstrate this calculation with the use of two examples:
Example 1: Assume, you and the other participants in your group correctly positioned 16 slider, each. Each participant in your group has received exactly one bonus point. Fix: "Your basic income is 80 Taler." Team: "Your income from the project is given by 16 + 16 + 16 = 48 Taler." Your income from the bonus payment is $(80 + 40 + 0) / 3 = 40$ Taler. Consequently, your income in this period will be:
Fix:
"80 Taler basic income + 40 Taler bonus payment = 80 + 40 = 120 Taler"
Team:
"48 Taler from the project + 40 Taler bonus payment = 48 + 40 = 88 Taler"

Example 2: Assume, you positioned 0 sliders correctly, while the two other participants in your group repositioned 20 slider, each. You received 0 bonus-points and the other two participants received 1 and 2 bonus-points, respectively. Fix: "Your basic income is 80 Taler." Team: "Your income from the project is given by 0 + 20 + 20 = 40 Taler." Your income from the bonus payment is 0 Taler. Consequently, your income in this period will be:
Fix:
"80 Taler basic income + 0 Taler bonus payment = 80 + 0 = 80 Taler"
Team:
"40 Taler from the project + 0 Taler bonus payment = 40 + 0 = 40 Taler"

Conclusion of the experiment and payment

The experiment ends after 10 periods. Subsequently, we will ask you to answer a few general questions on the computer. Your answers to these questions have no influence on how much money you will earn in the experiment. When all participants have filled out the questionnaire, payments will be made. Your total income from the 10 periods will be converted into Euro and paid to you in cash.

Do you have any questions? If so, please raise your hand.

5

The impact of self-selection on performance*

> *"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

5.1 Introduction

Basketball hall of famer Larry Bird used to motivate himself to train harder not by focusing on any player but rather by looking at his rival Magic Johnson's performance during the previous night's game. Similarly, seeing a specific classmate study long and continuously might also help to focus on one's own work. In various dimensions of life – ranging from students in educational settings (Sacerdote, 2001; Duflo et al., 2011) over cashiers in supermarkets (Mas and Moretti, 2009) and fruit pickers on strawberry fields (Bandiera et al., 2005, 2009) to fighter pilots during World War II (Ager et al., 2016) – we look at the behavior of our peers and compare our own performance and choices with theirs.[1]

[*] We thank Lorenz Götte, Sebastian Kube, Pia Pinger for their guidance and support. We also thank Viola Ackfeld, Philipp Albert, Thomas Dohmen, Ingo Isphording, Ulf Zölitz and audiences at Bonn, MBEPS 2017, VfS 2017, ESA Europe 2017, IZA Brown Bag, Rady Spring School in Behavioral Economics 2017, Bonn-Mannheim Ph.D. Workshop, 20th IZA Summer School in Labor Economics, 12th Nordic Conference on Behavioral and Experimental Economics, and Max Planck Institute for Research on Collective Goods for helpful feedback and comments.

[1] The influence of peers on our own behavior has long been recognized in the social sciences in general, as well as in economics more specifically. Such effects – commonly referred to as "peer effects" – are widely observed across a wide range of outcomes, not only for performance on the job or in school: indeed, other contexts include investment behavior (Bursztyn et al., 2014), consumption (Kuhn et al., 2011), program participation (Dahl et al., 2014), propensity to exercise (Babcock and Hartman, 2010; Aral and Nicolaides, 2017) and wages in a firm (Cornelissen et al., 2017). The settings across these studies differ enormously, as does the underlying mechanism

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

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

In order to study the effects of self-selection, we conducted a framed field experiment with over 600 students (aged 12 to 16) in physical education classes of German secondary schools. Students took part in two running tasks (suicide runs) – first alone, then with a peer – and filled out a survey in between that elicited preferences for peers, personal characteristics and the social network within each class. Our treatments exogenously varied the peer assignment in the second run using three different matching rules. We implemented a random matching of pairs (Random) as well as two matching rules that use the elicited preferences to implement self-selected peers. The specific setup of our experiment allows for two notions of self-selection, based on either social identity or

(e.g., peer pressure, learning, complementarities). Nonetheless, all of these have in common that the behavior or action of peers imposes an externality on the action or behavior of others. Most of the research on peer effects takes the peer group or a single peer as given or randomly assigned.

[2] Festinger (1954) already conjectured that people tend to compare their own performance on average with slightly better performing individuals. Similarly, performance leaderboards for sales representatives are widespread for motivational reasons. They allow employees to compare their performance with others despite not knowing them personally.

the relative performance of one's classmates. First, the classroom environment enabled students to state preferences for known peers (*name-based preferences*). Second, using a running task yields direct measures of performance and thus could be used to select peers based on their relative performance in the first run (*performance-based preferences*). Utilizing these two sets of preferences, we implemented two treatments with self-selection. The treatments matched students based on name-based preferences (Name) or preferences over relative performance (Performance), which we elicited in the survey.

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

While the impact of a peer and the resulting quantitative effect might be specific to this setting, the underlying motive of the results are of general interest. Students have not only been successfully used to analyze phenomena like favoritism (Belot and van de Ven, 2011, and references therein), but they are also a highly relevant subject group, given that social comparisons are important drivers of effort and performance in school and consequently may affect educational attainment. The process of self-selecting peers is potentially equally important for settings in which peer effects do not arise due to social comparisons or peer pressure, but rather where effort, task or skill complementarities exist (e.g., Mas and Moretti, 2009) or where learning from peers is important (among others Bursztyn et al., 2014; Kimbrough et al., 2017). In these settings, peer effects originate from different mechanisms than studied in our setting, although in principle peers can also be self-selected. This may affect interactions among peers and the motivation of individuals themselves in ways similar to this study. Our results also complement the findings by Bartling et al. (2014), who demonstrate that people value the opportunity to actively select relevant aspects of life, whereas we highlight the motivational benefits of subjects being able to self-select their environment (i.e., their peer).[3] As our paper shows, the direct effect of being able to self-select peers might be even more important than those induced by exogenous group assignment. Hence, studies analyzing interactions between peers and policies leveraging these insights need to take into account any selection of peers taking place within groups.

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

The existence of peer effects in educational settings motivated a small strand of the literature to focus on reassignment policies. Rather than assigning students randomly to classrooms, Carrell et al. (2013) systematically formed classes comprising only high- and low-ability students to increase the GPA of the latter. Instead of increasing their GPA as was predicted by estimates in Carrell et al. (2009), the GPA actually decreased. The authors suggest that subgroups of either high- or low-ability students emerged, with little interaction between them. Therefore, the exogenous formation of classrooms changed the class composition and thereby the set of potential peers, while within this group students self-selected their relevant peers. Booij et al. (2017) also present evidence on exogenous peer group manipulations. They manipulated the group composition based on their prior ability, which led to a change in the social interaction: low-ability students were more involved in classes and reported more positive interactions within classrooms. In this paper, rather than reassigning students into classrooms as in the previous studies, we take the classes as given and focus on the peer assignment and resulting interactions within classrooms.

Researchers have recently analyzed the potentially differential effects of friends and non-friends and thus have moved away from the paradigm that all peers influence an individual's performance similarly. Lavy and Sand (2015) analyze how reciprocal – in contrast to non-reciprocal – friends affect the test scores

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

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

The closest paper to ours is Chen and Gong (2018). The authors study self-sorting of students into teams for a group task with skill complementarities. Similar to us, they find that peers are selected based on the social network and that those groups perform better than randomly formed groups. In contrast to their setting, we focus on pairs with individual production as the unit of analysis to identify a peer's effect on individual performance.

Finally, we add to the literature on rank effects in peer interactions. Our results are in line with research documenting the importance of ranks for subsequent outcomes (Elsner and Isphording, 2017; Gill et al., 2017). If individuals have preferences over ranks, this can give rise to heterogeneous peer effects similar to our setting (Tincani, 2017). Relatedly, Cicala et al. (forthcoming) also use rank-dependent preferences to build a Roy-model of social interactions, where agents can select into certain groups based on their ability to carry out different tasks. In our experiment, subjects can indirectly select their rank in the second run by choosing a specific peer or relative time (e.g., a faster or slower peer).

The remainder of the paper is structured as follows. The next section presents our experimental design as well as procedural details. Section 5.3 presents the data and describes our sample of students. We outline our empirical framework in section 5.4. In section 5.5, we analyze how self-selected peers affect performance relative to randomly assigned peers and decompose this effect in direct effects of self-selection and indirect effects of a change in the peer composition. We then discuss heterogeneous responses and highlight potential policy implications. Finally, section 5.6 concludes.

5.2 Experimental design

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

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

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

5.2.1 Experimental design

Figure 5.1 illustrates the experimental design. Students participated in a running task commonly known as “suicide runs”, a series of short sprints to different lines of a volleyball court.[4,5] The first run – in which students run alone

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

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

– served two purposes: first, recorded times can be used as a measure of ability and to evaluate the time improvement between the two runs; and second, we used (relative) times from the first run in combination with students' preferences to create pairs for the second run in one of the treatments described below. The second run mirrored the first one aside from the fact that students did not run alone, but rather in pairs. This means that both students performed the task simultaneously, while their times were recorded individually. Feedback about performance in both runs was provided at the end of the experiment only.

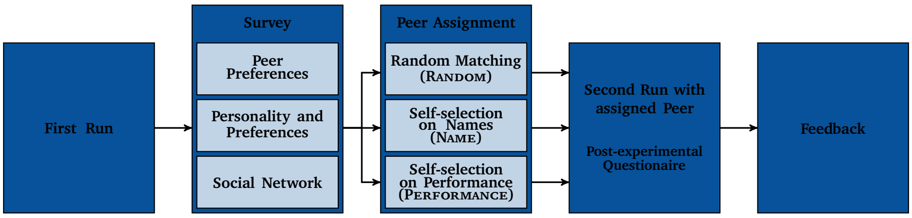

First Run
Survey
Peer Preferences
Personality and Preferences
Social Network
Peer Assignment
Random Matching (RANDOM)
Self-selection on Names (NAME)
Self-selection on Performance (PERFORMANCE)
Second Run with assigned Peer
Post-experimental Questionaire
Feedback

Figure 5.1. Experimental design

Between the two runs, students filled out a survey comprising three parts, eliciting preferences for peers, non-cognitive skills, and information about the social network within each class. We elicited two kinds of preferences: first, we asked subjects to state the names of those classmates with whom they would like to perform the second run; and second, we asked them to state the relative performance level of their most-preferred peers. Note that we elicited all preferences irrespective of the assigned treatment and used these preferences to match students for the second run in two of the three treatments.

In addition to these preferences, the survey included sociodemographic questions and measures of personality and preferences: the Big Five inventory as used in the youth questionnaire of the German socioeconomic panel (Weinhardt and Schupp, 2011), a measure of the locus of control (Rotter, 1966), competitiveness[6], general risk attitude (Dohmen et al., 2011), and a short version of the INCOM scale for social comparison (Gibbons and Buunk, 1999; Schneider and Schupp, 2011). The survey concluded by eliciting the social network within every class. Subjects were asked to state their six closest friends within the class and indicate the intensity of their friendship on a seven-point Likert scale.

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

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

5.2.2 Preference elicitation

We elicited two sets of peer preferences, independent of the treatment to which a subject is assigned. The first set elicited those for situations in which social information is available (*name-based preferences*). Accordingly, we asked each student to state his or her six most-preferred peers from the same gender within their class, i.e., those people with whom they would like to be paired in the second run. They could select any person of the same gender, irrespective of this person's actual participation in the study or their attendance in class.[7] These classmates had to be ranked, creating a partial ranking of their potential peers.

Second, we elicited preferences solely based on the relative performance in the first run, ignoring the identities of the potential running partners (*performance-based preferences*). For this purpose, we presented subjects ten categories consisting of one-second intervals starting from $(4, 5]$ seconds slower than their own performance in the first run, to $(0, 1]$ seconds slower and $(0, 1]$ seconds faster up to $(4, 5]$ seconds faster. They had to indicate from which time interval they would prefer a peer for the second run, irrespective of the potential peer's identity. Similar to the name-based preferences, we elicited a partial ranking for those performance-based preferences. Accordingly, subjects had to indicate their most-preferred relative time interval, second most-preferred relative time interval and so on.[8]

5.2.3 Treatments

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

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

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

The second matching rule used the elicited name-based preferences (NAME) and the third rule formed pairs based on the elicited performance-based preferences (PERFORMANCE). Note that the problem of matching pairs constitutes a typical roommate problem. We thus implemented the “stable roommate” algorithm proposed by Irving (1985) to form stable pairs using the elicited preferences.[9]

Subjects did not know the specific matching algorithm, but were only told that their preferences would be taken into account when forming pairs. We informed subjects about the existence of all three matching rules in the survey to elicit both sets of preferences irrespective of the implemented treatment. Just before the second run took place, they were informed about the specific matching rule employed in their class and the resulting pairs.

In addition, we conducted an additional control treatment (NOPEER) in which students ran alone twice and which featured a shortened survey but was otherwise identical to the other treatments.[10] As this only serves the purpose of excluding learning as a source of time improvements between the two runs, we exclude it from the main analysis and focus only on the evaluation of different peer assignment rules.

5.2.4 Procedures

We conducted the experiment in physical education lessons at three secondary schools in Germany.[11] All students from grades 7 to 10 (corresponding to age 12 to 16) of those schools were invited to participate in the experiment.

Approximately two weeks prior to the experiment, teachers distributed parental consent forms. These forms contained a brief, very general description of the experiment. Only those students who handed in the parental consent before the study took place participated in the study.

[9] Given the mechanism proposed by Irving (1985), it is a (weakly) dominant strategy for all participants to reveal their true preferences. The matching algorithm requires a full ranking of all potential peers to implement a matching. Since we only elicited a partial ranking, we randomly filled the preferences for each student to generate a full ranking. However, in most cases subjects were assigned a peer according to one of their first three preferences. Nonetheless, if groups were small, it could be the case that subjects were not assigned one of their most-preferred peers. This is especially the case for performance-based preferences. See also the discussion in section 5.3.1 below.

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

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

The experiment started with a brief explanation of the following lesson and demonstration of the experimental task. We informed students that their teacher would receive each student's times from both runs, but no information about the pairings during the second run.[12] The students themselves did not receive any information on their performance until the completion of the experiment. We did not incentivize students with monetary rewards. Instead, we stressed that the objective was to run as fast as possible in both runs. Moreover, teachers used the times in their own class evaluation and students themselves were also interested in their own times.[13] The introduction concluded with a short warm-up period. After this, the subjects were led to a location outside of the gym.

Students entered into the gym individually. Thus, any potential audience effects from classmates being present were ruled out by design. Students completed the first suicide run and subsequently were handed a laptop to answer the survey. Answering the survey took place in a separate room.[14] After the completion of the survey, subjects returned the laptop to the experimenter and waited with the other students outside the gym. Upon completion of the survey by all students, they returned to the gym to receive further instructions for the second run. In particular, we reminded the students of the existence of the three matching rules, announced which rule was implemented in their class and the resulting pairs from the matching process. Following these instructions, the entire group waited again outside the gym. Pairs were called into the gym and both students participated in the second run simultaneously on neighboring tracks.

After all pairs had finished their second suicide run, the experiment concluded with a short statement by the experimenters thanking the students for their participation. The teacher received a list of students' times in both runs and students were informed about their performance. We then asked the teacher to evaluate the general atmosphere within the class.[15]

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

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

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

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

5.3 Data description and manipulation check

We present summary statistics of the students in our sample in Table 5.1.[16] In total, 627 students participated in the treatments, with 66% being female.[17,18] This corresponds to a participation rate of 73%.[19]

On average, female students took 27.57 seconds (SD of 2.50 seconds) in the first run. Their performance is quite stable across all grades, with students from the seventh grade being somewhat slower. Male students' times decreased with age: while male students in grade 7 took on average 25.33 seconds in the first run, their performance improved to 23.27 seconds on average in grade 10. In the following, we control for these effects by including gender-specific grade fixed effects in all of our regressions. Independent of their treatment assignment, males and females improved their performance in the second run by .78 seconds and .85 seconds on average, respectively.

We randomized classes into treatment and check whether observable characteristics differ between our treatments in Appendix Table 5.A.1. There are no observable differences across treatments for most variables, except for a difference in the pre-treatment times in the first run. However, this gap can be explained entirely by variation in observables. Conditional on gender-specific grade fixed effects, school fixed effects and age, these differences are no longer significant.

5.3.1 Preferences for peers and manipulation check

Before turning to the results of the experiment, we briefly present the preferences for peers as elicited in the survey. Furthermore, we show that our peer assignment based on those preferences indeed changed the actual match quality, which we define as the rank of the assigned peer in the elicited preference

[16] We focus on the students in the three main treatments, namely Random, Name and Performance and do not include the students from the NoPeer treatment.

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

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

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

Table 5.1. Summary statistics

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

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.

rankings. This means that students in the self-selected treatments have a higher probability of being matched with someone who they prefer more, i.e., who ranks higher in their name- or performance-based preferences. Hence, our experimental variation of taking the preferences into account should have an effect on the rank of the assigned peers within a subject's preferences (i.e., the quality of that match) in the respective treatment with self-selection.

Table 5.2. Share of name-based preferences being friends

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

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.

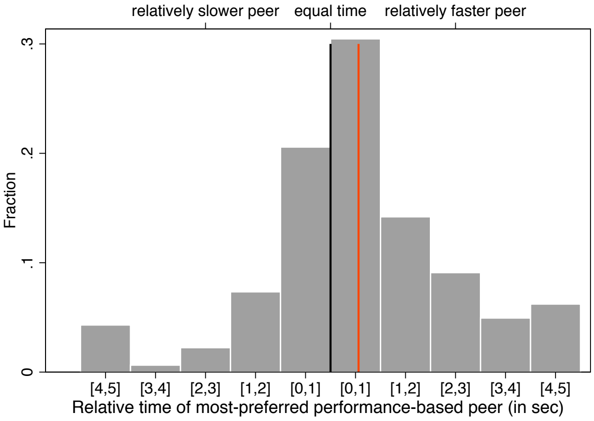


Figure 5.2. Most-preferred performance-based peer

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).

We summarize the preferences for peers according to name- and performance-based preferences in Table 5.2 and Figure 5.2, respectively. Two findings emerge: first, most students nominate friends as their most-preferred peer; and second, while students prefer to run on average with a slightly faster peer, there is strong heterogeneity in this preference. We analyze the determinants of these preferences as well as how these two preference measures relate to each other in more detail in Kiessling et al. (2018).[20]

Figure 5.3 shows the realized match quality for all three treatments with respect to the ranking of peers in the two sets of elicited preferences. The upper panel shows the realized match quality according to name-based preferences. We observe that some people are randomly matched to someone they would like to be paired with in Random and Performance. As expected, this share is rather low. While the median peer in Name corresponds to the most-preferred peer according to the elicited name-based preferences, the median peer is not part of the elicited preferences (i.e., not among the six most-preferred peers) for Random and Performance. A similar, albeit less pronounced picture arises when analyzing the match quality according to the preferences over relative performance as presented in the lower panel of Figure 5.3. We observe that students in Performance are paired with more preferred peers according to their preferences relative to the other two treatments. However, note that subjects may prefer other students or relative times that are not available to them,

[20] In Appendix 5.B, we also show that the rankings of preferred name- and performance-based peers measure two distinct sets of preferences mitigating concerns that the two peer measures correspond to the same underlying preference.

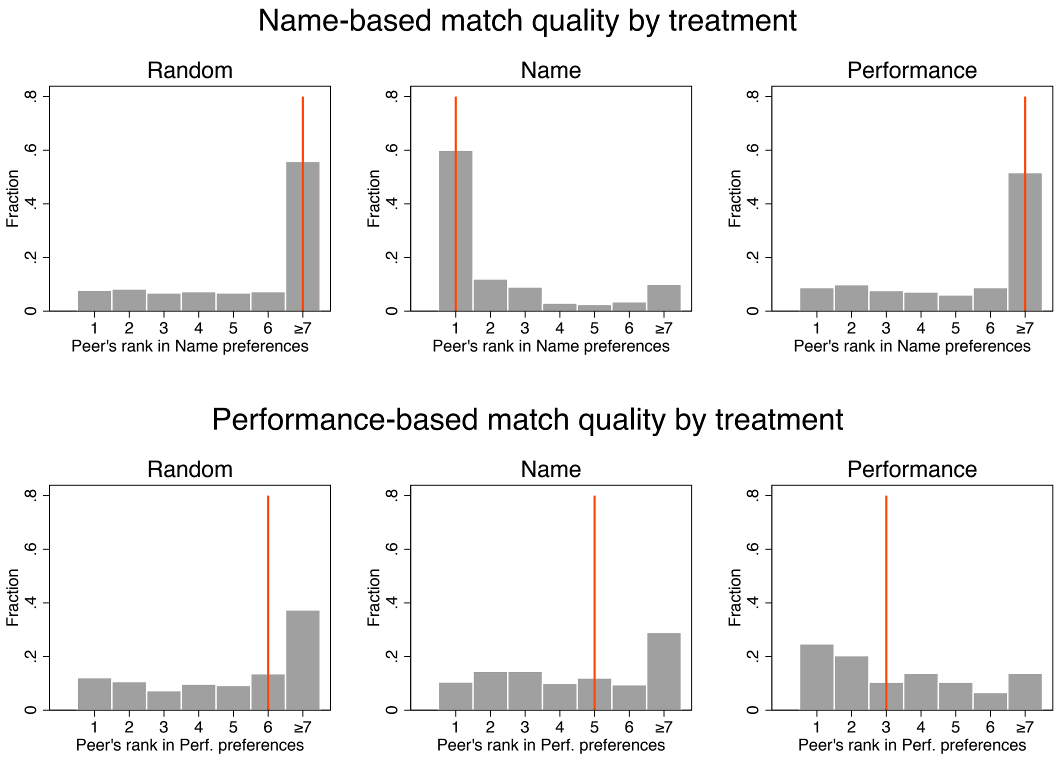


Figure 5.3. Match quality across treatments

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.

which mechanically affects the match quality. Moreover, to match students in PERFORMANCE, the preferences need to exhibit sufficient heterogeneity. We discuss these issues in more detail in Appendices 5.B and 5.C and show that sufficient heterogeneity in preferences exists to match students successfully.

5.4 Empirical Strategy

This section outlines our empirical framework. For this purpose, we first analyze the effect of being assigned to a particular peer assignment mechanism. In a second step, we decompose this change in performance into two effects – an indirect effect stemming from a change in the peer composition and a direct effect due to self-selection – before we show how to allow for heterogeneities in the direct effect depending on the rank within a pair. In Appendix 5.D, we show how to derive these estimation equations from an economic model similar to the mediation analysis as described in Heckman and Pinto (2015).

The random assignment of classes into treatments allows us to estimate the average effect of peer selection on performance. Let $D^d = 1$ with $d \in \{N, P\}$ denote treatment assignment to Name and Performance, respectively, and zero otherwise. Our baseline specification for an outcome y_{igs} of individual i in gender-specific grade g of school s is therefore given by:

$$y_{igs} = \tau + \tau^N D_i^N + \tau^P D_i^P + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{5.1}$$

The main parameters of interest are τ^N and τ^P, the effect of being assigned to one of our treatments relative to Random. School fixed effects, ρ_s, and gender-specific grade fixed effects, λ_g, control for variation due to different schools (i.e., due to different locations and timing of the experiment) and variation specific to gender and grades.[21] Finally, X_i is a vector of predetermined characteristics such as age as well as personality characteristics and in some specifications class-level control variables, and u_{igs} is a mean zero error term clustered at the class level.

Any change in outcomes can be attributed to one of two main sources: first, different peer-assignment mechanisms may affect peer interactions directly; and second, self-selection may change the peers and therefore the difference between the student's and his or her peer's characteristics. We therefore decompose the average treatment effect into a direct effect of self-selection as well as a pure peer composition effect.[22] This takes into account the change in relative peer characteristics across treatments. We implement this decomposition using the following specification:

$$y_{igs} = \bar{\tau} + \bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P + \beta \theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{5.2}$$

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

[21] See the section 5.3 for a discussion concerning why we include gender-specific grade fixed effects rather than gender and grade fixed effects separately.

