Peer Effects Under Different Relative Performance Feedback and Grouping Procedures

Kathrin Thiemann Niklas Wallmeier

January 25, 2018

Abstract

We conduct a laboratory experiment to test theoretical predictions about subjects' performance in an effort task conditional on their peer group's composition and relative performance feedback. Subjects are grouped either randomly or according to their ability, with the feedback being the maximum or average performance of their group. We are able to support theory-driven hypotheses on the influence of ability, competitiveness and loss aversion on optimal performance in the presence of peer effects. While random grouping is beneficial for male subjects it is detrimental for female subjects. With respect to the reference point we find male subjects performing significantly better when they compare themselves to the best peer instead of the average, while the opposite is true for females.

1 Introduction

In many areas of life, individuals that perform on a certain task find their performance evaluated relatively to that of other individuals of a reference group. For instance, in firms, employees can observe the performance of their team members, or get feedback on the "best salesperson of the month". Assuming that individuals have reference dependent preferences (see e.g. Kahneman and Tversky, 1979), with the relative performance feedback being the reference point, individual performance consequently depends on the kind of relative performance feedback and on the composition of the reference group. The kind of feedback that is actively provided by firms and institutions might vary with the organization's philosophy. For example, a firm might actively highlight only the top performers in order to motivate employees to perform better. The reference point can also vary with culture as acquired by groups of people that share a religion or ethnic origin. In more competitive cultures individuals are expected to compete for the top po-

sitions. Opposite, in less competitive cultures, social comparison plays a less emphasized role and individuals are expected to conform to the average. Therefore, the question arises whether group composition and performance feedback can be optimized in order to maximize group performance.

Thiemann (2017) addresses this issue theoretically, focusing on the question whether ability segregated classes (also referred to as *ability tracking* or *ability grouping*) or classes with heterogeneous ability students are to be preferred. Theory predicts that it depends on the culture of competitiveness of the student body, i.e. on the kind of the reference point and the importance of social comparison. The intuition is that a comprehensive school, i.e. a class with heterogeneous students, yields optimal incentives for highly competitive individuals, who want to be the best student in class. Here also subjects with very low ability are motivated by the best student and they exert effort in order to minimize the performance distance. In a system with ability grouping, where high-ability subjects are sorted into a high track and low-ability subjects into a low track, the low-ability students can only compare with the top performer in their class, which is less motivating. When students are less competitive and only compare their performance to that of the average student, the model predicts ability grouping to be optimal. This is driven by stronger motivation in a high-ability group due to the higher reference point compared to a more heterogeneous group. This effect may on average outweigh the negative effect of ability grouping for low-ability students.

In our experiment we also want to test whether gender plays a role for the optimal performance feedback and grouping policy, something not considered in Thiemann (2017). We believe that gender might be of importance, since women and men have been found to differ to a huge extent in their preferences for competitiveness (see e.g. Gneezy, Niederle, and Rustichini, 2003; Niederle and Vesterlund, 2007; Niederle, 2016). In our framework high reference points and pressure for social comparison will create a competitive environment and might cause different effort choices of male and female subjects. The existing research generally finds that men perform better in competitive environments (e.g. tournaments), whereas women's performance does not change in a tournament-based compensation scheme compared to a piece rate.

Hypotheses derived from Thiemann (2017) combined with gender differences in preferences for competition are tested in a laboratory experiment where subjects perform a real effort task and earn a piece rate. Subjects are grouped either randomly or according to ability. As performance feedback they receive either the best or the average performance of their group. We find support for subjects behaving according to the model prediction of optimal performance. While we cannot confirm hypotheses on aggregated treatment

effects, we find significant gender differences in the reaction to different reference points and grouping procedures. On the subject level regression analysis suggests that incentives differ conditional on whether the best or the average performance is available. These peer effects with respect to the reference point seem to be non-linear.

Our study contributes to two existing fields of economic literature. First, it is settled in the field of empirical evidence on peer effects (For an overview see Herbst and Mas, 2015). Of particular relevance to this study, Kuhnen and Tymula (2012) find the mere possibility of being evaluated relative to peers as performance enhancing. According to Beugnot, Fortin, Lacroix, and Villeval (2013) this performance-enhancing effect of relative performance feedback is larger for male than for female subjects. Furthermore, Gill and Prowse (2012) find that loss-averse individuals respond negatively to a rival's effort in a sequential-move tournament and Gill, Kissová, Lee, and Prowse (rthc), who conduct a real-effort experiment with rank-order feedback, find that peer effects are non-linear in the distance between a subject's performance and the reference point. In particular they find evidence for "last-place loathing" and "first-place loving". While these studies focus on a particular performance feedback, we contrast the effects of different relative performance feedback: the average peer achievement and the best peer performance.

Second, our study contributes to the literature that addresses the effect of grouping individuals according to their ability. These effects can arise from mutual learning or norm setting within the group. The latter corresponds to the pure peer effect analyzed in lab experiments. A number of field studies have analyzed the influence of ability tracking on student performance in school (e.g. see surveys by Slavin, 1990; Meier and Schütz, 2007). Effects of ability tracking on mean achievement are usually low and non-significant. Studies usually find that tracking harms low-ability students but benefits high-ability students (e.g. Argys, Rees, and Brewer, 1996; Duflo, Dupas, and Kremer, 2011). While the above mentioned field studies cannot disentangle whether different group compositions affect performance through mutual learning or through different group norms, our laboratory study can exclude mutual learning effects and focus on the latter.