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

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

$$\begin{aligned} y_{igs} \;=\; & \bar{\tau} + \bar{\tau}_h^N \mathbb{1}_{\{a_i \geq a_j\}} D_i^N + \bar{\tau}_l^N \mathbb{1}_{\{a_i < a_j\}} D_i^N \\ & + \bar{\tau}_h^P \mathbb{1}_{\{a_i \geq a_j\}} D_i^P + \bar{\tau}_l^P \mathbb{1}_{\{a_i < a_j\}} D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \end{aligned} \tag{5.3}$$

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

5.5 Results

Our experimental design allows us to study the causal effect of different peer assignment mechanisms on individual performance. Two of these assignment rules use the preferences for peers elicited in the survey to form pairs and therefore allow for the self-selection of peers. More specifically, the three treatments correspond to random matching (Random), matching based on self-selected peers using name-based peer preferences (Name) and using preferences over relative performance (Performance). As outlined in section 5.2, the random assignment of peers constitutes a natural starting point for at least two reasons: first, the pure presence of any peer might already improve performance; and second, randomly assigned peers are used to document peer effects in a wide range of settings (e.g., Falk and Ichino, 2006; Guryan et al., 2009). We contrast this baseline condition with two treatments that assign peers based on elicited preferences, i.e., in which subjects endogenously choose their peer.

5.5.1 Average effect of self-selection on performance

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

Figure 5.4 presents our main result. Subjects in Random improve on average by 1.93 percentage points during their second run. However, performance improves even more in Name and Performance by 3.22 and 3.58 percentage points, respectively. We present the corresponding estimates in Table 5.3. Columns (1)-(3) present the estimated percentage point improvements in time according to equation 5.1. Columns (3)-(7) express the results additionally in terms of (standardized) times in the second run controlling for times in the first

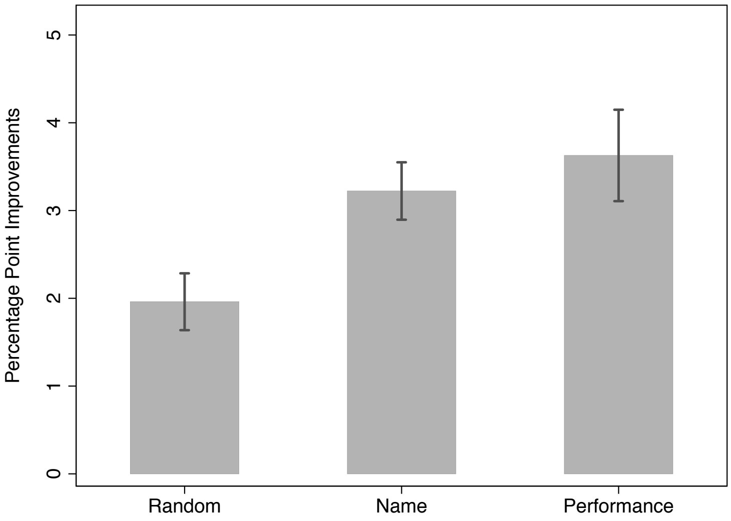


Figure 5.4. Average performance improvements

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

Table 5.3. Average treatment effects

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

Notes: This table presents least squares regressions according to equation 5.1 using percentage point improvements (panel (a)) and times of the second run controlling for times in the first run (panel (b)) as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own characteristics include the Big 5, locus of control, social comparison, competitiveness and risk attitudes. Class-level control variables in columns (3) and (6) include the share of participating students, three variables to capture the atmosphere within a class (missing for four classes), and indicators for the size of the matching group. Column (7) uses standardized times.

run to confirm these effects in times rather than percentage point improvements. Assigning peers based on name-based preferences results in an additional 1.26 percentage point improvement in performance relative to the random assignment of peers. The coefficient for performance-based matching is 1.67 percentage points and thus somewhat larger, but it does not differ significantly from Name. These effects persist when controlling for students' own personal characteristics (column (2)) as well as if we additionally control for class-level variables capturing the atmosphere within a class (column (3)). Our baseline effects correspond to additional time improvements of .38 to .41 seconds and account for 14% of a standard deviation in Name and 15% in Performance (cf. columns (4)-(7)).[23]

In Appendix 5.E, we show that the observed performance improvements are due to the presence of peers and not due to learning. We present the results of an additional control treatment (NoPeer) and its implementation details. In the control treatment, subjects run twice without any peer and we find that they do not improve their time from the first to the second run; in fact, individual performance decreases. The improvements that we observe here can therefore be attributed to the presence of peers rather than learning or familiarity with the task.

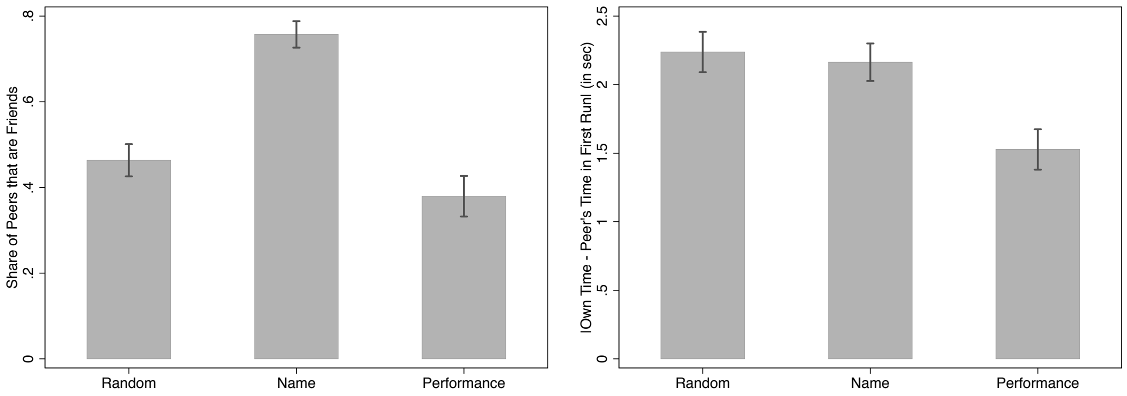
5.5.2 Changes in the peer composition and the direct effect of self-selection

As outlined in section 5.4, the estimated average treatment effects consist of a direct effect due to self-selection and an indirect effect. The latter captures changes in the relative characteristics of the peer (e.g., the time differences between the student and peer in the first run) due to the altered peer composition induced by our treatments.[24] In the following, we first document how Name and Performance change the peer composition relative to Random, before analyzing the extent to which this change in the peer composition can explain the average treatment effect.

It is important to check for a change in the composition and the resulting indirect effect as potentially not all peers are equally important. Suppose that only interacting and comparing yourself with a friend leads to a change in performance (e.g., Bandiera et al., 2009) and at the same time subjects only select their friends in Name. Alternatively, suppose that peers only matter if they have a sim-

[23] Appendix 5.F presents additional robustness checks using biased linear reduction standard errors, controlling for outliers, and presents the average treatment effects for different subgroups. Our results are robust to all of these checks.

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



(a) Share of Peers Being Friends

(b) Absolute Differences in Ability

Figure 5.5. Changes in peer composition

Notes: Figure 5.5a presents the share of all students who nominated their assigned peer as a friend for each of the three treatments including standard errors. Figure 5.5b shows the average absolute within-pair difference in ability (measured in times from the first run) and including standard errors for each treatment. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level. We present the corresponding regressions and highlight additional compositional differences of the treatments in Appendix Table 5.C.1.

ilar performance and at the same time subjects more commonly select someone with a similar performance in Performance. Potentially, our treatments would simply change the likelihood of interacting with such a person (i.e., change the peer composition between treatments) and these changes would explain the average treatment effect.

Figure 5.5 shows that our treatments indeed changed the peer composition with respect to two prime examples of relative peer characteristics, namely friendship ties and ability differences within pairs. Even though students could mainly target peers along these two dimensions, we present how our treatments affect the peer composition along various other characteristics in Appendix Table 5.C.1. More specifically, Figure 5.5a shows that students are predominantly paired with friends in Name (76% of all peers are friends), whereas the share of peers being friends in Random and Performance is 49% and 37%, respectively. As matching based on preferences over relative performance (Performance) allows for targeting other students with a similar or slightly higher ability, the students' absolute time differences in the first run might change. Panel B of Figure 5.5b confirms this by showing that the average absolute difference in times from the first run is 1.53 seconds in Performance, while it is greater than two seconds in the other two treatments (2.24 and 2.16 seconds in Random and Name).

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

Moreover, by having access to preferences over peers, we are able to include the match quality of a peer as a potential mediator. For this purpose, we define two indicators to measure whether the assigned peer is nominated among the first three peers for name-based preferences or falls into the three highest ranked categories for performance-based preferences.[26]

The results of the decomposition based on equation 5.2 are presented in Table 5.4. Column (1) replicates the baseline estimates from column (2) of Table 5.3 for means of comparison. In columns (2)-(5), we include different sets of characteristics, before we allow all of them to mediate the direct effects in column (6).

Only controlling for name-based and performance-based match quality or friendship ties (column (2) and (3)) has little to no effect as the variables themselves have only small and insignificant effects on performance improvement. Hence, the estimated direct effects closely resemble the average treatment effects. In column (4), we focus on ability differences and ranks within a pair. Since faster and slower students within a pair might be affected differentially, we allow the effect of ability differences, $|\Delta Time1|$, to differ by the rank within a pair. We find that ability differences have a significant effect on both faster and slower students within a pair. On the one hand, slower students within a pair benefit strongly from running with a faster student, whereby a one-second difference in ability leads to a 1.03 percentage point improvement in the second run. On the other hand, the performance of the relatively faster student suffers from ability differences and their performance declines by .39 percentage points per second. In sum, the average performance of a pair thus improves with increased ability differences. However, the impact of ability differences does not mediate the direct treatment effects. The estimated coefficient for Name remains stable, and the effect for Performance even increases, implying that the indirect effect for Performance is negative. This is partially a consequence of the smaller ability differences in Performance relative to Random as shown in Figure 5.5b and the overall positive impact of ability differences.

In column (5), we analyze the direct effects if we include the similarity in several personal characteristics of the two students of a pair. In contrast to ability differences and friendship ties, personality characteristics could not be tar-

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

[26] Appendix Table 5.G.3 also controls for match quality in a flexible way. The results remain qualitatively and quantitatively similar.

Table 5.4. Decomposition of treatment effects

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

Notes: This table presents least squares regressions according to equation 5.2 using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. High match quality is an indicator equaling one if the partner was ranked within the first three preferences according to his or her name- or performance-based preferences. Own characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. The last two rows quantify the indirect effect for Name and Performance given by the combining the change in peer composition across treatments (cf. Appendix Table 5.C.1) with the corresponding compositional effects of these characteristics in column (6). Further robustness checks are relegated to the Appendix.

geted easily in the preference elicitation. Nonetheless, subjects could have chosen peers with certain personality characteristics indirectly in both treatments. However, the treatment effects remain stable if we control for those characteristics.

Finally, we control for all of these mediators simultaneously in column (6). The effects of the peer characteristics are in line with what we have discussed above. In the last two rows of the table, we quantify the indirect effect as the change in the coefficient of Name and Performance when controlling for the peer composition (column (1) vs (6)). This corresponds to multiplying the coefficients from column (6) with the change in the peer composition across treatments. We describe these changes in Appendix Table 5.C.1.

In Name, we estimate a positive indirect effect of .10 percentage point improvements. This means that the altered peer characteristics have only a slightly positive effect on the students' performance. The direct effect is 1.26 percentage points and therefore somewhat smaller than the average effect, but not significantly different (Wald test, p-value = 0.66). For Performance, we observe an indirect effect of -.49 percentage points. Therefore, the change in the peer composition suppresses improvements in Performance. The direct effect is 2.18 percentage points and it significantly differs from the average effect (Wald test, p-value = 0.029). The magnitude of the direct effects is more than five times that of the indirect effects.[27]

Our analysis suggests that self-selection improves individual performance directly and not due to a change in the peer composition. This means that subjects react to observationally similar peers differently once they have chosen them actively. The direct effect could stem from an additional motivational value of self-selection, as the comparison and interaction with self-selected peers might become more important. In principle, a compositional change in unobserved characteristics – that is not measured by those included in our analysis and differs across treatments – could still account for the direct effects. However, the effect would have to be at least five times the size of the measured indirect effect.

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

[27] We present additional robustness checks in Appendix 5.G. In Table 5.G.1, we show that match quality itself has no influence in Random. Being paired randomly with a preferred peer does not increase performance. Furthermore, Table 5.G.2 presents the robustness of the direct effects to using only those subjects in Random who are matched in line with their preferences. These matches occurred by pure chance and not due to self-selection. Finally, we document in Table 5.G.3 that the piecewise-linear specification of ability differences and the definition of the high matching quality indicator are not restrictive by including interval fixed effects for each one-second interval of ability differences and fixed effects for each rank of the name- and performance-based preference ranking, respectively. Additionally, this table also shows that conditioning on class-level variables does not alter our results.

5.5.3 Markers for changed social interactions

In this section, we study the effects of our treatments on students' experience during the tasks. Our experiment features a small post-experimental questionnaire, in which we elicited how much peer pressure students experienced and how much fun they had during the second run.[28] In order to analyze the effects of the treatments on these two variables, Table 5.5 presents estimates for the direct effects of our treatments based on equation 5.2 using standardized measures of pressure and fun as outcome variables. Here, we control for times in the second rather than the first run for two reasons: first, these measures are elicited after the second run; and second, a tight race could increase pressure across all treatments.

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

Focusing on fun in the second run, we do not find any significant direct effects (see panel (b) of Table 5.5). However, we observe a significant negative effect on time differences in the second run for the slower student. Fun decreases for the slower peer with increasing distance to the peer. Combined with the zero effect of finishing second, we conclude that it is not losing per se that affects fun, but rather the margin of losing. Furthermore, the absence of direct effects alleviates a potential concern that knowledge of all three treatments leads to disappointment when students are assigned to Random, namely when they are unable to select their peer themselves.[29] If those students were more disappointed, this might lead to smaller improvements by students in Random compared to the two other treatments. If disappointment had driven the results, we would have expected students in Random to have significantly less fun.[30]

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

[29] One might also argue that this also describes a feature of real-world settings. Imagine that you are randomly assigned a partner from a group of available people. Even if you have not explicitly been asked with whom you would have liked to interact, you still have preferences about interacting with certain people. Therefore, disappointment could also play a role in these settings. This might be true for all settings that feature exogenous assignment and overrule the underlying preferences of the people involved.

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

Table 5.5. Post-experimental questions

	(a) Pressure (std.).		(b) Fun (std.)	
	(1)	(2)	(3)	(4)
Direct Effects				
Name	0.24	0.10	0.14*	-0.01
	(0.20)	(0.18)	(0.08)	(0.10)
Performance	0.32	0.46**	-0.10	-0.10
	(0.20)	(0.15)	(0.07)	(0.08)
Faster Student (2nd Run) × Match Quality (name-based)		0.28		0.30**
		(0.35)		(0.14)
Slower Student (2nd Run) × Match Quality (name-based)		0.33		0.07
		(0.28)		(0.17)
Faster Student (2nd Run) × Match Quality (perf.-based)		0.01		0.07
		(0.33)		(0.11)
Slower Student (2nd Run) × Match Quality (perf.-based)		-0.16		0.22*
		(0.27)		(0.12)
Faster Student (2nd Run) × Peer is friend		-0.14		0.12
		(0.44)		(0.13)
Slower Student (2nd Run) × Peer is friend		-0.03		0.15
		(0.35)		(0.15)
Faster Student (2nd Run) × $\lvert \Delta Time\ \ 2 \rvert$		-0.25**		-0.01
		(0.08)		(0.04)
Slower Student (2nd Run) × $\lvert \Delta Time\ \ 2 \rvert$		0.10		-0.14***
		(0.09)		(0.04)
Slower Student in Pair (2nd Run)		-0.25		0.04
		(0.26)		(0.18)
Gender/Grade/School FEs, Age	Yes	Yes	Yes	Yes
Own and Peer Characteristics	Yes	Yes	Yes	Yes
Abs. Diff. in Personality	No	Yes	No	Yes
N	161	161	582	582
R^2	.2	.32	.28	.34
p-value (Name vs. Performance)	.72	.17	.03	.46

Notes: This table presents least squares regressions according to equation 5.2 using the standardized survey measure of pressure (Panel (a)) or fun (Panel (b)) as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Match quality equals one if a student's peer is among his three most-preferred peers according to his name- or performance-based preferences. Note that the faster/slower student is defined according to relative times in the second run.

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

5.5.4 Do treatments change the within-pair interaction?

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

Column (1) of Table 5.6 replicates specification (6) of Table 5.4. In column (2), we allow the direct effect of our treatments to differ by rank according to equation 5.3. Self-selection yields a positive direct effect for all students independent of their rank in Performance. In Name, only slower students within a pair exhibit significant direct effects compared to Random. Faster students within a pair are unaffected in Name. This shows that selection on names motivates slower students to catch up with their faster peers. By contrast, selection on relative performance causes both students to improve their performance.

The observed within-pair interaction has direct consequences for the difference in performance levels across treatments. As the slower student within a pair drives the direct effect in Name, we expect a decrease in the within-pair difference in levels in Name. In Table 5.7, we analyze the absolute within-pair time difference in the second run. In column (1), we calculate the average treatment effect for these differences and show that they are significantly smaller for both treatments allowing for self-selection. In column (2), we decompose this effect again in a direct and indirect one using pair-level mediators, i.e., absolute time difference in the first run, friendship indicators and absolute differences in personality characteristics. We find that lower absolute differences in Performance are an artifact of the changed peer composition and therefore due to the selection mechanism (i.e., lower absolute differences in ability), while we observe a direct convergence effect for Name.

Although the direct effect of self-selection in both treatments is similar in sign and magnitude, the two treatments induce distinct interaction patterns within pairs. While in Name only the slower student within a pair drives the direct effect, all students improve due to self-selection in Performance. We also observe a similar convergence in performance levels across both treatments with self-selection. However, this result is due to the selection mechanism in Performance and due to the interaction in Name. In combination with the results in section 5.5.3, these heterogeneous effects show that our treatments work through different channels and thereby affect the subjects differently.

Table 5.6. Rank heterogeneity within pairs

	Percentage point imprv.	
	(1)	(2)
Direct Effects		
Name	1.24**	0.58
	(0.49)	(0.61)
Performance	2.22***	2.15***
	(0.68)	(0.69)
Name × Slower Student in Pair		1.35*
		(0.70)
Performance × Slower Student in Pair		0.15
		(0.65)
Faster Student × Match Quality (name-based)	0.53	0.82
	(0.43)	(0.49)
Slower Student × Match Quality (name-based)	0.49	0.16
	(0.66)	(0.65)
Faster Student × Match Quality (perf.-based)	0.49	0.51
	(0.52)	(0.52)
Slower Student × Match Quality (perf.-based)	-0.65	-0.62
	(0.65)	(0.63)
Faster Student × Peer is friend	-1.18**	-1.13**
	(0.49)	(0.48)
Slower Student × Peer is friend	0.11	0.06
	(0.66)	(0.67)
Faster Student × $\lvert \Delta Time\ \ 1\rvert$	-0.32**	-0.31*
	(0.15)	(0.15)
Slower Student × $\lvert \Delta Time\ \ 1\rvert$	1.02***	1.01***
	(0.20)	(0.20)
Slower Student in Pair	-0.21	-0.34
	(0.70)	(0.77)
Abs. Diff. in Personality	Yes	Yes
Own Characteristics	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes
N	582	582
R^2	.28	.28
p-value (Name vs. Performance)	.16	.028

Notes: This table presents least squares regressions according to equation 5.3 using percentage point improvements as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Own characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. Match quality equals one if a student's peer is among his three most-preferred peers according to his name- or performance-based preferences.

5.5.5 Implications for targeting individuals

Our results show that the process of self-selection has a heterogeneous impact on the subjects depending on the rank within a pair. However, a policy maker

Table 5.7. Convergence of performance within pairs

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

Notes: This table presents least squares regressions using absolute differences of times in the second run as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Additional peer composition controls include absolute differences of personality characteristics of subjects and their peers (Big Five, locus of control, social comparison, competitiveness, risk attitudes).

might not only be interested in the changed interaction within pairs, but rather they might target specific groups of individuals to improve their performance, irrespective of direct or indirect effects driving these improvements. For this purpose, we look at the heterogeneity in average treatment effects conditional on ability and simulate the effects of other rules employing exogenous peer assignment.

Figure 5.6 presents percentage point improvements of low-, medium- and high-ability subjects across the three assignment rules.[31] Across all treatments, the performance improvements decrease when ability increases but remain positive even for high-ability students. This mainly stems from the positive effect of ability differences for slower students within a pair and negative effect for faster ones.[32]

Although this decreasing pattern holds for all three treatments, there are some differences. Low-ability students in Random show large improvements of

[31] The corresponding regressions as well as alternative specifications are presented in Appendix Table 5.H.1. Low, medium and high ability are defined according to terciles of times in the first run within each school, grade and gender.

[32] Table 5.6 shows that a one-second ability difference improves performance by 1.06 percentage points for slower students within a pair and reduces the faster students' performance by .34 percentage points.

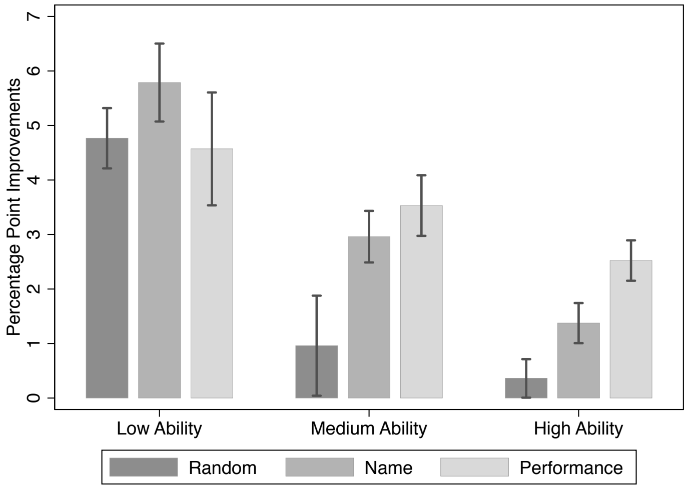


Figure 5.6. Heterogeneity by own ability

Notes: The figure presents percentage point improvements and standard errors for the three treatments Random (dark gray), Name (gray), and Performance (light gray) by ability terciles. We control for gender, grade and school fixed effects as well as for age and cluster standard errors at the class level. The corresponding regressions are presented in Appendix Table 5.H.1.

4.77 percentage points (p-value < 0.01), while medium- and high-ability students do not improve significantly (.96 and .36 percentage points with p-values of .30 and .31, respectively). All students across the ability distribution improve more in Name than in Random by 1.02 (p-value = 0.28), 2.00 (p-value = 0.05) and 1.01 percentage points (p-value = 0.04) for low-, medium- and high-ability students. By contrast, Performance does not help low-ability students relative to Random (.20 percentage points decrease, p-value = 0.86) but benefits students from the upper two terciles of the ability distribution by 2.57 (p-value = 0.02) and 2.16 (p-value < 0.01) percentage points. Overall, the performance improvements are more equally distributed across different levels of ability.

The treatments therefore target different groups of individuals. Low-ability students benefit most from name-based matching, whereas students with higher ability show the largest improvements when matched using preference over relative performance. Policy makers can therefore use different peer assignment rules to benefit specific groups of individuals.