2 Theory

In line with Thiemann (2017) we assume that subjects in our experiment maximize utility by choosing an effort level. Assume that effort translates linearly into performance and that subjects have reference-dependent preferences as in Kahneman and Tversky (1979) with relative performance feedback being the reference point. Subjects face the following optimization problem:

$$\underset{p_i}{\text{Max}}\ u_i(p_i) = (1-s)p_i + s \cdot v(p_i - r_i) - c(p_i, a) \tag{1}$$

$$\text{with} \quad v(p_i - r_i) = \begin{cases} \lambda \cdot (p_i - r_i) & if\ p_i < r_i \\ (p_i - r_i) & if\ p_i \geq r_i \end{cases} \tag{2}$$

$$\text{and} \quad c(p_i, a) = \frac{p_i^2}{2a} \tag{3}$$

Performance p_i is the number of correctly answered multiplication problems per period. Before each period, each subject is shown a reference point r_i, that yields information about the performance of the group members. Subjects' utility depends on a direct private component of utility and a comparison oriented component given by the value function $v(\cdot)$. In the experiment the direct private utility from performance is given because of direct remuneration of performance. The utility from the comparison oriented component is assumed to be larger the more competitive a subject is (s, with $s \geq 0$ is the degree of social comparison). For subjects performing below the reference point, the disutility from the difference to the reference performance is increasing with loss aversion, λ, with $\lambda > 1$. The cost of performance $c(p_i, a)$ increases in performance and decreases with ability a. A subject's optimal performance is then given by the following best response function:[1]

$$BR_i(r_i) = \begin{cases} (1 - s + \lambda s)a & if \quad p_i < r_i \\ a & if \quad p_i > r_i \end{cases} \tag{4}$$

Optimal performance depends positively on ability a. If the subject's performance is below the reference point, performance also depends positively on loss aversion (λ) and competitiveness (s).

The derived best response function is the basis to compare equilibrium performances across different regimes. First, we compare performances for different reference points: the *average* performance among the other group members and the *best* performance among the other group members. Second, we compare a regime where subjects are randomly grouped with a regime, where subjects are grouped according to ability. In the latter we have groups consisting only of low-ability subjects and groups only with high-ability subjects. We follow the theoretical analysis of Thiemann (2017), where proof is found for

[1]For simplification we ignore the case where $p_i = r_i$. See Thiemann (2017) for the full solution.

four main hypotheses:

H1 *When the best reference point is given, average performance is higher under random grouping than under ability grouping.*

H2 *When the average reference point is given, average performance is higher under ability grouping than under random grouping, if s and λ are sufficiently low.*

H3 *Low-ability individuals always perform lower under ability grouping than under random grouping.*

H4 *High-ability individuals benefit from ability grouping when the average reference point is given, and are not affected when the best reference point is given.*

In addition to the above mentioned hypotheses we also want to investigate the role of gender. Thiemann (2017) assumes that all individuals in a group have the same degree of social comparison and loss aversion. Past research, however, has shown that preferences for competitiveness differ to a high extent with gender (e.g. Niederle and Vesterlund, 2007). The degree of social comparison, the s in our framework, might therefore be higher for male subjects than for female subjects. From 4 we can then derive ***H5***:

H5 *Average performance is higher when the best reference point is given both for males and females, but the difference for males is larger.*

In terms of grouping regimes ***H1*** should still hold both for female and male students, since the comparison between random and ability grouping under a best reference point does not depend on the level of s and λ (Thiemann, 2017). However, when the average reference point is given, random grouping is only beneficial when s and λ are low. This leads to the following alternative hypothesis to ***H2***:

H2a *When the average reference point is given, female (male) subjects have a higher average performance under ability grouping (random grouping) than under random grouping (ability grouping).*

3 Experimental Design

3.1 Effort Task

Subjects were asked to solve as many multiplication problems as possible in five periods of four minutes each. In particular we asked subjects to multiply one-digit numbers (3-9) with two-digit numbers (11-99)(see Dohmen and Falk, 2011). By remunerating subjects with a piece rate per solved problem, they were linearly incentivized. Every subject was given the same problems in the same order to ensure that the difficulty of the problems was identical. Problems were purposefully designed such that the difficulty would vary

to the same extent within each period. In case subjects answered a problem incorrectly, the screen reported "false" and subjects had to repeat it instead of searching for easy problems.[2]

Multiplication problems were chosen as an effort task to ensure that performances during the experiment depend both on ability and effort. On the one hand the given task is a good proxy for cognitive ability and generates heterogeneous outputs that allow for grouping according to ability. On the other hand the task offers sufficient scope to vary effort, since solving the problems needs high concentration and is thus costly.

3.2 Treatments and Procedural Details

In order to test hypotheses ***H1*** and ***H2*** we implement a two-by-two design to compare mean group performances along the two major treatments: *best* vs. *average* reference point and *ability grouping* vs. *random grouping*. To test hypotheses ***H3*** and ***H4*** we will compare low and high-ability subjects between these four main groups. In addition, we have a baseline treatment that is used to group subjects according to ability. Subsequent to the experiment we measure individual loss aversion and competitiveness by survey questions in order to test the theoretical optimal performance.

(a) Baseline Treatment All subjects participated in the baseline treatment, taking place in the first period. They did not receive any information on other subjects' performance and were neither sorted into groups. They only received information on their own total number of solved tasks after the period.

(b) Best vs. Average Treatment The *best* vs. *average* treatments are modeled in a between-subject design, i.e. subjects are either shown the *best* reference point or the *average* throughout the session. Thereby we avoid a demand effect that could arise, if subjects are offered two different reference points subsequently. During the experiment subjects are sorted into groups of five. These groups serve the only purpose of providing the reference point. In the *best* treatment we provide subjects with information on the *best* performance of their group after every period. If the subject herself had the best performance we gave information on the second best performance. The subjects from the *average* treatment were given information about the *average* performance of their group, excluding the subject's own performance.

[2]For further details see Appendix 1.

(c) Ability Grouped vs. Randomly Grouped Treatment The grouping treatments are modeled in a within-subject design. All subjects went through two periods of the *randomly grouped* treatment and through two periods of the *ability grouped* treatment. We implemented a crossover design to account for ordering as well as learning effects. In the *randomly grouped* treatment subjects were randomly grouped with other subjects. This resulted in groups of subjects with more heterogeneous abilities. For the *ability grouped* treatment subjects were ranked according to their performance in the first period (*baseline*). All subjects that performed in the top 50% were sorted into a high track (high-ability type), and those that performed in the bottom 50% were sorted into a low track (low-ability type). Groups under the *ability grouped* treatment were then only randomly composed of subjects within these tracks. This resulted in groups of rather homogenous abilities.

Table 1: Session Designs

Reference Point: Average			**Reference Point: Best**		
(1)			(2)		
baseline →	random grouping →	ability grouping	baseline →	ability grouping →	random grouping
(1 period)	(2 periods)	(2 periods)	(1 period)	(2 periods)	(2 periods)
(3)			(4)		
baseline →	ability grouping →	random grouping	baseline →	random grouping →	ability grouping
(1 period)	(2 periods)	(2 periods)	(1 period)	(2 periods)	(2 periods)

Table 1 illustrates the composition of the sessions with respect to the reference point and the ordering of the grouping procedure. Altering the two possible reference point frameworks and switching the order of the grouping treatments allows observing all four possible setups. The crossover design with respect to the grouping treatments has two crucial advantages. First, we are able to deal with potential order effects. i.e. biases from being grouped by ability first and randomly later. In addition, we can also disentangle potential learning effects from treatment effects.

The experiment was programmed with zTree (Fischbacher, 2007) and four sessions with a total of 120 participants were conducted at the experimental laboratory of the University of Hamburg in June and July 2015. We used hroot for recruitment (Bock, Baetge, and Nicklisch, 2014). The subjects were students of the University of Hamburg of which 58 were female. One correct answer in the relevant periods was exchanged for 30 euro cent. On average, a participant received a payout of 14 euro, including the show up fee of 5 euro. The sessions took about 60 minutes each.

4 Results

4.1 Summary Statistics and Prima Facie Evidence

First, we highlight the data on the aggregate level. Performance is characterized under different grouping regimes and reference point settings. We test theoretical predictions on aggregate outcome[3] , before the analysis focuses on the individual level.

The distribution of output over the entire experiment shows quite heterogeneous performance in the effort task. Output has a range from no correct answer up to a total of 60 correctly solved multiplications with a mean of 21.4. It can be taken from Figure 1 (a) and (b) that performance is positively skewed around the median of 20.

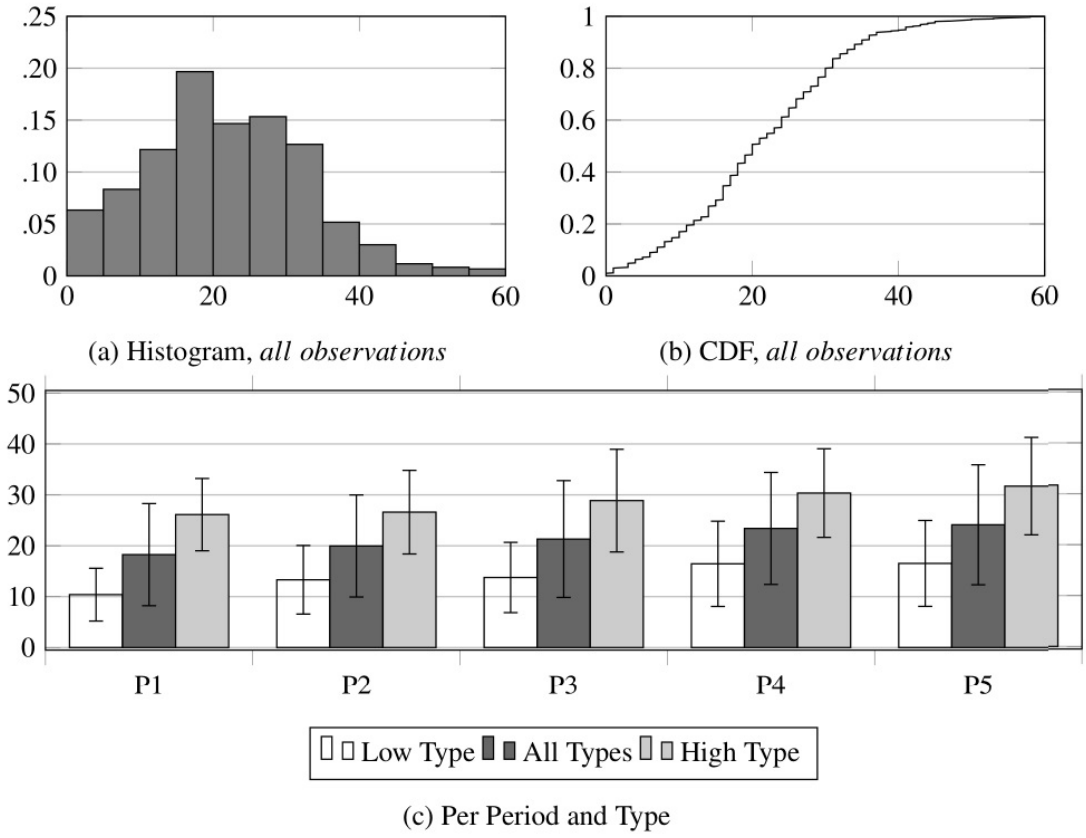

Figure 1: Distribution of Correct Answers

Figure 1 (c) illustrates mean performance and its standard deviation per period and type. The results for period one show substantial differences in mean performance of those subjects who perform above the 50th percentile (26.1) compared to those whose output is below the 50th percentile (10.4). Therefore, we argue that grouping subjects

[3]We use a Wilcoxon signed-ranks test for within-subject differences and a Mann-Whitney-U test for between-subject differences, respectively. Since each individual is observed twice in a treatment, we take the average of a subject over the two periods as an observational unit.

according to ability based on this first round performance is reasonable. Across periods, mean performance increases steadily (from 18.3 to 24.1, dark gray bars), indicating that subjects improve over time independently of the treatment. Further, evaluating learning separately for high-ability (light gray bars) and low-ability subjects (white bars), suggests that the improvement is similar for both types. The difference between the two types remains in the range between 13.3 and 15.8.

By contrasting the mean performance of the two grouping scenarios under a given reference point we test hypotheses ***H1*** and ***H2***. Figure 2 displays the mean outcome and standard deviation for both random (RG) and ability grouping (AG) given average group performance as reference point (AVRG) on the left-hand side, and the best group performance as reference point (BEST) on the right-hand side. Evaluating performance of all subjects (dark gray bars) under the *best* setting suggests that our experiment cannot confirm hypothesis ***H1*** (RG $\approx$ AG = 22.7). Also with respect to hypothesis ***H2***, we do not find a significant difference in performance under the *average* setting (21.8 vs. 21.7).