The previous sections and the patterns in Figure 5.6 imply that individual improvements are largely determined by the interplay of the peer – especially his or her relative ability – and the treatment. Table 5.6 shows that mainly the slower students within a pair improve in Name, while both improve similarly in Performance compared to the random assignment of peers. Low-ability students benefit most from this as they are more likely to be paired with faster

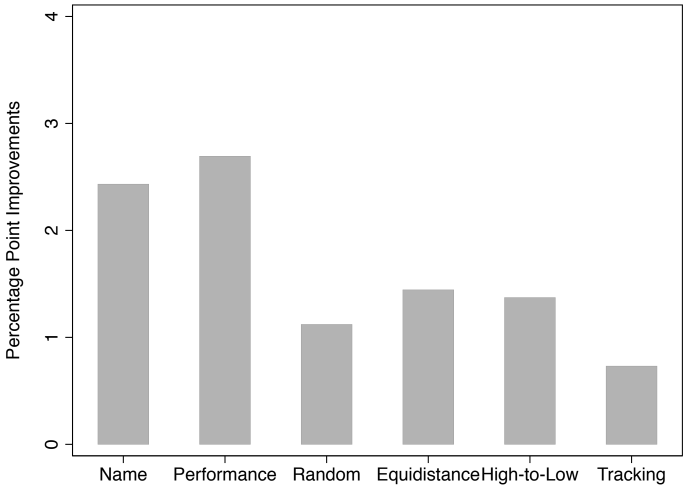


Figure 5.7. Simulation of other peer assignment rules

Notes: The figure presents predicted percentage point improvements for the three treatments (Name, Performance and Random) as well as three simulated peer-assignment rules (Equidistance, High-to-Low and Tracking). We fix the personal characteristics and other covariates not at the pair level to 0, whereby effect sizes are therefore not directly comparable to treatment effects above. More details are provided in the text and Appendix 5.I.

students. This effect is amplified compared to Performance as this treatment results in pairs with smaller ability differences relative to the other two treatments.[33] Note that this only results in a positive effect for low-ability students in Name if these students choose faster students and are subsequently matched with them, a condition that is satisfied in our setting (see Appendix Table 5.H.2). This implies that the choice of a peer by an individual carries greater weight for individual improvements in treatment Name than in Performance, as the former only benefits slower students in a pair whereas the latter benefits both students. This might also help to understand the absence of improvements for low-ability students in Carrell et al. (2013) as students in their setting might not have chosen high-ability students as relevant peers.

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

[33] Figure 5.6b shows that the ability differences are indeed lower in Performance. These lower ability differences translate into smaller indirect effects for the pair and especially for the slower peer. This is due to smaller ability differences reducing improvements for slower students in a pair, which are not compensated by the effects on faster students. See the coefficients on ability differences interacted with rank of a student in Table 5.6.

effect of such peer-assignment rules and compare their effect to the outcomes under self-selection. From our estimates obtained in section 5.5.2, we know that pairs with a higher difference in ability will improve their performance. If this is the only characteristic of a peer that affects performance, aggregate performance would be maximized as long as the sum of ability differences within a pair is maximized.[34] In order to compare the results of self-selection against exogenous assignment rules that promise the largest aggregate improvements, we consider two matching rules that maximize ability differences within pairs (Equidistance and High-to-Low). Additionally, we look at the effect of tracking (i.e., pairing the best student with the second best, third with the fourth, etc.; Tracking). We compare the predicted performance improvements for those rules with our estimated performance improvements for the three assignment rules used in the experiment.[35]

Figure 5.7 presents the simulated average performance improvements of each assignment rule. The results show that no other peer-assignment rule is able to reach similar performance improvements as those featuring self-selection. In fact, they are close to the results from our random matching, since these students under those peer assignment rules do not benefit from the additional motivational value of self-selection. More surprisingly, the reassignment rules that maximize ability differences in pairs – Equidistance and High-to-Low – do not improve average performance compared to the random assignment of peers. Although both rules increase the average ability difference in pairs by construction and affect performance through this channel, those rules also change other characteristics of the peer. The lack of any additional improvement implies that these other changes in peer characteristics offset the positive effect of increased ability differences.

In general, depending on the objectives such as targeting specific groups of individuals, a policy maker such as a teacher might want to implement different peer assignment mechanisms. While our treatments allowing for self-selected peers seem to induce similar performance improvements on average, they affect different individuals. Compared to Random, we observe performance improvements across the entire ability distribution in Name, but only for higher-ability students in Performance. Nonetheless, such peer assignments may come at a cost, such as increased pressure in Performance (as documented in section 5.5.3) or a large perturbation of individual ranks in Name.[36] Hence, a policy maker might not only look at the resulting outcomes but also how different assignment rules affect the individuals' overall well-being.

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

[35] We provide details on the prediction of performance improvements and the peer assignment rules in Appendix 5.I.

[36] We document this perturbation of ranks in the Appendix Table 5.H.3.

5.6 Conclusion

Peer effects are an ever-present phenomenon discussed in a wide range of settings across the social sciences. For many situations, identifying the effect of an actively self-chosen peer is important beyond estimating peer effects in general. Our framed field experiment introduces a novel way to study the self-selection of peers in a controlled manner and is able to separate the impact of a specific peer on a subject's performance from the overall effect of self-selection. The results of our experiment provide evidence that self-selecting peers yields performance improvements of .14-.15 SD. These cannot be explained by indirect effects of a differing peer composition; rather, they stem from a direct effect, corresponding to a changed social interaction since students are able to select their partner themselves. This implies that self-selected peers can serve as a substantial motivator to improve performance.

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

The results reported in this paper are also in line with earlier studies, which indicate that being paired with high-ability peers leads on average to higher performance (e.g., Carrell et al., 2009). Combined with the process of self-selecting high- or low-ability peers, this can set ex-ante similar individuals on divergent trajectories in classrooms and organizations. Repeatedly choosing higher-ability peers can lead to continuous improvements, whereas selecting lower-ability peers may stall individual development.

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

Our experimental design can easily be transferred to situations in which other production functions are used or where peer effects arise via other channels, e.g., implementing team production by reporting a function of both students' times to the teacher, or varying the task to allow for learning or skill complementaries as sources of peer effects. Self-selection of peers can often be observed in those settings. For example, study groups at universities often form endogenously (Chen and Gong, 2018), researchers select their co-authors and workers in firms increasingly form self-managed work teams (Lazear and Shaw, 2007).

In this paper, we highlight that self-selecting peers can serve as a complement to other established methods such as incentives and exogenous peer assignment policies aimed at increasing individual performance. However, further research on the interplay between endogenous group formation, social interactions and production environments remains imperative to understand how peer effects work.

References

Ager, Philipp, Leonardo Bursztyn, and Hans-Joachim Voth (2016): "Killer Incentives: Status Competition and Pilot Performance during World War II." NBER Working Paper Series. [139]

Aldashev, Gani, Georg Kirchsteiger, and Alexander Sebald (2017): "Assignment Procedure Biases in Randomised Policy Experiments." *Economic Journal*, 127 (602), 873–895. [161]

Allcott, Hunt and Judd Kessler (2015): "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons." *NBER Working Paper Series*. [169]

Aral, Sinan and Christos Nicolaides (2017): "Exercise Contagion in a Global Social Network." *Nature Communications*, 8 (14753). [139, 143]

Babcock, Philip and John Hartman (2010): "Networks and Workouts: Treatment Size and Status Specific Peer Effects in a Randomized Field Experiment." *NBER Working Paper Series*. [139]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2005): "Social Preferences and the Response to Incentives: Evidence from Personnel Data." *Quarterly Journal of Economics*, 120 (3), 917–962. [139]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2009): "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica*, 77 (4), 1047–1094. [139, 156]

Bartling, Björn, Ernst Fehr, and Holger Herz (2014): "The intrinsic value of decision rights." *Econometrica*, 82 (6), 2005–2039. [141]

Belot, Michèle and Jeroen van de Ven (2011): "Friendships and Favouritism on the Schoolground – A Framed Field Experiment." *Economic Journal*, 121 (557), 1228–1251. [141]

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. [142]

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. [142]

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009): “Identification of Peer Effects through Social Networks.” *Journal of Econometrics*, 150, 41–55. [142]

Brandts, Jordi, David Cooper, and Roberto Weber (2014): “Legitimacy, Communication, and Leadership in the Turnaround Game.” *Management Science*, 61 (11), 2627–2645. [142]

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. [139, 141]

Carrell, Scott, Richard Fullerton, and James West (2009): “Does Your Cohort Matter? Measuring Peer Effects in College Achievement.” *Journal of Labor Economics*, 27 (3), 439–464. [142, 169]

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. [142, 167, 169, 194]

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.” [143, 158]

Chen, Roy and Jie Gong (2018): “Can self selection create high-performing teams?” *Journal of Economic Behavior and Organization*, 148, 20–33. [143, 170]

Cicala, Steve, Roland Fryer, and Jörg Spenkuch (forthcoming): “Self-Selection and Comparative Advantage in Social Interactions.” *Journal of the European Economic Association*. [143]

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg (2017): “Peer Effects in the Workplace.” *American Economic Review*, 107 (2), 425–456. [139]

Dahl, Gordon B., Katrine V. Løken, and Magne Mogstad (2014): “Peer Effects in Program Participation.” *American Economic Review*, 104 (7), 2049–2074. [139]

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. [145]

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. [139, 167]

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. [143, 191]

Falk, Armin and Andrea Ichino (2006): “Clean Evidence on Peer Effects.” *Journal of Labor Economics*, 24 (1), 39–57. [154]

Feld, Jan and Ulf Zölitz (2017): “Understanding Peer Effects: On the Nature, Estimation, and Channels of Peer Effects.” *Journal of Labor Economics*, 35 (2), 387–428. [142]

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

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. [145]

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse (2017): “First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision.” [143, 191]

Gneezy, Uri and Aldo Rustichini (2004): "Gender and Competition at a Young Age." *American Economic Review*, 94 (2), 377–381. [144, 194]

Golsteyn, Bart, Arjan Non, and Ulf Zölitz (2017): "The Impact of Peer Personality on Academic Achievement." [158]

Guryan, Jonathan, Kory Kroft, and Matthew Notowidigdo (2009): "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments." *American Economic Journal: Applied Economics*, 1 (4), 34–68. [154]

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. [153, 180, 181]

Herbst, Daniel and Alexandre Mas (2015): "Peer Effects on Worker Output in the Laboratory Generalize to the Field." *Science*, 350 (6260), 545–549. [142]

Irving, Robert (1985): "An Efficient Algorithm for the Roommates"Problem." *Journal of Algorithms*, 6 (4), 577–595. [147]

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube (2018): "To whom may you compare: Preferences for peers." [151]

Kimbrough, Erik, Andrew McGee, and Hitoshi Shigeoka (2017): "How Do Peers Impact Learning? An Experimental Investigation of Peer-to-Peer Teaching and Ability Tracking." *IZA Discussion Paper Series*. [141]

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. [139]

Lavy, Victor and Edith Sand (2015): "The Effect of Social Networks on Student's Academic and Non-Cognitive Behavioral Outcomes: Evidence from Conditional Random Assignment of Friends in School." [142]

Lazear, Edward and Kathryn Shaw (2007): "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives*, 21 (4), 91–114. [170]

Manski, Charles (1993): "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60 (3), 531–542. [142]

Mas, Alexandre and Enrico Moretti (2009): "Peers at Work." *American Economic Review*, 99 (1), 112–145. [139, 141]

Rotter, Julian B. (1966): "Generalized Expectancies for Internal Versus External Control of Reinforcement." *Psychological Monographs: General and Applied*, 80 (1), 1–28. [145]

Sacerdote, Bruce (2001): "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics*, 116 (2), 681–704. [139]

Sacerdote, Bruce (2014): "Experimental and quasi-experimental analysis of peer effects: two steps forward?" *Annual Review of Economics*, 6 (1), 253–272. [142]

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*. [145]

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. [144]

Tincani, Michela (2017): "Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence." [143]

Weinhardt, Michael and Jürgen Schupp (2011): "Multi-Itemskalen im SOEP Jugendfragebogen." *DIW Data Documentation*. [145]

Appendix 5.A Randomization check

Table 5.A.1. Randomization check

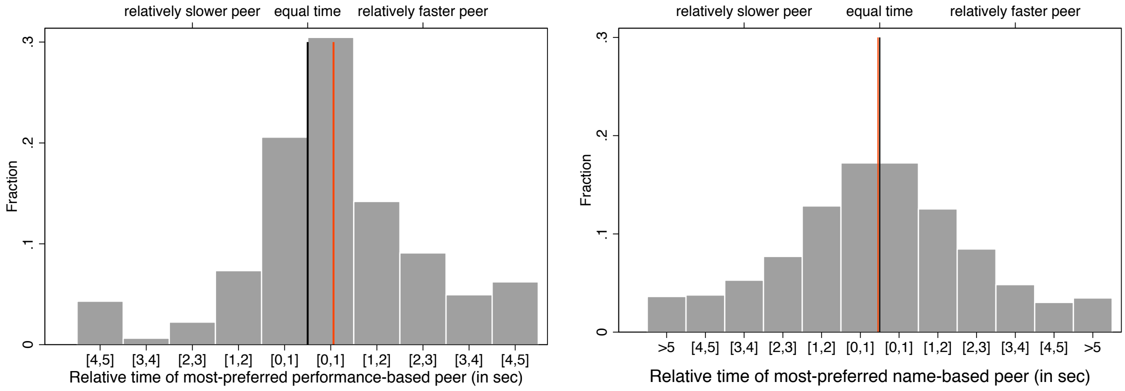
	Random	Name	Diff.	Performance	Diff.
Socio-Demographics					
Age	14.43	14.55	0.13	14.58	0.15
	(1.18)	(1.24)	(0.12)	(1.24)	(0.12)
Female	0.73	0.62	-0.11*	0.61	-0.12*
	(0.45)	(0.49)	(0.04)	(0.49)	(0.05)
Doing sports regularly	0.82	0.82	0.00	0.90	0.08
	(0.39)	(0.38)	(0.04)	(0.31)	(0.04)
Times (in sec)					
Time (First Run)	26.81	26.08	-0.73*	26.19	-0.62*
	(2.96)	(2.93)	(0.28)	(2.78)	(0.28)
Residual of Time (First Run)	0.25	-0.00	-0.25	-0.00	-0.25
	(2.96)	(2.93)	(0.28)	(2.78)	(0.28)
Class-level Variables					
# Students in class	26.01	25.39	-0.62*	25.61	-0.41
	(2.95)	(2.02)	(0.24)	(3.11)	(0.30)
Share of participating students	0.72	0.74	0.02	0.73	0.01
	(0.16)	(0.13)	(0.01)	(0.12)	(0.01)
Grade	8.68	8.76	0.08	8.75	0.07
	(1.07)	(1.12)	(0.11)	(1.13)	(0.11)
Observations	221	213	434	193	414

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 on school, grade and gender fixed effects as well as an indicator for the first or second run. We then use the residuals from this regression.

Appendix 5.B Description and comparison of peer preferences

In this section, we briefly describe the preferences elicited in the survey and then compare preference over relative performance and based on names. Suppose that all subjects want to be paired with a faster peer. Subsequently, we may not be able to match this most-preferred peer to half of the sample. This implies that we need a sufficient amount of heterogeneity in performance-based preferences to match pairs optimally given their preferences. Figure 5.B.1a presents a histogram of the most-preferred relative performance of a peer. It shows that – although subjects prefer a similar or slightly faster peer on average – preferences are still heterogeneous mitigating the concern that we are unable to provide subjects with peers according to their preferences. Moreover, Appendix 5.C presents a manipulation check of our treatments and shows that we are indeed able to form pairs based on these elicited preferences.

In Figure 5.B.1b, we present the corresponding histogram using the relative times of the most-preferred name-based peers. On average, students choose similar peers of similar ability, but the dispersion of preferences is much larger than for performance-based preferences. Moreover, the most-preferred name-based peer is a friend in 89% of all cases (see Table 5.2).



(a) Heterogeneity in perf.-based preferences **(b)** Heterogeneity in name-based preferences

Figure 5.B.1. Heterogeneity in preferences

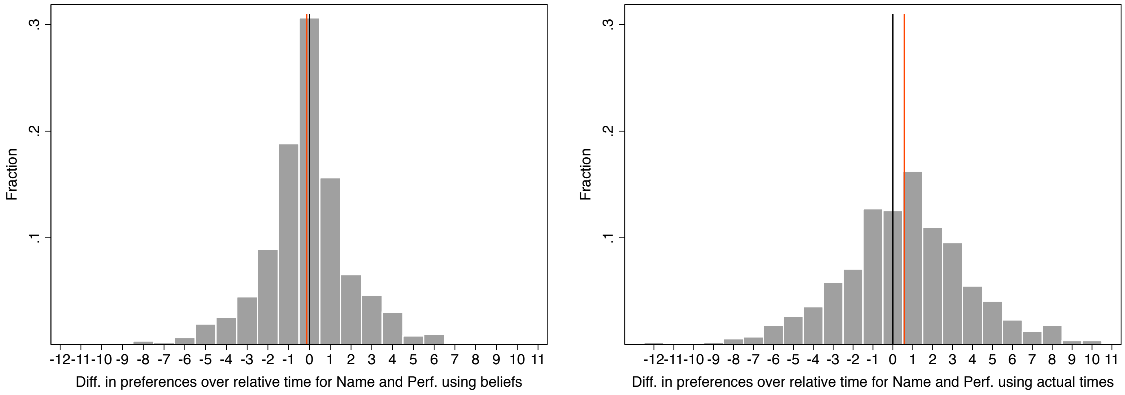
Notes: The figures present histograms of the most-preferred relative performances of the students in Performance (Panel (a), same as in Figure 5.2) and the relative time of the most-preferred name-based peers (Panel (b)). The intervals used here and in the survey are one-second intervals of relative times in the first run. Vertical lines indicate own time (black; equals zero by definition) and mean preference (red; 0.56 sec faster for performance-based preferences where we used the mean of each interval to calculate the mean, and 0.05 sec slower for name-based preferences).

In order to show the difference between name- and performance-based preferences, we make use of the elicited beliefs over the relative performance of peers nominated in the name-based preferences. As the elicitation procedure of those beliefs is identical to that of the preferences over relative performance, we can therefore check if subjects want to choose the same kind of peer in terms of relative performance. If only relative performance matters as a criterion for the selection process, subjects should choose a peer, which they believe has the same relative performance as they choose in the performance-based selection process. At least, this difference should be very small.[1] Since subjects beliefs might be noisy, we can repeat this exercise with the actual performance differences in the first run. Figure 5.B.2 shows that although on average subjects choose somebody with a similar performance (based on their belief or actual times), there is a lot of variation in those preferences.[2] Therefore, we can conclude that the two sets of preferences are distinct preferences and that not only relative performance matters for the name-based selection process.[3]

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

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

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



(a) Dissimilarity of preferences using beliefs **(b)** Dissimilarity of preferences using times

Figure 5.B.2. Dissimilarity of preferences

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

Appendix 5.C Manipulation checks

In section 5.3.1, we presented the resulting match qualities using the preferences as elicited in the survey. However, some subjects may prefer relative times, which are not available to them. For example, the fastest subject in the class might want to run with someone who is even faster, or a student wants to run with somebody else who is 1-2 seconds faster but by chance there is no one in the class with such a time. Similarly, subjects in Name may rank other students which were not present during the experiment or did not participate. We therefore present an alternative approach to evaluate the match quality by taking the availability of peers into account. This implies that the quality of a match does not correspond directly to the elicited preferences; rather, based on these preferences all available subjects (i.e., the students participating in the study) are ranked. The quality of the match is then calculated based on this new ranking and results in a realized feasible match quality.

Consequently, we determine the feasible match quality by calculating how high a classmate is ranked in a list of available classmates.[1] In Name, this can only increase the match quality. If someone nominates another student who is not available as her most-preferred peer and she received her second highest ranked choice, this means that she is matched with her most-preferred feasible peer. Similar arguments can increase the match quality for preferences over relative performance. However, the match quality in performance can also be lower. Suppose that a student ranks the category "1-2 seconds faster" highest and there are three students in that category. However, she is only matched with her second highest ranked category. There would have been three subjects whom she would have preferred more, generating a feasible match quality of 4. We present the corresponding histograms in Figure 5.C.1 and observe that the median of the feasible match quality is actually higher for both treatments relatively to the match qualities depicted in Figure 5.3.

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 5.C.1.

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

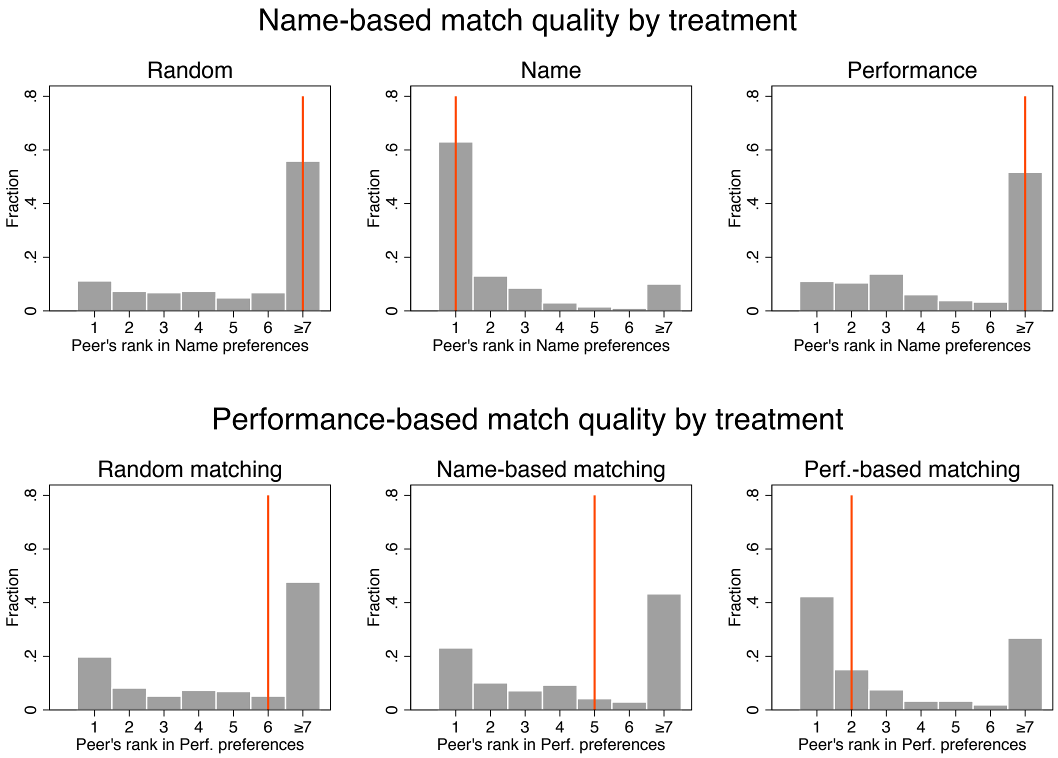


Figure 5.C.1. Feasible match quality across treatments

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.

Table 5.C.1. Effects of treatments on peer composition

	Match Qual. (name)	Match Qual. (time)	Friendship Ties	Time 1	
Name	0.49***	0.07	0.27***	-0.08	
	(0.06)	(0.04)	(0.06)	(0.19)	
Performance	-0.06	0.24***	-0.12*	-0.70***	
	(0.06)	(0.04)	(0.07)	(0.21)	
N	588	588	294	294	
R^2	.34	.083	.19	.09	
p-value: Name vs. Performance	1.0e-11	.0002	3.4e-07	.0037	
Mean in Random	.23	.3	.43	2.4	
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Name	-0.14	0.09	-0.15	0.11	-0.15
	(0.14)	(0.09)	(0.11)	(0.13)	(0.10)
Performance	0.01	0.14	-0.20	0.28**	0.12
	(0.17)	(0.09)	(0.12)	(0.13)	(0.11)
N	292	292	292	292	292
R^2	.05	.058	.047	.039	.03
p-value: Name vs. Performance	.19	.53	.63	.19	.031
Mean in Random	1.2	1	1.1	.98	1.1
	Locus of Control	Social Comparison	Competitiveness	Risk	
Name	0.12	0.00	0.03	0.07	
	(0.11)	(0.10)	(0.13)	(0.11)	
Performance	0.46***	-0.19**	0.12	0.05	
	(0.12)	(0.09)	(0.11)	(0.11)	
N	292	293	291	292	
R^2	.065	.033	.03	.019	
p-value: Name vs. Performance	.003	.079	.37	.76	
Mean in Random	.98	1.1	1.1	1.1	

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.