To investigate hypotheses ***H3*** and ***H4***, we compare the mean performances separately for high-ability subjects (light gray bars in Figure 2) and low-ability subjects (white bars). Hypothesis ***H3*** predicts a generally lower mean for low-ability subjects in an ability grouped setting compared to random grouping. This can neither be supported for the *best* setting (RG $\approx$ AG = 15.2), nor the *average* setting (RG: 15.2 vs. AG: 14.7) on the aggregate level. From hypothesis ***H4*** we expect an output-enhancing effect from ability grouping for high-ability subjects given average group performance as reference point. However, Figure 2 depicts the mean performance of high-ability subjects in the *average* setting as not significantly different across the grouping treatments (RG: 28.8 vs. AG: 28.4).

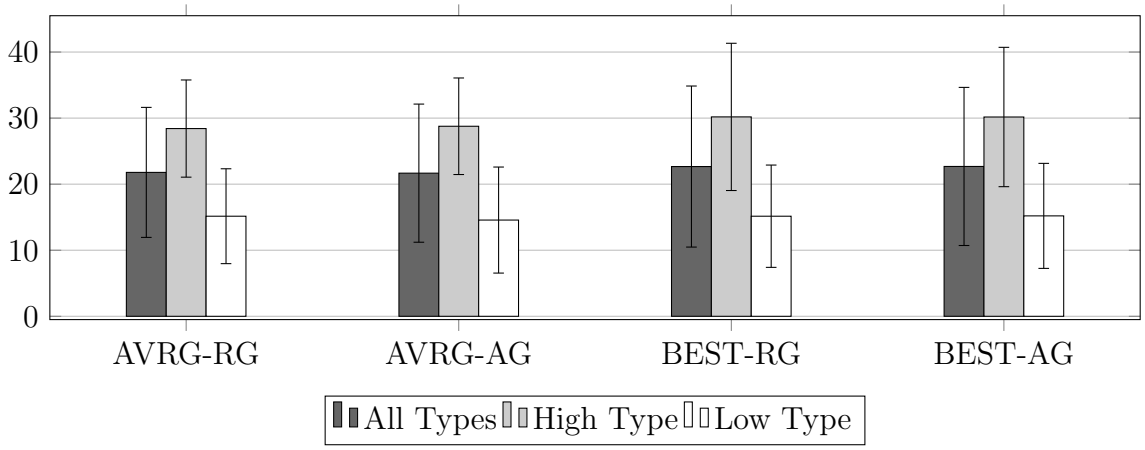

Figure 2: Performance per Reference Point and Grouping Treatment

This lack of support for the theoretical predictions on the aggregate level might result from systematic differences in performance by gender. From a question in the survey after our experiment in which participants had to choose between a tournament-based compensation scheme and a piece rate, we see that 24% of the female subjects chose the competitive alternative vs. 30% of male subjects.[4]

With evidence for a heterogeneous degree of social comparison between male and female subjects we turn the analysis to the hypotheses that predict differing behavior with respect to gender. In Figure 3 we plot average performance of male and female subjects under the average reference point regime and the best reference point regime. Comparing the respective solid lines to the dashed ones shows that male subjects perform mildly significantly better in the *best* treatment than in the *average* treatment (best: 26.5, avrg: 21.3, $p < 0.09$)[5], while female subjects perform higher in the *average* treatment, without significance (best: 19.7, avrg: 22.4, $p > 0.10$).[6] This confirms ***H5*** for male subjects benefiting more strongly from a competitive reference point than females.

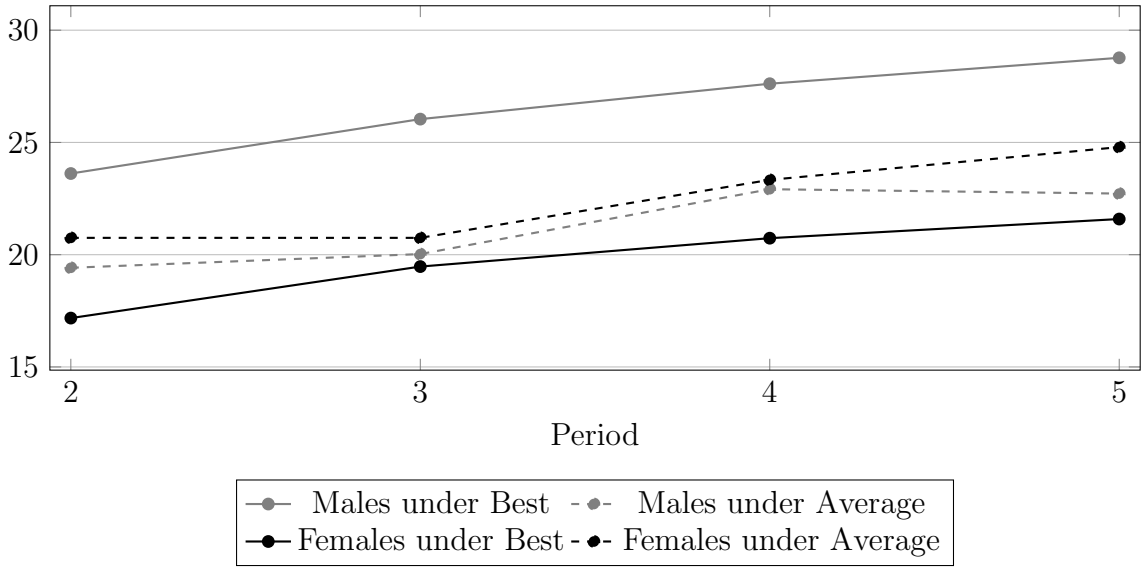

Figure 3: Average Performance by Reference Point Treatment and Gender over Time

To test hypothesis ***H2a*** we compare average performance by gender and reference point under the two grouping regimes. Figure 4 suggests that both grouping procedures have

[4]This choice was without monetary incentives. In terms of ability (measured by the math grade) and loss aversion, that was also elicited after the experiment, men and women do not differ substantially (see Appendix C).

[5]Since each individual is observed four times in a treatment, we take the average of a subject over the four periods as an observational unit and use the "bootstrap" technique of the two-sample t-test to calculate p-values (Efron and Tibshirani, 1993).

[6]Theory predicts an increasing effect also for female subjects. A possible explanation might be found in competition-aversion of females causing an adverse reaction to a competitive reference point (e.g. Niederle and Vesterlund, 2007).

an effect on performance, but differently by gender. Overall we find a weakly significant difference between the two grouping procedures for women (AG: 21.5, RG: 20.2, $p < 0.08$). The opposite is true for male subjects, who on average perform significantly better under *random grouping* (AG: 22.8, RG: 24.1, $p < 0.04$). Splitting this up by reference point regime, we find that there are no significant differences between random grouping and ability grouping under the best reference point. When the average reference point is given, i.e. testing ***H2a***, we find that female subjects perform indeed higher under ability grouping (AG: 23.2, RG: 21.6, $p > 0.10$) and male subjects under random grouping (AG: 20.6, RG: 21.9, $p < 0.05$), with only the latter being significant.

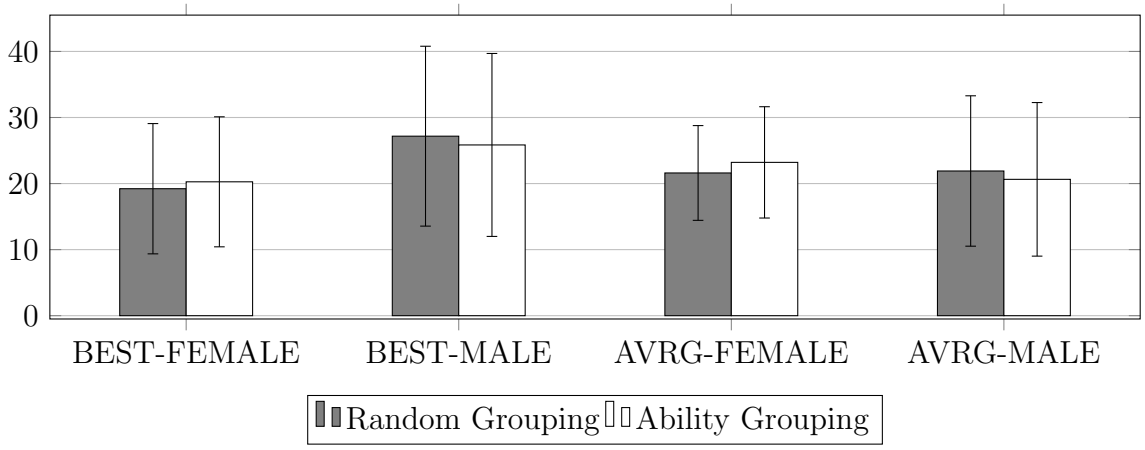

Figure 4: Average Performance under Random and Ability Grouping by Gender and Reference Point Treatment

Overall, we find significant results only when we acknowledge that male and female subjects differ in their preferences for social comparison. The first main result so far is that the performance of men is higher when the best reference point is given compared to an average reference point, whereas no significant difference is found for women (supporting hypothesis ***H5***). The second main result is that women perform better under ability grouping and men under random grouping, specifically when the average reference point is given (supporting hypothesis ***H2a***).

4.2 Testing Optimal Performance

The hypotheses tested in Section 4.1 were derived from the theoretical optimal performance as given in Section 2. Whether individual subjects behave according to the derived best response function can be tested directly in a system of regressions. If subjects behave optimally, their performance should depend positively on ability (a_i) and competitiveness ($comp_i$). If the subject's performance is below the reference point, performance should

also increase with the degree of loss aversion ($lossavers_i$).

The dependent variable is performance of subject i in period t. The first regression only includes subjects that performed below the average (or best) performance of their current group members in the last period. The second regression includes those that performed above. The three covariates of interest are derived from questions that subjects answered in the questionnaire subsequent to the experiment. Estimated coefficients of loss aversion (elicited by a method developed by Abdellaoui, Bleichrodt, and Haridon (2008)) had a mean of 3 and a standard deviation of about 3.5. As a control for ability we asked subjects for their last math grade at school (ranging from 1-6, with 1 being the best grade). The regression also includes period and session dummies (μ_t) to control for period and session specific effects, especially for learning effects. Results of Ordinary Least Squares (OLS) regressions with standard errors clustered at the individual level to control for serial correlation in the error term are reported in Table 2, separately for the *best* and the *average* treatment.

Table 2: Testing Theory-Derived Optimal Performance

Variables	Average		Best	
	Below (1)	Above (2)	Below (3)	Above (4)
Loss Aversion	0.458**	0.138	0.980**	-0.418
	(0.180)	(0.166)	(0.415)	(0.691)
Competitiveness	-3.145	3.859	-0.995	9.896*
	(3.162)	(2.600)	(3.392)	(5.534)
Math Grade	-2.084**	-1.346	-0.185	-7.832***
	(0.874)	(1.248)	(1.288)	(1.684)
Constant	20.630***	25.125***	18.343***	44.431***
	(3.416)	(3.763)	(4.722)	(6.595)
Period FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
R^2	0.22	0.20	0.13	0.56
Adj. R^2	0.15	0.13	0.08	0.46
N	85	91	130	38

Notes: Ordinary least squares regressions. Dependent variable: Number of correct answers. Regressions include periods 2-5. Robust standard errors in paranthesis are clustered at the individual level. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

As predicted the coefficient of loss aversion has a positive and significant impact on performance only for subjects whose past performance was below the reference point both in the *best* and the *average* treatment. Precisely, for subjects below the reference point in the *average* treatment an increase of the coefficient of loss aversion by 1 induces on average an increase of solved tasks by 0.5 and almost by 1 in the *best* treatment. The indicator for competitiveness has no positive impact on performance. Taken altogether,

especially the estimates for loss aversion that drives performance below the reference point can confirm the theoretical prediction.

4.3 Linear Peer Effects

In the preceding section we have shown that performance increases in loss aversion if the subject's performance is below the reference point. Here we estimate the size of the average effect of the reference point on performance. Typically these *peer effects* are empirically modeled by the linear-in-means-model, meaning that performance of a single subject is regressed on the average performance of the subjects' reference group (see e.g. Brock and Durlauf, 2001). We proceed in this way for the *average* treatment, while for the *best* treatment we regress individual performance on the best performance of each group. The following regression with period fixed effects μ_t and covariates $\mathbf{X}_i$ is estimated separately for the *best* and *average* treatment.

$$p_{it} = \alpha + \beta \; refpoint_{it} + \mathbf{X}_i\gamma + \mu_t + \epsilon_i \tag{5}$$

The variable $refpoint$ is the average (best) performance of the current group members from the last period that was shown to the subjects before each period. If performance below the reference point increases linearly in loss aversion, the size of the peer effect should be larger in the *best* treatment than in the *average* treatment. The way in which subjects react to a reference point should strongly depend on subject specific characteristics, as suggested by theory e.g. on factors like loss aversion, competitiveness and ability. These factors again might vary, for instance, with the cultural background or the gender of the individual subject.