Appendix 5.D Econometric Framework

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

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

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

$$Y \;=\; D^N Y^N + D^P Y^P + (1-D^P)(1-D^N)Y^R \tag{5.D.1}$$

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

$$Y_d \;=\; \alpha_d + \beta_d\theta + \gamma X + \epsilon_d \tag{5.D.2}$$

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

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

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

Potentially, there are unobserved factors in θ. We therefore split θ in a vector with the observed inputs ($\bar{\theta}$) and unobserved inputs ($\tilde{\theta}$)[3] with corresponding effects $\bar{\beta}_d$ and $\tilde{\beta}_d$ and can rewrite equation 5.D.2 as follows:

$$Y_d = \alpha_d + \bar{\beta}_d \bar{\theta} + \tilde{\beta}_d \tilde{\theta} + \gamma X + \epsilon_d \tag{5.D.3}$$

$$= \tau_d + \bar{\beta}_d \bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{5.D.4}$$

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

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

$$= \hat{\tau}_d + \bar{\beta}\bar{\theta} + \gamma X + \tilde{\epsilon}_d \tag{5.D.6}$$

In what follows, we are interested in $\bar{\tau}_d = \mathbb{E}[\hat{\tau}_d - \hat{\tau}_R]$ ($d \in \{N, P\}$; $\hat{\tau}_d = \tau_d + \bar{\Delta}_{R,d}\bar{\theta}$), i.e., the direct treatment effect of Name and Performance conditional on indirect effects from changes in the peer composition captured in $\bar{\theta}$. This direct effect subsumes the effect of the treatment itself ($\alpha_d - \alpha_R$), the changed impact of the same peer's observables ($\bar{\Delta}_{R,d}\bar{\theta}$), and changes in unmeasured inputs as well as their effect (($\tilde{\beta} + \tilde{\Delta}_{R,d})\tilde{\theta}$). We interpret this direct effect as an additional motivation due to being able to self-select a peer. This focus on the direct effect is a key difference compared with Heckman and Pinto (2015), who are mainly interested in the indirect effects of the mediating variables. The empirical specification of 5.D.6 is given by

$$y_{igs} = \bar{\tau} + \bar{\tau}^N D_i^N + \bar{\tau}^P D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \tag{5.D.7}$$

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

Finally, we analyze heterogeneous direct effects of ranks within pairs using equation 5.D.8:

$$\begin{aligned} y_{igs} = {} & \bar{\tau} + \bar{\tau}_h^N \mathbb{1}_{\{a_i \geq a_j\}} D_i^N + \bar{\tau}_l^N \mathbb{1}_{\{a_i < a_j\}} D_i^N \\ & + \bar{\tau}_h^P \mathbb{1}_{\{a_i \geq a_j\}} D_i^P + \bar{\tau}_l^P \mathbb{1}_{\{a_i < a_j\}} D_i^P + \beta\theta_i + \gamma X_i + \rho_s + \lambda_g + u_{igs} \end{aligned} \tag{5.D.8}$$

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

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

Appendix 5.E Control treatment to disentangle peer effects from learning

Table 5.E.1 and Figure 5.E.1 present the estimated average treatment effects and the margins including an additional control treatment. The NoPeer treatment featured the same design as all other treatments. The only difference was that students participated in the running task twice without a peer. Moreover, we shortened the survey for this treatment by removing the questionnaires on personal characteristics. The control treatment was conducted to show that the observed performance improvements are not due to learning. If learning drives our effects, we should observe performance improvements in NoPeer, which is not the case. Even if this control treatment had yielded performance improvements, this would not affect any of our results. To see this, note that we are interested in a between treatment comparison of performance improvements. Learning effects between the runs should therefore be constant across treatments.

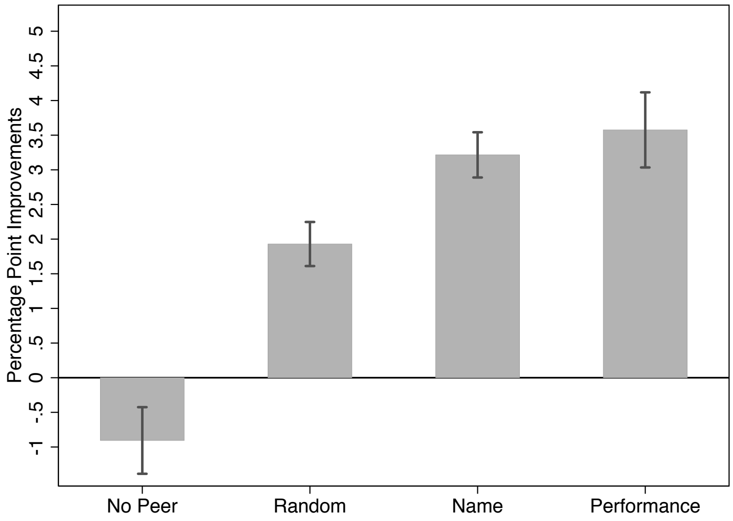


Figure 5.E.1. Average treatment effects

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 5.E.1 for the corresponding regression. We control for gender, grade and school fixed effects as well as age and cluster standard errors at the class level.

Table 5.E.1. Robustness checks

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

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

Appendix 5.F Robustness checks for average treatment effects

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

Table 5.F.1. Robustness checks

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

Notes: This table presents least squares regressions using times (Panel (a)) or percentage point improvements (Panel (b)) as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Column (1) presents the baseline specifications as used in Table 5.3. Columns (2) uses biased-reduced linearization to account for the limited number of clusters. Column (3) uses matching group-specific means as the unit of observation. Finally, columns (4) and (5) apply a 90% winsorization and truncation, respectively.

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 5.F.2 and by schools as well as gender in the lower panel and estimate the treatment effects separately for those samples. The table shows the robustness of the estimated treatment effects as these effects persists for all subsamples with similar magnitude.

Table 5.F.2. Robustness checks – Subsample analyses

	Percentage Point Improvements				
	(1)	(2)	(3)	(4)	(5)
	Baseline	7th grade	8th grade	9th grade	10th grade
Name	1.26***	1.95***	2.60***	1.53**	1.08*
	(0.43)	(0.08)	(0.35)	(0.59)	(0.61)
Performance	1.67**	2.78***	2.51***	2.53***	1.32
	(0.62)	(0.63)	(0.15)	(0.62)	(0.88)
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	588	116	116	174	182
R^2	.056	.073	.064	.16	.039
p-value: Name vs. Performance	.51	.21	.82	.19	.82
	(6)	(7)	(8)	(9)	(10)
	Female	Male	School 1	School 2	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. Performance	.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 5.3. Columns (2)-(5) restrict the sample to one grade, columns (6) and (7) to each gender and columns (8)-(10) to one school.

Appendix 5.G Peer composition robustness checks

We run three robustness checks for the results presented in Table 5.4. First, to provide further evidence that it is not the quality of the match itself that drives our results, we estimate the effect of match quality within Random (cf. Table 5.G.1). As subjects in Random are matched with someone they prefer by pure chance, this allows us to estimate the impact of match quality itself. The estimates show that match quality itself has no significant effect on the performance in Random. Second, in Table 5.G.2 we restrict our estimation sample to subjects with a high match quality only to show that the treatment effects persist for these subjects and the coefficients on peer compositional effects do not substantially change. Third, we control for differences in ability and matching quality in a more flexible way in Table 5.G.3 by including interval fixed effects for ability differences and fixed effects for every rank of the preferences. More specifically, we include an indicator for each one-second interval of ability differences between subjects within a pair. Similarly, we include indicators for each rank in the two sets of preferences to check whether the high match quality indicators are restrictive. This allows for a potential non-linear influence of ability differences and match quality on our estimates. Comparing the estimates shows that neither the piecewise-linear functional form of ability differences nor using high match quality indicators is restrictive. Finally, this table shows that the decomposition presented in Table 5.4 is robust to the inclusion of additional class-level controls.

Table 5.G.1. Effect of match quality within Random

	Percentage Point Improvements			
	(1) Only Name MQ.	(2) Only Perf. MQ.	(3) with Controls	(4) Baseline
Direct Effects				
Name				1.24**
				(0.50)
Performance				2.21***
				(0.68)
Peer Characteristics				
Faster Student × High match quality (Name)	1.00		0.89	0.52
	(0.85)		(0.95)	(0.43)
Slower Student × High match quality (Name)	-0.39		0.15	0.46
	(1.53)		(1.10)	(0.66)
Faster Student × High match quality (Perf.)		1.02	0.06	0.43
		(1.08)	(1.08)	(0.53)
Slower Student × High match quality (Perf.)		-1.87	-0.51	-0.71
		(1.42)	(1.22)	(0.66)
Faster Student × Peer is friend			0.10	-1.15**
			(0.74)	(0.53)
Slower Student × Peer is friend			0.01	0.13
			(1.15)	(0.67)
Faster Student × $\lvert \Delta Time\ \ 1 \rvert$			-0.54**	-0.35**
			(0.25)	(0.16)
Slower Student × $\lvert \Delta Time\ \ 1 \rvert$			0.73**	1.04***
			(0.32)	(0.20)
Slower Student in Pair	3.14***	3.72***	0.43	-0.15
	(0.43)	(0.74)	(1.15)	(0.68)
Abs. Diff. in Personality	No	No	Yes	Yes
Peer Characteristics	No	No	Yes	Yes
Own Characteristics	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes
N	205	205	204	582
R^2	.12	.13	.28	.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 characteristics include the Big Five, locus of control, social comparison, competitiveness and risk attitudes. Absolute differences in personality include the difference in those. We use only observations within Random. If we restrict the sample to students in Random, the explanatory power of the match quality (MQ) is not significant.

Table 5.G.2. Only high match quality sample as comparison group

	Percentage Point Improvements				
	(1) All	(2) RANDOM&NAME	(3) with Controls	(4) RANDOM&PERF.	(5) with Controls
Direct Effects					
NAME	1.24**	1.83***	1.93***		
	(0.50)	(0.55)	(0.47)		
PERFORMANCE	2.21***			2.38***	1.75**
	(0.68)			(0.71)	(0.64)
Peer Characteristics					
Faster Student × High match quality (NAME)	0.52				-0.47
	(0.43)				(1.28)
Slower Student × High match quality (NAME)	0.46				-0.56
	(0.66)				(1.15)
Faster Student × High match quality (PERF.)	0.43		-0.51		
	(0.53)		(0.65)		
Slower Student × High match quality (PERF.)	-0.71		-1.21		
	(0.66)		(0.86)		
Faster Student × Peer is friend	-1.15**		-1.53		-0.98
	(0.53)		(1.05)		(1.87)
Slower Student × Peer is friend	0.13		-1.18		-1.38
	(0.67)		(1.06)		(1.13)
Faster Student × $\|\Delta Time \;\; 1\|$	-0.35**		-0.72**		-0.07
	(0.16)		(0.29)		(0.51)
Slower Student × $\|\Delta Time \;\; 1\|$	1.04***		1.25***		1.08**
	(0.20)		(0.38)		(0.47)
Slower Student in Pair	-0.15		-0.44		-0.97
	(0.68)		(1.70)		(1.47)
Abs. Diff. in Personality	Yes	No	Yes	No	Yes
Peer Characteristics	Yes	No	Yes	No	Yes
Own Characteristics	Yes	Yes	Yes	Yes	Yes
Gender-Grade/School FEs, Age	Yes	Yes	Yes	Yes	Yes
N	582	208	207	162	160
R^2	.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 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 5.4 for reference. Columns (2) to (5) show that even if we restrict the comparison group to the sample of individuals in random that received a peer with high match quality according to name- (columns (3) and (4)) or performance-based preferences (columns (5) and (6)), respectively, our treatment effects persist and the coefficients on peer compositional effects do not change much.

Table 5.G.3. Robustness Check

	Percentage Point Improvements		
	(1) Linear	(2) Time Int. FE	(3) Class Controls
Direct Effects			
Name	1.24**	1.20**	1.46***
	(0.50)	(0.52)	(0.46)
Performance	2.21***	2.25***	1.73**
	(0.68)	(0.74)	(0.68)
Peer Characteristics			
Faster Student × High match quality (Name)	0.52	0.71	0.69
	(0.43)	(0.44)	(0.45)
Slower Student × High match quality (Name)	0.46	0.27	0.62
	(0.66)	(0.65)	(0.74)
Faster Student × High match quality (Perf.)	0.43	0.41	0.12
	(0.53)	(0.49)	(0.59)
Slower Student × High match quality (Perf.)	-0.71	-0.72	-1.15
	(0.66)	(0.58)	(0.73)
Faster Student × Peer is friend	-1.15**	-1.08**	-1.03**
	(0.53)	(0.51)	(0.47)
Slower Student × Peer is friend	0.13	0.07	0.45
	(0.67)	(0.73)	(0.79)
Faster Student × $\|\Delta Time\ \ 1\|$	-0.35**		-0.36**
	(0.16)		(0.16)
Slower Student × $\|\Delta Time\ \ 1\|$	1.04***		0.84***
	(0.20)		(0.19)
Slower Student in Pair	-0.15		0.11
	(0.68)		(0.76)
Time Diff. FEs	No	Yes	No
Class-level Controls	No	No	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	512
R^2	.29	.3	.29
p-value: Name vs. Performance	.17	.14	.72

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

Appendix 5.H Additional material for implications

Table 5.H.1 shows the regressions underlying Figure 5.6. In particular, in column (1) we estimate equation 5.1 but interact treatment indicators with ability terciles (low, medium, or high). Ability terciles are defined according to tercile splits of times in the first run within each school, grade and gender. Column (2) repeats the exercise using quintiles rather than terciles to show that the pattern holds for finer splits.

As argued in section 5.5.5, low-ability students in Name need to prefer to be and subsequently are matched with faster students on average. We present the shares of students in Name who prefer a faster student (based on their name-based preferences) and who are matched to a faster student for the three ability terciles defined above in Table 5.H.2. Indeed, low-ability students in Name are more likely to prefer a faster peer and on average are matched to faster peers than students of higher ability.

Our treatments also have implications for individual ranks of students within a class since slower students improve more than faster ones. As ranks are important in determining subsequent outcomes (Elsner and Isphording, 2017; Gill et al., 2017), a policy maker has to take the distributional effects of peer assignment mechanisms into account.[1] Since low-ability students improve relatively more than high-ability students in Name and Random, these treatments yield potentially large changes of a student's rank within the class between the two runs. By contrast, Performance will tend to preserve the ranking of the first run as improvements are distributed more equally relative to the two other treatments. We confirm this intuition in Table 5.H.3, where we regress the absolute change in percentile scores from the first to the second run on treatment indicators. The outcome variable measures the average perturbation of ranks within in a class across the two runs. The results show that Performance shuffles the ranks of students less in comparison to Random and Name. While in Random students change their position by about 15 out of 100 ranks, we find significantly less changes in the percentile score in Performance relative to Random. This change corresponds to a 27% reduction in reshuffling. However, in Name we do not find any effect compared to Random.

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

Table 5.H.1. Heterogeneous treatment effects by own ability

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

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) assigns one of three ability levels – low, medium or high – according to tercile splits of times in the first run within each school, grade and gender and presents the underlying regression for Figure 5.6. Column (2) uses quintiles rather than terciles to show that the pattern is robust to other definitions of ability quantiles.

Table 5.H.2. Share of students preferring and receiving a faster peer in Name

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

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

Table 5.H.3. Absolute change in percentile scores

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

Notes: This table presents least squares regressions using absolute change in percentile scores as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered at the class level. Absolute changes in percentile scores within matching groups are calculated based on the change of individual ranks of students in the their class and gender from the first to the second. Percentile scores within treatment are calculated for all students within the same treatment and gender (i.e., across classrooms).

Appendix 5.I Simulation of matching rules

We simulate three matching rules and predict their impact on performance improvements using our estimates from Table 5.4. In a first step, we create artificial pairs, based on the employed matching rules described below. In a second step, we then calculate the vector θ of differences for the artificial pairs as well as the matching quality of artificial peers. Finally, we use the estimated coefficients from the column (6) of Table 5.4 to predict the performance improvements we would observe for the artificial pairs. As peer-assignment rules only change θ, we are interested in the difference in the respective sums of the indirect effect and direct effect, that is between $\bar{\tau} + \beta\theta_i^{sim}$ and $\bar{\tau} + \beta\theta_i^{obs}$ from equation 5.2, where sim and obs denote simulated and observed pair characteristics, respectively. Furthermore, we assume that the direct effect of the simulated policies equals the one in Random. We additionally fix the covariates X to 0 and leave out the fixed effects for the simulations and predictions. This means, we calculate the performance improvements for a particular baseline group for our treatments as well as the simulations. This enables us to compare our results of the simulations directly to the peer-assignment rules using self-selection implemented in the experiment, as we compare the performance improvements for the same group.

We simulate the following three peer assignment rules. First, we implement an ability tracking assignment rule, Tracking, in the spirit of the matching also employed in Gneezy and Rustichini (2004). Students are matched in pairs, starting with the two fastest students in a matching group and moving down the ranking subsequently. This rule minimizes the absolute distance in pairs. Second, we employ a peer assignment rule that fixes the distance in ranks for all pairs (Equidistance). We rank all students in a matching group and match the first student with the one in the middle and so forth. More specifically, if G denotes the group size, the distance in ranks is $G/2 - 1$ for all pairs. This rule is one way to maximize the sum of absolute differences in pairs, but keeps the distance across pairs similarly. Third, we match the highest ranked student with the lowest one, the second highest ranked with the second lowest one and so forth (High-to-Low). This is similar to Carrell et al. (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 5.I.1 summarizes the distance in ability of the experimental treatments as well as the simulated assignment rules.

Table 5.I.1. Overview of simulated peer assignment rules

Peer Assignment Rule	Simulated?	Mean Ability Distance (in sec)	Description
NAME	No	2.09	Self-selected peers based on names
PERFORMANCE	No	1.41	Self-selected peers based on relative performance
RANDOM	No	2.42	Randomly assigned peers
EQUIDISTANCE	Yes	3.11	Same distance in ranks across pairs
HIGH-TO-LOW	Yes	3.11	First to last, second to second to last etc.
TRACKING	Yes	0.90	First to second, third to fourth etc.

6

To whom you may compare - Preferences for peers[⋆]

6.1 Introduction

In many situations, individuals compare their own performance, status, or outcome to their peers impacting their behavior. These social comparisons impact job satisfaction (Card et al., 2012), performance (Ashraf et al., 2014; Cohn et al., 2014) or consumption (Kuhn et al., 2011), individual happiness as well as overall well-being (A. Clark and Senik, 2010), and there is evidence that firms take these tendencies to engage in comparisons into account when designing incentive schemes (Frank, 1984; Nickerson and Zenger, 2008). But these people to whom one compares one's own performance and income are not randomly chosen. Rather, people select their peers who in turn serve as their social reference points. For instance, a student might avoid her very successful classmate, because seeing her makes her feel bad about herself; nonetheless, she seeks out for her sporty friend to motivate her for a marathon.

This selection of social reference points can differ across environments, due to varying available information as well as changing relevant selection criteria and motives. In some cases, individuals have lots of information about others and potentially can select peers based on particularly important personality characteristics. In such situations individuals may be able to select friends or consciously avoid them. In other situations, information might be limited and preferences can only be conditioned on very specific characteristics such as past performance. This latter situation might be important for people preferring workplaces with high ability peers or when selecting into specific schools to be "the

[⋆] We thank Lorenz Götte, Sebastian Kube, Pia Pinger for their guidance and support. We also thank Viola Ackfeld, Philipp Albert, Thomas Dohmen, Ingo Isphording, Ulf Zölitz and audiences at Bonn, IZA Brown Bag and the CRCTR-224 Workshop for Young Researchers for helpful feedback and comments. We also thank the schools and students that participated in the experiments.

big fish in the small pond".[1] This highlights that individuals have preferences for specific peer environments and even within a given setting or group only some peers might serve as social reference points. At the same time, there are several motives that can rationalize different peer preferences: in many environments, high ability peers have a positive impact on individual performance. However, individuals may simultaneously suffer from disadvantageous social comparisons. Lastly, some individuals might receive special pleasure simply from interacting with peers they know well – their friends. This suggests that peer choice processes might be more complex than previously assumed.

In this paper, we study preferences for peers for a strenuous task. We develop a theoretical framework that combines social reference points and time inconsistent preferences to derive predictions for the selection of specific peers. Given this framework, we describe preference for peers based on two dimensions: First, based on limited information corresponding to relative past performance, and second, based on the full set of personal characteristics (e.g., friendship ties, personality, network characteristics, etc.). After documenting large heterogeneities in these preferences, we analyze the determinants of those preferences. We demonstrate how these preferences depend on a person's personality and that the students in our sample exhibit homophily in their peer choice, i.e., that individuals choose others that are similar to them in several dimensions. Lastly, we explore the relationship between the two different kinds of preferences and study whether these are two measures of the same underlying preference or whether preferences for peers are multidimensional in nature.

In order to study the selection of peers and the corresponding preferences, we utilize the dataset of a framed field experiment with over 600 students (aged 12 to 16) in physical education classes of German secondary schools. Students took part in two running tasks – first alone, then simultaneously with a peer. In between the two runs, students filled out a survey that asked them with whom they would like to run in that second run, i.e., their preferences for peers, and elicited their personal characteristics as well as the social network within each class. The elicited preferences were used to form pairs in the second run, where the peer assignment rule was varied across classes; in some classes, students were randomly assigned to pairs, in others we allowed for self-selected peers implementing two notions of self-selection of peers. More specifically, students could self-select their peer either based on the relative performance or the identity of their classmates. Using this setup has three crucial features for the questions at hand: First, using a running task yields direct measures of performance and thus could be used to select peers based on their relative performance in the first run (*performance-based preferences*), where students indicated their pre-

[1] This is important in models where individuals have preferences over their ordinal ranks in their peer group (Cicala et al., forthcoming; Tincani, 2017).

ferred relative performance in the ±5 seconds range. Second, the classroom environment enabled students to state preferences for known peers (*name-based preferences*). Third, focusing on a single peer in the second run, we circumvent issues associated with multiple reference points (Kahneman, 1992). We analyze the effects of these peer assignment mechanisms in depth in Kiessling et al. (2018).

For performance-based preferences, we find that students prefer on average slightly faster peers (0.56 sec or 0.20 SD in terms of times in the first run). Yet, this masks large heterogeneities in these preferences. In fact, half of the students prefer peers differing by more than one second, both, faster and slower. Analyzing the relationship between students' personality and their preferences, we document strong effects of three characteristics: more competitive students and those with a higher internal locus of control tend to prefer faster peers, while extraversive individuals select relatively slower peers. For name-based preferences, we show that friendship ties play a crucial role to understand peer preferences based on names as 80% of the three most preferred name-based peers are also friends. We document the importance of other characteristics as well: students prefer peers that are similar in ability, score similar on measures of agreeableness as well as attitudes towards social comparisons, and have a similar influence in the social network as themselves. Given the finding that students prefer similar strong peers in both preference measures begs the question whether these two are different measures of the same underlying preference. When analyzing the relationship between the preferences, we find that while performance-based preferences are important to understand name-based preferences, other homophily dimensions are also important, highlighting the multidimensionality of preferences for peers.

The results in this paper relate to several recent strands of the literature on social comparisons and (social) reference points.