Thus, we estimate a model that only includes $refpoint$ as a first step. The estimated coefficient gives the total impact of the reference point on performance, including any effect that might work through different subject characteristics such as culture, gender or ability. In a second step we include control variables for subject background factors gathered in the questionnaire subsequent to the experiment to see how this changes the impact of the reference point (these are: female, years since Abitur[7], studies math[8], income[9]). To analyze which factors drive the sensitivity to the reference point, we include

[7] *Abitur* is the name of the diploma awarded to students at the end of secondary schooling in Germany.

[8] The variable *studies math* is a dummy that takes on the value 1 if the subject studies a course that includes mathematics as a major component, such as information systems, economics, business, physics or mathematics.

[9] The variable *income* is an ordered categorical variable taking on the following values of disposable income per months (in Euros): 1 = less than 400, 2 = 400-600, 3 = 600-800, 4 = 800-1000, 5 = 1000-1200, 6 = more than 1200.

Table 3: Linear Peer Effects

Variables	Average		Best	
	(1)	(2)	(3)	(4)
Reference Point	0.569***	0.475***	0.298***	0.216**
	(0.115)	(0.109)	(0.079)	(0.087)
Math Grade		-2.417***		-2.662**
		(0.816)		(1.181)
Female		-2.309		-7.178***
		(2.408)		(2.624)
Years since Abitur		-0.601		0.147
		(0.412)		(0.246)
Studies Math		1.353		9.086***
		(2.014)		(2.695)
Income		1.055		0.300
		(0.746)		(1.016)
Period FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
R^2	0.16	0.31	0.10	0.35
Adj. R^2	0.14	0.28	0.08	0.32
N	240	236	240	228

Notes: Dependent variable: Number of correct answers per period. Robust standard errors in paranthesis are clustered at the individual level. Regressions include period 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

some interactions of $refpoint$ with subject characteristics in a third step. We use an OLS approach with clustered standard errors at the individual level. We expect β to be positive in specifications (1), (2), (3) and (4).

From the results reported in Table 3 we see that in both treatments individual performance increases in the reference point. However, the effect is almost twice as large in the *average* treatment. When the reference point is one correct answer higher, individual performance increases on average by more than half a correct answer in the *average* treatment and only by 0.3 correct answers in the *best* treatment. In both treatments the impact of the reference point decreases once we control for subject characteristics, but it remains positive and significant. Including interactions does not shed any light on what drives the sensitivity to the reference points in the *best* treatment. In the *average* treatment, however, we find that subjects that have a better math grade and older subjects (subjects whose graduation from school is longer ago) react more strongly to the reference point. Unlike Beugnot, Fortin, Lacroix, and Villeval (2013) we find no difference in the reaction to reference points between male and female subjects. This effect might be taken up by the math grade, which is significantly better for female subjects (pairwise correlation: -0.128***).

4.4 Non-Linear Peer Effects

Unlike suggested by theory we have seen in the last section that an average reference point has a higher impact on individual performance than the best reference point. A reason for this could be nonlinear effects and diminishing sensitivity with respect to the reference point as suggested by Kahneman and Tversky (1979). The motivating effect of the reference point might become smaller the further away a subjects' performance is from the reference point. To find the effect of the distance to the reference point in our sample we use a differencing method, i.e. the dependent variable is the change in correctly answered problems compared to the period before. With this approach we can avoid multicollinearity of the subjects' performance and the distance to the reference point. We can also eliminate time-invariant factors like subject ability and concentrate on what causes the change in performance between periods. The following regression is estimated separately for the *best* and *average* treatment:

$$\begin{aligned}\Delta p_{it} =& \alpha + \beta_1 below_{it-1} + \beta_2 absdist_{it-1} + \beta_3 absdist_{it-1} \times below_{it-1} \\ &+ \beta_4 trackdec_{it} + \beta_5 trackdec_{it} \times lowtype_i + \mu_t + \mu_i + \Delta\epsilon_{it} \end{aligned} \tag{6}$$

The variable *absdist* is the absolute distance in points of the subjects last period performance to the reference point. The variable *below* indicates whether the subject had performed below the reference point in the last period. The only other thing that changes with t is that subjects are told before the *ability grouped* treatment whether they were sorted into the low or high track. This is controlled for by a dummy (*trackdec*). We also include an interaction of *trackdec* with *lowtype*, which indicates whether subjects were sorted into the low track. At the cost of explanatory power, we estimate fixed effects models with subject and period fixed effects to eliminate biases due to unobserved subject characteristics and learning effects.

To find proof of a peer effect that is larger below the reference point, we would expect $\beta_1 > 0$. In order to find support for diminishing sensitivity as suggested by Kahneman and Tversky (1979), we would expect $\beta_2 < 0$ and $\beta_2 + \beta_3 < 0$. Results are reported in Table 4.

Specification (1) shows that, while there is no increase in performance for subjects above the reference point (see constant), subjects who were told that they performed below the average improve their output by more than four in the following period. In contrast, no significant difference can be found for the *best* treatment. Since the output of those below the average performance is on average clearly lower than of those who only failed to make the top position, this result suggests that the effect not only depends on being below

Table 4: Effect of Distance to Reference Point

Variables	Average		Best	
	FE (1)	FE (2)	FE (3)	FE (4)
Below the Reference Point	4.381***	1.810	0.409	0.549
	(0.861)	(1.315)	(0.671)	(1.729)
Absolute Distance to the Reference Point		-0.261**		-0.180
		(0.110)		(0.181)
Absolute Distance to the Reference Point				
× Below the Reference Point		0.505***		0.359*
		(0.176)		(0.190)
Period of Tracking Decision	-0.089	-1.980*	-0.340	-2.950**
	(0.902)	(1.193)	(1.092)	(1.411)
Period of Tracking Decision				
× Low Type		3.138**		6.097***
		(1.535)		(1.898)
Constant	-0.046	1.457	1.243	-0.920
	(0.913)	(1.197)	(1.207)	(1.667)
Period FE	Yes	Yes	Yes	Yes
Subject FE	Yes	Yes	Yes	Yes
R^2	0.17	0.22	0.01	0.11
Adj. R^2	-0.13	-0.08		-0.23
N	240	240	240	240

Notes: Dependent variable: Change in performance compared to last period. Standard errors in paranthesis. Regressions include periods 2-5. Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

the reference, but also on the size of the gap. Including the variable on the distance, specification (2) shows that for the *average* treatment there is evidence for diminishing sensitivity above the reference point, but for increasing sensitivity below the reference point. Also in the *best* treatment we find weak evidence for increasing sensitivity with growing distance to the best performance (see specification (4)).

Evaluating the output subsequent to the tracking information, we find patterns that also have been found in previous literature (Kuhnen and Tymula, 2012; Gill, Kissová, Lee, and Prowse, rthc). Subjects that are told that they were sorted into the low track do significantly improve in the following period, especially in the *best* treatment.

5 Conclusion

We tested theoretical predictions about subjects' performance conditional on their peer group's composition and relative performance feedback in a laboratory experiment. Support is found for subjects behaving according to the theoretically derived optimal performance. While hypotheses on treatment differences cannot be confirmed on the aggregated level, we find evidence when gender differences are taken into account. Male subjects per-

form significantly better than women in response to the best performance. With respect to the grouping treatments we find a significant negative effect of random grouping for female subjects and a positive significant effect for male subjects. Considering that men are on average more competitive than women, this result supports the theoretical prediction that random grouping is beneficial when subjects have a competitive mindset and detrimental when subjects are non-competitive. In addition, the reaction to the reference point *does* depend on individually differing factors like loss aversion. Our findings imply that feedback technologies and grouping procedures should be purposefully designed with respect to the individuals' background.

6 Acknowledgements

We would like to thank Olaf Bock, Igor Legkiy, Berno Büchel, and Gerd Mühlheußer for helpful comments, the team of the WiSo experimental laboratory in Hamburg for its help conducting the experiment. The authors gratefully acknowledge the financial support of the WiSo graduate school in Hamburg.

References

Abdellaoui, M., H. Bleichrodt, and O. Haridon (2008). A tractable method to measure utility and loss aversion under prospect theory. *Journal of Risk and Uncertainty 36*(3), 245–266.

Argys, L. M., D. I. Rees, and D. J. Brewer (1996). Detracking america's schools: Equity at zero cost? *Journal of Policy Analysis and Management 15*(4), 623–645.

Beugnot, J., B. Fortin, G. Lacroix, and M. C. Villeval (2013). Social networks and peer effects at work. *IZA Discussion Paper 7521*.

Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg registration and organization online tool. *European Economic Review 71*, 117–120.

Brock, W. A. and S. N. Durlauf (2001). Interactions-based models. *Handbook of Econometrics 5*, 3297–3380.

Dohmen, T. and A. Falk (2011). Performance pay and multidimensional sorting: Productivity, preferences, and gender. *The American Economic Review 101*(2), 556–590.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review 101*(5), 1739–74.

Efron, B. and R. J. Tibshirani (1993). An introduction to the bootstrap: Monographs on statistics and applied probability, vol. 57. *New York and London: Chapman and Hall/CRC*.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Gill, D., Z. Kissová, J. Lee, and V. L. Prowse (forthc.). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science*.

Gill, D. and V. Prowse (2012). A structural analysis of disappointment aversion in a real effort competition. *The American Economic Review 102*(1), 469–503.

Gneezy, U., M. Niederle, and A. Rustichini (2003). Performance in competitive environments: Gender differences. *Qarterly Journal of Economics 118*(3), 1049–1074.

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

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263 – 291.

Kuhnen, C. M. and A. Tymula (2012). Feedback, self-esteem, and performance in organizations. *Management Science 58*(1), 94–113.

Meier, V. and G. Schütz (2007). The economics of tracking and non-tracking. *Ifo Working Paper 50*.

Niederle, M. (2016). Gender. In J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*, Volume 2. Princeton University Press.

Niederle, M. and L. Vesterlund (2007). Do women shy away from competition? Do men compete too much? *Quarterly Journal of Economics 122*(3), 1067–1101.

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research 60*(3), 471–499.

Thiemann, K. (2017). Ability tracking or comprehensive schooling? A theory on peer effects in competitive and non-competitive cultures. *Journal of Economic Behavior & Organization 137*, 214–231.

A Translated Instructions

Welcome to today's experiment!
Today you are taking part in an economic experiment. Please note, that from now on and during the whole experiment no communication is allowed. If you have any questions during the experiment, please raise your hand and one of the experimenters will come to your cabin. In this experiment you can earn money by solving multiplication tasks. To solve the tasks you are not allowed to use any helping device, in particular no paper, pencil, calculator or mobile telephone. If you use any such helping device, you will be immediately excluded from the experiment and will get no remuneration. This experiment consists of five multiplication periods of four minutes each (240 seconds). We ask you to solve as many multiplication tasks as possible in one period. The tasks always consist of the multiplication of a one-digit number and a two-digit number. A task will be displayed as long as you need to answer the task correctly. Your remaining time will be displayed at the top of the screen. At the end of the experiment one of the five periods will be randomly chosen for the remuneration. The number of correctly answered problems in that period will be converted into Euros according to the following exchange rate:

1 solved problem = 30 euro cent

In addition everyone receives 5 Euros for attendance. At the beginning of the experiment you will have the possibility to test the input-screen in a 30 seconds trial period. After going through the five multiplication periods, we ask you to fill in a short questionnaire. The experiment is divided into three parts. Part 1 consists of one of the above described multiplication periods.

[The order of the following two paragraphs was changed depending on the treatment]

Part 2 [3] consists of periods 2 and 3 [4 and 5]. Here, you will be randomly allocated to a group of five. Your identity will at no point be published to your group members. Before each period you will receive information about the average [best] performance (in correctly answered problems) of your group members in the last period.