In our experiment the students may expect an additional motivation through a faster peer. This motive has been modeled theoretically through endogenous reference points or goals as in Koch and Nafziger (2011). They consider goal-setting as a way for time-inconsistent individuals to commit their future self to exert higher amounts of effort. In contrast to our framework, they consider solely an individual's choice, but no further interaction between peers. In a similar spirit, Falk and Knell (2004) build a model of endogenous reference points. Here an individual can partially determine her reference points. They assume that higher reference points decrease effort costs, providing an incentive to actively set high goals. This mechanism likely affects individual performance as in Allen et al. (2017), who find that marathon-performances tend to bunch around round times that were likely aimed at previously. In a university setting, D. Clark et al. (2017) report a positive effect of self-set goals on effort and performance by mitigating self-control problems. Relatedly, Brookins et al. (2017) add to the growing literature on goal setting (e.g., Heath et al., 1999; Locke and Latham, 2002; Wu et al., 2008; Hsiaw, 2013; Goerg, 2015; Corgnet et al., 2015) and find that self-chosen goals might improve performance even in the presence of monetary incentives.

More generally, our study adds to the literature on specifying reference points in models with reference-dependence preferences (e.g., Kőszegi and Rabin, 2006, 2007; Abeler et al., 2011). While the existing literature focuses on predominantly on static reference points, we introduce peers as social reference points and study how these reference points are selected. Introducing this social dimensions into reference points – and thereby specific characteristics of peers – might have profound consequences on how individuals perceive a situation. Performing a task jointly and comparing the outcome might induce a tournament-like feeling amongst peers even when no explicit tournament incentives are present. The consequences of these competitive environments for individual behavior have sparked a broad literature over the last decade. Most prominently, Niederle and Vesterlund (2007) find that females enter tournaments less frequently than males and others have examined how to overcome this gender gap (e.g., Healy and Pate, 2011). Additionally, the willingness to compete in these settings has been identified as a potential key component in different career choices and life outcomes such as schooling decisions and wages (Buser et al., 2014; Almås et al., 2015). In contrast to that literature, we highlight a different aspect of competitiveness. We fix the tournament entry decision often used as a measure of competitiveness in the experimental design and allow students to select a peer, which consequently affects the chances of being faster or slower in the second run. Hence, we allow students to select their competitors and not to decide on their participation in the competition.[2] This relates to Niederle and Yestrumskas (2008), who analyze preferences for challenging tasks for given ability levels rather than preferences for peers of higher or lower ability.

Social psychologists and sociologists have studied social comparisons at least since Festinger (1954). In general, they find that on average, there is a slight upwards tendency in social comparisons (e.g., Blanton et al., 1999; Huguet et al., 2001); i.e., individuals tend to prefer somewhat better performing peers. These findings are in line with our results. However, similar to the economic literature on goal setting, these studies focus on peer preferences solely based on relative performance and neglect other potential characteristics of peers. The direction of social comparisons has been explored predominantly in the income dimension. A. Clark and Senik (2010) find that employees tend to compare their income with that of their colleagues and to a lesser degree with that of

[2] Although by choosing much faster or much slower peers, subjects could basically remove any competitive element from the task.

their friends. Moreover, comparing with a person that has a higher income is associated with lower levels of satisfaction. Kuegler (2009) complements these findings using siblings as reference points. In contrast to comparisons in the income dimension, comparison points in a performance dimension might not necessarily be selected for the purpose of status creation or self-assurance (for these motives see, e.g., Wills, 1981; Markman and McMullen, 2003). Due to the positive effect of the social interactions this could create incentives to spend additional effort to pursue higher performance levels (Collins, 1996).

Since students in our setting are able to select their peers from their social network, we also relate to the research on the determinants of link formations in networks (for a recent overview of the literature on network formation see Jackson et al., 2017). Similar to one strand of the literature, we focus on empirical correlates of link formation and dimensions of homophily (see, e.g., Girard et al., 2015; Lewis et al., 2012; Marmaros and Sacerdote, 2006; Mayer and Puller, 2008, for the role of personality, homophily in the taste for music and movies, geographic proximity, and race, respectively, in the formation of friendship ties). Yet, our study takes the social network as given and analyzes a nomination process on this social network. Our results highlight that for a given task individuals may choose specific persons that are or are not part of their social network as peers.[3] Related to this literature, but more similar to our paper is the study by Cicala et al. (forthcoming), which presents a Roy model where agents self-select into different peer groups based on their comparative advantage within a given environment.

Finally, this paper relates to the extensive literature on peer effects, by asking the question who the relevant peers actually are. While peer effects in performance have been studied extensively (for an overview see Herbst and Mas, 2015), there is growing evidence that different kind of peers differentially affect performance. For instance, individuals may change their effort provision once they are working with friend, even forgoing additional earnings (Bandiera et al., 2010). This is in line with with our results that many students are preferring their friends as peers. Likewise peers affect behavior differentially in competitive situations (e.g., Gneezy et al., 2003; Gneezy and Rustichini, 2004). More importantly, most studies take the reference group or peer group as given, or use exogenous re-assignment policies to sort subjects into peer groups (e.g., Carrell et al., 2013). The design of optimal policies though hinges on a good understanding of the endogenous group formation process within those social groups. Our study closes this gap by describing, whom students prefer and therefore choose as their peer within a group.

[3] Thus, one can think of our setting as an additional network formation process on top of an existing network.

While the importance of peers for educational attainment, consumption, or performance on the job is undisputed, evidence on to whom people compare their performance and how these peers are selected remains scarce. Yet, already Manski (1993, p. 536) noted that "*informed specification of reference groups is a necessary prelude to [the] analysis of social effects*". Therefore, we take a first step by describing and analyzing the selection of peers across two distinct dimensions within one setting. This allows us to provide novel evidence on the determinants of the underlying preferences. By studying them in a single setting, we can shed light on their relationship and highlight the multidimensionality of preferences for peers.

Notably, individual personality has been shown to correlate with labor market outcomes (e.g., Groves, 2005; Heckman et al., 2006; Almlund et al., 2011). Some of the characteristics that were found to be predictive for educational attainment and labor market success (see, e.g., Buser et al., 2014; Piatek and Pinger, 2016, for evidence on competitiveness and locus on control, respectively) also correlate with individuals selecting into peer groups of higher ability. As these environments in turn have positive impacts on individual performance, the selection of specific peer environments could potentially serve as an important link between personality on the one side and schooling decisions as well as labor market outcomes on the other.

The preferences for peers analyzed in this paper and their link to personal characteristics might be specific to situations in which only own performance matters and that have competitive components. Other peers might be selected in cooperative settings or situations where the positive impact of high ability peer becomes more apparent and some other personality traits may be important. Yet, it is reassuring that we document the importance for the peer selection process of those personality traits that are also predictive for labor market success in general. As this paper demonstrates, individuals are highly heterogeneous with respect to whom they select as social reference points or relevant peers. Policies trying to leverage the influence of peers to boost educational attainment or job performance therefore might be well advised to take this heterogeneity into account.[4]

The remainder of the paper is structured as follows. The next section (6.2) presents a theoretical framework of effort provision and peer choice. We then present the data and describe our sample in section 6.3. Section 6.4 describes two kinds of preferences for peers – based on relative performance and based on names – and analyzes the determinants of these preferences. We analyze the relationship of these preferences in section 6.5. Finally, section 6.6 concludes.

[4] The peer preferences and associated heterogeneities might help to explain why reassigning students into different classrooms as in Carrell et al. (2013) did not have the intended effects of increasing the GPA of low ability students.

6.2 Theoretical Framework

In this section, we present a simple model of peer choice and effort provision to broadly structure the following analyses. We model an individual's problem as a two stage process closely following our experimental design: On the first stage, an individual decides over her peer choice, while on the second stage the resulting pairs then perform a strenuous effort task simultaneously and compare their performances. Thus, the peer's performance will serve akin to a goal or (social) reference point. Given this similarity, the theoretical framework presented here corresponds to the models of Koch and Nafziger (2011) and Falk and Knell (2004).[5] Yet, we extent their reasoning to a setting with interaction between individuals and their peers.[6]

We assume that each individual has some intrinsic motivation p to provide effort e_i.[7] We allow this motivation to differ for the peer choice (p_{PC}) and the effort provision (p_{EP}). This is similar to dual-self models of self-control (Fudenberg and Levine, 2006; Banerjee and Mullainathan, 2010) or to models where agents are time inconsistent choices due to different discount rates (Laibson, 1997). If the individual provides effort, she incurs costs c, where we assume for simplicity that these costs are given by $c(e_i) = \theta_i e_i^2/2$. Here $1/\theta_i$ denotes the individual's ability at the running task; higher ability therefore is associated with lower marginal costs. Moreover, we assume that peers affect an individual's utility as follows: whenever an individual runs with another person, she will compare her own performance with her peer's. Being slower induces a dis-utility, while being ahead affects utility positively. We allow this part of the utility function to depend on a peer's ability to capture the idea that being better than a high performing peer should matter relatively more than being ahead of a low performing peer. We introduce these social reference points by including an additional utility term $\alpha e_j\left(e_i - e_j\right)$, where α denotes the weight of this peer effect. Additionally, we assume that individuals are loss averse if they fall behind, which we capture by a loss aversion parameter $\lambda > 1$ that multiplies the preceding peer effect. The individual's overall utility is then given by

$$u_i^t(e_i, e_j|\theta_i) = \begin{cases} p_t e_i - \theta_i \frac{e_i^2}{2} + \alpha e_j\left(e_i - e_j\right) & \text{if } e_i \geq e_j \\ p_t e_i - \theta_i \frac{e_i^2}{2} + \alpha\lambda e_j\left(e_i - e_j\right) & \text{if } e_i < e_j \end{cases} \tag{6.1}$$

[5] Koch and Nafziger (2011) analyze the goal setting behavior of a time-inconsistent agent with $\beta\delta$-preferences. The agent might be inclined to set higher goals to motivate her future self. In our model, individuals choose peers and these individuals interact with each other in the effort provision stage.

[6] Thus, one could think of peers corresponding to reference points that adjust given the effort individuals provide.

[7] We implicitly assume that the production function is linear in effort.

for $t = EP, PC$. In the following, we first solve for the optimal effort levels when a peer is present or when running alone. In a second step, we then characterize the optimal peer choice and highlight comparative statics.

6.2.1 Optimal effort

First, consider the situation where individuals run alone, that is without a peer being present. Then $p_{EP} = c'\left(e^*_{i,NoPeer}\right)$ and thus $e^*_{i,NoPeer} = {}^{p_{EP}}/_{\theta_i}$, that is those individuals with a higher intrinsic motivation p_{EP} or higher ability ${}^{1}/_{\theta_i}$ provide more effort and thus perform better.

Now consider the situation where a peer with ability ${}^{1}/_{\theta_j}$ is doing the task simultaneously, where – without loss of generality – we assume ${}^{1}/_{\theta_j} < {}^{1}/_{\theta_i}$, i.e., j has a lower ability than i. Then, i's reaction function is given by

$$e_i\left(e_j|\theta_i\right) = \frac{p_{EP}}{\theta_i} + \frac{\alpha}{\theta_i}e_j, \tag{6.2}$$

while the corresponding function for j is

$$e_j\left(e_i|\theta_i\right) = \frac{p_{EP}}{\theta_j} + \frac{\alpha\lambda}{\theta_j}e_i \tag{6.3}$$

as long as $e_i \geq \frac{p_{EP}}{\theta_j - \alpha\lambda}$. Otherwise the two will select the same effort level. Note that due to the presence of loss aversion, the peer effect (${}^{\alpha\lambda}/_{\theta_j}$) is more pronounced for j, the slower of the two, than for i (${}^{\alpha}/_{\theta_i}$) as long as the ability level is not too different, i.e., $\lambda > {}^{\theta_j}/_{\theta_i}$. The optimal effort provision by both individuals is therefore given by:

$$e^*_i\left(\theta_i, \theta_j\right) = p_{EP}\frac{1 + \frac{\alpha}{\theta_j}}{\theta_i - \frac{\alpha^2}{\theta_j}} \tag{6.4}$$

$$e^*_j\left(\theta_j, \theta_i\right) = p_{EP}\frac{1 + \frac{\alpha\lambda}{\theta_i}}{\theta_j - \frac{(\alpha\lambda)^2}{\theta_i}} \tag{6.5}$$

Hence, effort increases in both intrinsic motivation (e_{EP}) and own ability (${}^{1}/_{\theta_i}$) as in the case of no peer being present. Moreover, effort provision with a peer present increases in the ability of the peer; the higher the ability of a peer, the more effort an individual provides, and this peer effect is amplified by loss aversion if the peer has a higher ability.[8]

[8] These effort levels hold as long as the ability difference between the two individuals is not too small, that is $\theta_j - \theta_i \geq \alpha(\lambda - 1) + {}^{\alpha^2}/_{\theta_j} - {}^{\alpha^2\lambda^2}/_{\theta_i} + {}^{\alpha^3\lambda}/_{\theta_i\theta_j}(1 - \lambda)$. Otherwise they will exert the same effort: $e^*_{i,j} = p_{EP}\left(1 + {}^{\alpha}/_{\theta_j}\right)/\left(\theta_i - {}^{\alpha^2}/_{\theta_j}\right)$.

6.2.2 Optimal Peer Choice

We now consider an individual's optimal peer choice. Given our experimental design, we model this stage as choosing an effort level in the first run, and assume that an individual i maximizes her utility function 6.1 given the reaction function 6.2 over the peer's effort. This means that the selection ignores the own effect on the peer's effort choice and treats it as fixed implying that students are naive about their own influence on their peer. The optimal peer choice of i then yields $e^*_{j,PC} = \frac{p_{PC}}{2\theta_i - \alpha}$ if $e_i > e^*_{j,PC}$ and $e^*_{j,PC} = \frac{p_{PC}}{2\theta_i - \alpha\lambda}$ otherwise, or in terms of ability:

$$\frac{1}{\theta^*_{j,PC}} = \begin{cases} \frac{p_{PC}/p_{EP}}{2\theta_i - \alpha} & \text{if } \frac{1}{\theta_i} \geq \frac{1}{\theta^*_{i,PC}} \\ \frac{p_{PC}/p_{EP}}{2\theta_i - \alpha\lambda} & \text{if } \frac{1}{\theta_i} < \frac{1}{\theta^*_{i,PC}} \end{cases} \tag{6.6}$$

Given this optimal peer choice and the bounds, we can conclude that

$$\frac{1}{\theta^*_{j,PC}} = \begin{cases} \frac{p_{PC}/p_{EP}}{2\theta_i - \alpha} & \text{if} \quad \frac{1}{\theta_i} < \frac{2 - p_{PC}/p_{EP}}{\alpha(1 - p_{PC}/p_{EP})} \\ \frac{1}{\theta_i} & \text{if} \quad \frac{2 - p_{PC}/p_{EP}}{\alpha(1 - p_{PC}/p_{EP})} \leq \frac{1}{\theta_i} \leq \frac{2 - p_{PC}/p_{EP}}{\alpha\lambda(1 - p_{PC}/p_{EP})} \\ \frac{p_{PC}/p_{EP}}{2\theta_i - \alpha\lambda} & \text{if} \quad \frac{2 - p_{PC}/p_{EP}}{\alpha\lambda(1 - p_{PC}/p_{EP})} < \frac{1}{\theta_i}. \end{cases} \tag{6.7}$$

The expressions in equations 6.6 and 6.7 show that individuals with a sufficient low ability choose even slower individuals as peers, and high ability individuals choose peers with an even higher ability. Moreover, between those two extremes there is a range of individuals who prefer peers that have a similar ability to them. Additionally, the peer's preferred ability level depends on the strength of the comparison parameter α and the ratio p_{PC}/p_{EP} (i.e., the ratio of motivation during the peer choice and the running task).[9] In our empirical analysis, we therefore investigate how personality characteristics and preferences affect the peer choice.

The theoretical framework presented here mainly applies to the selection of peers based on relative performance. If that would be the only motive determining peer choice, we would observe that the preferences over relative performance would also be a main determinant for selection based on names. While we will show in section 6.5 that this is indeed the case, other factors – most prominently friendships – play an important role, as well. Nonetheless small extensions to our framework can rationalize such deviations from the preference over relative performance. Individuals might receive an additional flat utility

[9] This ratio can be interpreted as a proxy for a self-control problem. If individuals are aware of their self-control problem and thus know that they will exert only low effort during the second run, they can choose a faster peer knowing that she can serve as a commitment to exert more effort.

from running with their friends, leading the long-term self to trade off the desired effort level and utility from interaction with friends.

6.3 Data

In most environments it is difficult to observe to whom people want to compare their own performance. This is especially difficult when there is not a single peer available as objective standard but rather several peers are observed at the same time. Additionally, while most models such as the one presented in section 6.2 assume peer selection takes only place based on preferences over relative performance, any observed selection of peers is potentially based on a much broader set of these peers' characteristics. Hence, it is even harder to determine preferences for relative performance comparisons based on selected peers using observational data.

In this paper, we use the dataset of a framed field experiment presented in Kiessling et al. (2018) to analyze multiple dimensions of preferences for peers. The field experiment featured three treatments, which allowed for the self-selection of peers and studied the impact of self-selection on performance. As the experiment allowed for (controlled) self-selection of peers, it elicited preferences for peers for a population of over 600 students in Germany. Additionally, the experiment obtained the social network and several personal characteristics, which we use in this study to analyze preferences for peers. In the following, we describe in detail how preferences for peers were elicited. We refer the reader to Kiessling et al. (2018) for a detailed description of the experiment itself.

6.3.1 Experiment

The experiment was embedded into physical education classes in German secondary schools. Subjects participated in two suicide runs, each consisting of a series of short sprints along the lines of a volleyball court[10]: First, at the beginning of the experiment alone; then at the end of the experiment simultaneously with a peer, where the treatments implemented different peer-assignment rules (random assignment, self-selection based on names, or self-selection based on relative performance). Between the two runs, subjects participated in a survey. In addition to socio-demographics, the survey asked students to reveal their preferences for peers according to two dimensions and obtained several personal

[10] 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 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.

characteristics as well the social network of the class. In the following, we describe each of these survey elements in more detail.

6.3.2 Preference elicitation

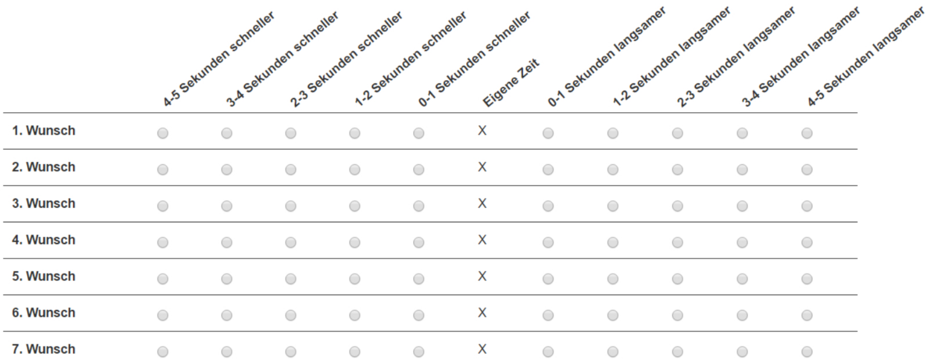
The core element of this paper is to describe who people select as peers. The survey elicited peer preferences for two situations, which were used to implement self-selected peers in the experiment. First, we elicited preferences for situations solely based on relative performance (*performance-based preferences*). Second, we asked for preferences for those settings where social information is available (*name-based preferences*). These preferences were elicited independent of the treatment, as the treatment was only assigned after the survey took place. Note that these preferences are revealed rather than stated preferences as there was a positive probability that these preferences were taken into account when forming pairs due to the random assignment of treatments to classes.

We first discuss the elicitation of preferences for peers based on relative performance. For this purpose, the survey presented subjects ten categories consisting of one-second intervals starting from $(4,5]$ seconds slower than their own performance in the first run, to $(0,1]$ seconds slower and $(0,1]$ seconds faster up to $(4,5]$ seconds faster. Subjects indicated from which time 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, i.e., the most-preferred peer's relative performance. In the second row, they indicated their second most preferred interval; and so forth. We present a screenshot of the elicitation procedure in Figure 6.1. We asked students to rank their seven most preferred time intervals and therefore generated a partial ranking of potential peers for performance-based preferences. Naturally, each time interval could only be chosen once, but included potentially several peers. Similarly, some intervals might have been empty.[11]

The second preference measure elicited preferences for situations, where 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, i.e., those people with whom they would like to be paired in the second run.[12] They could select any person of the same gender, irrespective of this person's actual participation in the study or their attendance

[11] Since we elicit preferences over the relative performance of peers and not whether these preferences can be satisfied, having multiple potential peers is no concern.

[12] In the experiment, pairs were only formed within gender. Hence, preferences were also restricted to peers from the same gender.



	4-5 Sekunden schneller	3-4 Sekunden schneller	2-3 Sekunden schneller	1-2 Sekunden schneller	0-1 Sekunden schneller	Eigene Zeit	0-1 Sekunden langsamer	1-2 Sekunden langsamer	2-3 Sekunden langsamer	3-4 Sekunden langsamer	4-5 Sekunden langsamer
1. Wunsch	○	○	○	○	○	X	○	○	○	○	○
2. Wunsch	○	○	○	○	○	X	○	○	○	○	○
3. Wunsch	○	○	○	○	○	X	○	○	○	○	○
4. Wunsch	○	○	○	○	○	X	○	○	○	○	○
5. Wunsch	○	○	○	○	○	X	○	○	○	○	○
6. Wunsch	○	○	○	○	○	X	○	○	○	○	○
7. Wunsch	○	○	○	○	○	X	○	○	○	○	○

Figure 6.1. 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 faster to 4 to 5 seconds slower.

in class.[13] These classmates had to be ranked, creating a partial ranking of their potential 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 the one of the performance-based preferences described above: subjects could indicate their beliefs about the performance of the potential peer in the first run using the same ten categories in the same layout.

6.3.3 Personal characteristics and social network

After the preference elicitation, two further survey elements asked for personal characteristics and the social network of the class. First, the survey 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[14], general risk attitude (Dohmen et al., 2011), and a short version of the INCOM scale for social comparison (Gibbons and Buunk, 1999; Schneider and Schupp, 2011). For each of multiple item characteristics, we use a factor

[13] 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.

[14] Rather than using tournament entry decisions as measures of competitiveness, we introduced a continuous measure based on a subject'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.

analysis to retain one underlying principle components factor with mean zero and standard deviation of one.

Second, we obtained the social network of the class: we asked every student to name up to six friends in their class. Since the maximum number of friends in the survey is six, we focus on undirected links. Therefore, 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 than more than six friends if they were nominated by participants that they did not nominate themselves.[15]

6.3.4 Summary statistics

We present summary statistics of our sample in Table 6.1. Overall, we have preference measures and the social network for 745 individuals from 48 classes of grades 7 to 10 (aged 12 to 16) with 65% of students being female.[16] This amounts to 73% of all students in a class participating in the experiment.[17] The average class size is 25.44 and students have 6.81 friends on average with 78% of those friends being from a student's own gender. Turning to the performance of individuals in the first run without a peer being present, we observe that on average females took about 27.50 seconds to finish the running task, which does not vary by age. Males, in contrast, improve their performance with age: while the average performance of males in grade 7 is 25.29 seconds, it improves to 23.07 seconds in grade 10.

6.4 Determinants of preferences for peers

In this section, we analyze subjects' preferences for peers. These preferences correspond to two dimensions along which peer selection can occur more generally. Individuals may have limited information such as the relative performance about their potential peers. Alternatively, they may know their peer group well and therefore can condition their peer choice on many characteristics of potential peers. The preferences elicited in our survey correspond to these two cases: first, we asked subjects what (relative) performance their peer should have (*performance-based preferences*), corresponding to the former dimension;

[15] About 79% of the participants nominated six friends. Thus, we were worried that a maximum of six friends might be restrictive and accordingly define friendships as undirected rather than directed links.

[16] These classes are from three Germany secondary schools from the highest track preparing students for university entry after grade 12 (*Gymnasien*).

[17] Only those students who handed in 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.

Table 6.1. Summary statistics

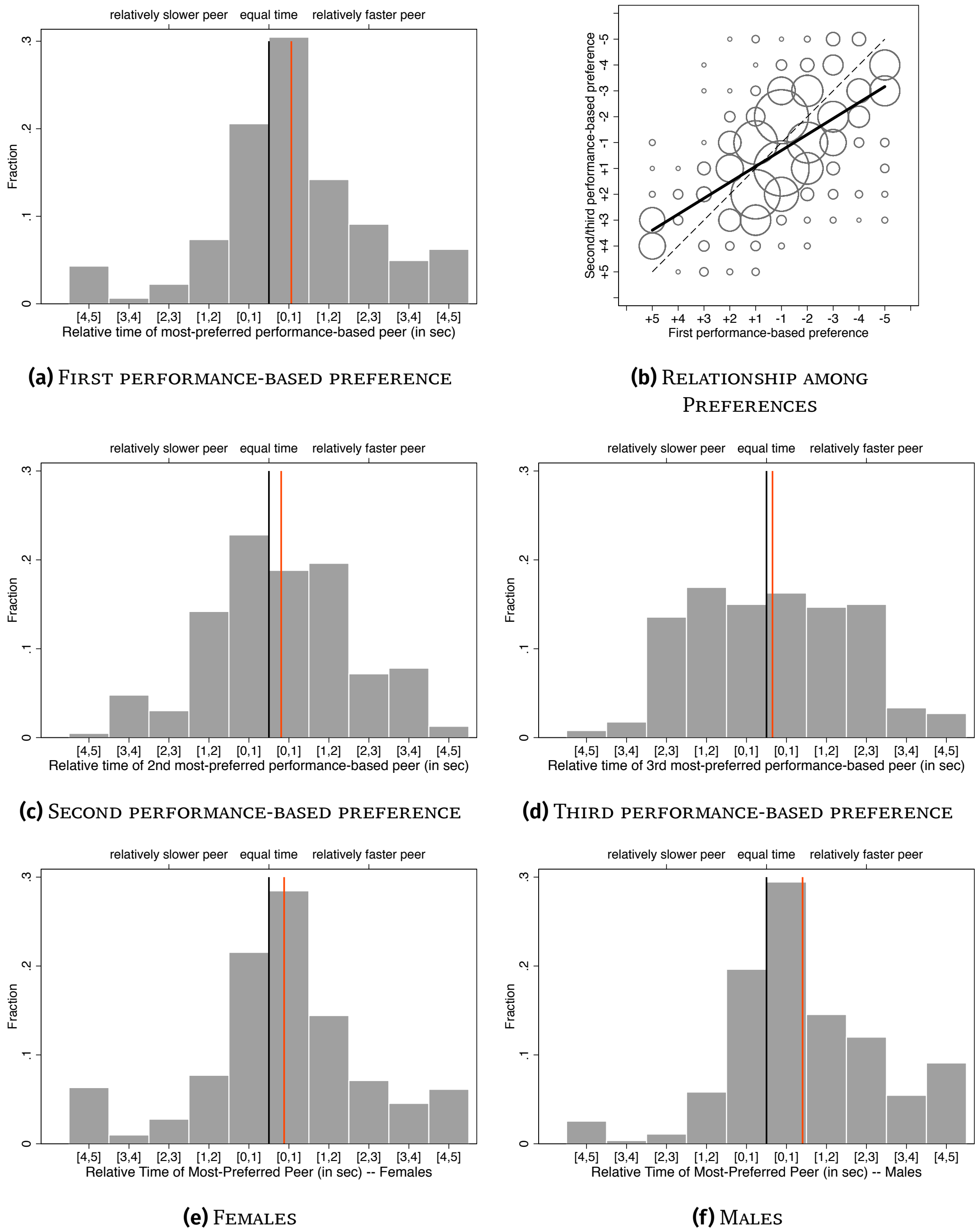
	7th grade	8th grade	9th grade	10th grade	Total
Socio-Demographic Variables					
Age	12.76	13.76	14.77	15.84	14.39
	(0.44)	(0.45)	(0.39)	(0.53)	(1.24)
Female	0.59	0.63	0.67	0.70	0.65
	(0.49)	(0.48)	(0.47)	(0.46)	(0.48)
Number of friends	6.90	6.99	6.94	6.49	6.81
	(1.34)	(1.61)	(1.60)	(1.67)	(1.58)
Share of friends of own gender	0.82	0.73	0.84	0.75	0.78
	(0.21)	(0.24)	(0.20)	(0.26)	(0.23)
Times (in sec)					
Time 1 (Females)	27.71	27.13	27.35	27.78	27.50
	(2.65)	(1.98)	(2.25)	(2.72)	(2.44)
Time 1 (Males)	25.29	24.54	23.60	23.07	24.15
	(2.02)	(2.50)	(1.79)	(2.01)	(2.26)
Class-level Variables					
# Students in class	25.12	25.64	26.01	25.01	25.44
	(2.50)	(2.16)	(2.76)	(3.02)	(2.68)
Share of participating students	0.73	0.66	0.76	0.72	0.72
	(0.12)	(0.13)	(0.17)	(0.12)	(0.14)
Observations	165	177	189	214	745

second, we elicited preferred peers based on names (*name-based preferences*), where students in principle could condition their peer choice on all known characteristics.

These two distinct preference measures allow us to describe to whom students want to compare themselves and how they choose their social reference point depending on their own personality. In the following, we begin by describing the performance-based preferences. In a second step, we use the information gathered in the survey to analyze their determinants. That is, we ask which of one's own personality traits predict the relative ability of the most-preferred peers. We then turn to name-based preferences. Similar to models in the network formation literature, we lever the peers' personality and social network to analyze homophily in these characteristics.

6.4.1 Determinants of peer selection – Performance-based preferences

As described in section 6.3, we generated a partial ranking over ten categories, each category consisting of a one second time interval. In the following, we first describe patterns in the preferences. Afterwards, we analyze which personality traits are associated with the respective peer preferences.



(a) First performance-based preference

(b) Relationship among Preferences

(c) Second performance-based preference

(d) Third performance-based preference

(e) Females

(f) Males

Figure 6.2. Distribution of performance-based peer preferences

Notes: The figures (a) and (c) through (f) present histograms of students' preferences over relative performance. Panel (a) presents the distribution of relative times of most-preferred peers, Panel (c) and (d) present the second and third performance-based preference, respectively, while Panel (e) and (f) show the distribution for females and males. The intervals used here and in the survey are one second intervals of relative times in the first run. Vertical lines indicate own time (black; equals zero by definition) and mean preference (red; where we used the mean of each interval to calculate the mean). Panel (b) presents the relationship of the first performance-based preference and the second/third preference.

Figure 6.2 presents the preferences for peers over relative performance. First, turning to the distribution of the most-preferred performance differential (Figure 6.2a), we find that students prefer peers from the entire possible set. In other words, students' first preference ranges from 4 to 5 seconds slower to 4 to 5 seconds faster peers, spanning every one second interval in between. Second, around half of the students prefer similarly performing peers, i.e., their first performance-based preference lies within one second of their own time in the first run. Finally, students prefer on average slightly faster peers, i.e., they select peers who were .56 seconds faster in the first run corresponding to .20 SD in terms of times in the first run. As Figures 6.2e and 6.2f show, males prefer slightly faster peers than females.[18] This tendency for males to select faster, potentially more challenging peers is in line with previous findings that men are more likely to select into competitive environments (for an overview of this literature see Dariel et al., 2017) or choose harder challenges (Niederle and Yestrumskas, 2008).[19]

In Figures 6.2c and 6.2d, we present the distributions of the second and third highest ranked interval. While the probability mass in these histograms is shifted towards a relative time of zero, this is just an artifact of the limited amount of categories as can be seen in Figure 6.2b. The figure shows the relationship between the first performance-based preference and the second as well as the third one. We observe that the second and third preference are centered around the first performance-based preference. Using tobit regressions, we show in Appendix Table 6.A.1 that the second and third preferences are indeed clustered around the most-preferred relative time.[20] In the following analyses, we therefore focus on the first performance-based preference only as this preference captures the same information as the lower-ranked ones.

In general, the histograms support the conjecture of Festinger (1954, p. 121) that people compare themselves to others who are "close to [their] own ability" and is in line with evidence from other disciplines noting also tendencies to engage in upward comparisons (e.g., Huguet et al., 2001). Yet, this does not hold for all of our subjects. In particular, there is a sizable share of subjects who prefer peers that do not have a similar ability as measured by their time in the first run.

[18] The difference corresponds to .61 seconds. Moreover, in Appendix 6.A, we present additional splits by age group and do not find any differences across cohorts.

[19] Yet, as our analysis below shows, this difference can be explained by gender differences in personality characteristics.

[20] The categories in which subjects 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. Tobit regressions in Appendix Table 6.A.1 confirm this intuition: allowing for censoring at the lower and upper limit, the regression coefficient on the second preferences is .97 and we cannot reject the hypothesis that the coefficient equals one.

In order to further investigate the choice of a preferred peer and analyze which characteristics predict the choice, we divide subjects preferences into three categories – preferring slower peers (i.e., preferring more than one second slower peers), similar peers (preferring up to one second slower or faster peers), and faster peers (preferring peers more than a second faster). These categories correspond to 15%, 50% and 35% of all preferences, respectively.[21]

First, we analyze how own ability measured by an individual's time in the first run is related to peer choice. In order to do this, we estimate a multinomial logit model of peer choice categories. To begin with, we explain individual i's of class c chosen peer categories k – where k may be slower, similar, or faster – with time in the first run, gender and class fixed effects:

$$Pr[PC_{ic} = k] = \frac{\exp\left\{\alpha_k + \beta_k X_i + \gamma_c + \delta_i\right\}}{\sum_l \exp\left\{\alpha_l + \beta_l X_i + \gamma_c + \delta_i\right\}} \quad k = \text{slower,similar,faster} \tag{6.8}$$

Panel (a) of Table 6.2 presents the corresponding marginal effects. We find directional evidence on the 10% significance level that the slower students are in the first run, the less likely they are to choose faster peers. Interpreting the time in the first run as a measure of ability, we find that this is broadly consistent with the prediction of the model outlined in section 6.2. More specifically, being one second slower in the first run decreases the probability of choosing a faster peer by 1.75 percentage points and increases the probability of choosing peers with a similar ability.

In the following, we analyze how individuals' personality measured by the Big 5, locus of control, attitudes towards social comparison, competitiveness and risk attitudes shape these preferences. In order to do so, we enrich our model of peer choice by including personality traits and preferences as additional explanatory variables in the vector of covariates X_i. Panel (b) of Table 6.2 presents the corresponding marginal effects. First, note that times in the first run do not help to explain peer choices conditional on the rich set of personality characteristics. Rather, we find that three traits are significant related to the peer choice – extraversion, locus of control and competitiveness, of which competitiveness is strongly related to performance. In order to capture the magnitude as well as the overall patterns of these effects more easily, we present the resulting marginal effects for different levels of these three skills in Figure 6.3.[22]

[21] We also estimate specifications, where we split up similar peers into "similar but slower" and "similar but faster" peers. Effects are in line with the specification using three categories. Since, we do not find differential effects on those two categories, we focus on the specification with three categories for the ease of exposition. We report these results in Appendix Table 6.A.2.

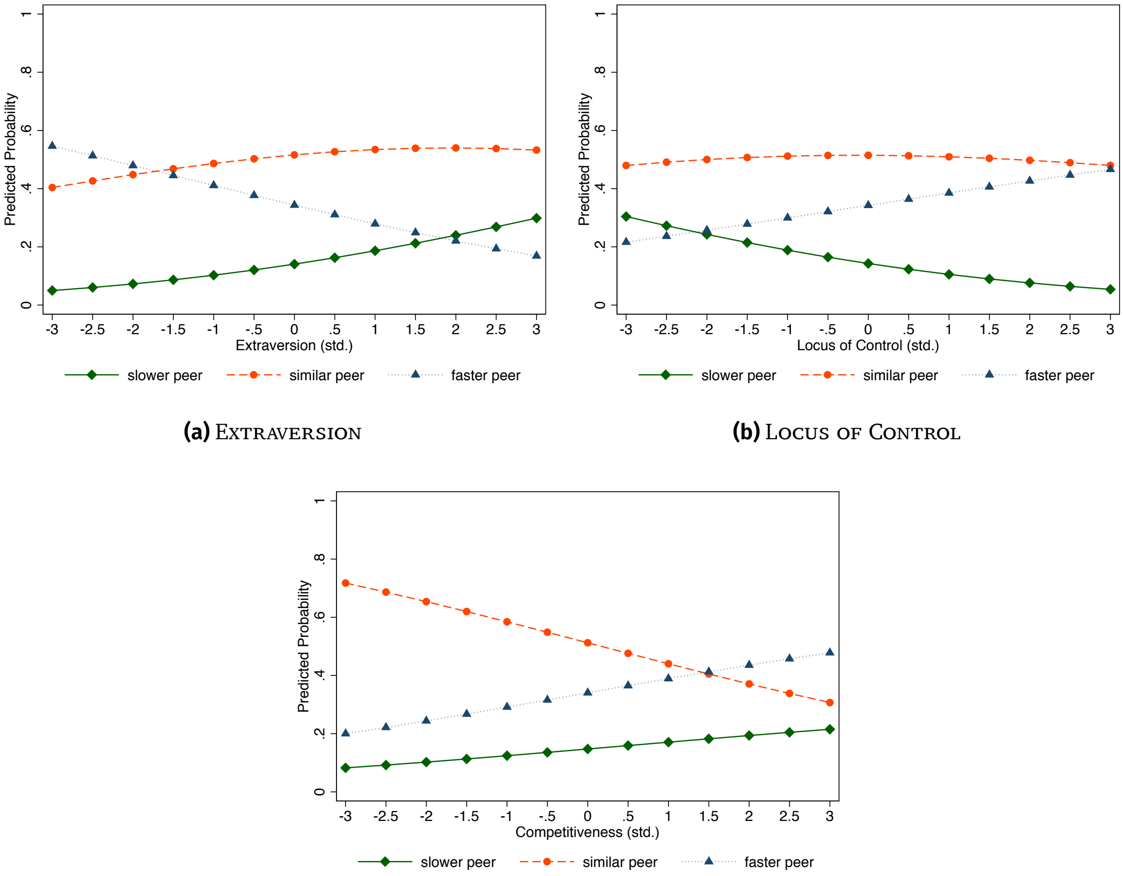
[22] These figures plot the marginal effects of the three non-cognitive skills using the same model as in Table 6.2 evaluated at levels of the personality characteristics ranging from -3 SD to +3 SD. All other variables are hold constant at their mean value.

Table 6.2. Marginal effects of choosing different peers

	Peer category		
	slower	similar	faster
Panel (a): Only time			
Time (First Run)	0.0026	0.0149*	-0.0175*
	(0.0071)	(0.0082)	(0.0091)
Female	0.0558	-0.0282	-0.0276
	(0.0379)	(0.0499)	(0.0446)
Panel (b): Including personality			
Time (First Run)	0.0110	-0.0056	-0.0054
	(0.0074)	(0.0095)	(0.0105)
Agreeableness	0.0206	-0.0006	-0.0200
	(0.0148)	(0.0223)	(0.0214)
Conscientiousness	-0.0187	0.0035	0.0151
	(0.0161)	(0.0207)	(0.0173)
Extraversion	0.0421***	0.0237	-0.0658***
	(0.0119)	(0.0188)	(0.0155)
Openness to Experience	-0.0115	-0.0024	0.0140
	(0.0128)	(0.0225)	(0.0204)
Neuroticism	-0.0034	0.0193	-0.0159
	(0.0221)	(0.0264)	(0.0248)
Locus of Control	-0.0417**	-0.0008	0.0425*
	(0.0177)	(0.0259)	(0.0250)
Social Comparison	0.0031	0.0049	-0.0081
	(0.0180)	(0.0228)	(0.0184)
Competitiveness	0.0231	-0.0720**	0.0489*
	(0.0177)	(0.0302)	(0.0278)
Risk Attitudes	0.0018	0.0050	-0.0069
	(0.0161)	(0.0229)	(0.0207)
Female	0.0254	-0.0272	0.0019
	(0.0498)	(0.0555)	(0.0552)
Observations	623	623	623
Class Fixed Effects	Yes	Yes	Yes
Baseline Probability	.15	.5	.35

Notes: This table presents marginal effects from multinomial logistic regressions using indicators for three categories of performance-based preferences as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Panel (a) uses only time in the first run as an independent variable, while panel (b) also includes personality measures. Standard errors in parentheses and clustered on class-level.

Increasing extraversion is related to substituting away from faster peers towards slower peers. A one standard deviation increase in extraversion reduces the probability to choose a faster peer by 6.58 percentage points and increases the probability to choose a relatively slower peer by 4.21 percentage points. This corresponds to a 19% decrease relative to the baseline of choosing a faster peer and a 28% increase for slower peers. In general, higher levels of extraversion are



(a) Extraversion

(b) Locus of Control

(c) Competitiveness

Figure 6.3. Marginal effects of personality characteristics on peer choice

Notes: These figures present marginal effects of the results in Table 6.2 for three personality characteristics, namely extraversion, locus of control, and competitiveness. Solid green lines with diamonds (⋄) correspond to preferring slower peers, dashed orange lines with circles (•) indicate peers of similar ability (±1 sec.), and dotted blue lines with triangles (△) correspond to faster peers.

associated with a lower probability of preferring faster peers but rather favoring slower ones, while leaving the probability of subjects preferring similar peers mainly unaffected.

Having a more internal locus of control has the opposite pattern as extraversion. A one standard deviation increase of locus of control increases the probability of choosing a faster peer by 4.25 percentage point, while it decreases the probability of choosing slower peers by 4.17 percentage points. As presented in Figure 6.3b, high levels are associated with subjects switching away from slower peers to faster peers. Again the share of subjects preferring similar peers is nearly unaffected. This behavior matches well with the core concept that people with a higher internal locus of control have higher aspirations and expectations for themselves (Phillips and Gully, 1997; Yukl and Latham, 1978; Ng et al., 2006).

Finally, the pattern for competitiveness is quiet different. More competitive individuals have a 4.89 percentage points higher probability of choosing a faster peer, while their share of similar peers is reduced by 7.20 percentage points.[23] In general, subjects scoring very low on the competitiveness scale choose predominantly peers that have a similar performance in the first run. This changes as individuals get more and more competitive; they prefer mainly faster and some slower peers. Only 20% of the most competitive subjects select similar peers. This relationship suggests that the influence of competitiveness on behavior is non-trivial. The previous literature related this trait mainly to tournament entry decisions (e.g., Niederle and Vesterlund, 2007). Here, we highlight the dual nature of competitiveness: More competitive individuals either seem to choose situations were "winning" their second run is very likely, or they choose especially challenging situations, i.e., competitiveness seems to indicate a preference for unambiguous results. Other characteristics do not seem to affect peer preferences over relative performance. Their point estimates are close to zero and non-significant at any conventional level.

In conclusion, we observe significant relationships between an individual's personality and her peer preferences. On the one hand, we find that more extraversive individuals and those scoring low on locus of control (i.e., with a more external locus) switch from choosing faster peers to choosing slower peers when increasing the respective characteristics. On the other hand, competitive individuals avoid similar strong peers, but rather prefer either slower or faster peers with more preferring the latter. These findings highlight that an individual's characteristics are important for the selection of peers or (social) reference points more generally.

6.4.2 Determinants of peer selection – Name-based preferences

The second set of preferences elicited in the survey allows students to state their preferences based on the identity of their classmates by selecting peers from a list of their classmates' names. In contrast to performance-based preferences, which stripped away all considerations that pertain to the social dynamics of the observed group or the specific peer (i.e., individuals could condition their peer choice only on information about relative past performance), students could in

[23] Note that these estimates show that it is not appropriate to estimate an alternative model such as an ordered logit since the estimates show that there is no clear ordering of slower, similar and faster peers. Intuitively, the same trait can lead to an increase in the likelihood to choose faster or slower peers as the pursuit of winning will increase the likelihood to choose slower peers, whereas the motive of competing with the best will increase the likelihood to choose faster peers. Hence, there is no natural ordering of these categories.

principle take all known information about their potential peers into account when selecting their social reference point. In our analysis, we proxy the information by a rich set of personality measures and individual specific network characteristics. In the following, we first present summary statistics of these preferences before we analyze their determinants.

Table 6.3. Share of name-based preferences being friends

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

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

Table 6.3 presents the share of name-based peers that are friends of an individual. While 89% of all individuals choose a friend as their most preferred peer, this number decreases by about 10 percentage points for each of the following preferences.[24] Thus, this finding shows that friendship ties are a good proxy for peers in general, which validates their use for other studies on peer effects. However, note that there are some students who do not solely choose their peers based on friendships. This suggests that some subjects – potentially strategically – avoid some of their friends in favor of other class members. Hence, two questions arise: First, if students do not nominate solely their friends, whom do they nominate? Second, given they nominate a friend, how do they decide which of their friends to nominate? We answer these questions using a nomination model similar to those used in the network formation literature (e.g., Graham, 2015).

In order to analyze the determinants of peer nomination in a structured way, we assume the following. Let y_{ij} equal one if individual i nominates individual j and zero otherwise. We allow the nomination to depend on unobserved heterogeneity captured by individual fixed effects ν_i and ν_j as well as a measure of similarity between those two individuals in terms of K observables $X_i = (x_{i1}, \dots, x_{iK})$ and $X_j = (x_{j1}, \dots, x_{jK})$ denoted by $\delta(\cdot, \cdot)$, potential existing friendship ties F_{ij}, and an idiosyncratic shock ϵ_{ij} for each nomination clustered on the class-level. This yields the following linear probability model:

$$y_{ij} \;=\; \alpha + \nu_i + \nu_j + \delta(X_i, X_j) + \epsilon_{ij} \tag{6.9}$$

In our application, we measure similarity in terms of the absolute difference in each observable characteristics defined as $\delta(x_{ik}, x_{jk}) = \beta_k |x_{ik} - x_{jk}|$. Ho-

[24] This pattern might be partially driven by the fact that students do not have enough friends of the same gender in the class, which they can nominate. Yet, our data shows that 78% of friends are of the same gender and thus there are sufficiently many friends to potentially nominate friends in one of their first preferences (see Table 6.1).

mophily – the tendency of individuals to nominate others with similar characteristics (McPherson et al., 2001) – is then defined as $\beta_k < 0$.[25]

We consider three sets of observable characteristics in which students potentially exhibiting homophily. First, we allow for homophily in ability. That is, higher differences in ability measured by times in the first run are likely to decrease the nomination probability. As we saw above, students have preferences to run with peers of similar ability. Therefore, it is likely to play a role also for name-based preferences. If one of their friends is much faster than they are, they potentially want to avoid that person. Second, we consider homophily in several personality characteristics, namely the Big Five, locus of control, competitiveness, attitudes to engage in social comparisons and risk attitudes. Several of these characteristics may be important, similar to what has been found in the network formation literature.[26] Finally, the position of the peer within the social network itself may be important. Popular individuals might be more likely to get nominated if other students would like to interact with a popular student or if those students are just more visible. We therefore consider the effects of the presence of friendship ties between i and j as well as other homophily terms in network characteristics: degree (the number of friends), eigencentrality (influence in a network), and clustering (embeddedness of an individual in a given network).[27]

In our analysis, we define a person to be nominated 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 somebody would be most willing to be paired with in the second run.[28] Peers according to this definition obviously constitute the most important nominations. Moreover, our matching rule to form pairs – a stable roommate algorithm – ensures that students are most likely to be paired with one of these three preferred peers.

Table 6.4 presents the results of the estimation of the linear probability model (equation 6.9). Column (1) restricts attention to homophily in person-

[25] Similarly, heterophily – the tendency to avoid others who are similar – is then defined as $\beta_k > 0$.

[26] A common finding across these studies is that extraversion and agreeableness are important for forming a link between two nodes of a social network. Moreover, many studies find homophily in one or several dimensions, i.e., similar individuals are more likely to be linked to each other (see, e.g., McPherson et al., 2001; Jackson, 2014; Girard et al., 2015). In our setting we take the social network as given, yet a similar logic may drive the nomination process.

[27] Eigencentrality is a measure of influence of an individual in a given social network. It measures whether an individual is connected to other influential individuals. Clustering describes the share of friends that are also friends of each other. In order to facilitate interpretation of these two measures, we standardize them such that we can interpret differences in units of standard deviations.