Part 3 [2] consists of periods 4 and 5 [2 and 3]. Before period 4 [2] you will be sorted either into track 1 or track 2 based on your performance in part 1. All the participants that performed higher than the median performance in the first period are allocated to track 1. Every subject that performed below median performance is allocated to track 2. Within

these tracks again groups of five will be formed randomly before each period. At the beginning of part 3 [2] you will be told into which track you have been sorted. In addition you will again be informed before each period about the average [best] performance of your group members.

If you have questions about these instructions, please raise your hand out of your cabin. One of the experimenters will come to you.

Good luck!

B Questionnaire

1. How old are you? _____
2. What is your sex? □ Male □ Female
3. What are you studying? _____
4. What was your last math grade at your last school? □ 1 □ 2 □ 3 □ 4 □ 5 □ 6
5. When did you graduate from secondary school? _____
6. How much money do you have at your disposal per month? (including rent) □up to 400 Euro □ 400-600 Euro □ 600-800 Euro □ 800-1000 Euro □ 1000-1200 Euro □ more than 1200 Euro
7. Is German your native language? □ Yes □ No
8. If no, please indicate your native language? _____
9. Do you have the feeling that you could answer the multiplication problems faster over time due to practice? □ Yes, very much □ Yes, a little □ No
10. Did you get exhausted as time in the experiment went by, so that you could concentrate less? □ Yes, very much □ Yes, a little □ No
11. Imagine you are playing a quiz with 10 questions. Which possibility of earning money would you prefer? A: You get 4 Euro for each correct answer. B: You get 60 Euro , if you give more correct answers than another unknown person. How do you decide? □ A □ B

B.1 Loss Aversion

Loss aversion of subjects was assessed by a method developed by Abdellaoui, Bleichrodt, and Haridon (2008). Subjects were asked the following three questions subsequent to the experiment:

1. Imagine a fair coin is flipped. You are offered a lottery, in which you can win 100 Euro if Head appears and nothing if Tails appears. Instead of playing the lottery

you can accept a certain gain. Which of the following gains would you accept?

	reject	accept
10 Euro	☐	☐
20 Euro	☐	☐
30 Euro	☐	☐
40 Euro	☐	☐
50 Euro	☐	☐
60 Euro	☐	☐
70 Euro	☐	☐
80 Euro	☐	☐
90 Euro	☐	☐
100 Euro	☐	☐

2. The coin is flipped again. You are offered a game in which you lose 150 Euro if Head appears and lose 50 Euro if Tails appears. Alternatively you can accept a certain loss. Which of the following certain losses would you accept?

	reject	accept
-140 Euro	☐	☐
-130 Euro	☐	☐
-120 Euro	☐	☐
-110 Euro	☐	☐
-100 Euro	☐	☐
-90 Euro	☐	☐
-80 Euro	☐	☐
-70 Euro	☐	☐
-60 Euro	☐	☐
-50 Euro	☐	☐

3. The coin is flipped again. You can either reject the game and earn/lose nothing, or you can accept the proposed game. Which of the following games would you accept?

	reject	accept
If Head appears, you earn 30 Euro. If Tails appears you lose 50 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 45 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 40 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 35 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 30 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 25 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 20 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 15 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 10 Euro.	☐	☐
If Head appears, you earn 30 Euro. If Tails appears you lose 5 Euro.	☐	☐

The first question is used to elicit the participants' utility in the domain of gains. By presenting a gain prospect x_i its certainty equivalent G_i is elicited. From $u(G_i) = \delta^+ u(x_i)$ the δ^+ can be determined. The second question is used to elicit the certainty equivalent for losses L_i for a prospect of losses (x_i, y_i). With $u(L_i) = \delta^-(u(x_i) - u(y_i)) + u(y_i)$ the δ^- is determined. The third question serves the elicitation of an indifference loss L^* for a given gain G^*. Then the coefficient of loss aversion λ was determined from the following equation: $\delta^+ u(G^*) + \lambda\delta^- u(L^*) = u(0) = 0$. Throughout the elicitation linear utility functions were assumed. For a more detailed description of the procedure see Abdellaoui, Bleichrodt, and Haridon (2008).

C Descriptive Statistics

Table C.1: Summary Statistics

	Male			Female		
Variable	Mean	Std. Dev.	N	Mean	Std. Dev.	N
Number of Correct Answers	22.784	12.574	310	19.945	8.972	290
Refpoint	25.924	11.05	248	26.158	10.799	232
Loss Aversion	2.995	3.784	250	3.032	3.023	180
Competitiveness	0.306	0.462	310	0.241	0.429	290
Math Grade	2.726	1.299	310	2.397	1.247	290
Years since Abitur	7.21	5.915	310	5.228	2.573	285
Studies Math	0.583	0.494	300	0.439	0.497	285
Age	26.355	6.529	310	24.086	3.069	290
Income	2.565	1.49	310	2.707	1.315	290
German Native Speaker	0.774	0.419	310	0.724	0.448	290

Table C.2: Pairwise Correlations

Variable	NumberAns	Refpoint	LossAv.	Compet.	Female	Grade	Abitur
Number Ans.	1.000						
Refpoint	0.296***	1.000					
Loss Aversion	0.095**	0.038	1.000				
Competitiveness	0.079*	0.067	-0.172***	1.000			
Female	-0.128***	0.011	0.005	-0.073*	1.000		
Math Grade	-0.297***	-0.090**	0.116**	0.004	-0.128***	1.000	
Years since Abitur	-0.049	0.039	-0.072	-0.046	-0.210***	0.178***	1.000
Studies Math	0.187***	-0.069	0.081*	0.099**	-0.145***	0.009	-0.048
Age	-0.082**	0.005	-0.068	-0.076*	-0.215***	0.222***	0.932***
Income	0.055	-0.027	-0.060	0.147***	0.051	-0.019	0.174***
German Native	-0.050	-0.104**	-0.130***	-0.162***	-0.058	0.135***	0.131***

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table C.3: Pairwise Correlations continued

Variable	StudMath	Age	Income	GermanNat.
Studies Math	1.000			
Age	-0.059	1.000		
Income	-0.024	0.151***	1.000	
German Native	-0.181***	0.116***	0.191***	1.000

Notes: Significance levels: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$