[28] That is, 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.

Table 6.4. Linear probability model of nominating a peer

	Nominated Peer					
	(1)	(2)	(3)	(4)	(5)	(6)
Abs. Diff. in Time of First Run		-0.0408***	-0.0292***	-0.0297***		-0.0580***
		(0.0066)	(0.0058)	(0.0058)		(0.0156)
Abs. Diff. in Beliefs over Times in First Run					-0.0443***	
					(0.0162)	
Homophily in Personality						
Abs. Diff. in Agreeableness	-0.0380***	-0.0365***	-0.0334***	-0.0347***	-0.0551**	-0.0713**
	(0.0091)	(0.0094)	(0.0100)	(0.0103)	(0.0229)	(0.0300)
Abs. Diff. in Conscientiousness	-0.0125	-0.0124	-0.0017	-0.0044	0.0013	-0.0120
	(0.0124)	(0.0114)	(0.0093)	(0.0096)	(0.0209)	(0.0247)
Abs. Diff. in Extraversion	-0.0265**	-0.0297**	-0.0146	-0.0155	-0.0111	-0.0332
	(0.0126)	(0.0129)	(0.0112)	(0.0105)	(0.0248)	(0.0323)
Abs. Diff. in Openness	-0.0050	-0.0080	-0.0038	-0.0022	0.0052	-0.0243
	(0.0128)	(0.0131)	(0.0112)	(0.0114)	(0.0253)	(0.0326)
Abs. Diff. in Neuroticism	-0.0148	-0.0217*	-0.0142	-0.0155	-0.0451	-0.0498*
	(0.0119)	(0.0122)	(0.0104)	(0.0096)	(0.0281)	(0.0285)
Abs. Diff. in Locus of Control	-0.0108	-0.0101	-0.0035	-0.0012	-0.0203	-0.0154
	(0.0116)	(0.0123)	(0.0097)	(0.0104)	(0.0283)	(0.0269)
Abs. Diff. in Social Comparison	-0.0347***	-0.0319**	-0.0223***	-0.0246***	-0.0512**	-0.0561**
	(0.0108)	(0.0118)	(0.0081)	(0.0080)	(0.0219)	(0.0256)
Abs. Diff. in Competitiveness	-0.0374**	-0.0253*	-0.0179*	-0.0174*	-0.0355	-0.0385
	(0.0144)	(0.0129)	(0.0102)	(0.0102)	(0.0307)	(0.0275)
Abs. Diff. in Risk Preferences	-0.0133	-0.0185	-0.0050	-0.0026	0.0194	-0.0151
	(0.0158)	(0.0157)	(0.0102)	(0.0106)	(0.0325)	(0.0308)
Friendship Ties						
Friendship Indicator			0.3935***	0.3745***	0.4196***	
			(0.0189)	(0.0185)	(0.0384)	
Homophily in Network Characteristics						
Abs. Diff. in Degree				0.0080	-0.0074	-0.0006
				(0.0081)	(0.0250)	(0.0190)
Abs. Diff. in Eigencentrality (std.)				-0.0508***	-0.0410	-0.0800
				(0.0182)	(0.0457)	(0.0570)
Abs. Diff. in Clustering (std.)				-0.0178	-0.0110	-0.0957**
				(0.0184)	(0.0479)	(0.0422)
Own Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Peer Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Using Beliefs	No	No	No	No	Yes	No
Using Friends only	No	No	No	No	No	Yes
N	7084	6654	6654	6572	2920	2894
R^2	0.20	0.21	0.37	0.37	0.71	0.67

Notes: This table presents the results of the linear probability model according to equation 6.9 using an indicator of being nominated as one of the three most-preferred name-based peers as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Column (5) uses beliefs over relative performance rather actual relative performance and thus restricts the sample to those observations with information over the beliefs. Column (6) restricts the set of potential peers to friends only. Standard errors in parentheses and clustered on the class-level.

ality measures. Similar to findings from the network formation literature (e.g., Selfhout et al., 2010), we find significant homophily in two of the Big Five, namely agreeableness and extraversion. A one standard deviation difference in those characteristics decreases the probability of nominating a particular peer by 3.80 and 2.65 percentage points for agreeableness and extraversion, respectively. Moreover, differences in the attitudes towards social comparisons and in competitiveness also decrease the nomination probabilities. More precisely, a one standard deviation increase in the absolute difference in social comparison attitudes and competitiveness decreases the probability by 3.47 and 3.74 percentage points. One possible interpretation of the latter effects is the following: individuals who tend to dislike social comparisons or are rather uncompetitive avoid peers that are eager to compare themselves as these students create a more competitive atmosphere or enforce social comparisons. Students, however, that gravitate towards social comparisons might actively seek out these peers to motivate themselves.

The second column additionally includes absolute differences in ability as measured by times in the first run. We observe a strong and significant homophily factor of 4.08 percentage points for these differences. Notably, except from competitiveness the influence of all the characteristics discussed above remains constant in magnitude and precision when adding absolute differences in performances to the model. The coefficient for homophily in competitiveness decreases to 2.53 percentage points since performance in the first run and competitiveness are strongly associated.[29]

Adding an indicator for friendship ties in column (3) shows that these links are highly predictive for nominations. Existing friendship ties increase the nomination probability of potential peers by 39.35 percentage. While this effect is hardly surprising given the share of friends amongst nominated peers (as displayed in Table 6.3), other dimensions remain important for the nomination process even conditional on friendship ties. In particular, we still observe homophily in ability as measured by times in the first run, agreeableness and attitudes for social comparisons.

We additionally control for the absolute differences in network characteristics in column (4). This reveals that the position in the social network is important, as well. Individuals are much more likely to be nominated as peers if they have a similar eigencentrality, i.e., have a similar influence in the network. More specifically, a one standard deviation difference in eigencentrality between two individuals decreases the nomination probability by 5.08 percentage points. This effect is larger than the remaining homophily terms on ability, agreeableness and social comparison and highlights the importance of having detailed information on entire social networks for understanding peer preferences.

Summarizing our results so far highlights that information on friendship ties is by far the most important determinant of peer nominations. However,

[29] The correlation between time in the first run and competitiveness is -.46 implying that more competitive individuals are on average faster (have a lower time) than less competitive ones.

this information does not reveal which of your friends you nominate. Table 6.4 conveys that differences in ability and two personality measures – agreeableness and attitudes to engage in social comparisons – as well as the difference in one's influence in a social network measured by eigencentrality are all important for understanding the peer nomination process. If a friend has, for instance, a highly different ability or does engage in more comparisons with others, the likelihood of nominating this friend decreases strongly.

For the preceding analysis, we used the time in the first run as a measure of ability. However, students in the experiment neither did know about their own performance in the first run nor about that of their peers. For the nomination process, they have to rely on their beliefs about the relative performance of their peers. We therefore check our results by including the beliefs over relative performance rather than actual relative performance in column (5).[30] The results – except the effect of homophily in eigencentrality, whose coefficient remains similar in size but is now insignificant – are robust to using beliefs rather than actual relative performance. In fact, homophily in these characteristics is even more pronounced. It amounts to 4.43 percentage points for a one second difference in believed ability differences, 5.51 and 5.12 percentage points for one standard deviation differences in agreeableness and social comparison attitudes. Yet, the analysis using beliefs is based on a much smaller sample than the one presented here. This is due to the fact that beliefs are only measured for those individuals which are nominated as one of their first three name-based preferences.[31]

Our analysis, so far, has looked at the nomination process across the whole class of an individual. It is also interesting to analyze whether the observed patterns hold when looking at nominations solely among friends. Studying this restricted sample of potential peers is meaningful as friends often share many similar characteristics due to homophily. Column (6) presents the results and shows that these are similar with one exception: rather than eigencentrality, clustering seems to be important for peer choice.[32] A clustering coefficient is high if friends of an individual are also friends with each other – that is, those individuals likely belong to the same clique given that we have restricted the

[30] In Appendix Table 6.C.1, we present additional estimates corresponding to columns (1) through (3) of Table 6.4.

[31] In Appendix 6.B, we show that beliefs and actual relative performance are strongly related to each other. Moreover, we show the consistency of the beliefs by validating that they are stable. In particular, 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.

[32] Similarly to the robustness check using beliefs, Appendix Table 6.C.1 presents additional specifications corresponding to the other columns of Table 6.4 for the restricted sample of friends only.

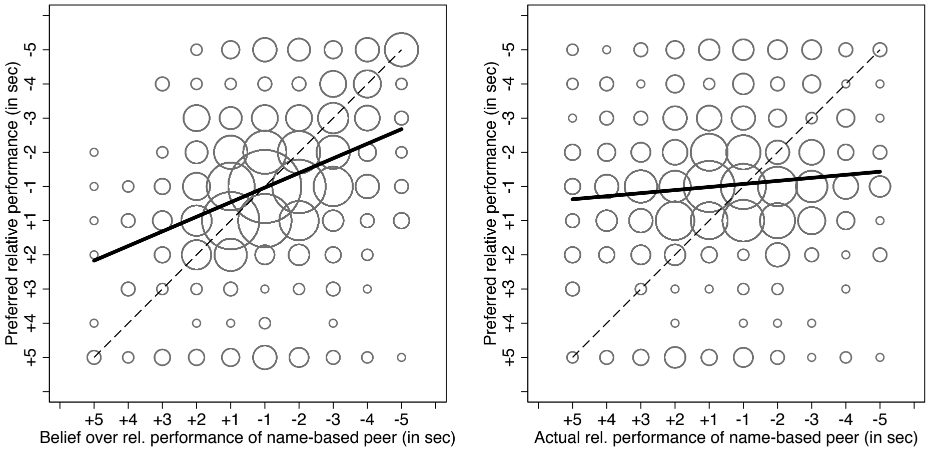
sample to friends only. Thus, when focusing on friends, not the relative influence of a given person matters, but rather whether she is as embedded as oneself in the network.

6.5 Relationship between performance- and name-based preferences

The preceding sections have analyzed both performance- and name-based preferences in isolation. A natural next step is to ask whether and, if yes, to what extent the two preference dimensions are related. Until now, we have established that students, on average, prefer to be paired with somebody slightly faster than themselves. At the same time, this preference is significantly related to several personality measures, namely extraversion, locus of control, and competitiveness. Second, we analyzed whom do students choose if they in principle can condition their peer choice on a larger set of information. We found that although friendship ties are crucial for understanding peer preferences, there exist non-negligible homophily in ability, agreeableness and the tendency to engage in social comparisons. We now study the relation of the two preferences and analyze whether students indeed target peers with ability levels similar to their own as indicated by the homophily term in Table 6.4, or whether they actually try to target their preference over relative performance when nominating peers based on names.

Figure 6.4 and Table 6.5 show the relation of performance- and name-based preferences. We plot the relative performance level of the most preferred name-based peer against the most preferred relative performance level. More specifically, Figure 6.4a and Panel (a) of Table 6.5 use beliefs over the relative performance of peers nominated in the name-based preferences. We observe a positive relationship: If subjects nominate a peer that they belief is one second faster than themselves, they choose a .44 seconds faster peer in their performance-based preferences (column (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. However, the relationship between name- and performance-based preferences is not perfect as it would be the case if students just try to select a name-based peer corresponding solely to their performance-based preferences. If this would be the case, we observed regression coefficients of unity. A similar, although less pronounced pattern holds if we look at actual time differences rather than beliefs in Panel (b) of Table 6.5.

A possible interpretation to explain this imperfect relation is the fact that preferences for peers are multidimensional in nature. In this case the coefficients smaller than one would result from other dimensions being important



(a) Perf.-based preference and belief

(b) Perf.-based preference and actual performance

Figure 6.4. Relationship of performance- and name-based preferences for peers

Notes: The figures present the relationship between performance- and name-based preferences using either beliefs over peer's performance (Panel (a)) or peer's actual performance (Panel (b)). Corresponding regressions are presented in Table 6.5.

and being used when nominating peers based on names.[33] In order to show that preferences for peers are indeed multidimensional and that name-based preferences in contrast to performance-based preferences allow students to take other dimensions into account, we enrich the linear probability model from the previous section, equation 6.9, and additionally include a peer's absolute deviation from the most-preferred relative performance in the nomination model. Table 6.6 presents the results of this analysis.

In column (1), our main specification, we observe homophily between the nominated peers and the most preferred performance of 2.37 percentage points, similar to the previous effect of homophily in ability as reported in Table 6.4. The direct homophily in ability is now much smaller and not significant at all conventional levels. Thus, students indeed target individuals that are close to their most-preferred performance rather than those who are close to their own performance. Importantly, the other homophily terms that were found to be important for the peer choice in the previous analysis of Table 6.4 remain unaffected: both agreeableness and attitudes to engage in social comparisons exhibit strong ho-

[33] A second possible explanation is that the true relation is indeed perfect and measurement error attenuates this association. We discuss this in Appendix 6.D, where we show that the preference measures would need more noise than actual signal components to explain the coefficients smaller than one. This makes measurement error as sole cause for this imperfect relation rather unlikely.

Table 6.5. Relationship between performance- and name-based preferences

	(a) Peer's relative time		(b) Peer is faster (binary)	
	(1)	(2)	(3)	(4)
Panel A: Perf.-based pref. and name-based beliefs				
Belief over name-based peer's performance	0.43*** (0.06)	0.44*** (0.06)		
Belief over name-based peer's performance (0/1)			0.29*** (0.03)	0.29*** (0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
N	781	648	781	648
R^2	.23	.28	.17	.2
Panel B: Perf.-based pref. and name-based actual performance				
Relative Time of most-preferred name-based peer	0.09*** (0.03)	0.09*** (0.03)		
Preferred name-based peer is faster			0.04 (0.03)	0.03 (0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
N	662	562	662	562
R^2	.11	.13	.095	.12

Notes: This table presents least squares regressions using a peer's relative time in one second intervals or an indicator for preferring a faster peer according to the performance-based preferences as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on class-level. Figure 6.4 presents the results graphically.

mophily of 3.50 and 2.35 percentage points, respectively, and clustering has a coefficient of 5.17 percentage points.

Columns (2) and (3) of Table 6.6 confirm these results using beliefs rather than actual performance and restricting the sample to friends only. As previously, the effects are even more pronounced for these subsets.

These results highlight that there exist a non-trivial relation between performance- and name-based preferences; that is, the most preferred relative ability and whom individuals choose as peers are indeed strongly related. This implies that reference points over relative performance matter in natural environments, in which these reference points are not induced by experimenters. Yet, these measures do not coincide completely. Instead, being able to select peers based on their names enables students to condition on a richer information set and thus introduces additional dimensions to the peer choice process. Moreover, while information on existing friendship ties play a crucial role, there exist significant and robust homophily in other dimensions such as agreeableness and attitudes towards social comparisons. These associations remain even if taking performance-based preferences and social networks into account.

Table 6.6. Linear probability model of nominating a peer II

	Nominated Peer		
	(1)	(2)	(3)
Abs. Diff. in Time of First Run	-0.0096		-0.0178
	(0.0063)		(0.0188)
Abs. Diff. from Perf.-based Preference	-0.0237***		-0.0465***
	(0.0066)		(0.0137)
Abs. Diff. in Beliefs over Times in First Run		-0.0061	
		(0.0180)	
Abs. Diff. from Perf.-based Preference (using beliefs)		-0.0794***	
		(0.0162)	
Homophily in Personality			
Abs. Diff. in Agreeableness	-0.0350***	-0.0493**	-0.0712**
	(0.0105)	(0.0239)	(0.0301)
Abs. Diff. in Conscientiousness	-0.0045	0.0042	-0.0131
	(0.0094)	(0.0219)	(0.0240)
Abs. Diff. in Extraversion	-0.0159	-0.0130	-0.0329
	(0.0105)	(0.0238)	(0.0317)
Abs. Diff. in Openness	-0.0016	0.0065	-0.0216
	(0.0114)	(0.0284)	(0.0327)
Abs. Diff. in Neuroticism	-0.0156	-0.0443*	-0.0501*
	(0.0094)	(0.0259)	(0.0284)
Abs. Diff. in Locus of Control	-0.0012	-0.0211	-0.0152
	(0.0104)	(0.0284)	(0.0271)
Abs. Diff. in Social Comparison	-0.0235***	-0.0499**	-0.0542**
	(0.0078)	(0.0216)	(0.0253)
Abs. Diff. in Competitiveness	-0.0168	-0.0310	-0.0399
	(0.0104)	(0.0300)	(0.0263)
Abs. Diff. in Risk Preferences	-0.0032	0.0140	-0.0202
	(0.0105)	(0.0306)	(0.0306)
Friendship Ties			
Friendship Indicator	0.3743***	0.4120***	
	(0.0183)	(0.0374)	
Homophily in Network Characteristics			
Abs. Diff. in Degree	0.0080	-0.0071	0.0013
	(0.0080)	(0.0236)	(0.0184)
Abs. Diff. in Eigencentrality (std.)	-0.0517***	-0.0374	-0.0811
	(0.0183)	(0.0466)	(0.0576)
Abs. Diff. in Clustering (std.)	-0.0187	-0.0040	-0.0971**
	(0.0181)	(0.0449)	(0.0418)
Own Fixed Effects	Yes	Yes	Yes
Peer Fixed Effect	Yes	Yes	Yes
Using Beliefs	No	Yes	No
Using Friends only	No	No	Yes
N	6572	2920	2894
R^2	0.38	0.72	0.67

Notes: This table presents the results of the linear probability model according equation 6.9 using an indicator of being nominated as one of the three most-preferred name based peers as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on the class-level.

6.6 Discussion and conclusion

As peers influence behavior in various aspects of our life as consumption, performance on the job, or even financial investments, the question how and whom individuals select as these peers is crucial for understanding peer effects in general and for designing policies aimed at leveraging them specifically. In this paper, we study preferences for peers across two dimensions – either based on relative performance levels or on the identity of peers – and analyze their relation. We find that many subjects tend to favor comparisons with people who have a similar or slightly higher ability level, or with their friends. Still, a large degree of heterogeneity in preferences persists. Individual preferences for peers are related to individual characteristics, as well as relative characteristics of the potential peer. This highlights the importance of personality for selecting social reference points as well as for the selection of more or less challenging environments more generally. While students target a preferred relative ability, even if selecting peers based on their social identity, this is only one amongst several factors determining peer choices. A peer's personality as well as existing friendship ties are also crucial to understand who serves as a peer. These findings therefore stress the multidimensionality of preferences for peers and validate using friendship ties as proxies for peers in other peer studies.

Understanding the heterogeneity of preferences for peers might not only be useful from an academic point of view, but also for many practical settings. Our results can help teachers or supervisors to figure out how peer groups emerge endogenously in schools or in the workplace. Even more importantly, we highlight that an individual's personality is crucial for the selection of social reference points. In order to design institutions that lever the positive impact of peers optimally and foster performance, it is not only necessary to understand how individuals select into different environments (e.g., Müller and Schwieren, 2012; Dohmen and Falk, 2011), but also how individuals choose relevant peers within these environments. This helps to realize the difficulties that lie in designing optimal re-assignment policies as in Carrell et al. (2013), and complements the theoretical discussion of Ederer and Patacconi (2010) who emphasize the role of relevant reference groups in tournaments.

Our descriptive analysis of preferences for peers is in line with the previous literature (e.g., Blanton et al., 1999; Huguet et al., 2001): students prefer similar, but slightly more able peers. Combined with our results on the importance of personality for peer selection, this process can potentially have far reaching effects. A student's personality might indirectly impact educational attainment or performance on the job, for instance, by selecting repeatedly into high performing environments that accelerate individual growth. Agostinelli (2018) provides evidence that the peer environment of adolescents indeed affects skill development and life outcomes. Our results demonstrate how personality traits help to understand the formation of peer groups and thus contribute to our knowledge of the link between personality and life outcomes. These potentially longlasting effects create a new avenue for interventions: How can individuals of different personality be encouraged to select themselves into peer groups that help them to unfold their full potential? Moreover, students might not be perfectly aware how their peers affect their own performance raising the question whether these preferences would change if they are informed about these effects or even "nudged" to select specific peers.

Our results can be seen as a first step to document the multidimensionality of preferences for peers. Nevertheless, further research on the interaction of personality, selection into specific (peer) environments, and the influence of peers is needed to improve our understanding of social comparison processes and the endogenous formation of peer groups as well as their consequences for important outcomes later in life.

References

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman (2011): "Reference Points and Effort Provision." *American Economic Review*, 101 (2), 470–492. [200]

Agostinelli, Francesco (2018): "Investing in Children's Skills: An Equilibrium Analysis of Social Interactions and Parental Investments." [226]

Allen, Eric J., Patricia M. Dechow, Devin G. Pope, and George Wu (2017): "Reference-Dependent Preferences: Evidence from Marathon Runners." *Management Science*, 63 (6), 1657–1672. [199]

Almås, Ingvild, Alexander W. Cappelen, Kjell G. Salvanes, Erik Ø. Sørensen, and Bertil Tungodden (2015): "Willingness to Compete: Family Matters." *Management Science*, 62 (8), 2149–2162. [200]

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz (2011): "Personality Psychology and Economics." In *Handbook of the Economics of Education*. Ed. by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Vol. 4. Elsevier, 1–181. [202]

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee (2014): "Awards unbundled: Evidence from a natural field experiment." *Journal of Economic Behavior and Organization*, 100, 44–63. [197]

Bandiera, Oriana, Iwan Barankay, and Imran Rasul (2010): "Social Incentives in the Workplace." *Review of Economic Studies*, 77 (2), 417–458. [201]

Banerjee, Abhijit and Sendhil Mullainathan (2010): "The shape of temptation: Implications for the economic lives of the poor." [203]

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. [200, 226]

Brookins, Philip, Sebastian Goerg, and Sebastian Kube (2017): "Self-chosen goals, incentives, and effort." [199]

Buser, Thomas, Muriel Niederle, and Hessel Oosterbeek (2014): "Gender, Competitiveness, and Career Choices." *Quarterly Journal of Economics*, 129 (3), 1409–1447. [200, 202]

Cameron, Colin and Pravin Trivedi (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press. [239]

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. [197]

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. [201, 202, 226]

Cicala, Steve, Roland Fryer, and Jörg Spenkuch (forthcoming): "Self-Selection and Comparative Advantage in Social Interactions." *Journal of the European Economic Association*. [198, 201]

Clark, Andrew and Claudia Senik (2010): "Who Compares to Whom? The Anatomy of Income Comparisons in Europe." *Economic Journal*, 120 (544), 573–594. [197, 200]

Clark, Damon, David Gill, Victoria Prowse, and Mark Rush (2017): "Using Goals to Motivate College Students: Theory and Evidence from Field Experiments." [199]

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. [197]

Collins, Rebecca L. (1996): "For Better or Worse: The Impact of Upward Social Comparison on Self-Evaluations." *Psychological Bulletin*, 119 (1), 51–69. [201]

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. [200]

Dariel, Aurelie, Curtis Kephart, Nikos Nikiforakis, and Christina Zenker (2017): "Emirati women do not shy away from competition: evidence from a patriarchal society in transition." *Journal of the Economic Science Association*, 3 (2), 121–136. [212]

Dohmen, Thomas and Armin Falk (2011): "Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender." *American Economic Review*, 101 (2), 556–590. [226]

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. [208]

Ederer, Florian and Andrea Patacconi (2010): "Interpersonal Comparison, Status and Ambition in Organizations." *Journal of Economic Behavior and Organization*, 75 (2), 348–363. [226]

Falk, Armin and Markus Knell (2004): "Choosing the Joneses: Endogenous Goals and Reference Standards." *Scandinavian Journal of Economics*, 106 (3), 417–435. [199, 203]

Festinger, Leon (1954): "A Theory of Social Comparison Processes." *Human Relations*, 7 (2), 117–140. [200, 212]

Frank, Robert H. (1984): "Are Workers Paid their Marginal Products?" *American Economic Review*, 74 (4), 549–571. [197]

Fudenberg, Drew and David K. Levine (2006): "A dual-self model of impulse control." *American Economic Review*, 96 (5), 1449–1476. [203]

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. [208]

Girard, Yann, Florian Hett, and Daniel Schunk (2015): "How Individual Characteristics Shape the Structure of Social Networks." *Journal of Economic Behavior and Organization*, 115. Behavioral Economics of Education, 197–216. [201, 218]

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini (2003): "Performance in Competitive Environments: Gender Differences." *Quarterly Journal of Economics*, 118 (3), 1049–1074. [201]

Gneezy, Uri and Aldo Rustichini (2004): "Gender and Competition at a Young Age." *American Economic Review*, 94 (2), 377–381. [201]

Goerg, Sebastian (2015): "Goal setting and worker motivation." *IZA World of Labor* (178). [200]

Graham, Bryan S. (2015): "Methods of Identification in Social Networks." *Annual Review of Economics*, 7 (1), 465–485. [217]

Groves, Melissa Osborne (2005): "How important is your personality? Labor market returns to personality for women in the US and UK." *Journal of Economic Psychology*, 26 (6), 827–841. [202]

Healy, Andrew and Jennifer Pate (2011): "Can Teams Help to Close the Gender Competition Gap?" *Economic Journal*, 121 (555), 1192–1204. [200]

Heath, Chip, Richard Larrick, and George Wu (1999): "Goals as Reference Points." *Cognitive Psychology*, 38 (1), 79–109. [200]

Heckman, James J., Jora Stixrud, and Sergio Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics*, 24 (3), 411–482. [202]

Herbst, Daniel and Alexandre Mas (2015): "Peer Effects on Worker Output in the Laboratory Generalize to the Field." *Science*, 350 (6260), 545–549. [201]

Hsiaw, Alice (2013): "Goal-setting and self-control." *Journal of Economic Theory*, 148 (2), 601–626. [200]

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. [200, 212, 226]

Jackson, Matthew O. (2014): "Networks in the Understanding of Economic Behaviors." *Journal of Economic Perspectives*, 28 (4), 3–22. [218]

Jackson, Matthew O., Brian W. Rogers, and Yves Zenou (2017): "The Economic Consequences of Social-Network Structure." *Journal of Economic Literature*, 55 (1), 49–95. [201]

Kahneman, Daniel (1992): "Reference points, anchors, norms, and mixed feelings." *Organizational Behavior and Human Decision Processes*, 51 (2). Decision Processes in Negotiation, 296–312. [199]

Kiessling, Lukas, Jonas Radbruch, and Sebastian Schaube (2018): "The impact of self-selection on performance." [199, 206]

Koch, Alexander K. and Julia Nafziger (2011): "Self-regulation through Goal Setting*." *Scandinavian Journal of Economics*, 113 (1), 212–227. [199, 203]

Kőszegi, Botond and Matthew Rabin (2006): "A Model of Reference-Dependent Preferences." *Quarterly Journal of Economics*, 121 (4), 1133–1165. [200]

Kőszegi, Botond and Matthew Rabin (2007): "Reference-Dependent Risk Attitudes." *American Economic Review*, 97 (4), 1047–1073. [200]

Kuegler, Alice (2009): "A Curse of Comparison? Evidence on Reference Groups for Relative Income Concerns." [201]

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. [197]

Laibson, David (1997): "Golden eggs and hyperbolic discounting." *Quarterly Journal of Economics*, 112 (2), 443–478. [203]

Lewis, Kevin, Marco Gonzalez, and Jason Kaufman (2012): "Social selection and peer influence in an online social network." *Proceedings of the National Academy of Sciences*, 109 (1), 68–72. [201]

Locke, Edwin and Gary Latham (2002): "Building a Practically Useful Theory of Goal Setting and Task Motivation." *American Psychologist*, 57 (9), 705–717. [200]

Manski, Charles (1993): "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies*, 60 (3), 531–542. [202]

Markman, Keith D. and Matthew N. McMullen (2003): "A Reflection and Evaluation Model of Comparative Thinking." *Personality and Social Psychology Review*, 7 (3), 244–267. [201]

Marmaros, David and Bruce Sacerdote (2006): "How Do Friendships Form?" *The Quarterly Journal of Economics*, 121 (1), 79–119. [201]

Mayer, Adalbert and Steven L. Puller (2008): "The old boy (and girl) network: Social network formation on university campuses." *Journal of Public Economics*, 92 (1), 329–347. [201]

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. [218]

Müller, Julia and Christiane Schwieren (2012): "Can personality explain what is underlying womenunwillingness to compete?" *Journal of Economic Psychology*, 33 (3), 448–460. [226]

Ng, Thomas W. H., Kelly L. Sorensen, and Lillian T. Eby (2006): "Locus of control at work: a meta-analysis." *Journal of Organizational Behavior*, 27 (8), 1057–1087. [216]

Nickerson, Jack A. and Todd R. Zenger (2008): "Envy, comparison costs, and the economic theory of the firm." *Strategic Management Journal*, 29 (13), 1429–1449. [197]

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. [200, 216]

Niederle, Muriel and Alexandra H. Yestrumskas (2008): "Gender Differences in Seeking Challenges: The Role of Institutions." [200, 212]

Phillips, Jean M. and Stanley M. Gully (1997): "Role of Goal Orientation, Ability, Need for Achievement, and Locus of Control in the Self-Efficacy and Goal-Setting Process." *Journal of Applied Psychology*, 82 (5), 792–802. [216]

Piatek, Rémi and Pia Pinger (2016): "Maintaining (Locus of) Control? Data Combination for the Identification and Inference of Factor Structure Models." *Journal of Applied Econometrics*, 31 (4). jae.2456, 734–755. [202]

Rotter, Julian B. (1966): "Generalized Expectancies for Internal Versus External Control of Reinforcement." *Psychological Monographs: General and Applied*, 80 (1), 1–28. [208]

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*. [208]

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. [219]

Tincani, Michela (2017): "Heterogeneous Peer Effects and Rank Concerns: Theory and Evidence." [198]

Weinhardt, Michael and Jürgen Schupp (2011): "Multi-Itemskalen im SOEP Jugendfragebogen." *DIW Data Documentation*. [208]

Wills, Thomas A. (1981): "Downward Comparison Principles in Social Psychology." *Psychological Bulletin*, 90 (2), 245–271. [201]

Wu, George, Chip Heath, and Richard Larrick (2008): "A prospect theory model of goal behavior." [200]

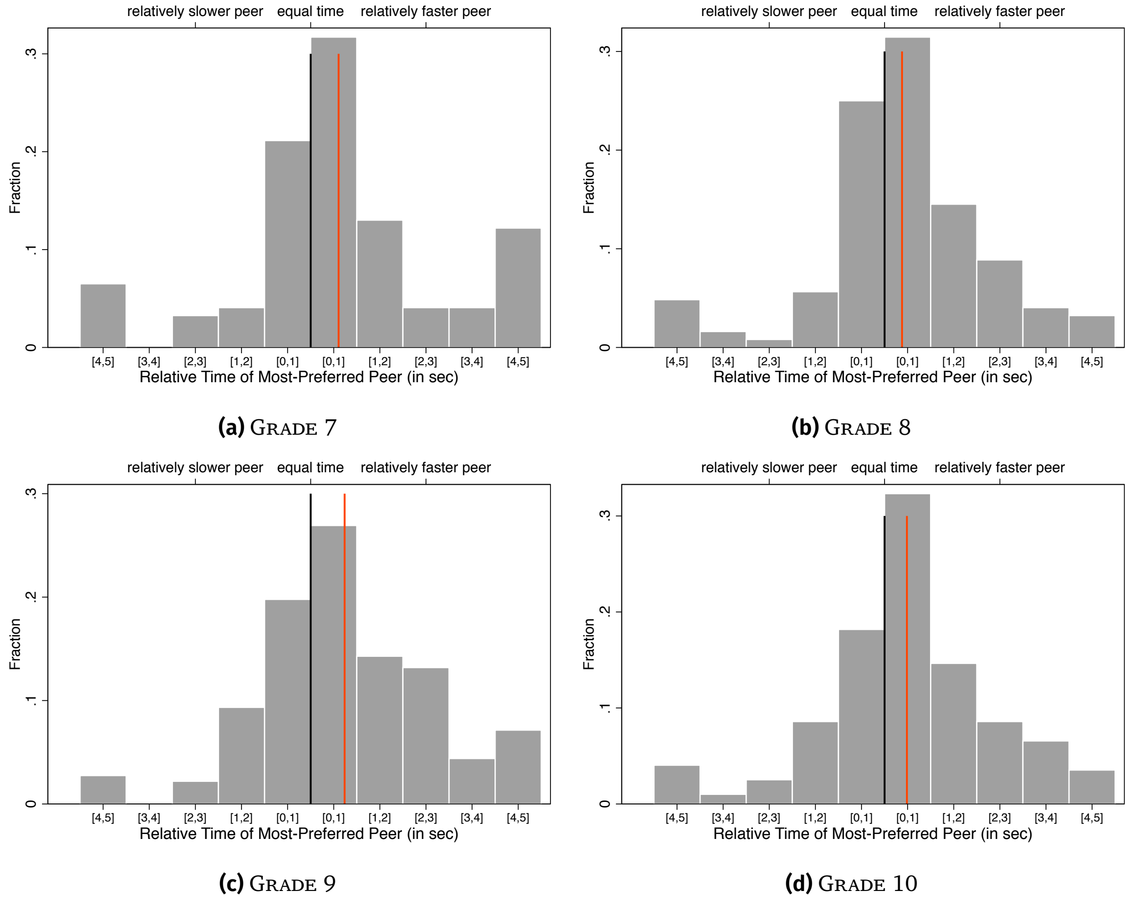
Yukl, Gary and Gary Latham (1978): "Interrelationships among employee participation, individual differences, goal difficulty, goal acceptance, goal instrumentality, and performance." *Personnel Psychology*, 31 (2), 305–323. [216]

Appendix 6.A Additional material for performance-based preferences

Figure 6.A.1 presents histograms similar to Figure 6.2 by cohort. More specifically, each histogram corresponds to the most-preferred relative times for each of grades 7 to 10 corresponding to ages 12 to 16. The resulting histograms look similar across cohorts. In unreported regressions we confirm this and conclude that there does not exist a gradient in these preference measures by age.

In Table 6.A.1, we analyze the relationship between first- and second-/third performance-based preference. Using Tobit regressions we allow for censoring in the dependent variable at relative times of ±5 seconds. In columns (2) and (4) we additionally control for censoring in the independent variables (the second and third performance-based preference, respectively). We observe that we cannot reject that the coefficient of the second-performance-based preference equals one indicating that subjects indeed try to mimic their most-preferred relative time with their second preference. A similar pattern holds for the third performance-based preference. We observe a significant association between third- and first preference. However, due to the additional dependence introduced by the second preference, the relationship is weaker.

Finally, we check whether the specification of Table 6.2 is restrictive by splitting the similar category in similar/slower and similar/faster corresponding to the intervals of preferring a peer who is up to one second slower or one second faster, respectively. Qualitatively, the patterns are similar to the results using three categories.



(a) Grade 7

(b) Grade 8

(c) Grade 9

(d) Grade 10

Figure 6.A.1. Distribution of performance-based peer preferences by grade

Notes: The figures present histograms of students' preferences over relative performance by grade. The intervals used here and in the survey are one second intervals of relative times in the first run. Vertical lines indicate own time (black; equals zero by definition) and mean preference (red; where we used the mean of each interval to calculate the mean).

Table 6.A.1. Relationship between performance-based preferences

	First perf.-based preference			
	(1)	(2)	(3)	(4)
model				
Second Performance-based Preference	0.92*** (0.04)	0.97*** (0.05)		
Third Performance-based Preference			0.69*** (0.05)	0.69*** (0.05)
Censoring in independent Variable	No	Yes	Yes	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
N	781	781	781	781
Pseudo R^2	.21	.22	.11	.11
p-value: Coefficient=1	0.03	0.47	0.00	0.00

Notes: This table presents tobit regressions using the first performance-based preference as the dependent variable with censoring at relative times of ±5 seconds. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on class-level.

Table 6.A.2. Marginal effects of choosing different peers

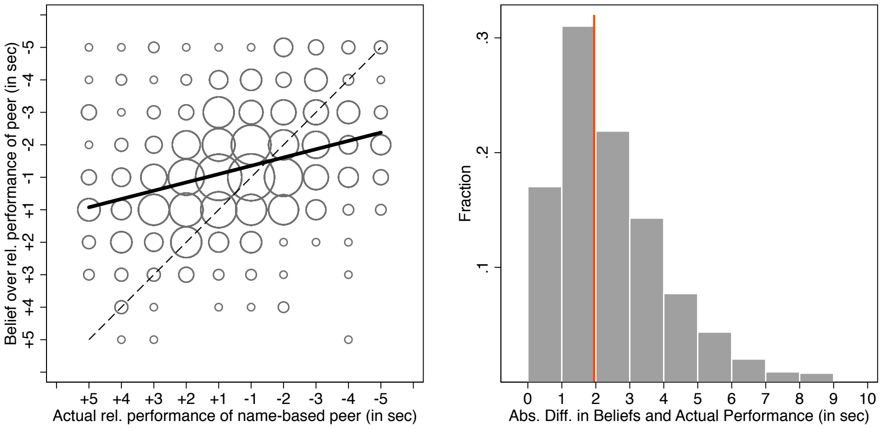
	Peer category			
	slower	similar/slower	similar/faster	faster
Panel (a): Only time				
Time (First Run)	0.0027	0.0153***	-0.0002	-0.0178*
	(0.0072)	(0.0058)	(0.0073)	(0.0093)
Female	0.0561	-0.0098	-0.0173	-0.0290
	(0.0385)	(0.0320)	(0.0473)	(0.0454)
Panel (b): Including personality				
Time (First Run)	0.0108	0.0074	-0.0125	-0.0057
	(0.0074)	(0.0068)	(0.0089)	(0.0107)
Agreeableness	0.0196	0.0208	-0.0193	-0.0210
	(0.0151)	(0.0187)	(0.0215)	(0.0214)
Conscientiousness	-0.0185	0.0349**	-0.0315*	0.0150
	(0.0164)	(0.0149)	(0.0182)	(0.0175)
Extraversion	0.0416***	0.0169	0.0085	-0.0670***
	(0.0121)	(0.0207)	(0.0160)	(0.0153)
Openness to Experience	-0.0121	0.0056	-0.0056	0.0121
	(0.0130)	(0.0135)	(0.0212)	(0.0203)
Neuroticism	-0.0039	0.0139	0.0064	-0.0164
	(0.0223)	(0.0195)	(0.0218)	(0.0248)
Locus of Control	-0.0417**	-0.0270	0.0259	0.0428*
	(0.0179)	(0.0186)	(0.0205)	(0.0252)
Social Comparison	0.0030	-0.0099	0.0151	-0.0083
	(0.0183)	(0.0183)	(0.0224)	(0.0186)
Competitiveness	0.0228	-0.0222	-0.0494**	0.0487*
	(0.0179)	(0.0213)	(0.0244)	(0.0280)
Risk Attitudes	0.0015	0.0136	-0.0074	-0.0077
	(0.0162)	(0.0202)	(0.0186)	(0.0205)
Female	0.0257	-0.0255	-0.0026	0.0024
	(0.0500)	(0.0402)	(0.0541)	(0.0556)
Observations	623	623	623	623
Class Fixed Effects	Yes	Yes	Yes	Yes
Baseline Probability	.15	.21	.29	.35

Notes: This table presents marginal effects from multinomial logistic regressions using indicators for four categories of performance-based preferences as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on class-level.

Appendix 6.B Relationship of actual performance and beliefs

In this section, we first describe the relationship between beliefs and actual performance. Afterwards, we provide evidence that the beliefs are meaningful, that is they are 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 6.B.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 6.B.1. Figure 6.B.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.



(a) Relationship of beliefs and actual performance

(b) Abs. Diff. in beliefs and actual performance

Figure 6.B.1. Relationship of actual performance and beliefs

Notes: Figure (a) presents the relationship beliefs over and actual relative performance of the name-based peers. The corresponding regression is presented in Table 6.B.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). Intervals used here and in the survey are one second intervals of relative times in the first run.

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

Table 6.B.1. Relationship between beliefs over and actual relative performance

	(a) Peer's relative time		(b) Peer is faster (binary)	
	(1)	(2)	(3)	(4)
Relative Time of most-preferred name-based peer	0.24*** (0.03)	0.24*** (0.04)		
Preferred name-based peer is faster			0.26*** (0.04)	0.25*** (0.05)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
N	662	562	662	562
R^2	.21	.23	.16	.17

Notes: This table presents least squares regressions using a peer's relative time according to the beliefs of the name-based preferences as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on class-level. Figure 6.B.1 presents the results graphically.

columns of Table 6.B.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., of whom we have beliefs) and that are matched as a peer in the second run (i.e., as only for those 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 6.B.2. Consistency of beliefs

	(a) Continuous belief		(b) Binary belief	
	(1)	(2)	(3)	(4)
Believe peer is faster	1.96*** (0.23)		0.58*** (0.05)	
RANDOM × Believe peer is faster		2.00*** (0.27)		0.53*** (0.06)
NAME × Believe peer is faster		1.92*** (0.23)		0.59*** (0.05)
PERFORMANCE × Believe peer is faster		2.01*** (0.23)		0.58*** (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. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on class-level. 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.

Appendix 6.C Additional material for name-based preferences

This table checks the robustness of the findings from Table 6.4 using absolute difference of the beliefs over the times in the first run rather than actual absolute differences. This reduces the number of observations because we need to restrict our sample to individuals that are nominated in the name-based preferences.

Table 6.C.1. Linear probability model of nominating a peer

	Nominated Peer								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Abs. Diff. in Beliefs over Times in First Run		-0.0385**	-0.0448***	-0.0443***	-0.0061				
		(0.0165)	(0.0153)	(0.0162)	(0.0180)				
Abs. Diff. from Perf.-based Preference (using beliefs)					-0.0794***				
					(0.0162)				
Abs. Diff. in Time of First Run							-0.0566***	-0.0580***	-0.0178
							(0.0148)	(0.0156)	(0.0188)
Abs. Diff. from Perf.-based Preference									-0.0465***
									(0.0137)
Homophily in Personality									
Abs. Diff. in Agreeableness	-0.0380***	-0.0433*	-0.0505**	-0.0551**	-0.0493**	-0.0617**	-0.0665**	-0.0713**	-0.0712**
	(0.0091)	(0.0222)	(0.0222)	(0.0229)	(0.0239)	(0.0294)	(0.0278)	(0.0300)	(0.0301)
Abs. Diff. in Conscientiousness	-0.0125	-0.0169	0.0046	0.0013	0.0042	-0.0188	-0.0119	-0.0120	-0.0131
	(0.0124)	(0.0241)	(0.0208)	(0.0209)	(0.0219)	(0.0251)	(0.0244)	(0.0247)	(0.0240)
Abs. Diff. in Extraversion	-0.0265**	-0.0279	-0.0092	-0.0111	-0.0130	-0.0179	-0.0281	-0.0332	-0.0329
	(0.0126)	(0.0245)	(0.0251)	(0.0248)	(0.0238)	(0.0300)	(0.0329)	(0.0323)	(0.0317)
Abs. Diff. in Openness	-0.0050	0.0059	0.0064	0.0052	0.0065	-0.0139	-0.0211	-0.0243	-0.0216
	(0.0128)	(0.0287)	(0.0246)	(0.0253)	(0.0284)	(0.0318)	(0.0339)	(0.0326)	(0.0327)
Abs. Diff. in Neuroticism	-0.0148	-0.0480	-0.0405	-0.0451	-0.0443*	-0.0279	-0.0443	-0.0498*	-0.0501*
	(0.0119)	(0.0299)	(0.0290)	(0.0281)	(0.0259)	(0.0279)	(0.0289)	(0.0285)	(0.0284)
Abs. Diff. in Locus of Control	-0.0108	-0.0245	-0.0204	-0.0203	-0.0211	-0.0133	-0.0158	-0.0154	-0.0152
	(0.0116)	(0.0260)	(0.0281)	(0.0283)	(0.0284)	(0.0264)	(0.0261)	(0.0269)	(0.0271)
Abs. Diff. in Social Comparison	-0.0347***	-0.0688**	-0.0488**	-0.0512**	-0.0499**	-0.0561**	-0.0525*	-0.0561**	-0.0542**
	(0.0108)	(0.0255)	(0.0218)	(0.0219)	(0.0216)	(0.0272)	(0.0266)	(0.0256)	(0.0253)
Abs. Diff. in Competitiveness	-0.0374**	-0.0438	-0.0398	-0.0355	-0.0310	-0.0499	-0.0406	-0.0385	-0.0399
	(0.0144)	(0.0371)	(0.0307)	(0.0307)	(0.0300)	(0.0305)	(0.0277)	(0.0275)	(0.0263)
Abs. Diff. in Risk Preferences	-0.0133	0.0122	0.0140	0.0194	0.0140	-0.0084	-0.0232	-0.0151	-0.0202
	(0.0158)	(0.0296)	(0.0307)	(0.0325)	(0.0306)	(0.0293)	(0.0289)	(0.0308)	(0.0306)
Friendship Ties									
Friendship Indicator			0.4329***	0.4196***	0.4120***				
			(0.0379)	(0.0384)	(0.0374)				
Homophily in Network Characteristics									
Abs. Diff. in Degree				-0.0074	-0.0071			-0.0006	0.0013
				(0.0250)	(0.0236)			(0.0190)	(0.0184)
Abs. Diff. in Eigencentrality (std.)				-0.0410	-0.0374			-0.0800	-0.0811
				(0.0457)	(0.0466)			(0.0570)	(0.0576)
Abs. Diff. in Clustering (std.)				-0.0110	-0.0040			-0.0957**	-0.0971**
				(0.0479)	(0.0449)			(0.0422)	(0.0418)
Own Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Peer Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Using Beliefs	No	Yes	Yes	Yes	Yes	No	No	No	No
Using Friends only	No	No	No	No	No	Yes	Yes	Yes	Yes
N	7084	2957	2957	2920	2920	3152	2934	2894	2894
R^2	0.20	0.68	0.71	0.71	0.72	0.64	0.66	0.67	0.67

Notes: *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses. Linear probability model using being nominated as one of the three most-preferred name-based peers as an outcome.

Appendix 6.D Additional material for relationship of preferences

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

$$p \lim \hat{\beta} \;=\; \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2}\beta \;=\; \lambda\beta \;=\; \lambda \tag{6.D.1}$$

as $\beta = 1$ and where λ is the attenuation factor.[1] Thus the regression coefficients in Table 6.D.1 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 6.5 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.

[1] For the multivariate case the formula is slightly different, but the basic idea remains the same.

Table 6.D.1. Relationship between performance- and name-based preferences

	(a) Peer's relative time		(b) Peer is faster (binary)	
	(1)	(2)	(3)	(4)
Panel A: Perf.-based pref. and name-based beliefs				
Belief over name-based peer's performance	0.43*** (0.06)	0.44*** (0.06)		
Belief over name-based peer's performance (0/1)			0.29*** (0.03)	0.29*** (0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
N	781	648	781	648
R^2	.23	.28	.17	.2
Noise-to-Signal Ratio needed for $\beta = 1$	1.3	1.3	2.4	2.5
Panel B: Perf.-based pref. and name-based actual performance				
Relative Time of most-preferred name-based peer	0.09*** (0.03)	0.09*** (0.03)		
Preferred name-based peer is faster			0.04 (0.03)	0.03 (0.04)
Personality	No	Yes	No	Yes
Class FEs, Gender, Age	Yes	Yes	Yes	Yes
N	662	562	662	562
R^2	.11	.13	.095	.12
Noise-to-Signal Ratio needed for $\beta = 1$	9.7	10	25	28

Notes: This table presents least squares regressions using a peer's relative time in one second intervals or an indicator for preferring a faster peer according to the performance-based preferences as the dependent variable. *, **, and *** denote significance at the 10, 5, and 1 percent level. Standard errors in parentheses and clustered on class-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 ($\beta < 1$). That is, 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 6.4 presents the results graphically.