CESifo Working Papers
www.cesifo.org/wp

Free-Riding in the Lab and in the Field

Florian Englmaier
Georg Gebhardt

CESIFO WORKING PAPER NO. 3612
CATEGORY 13: BEHAVIOURAL ECONOMICS
OCTOBER 2011

CESifo *Working Paper No. 3612*

Free-Riding in the Lab and in the Field

Abstract

To investigate the external validity of laboratory results, we combine a public good experiment with three treatments in a field experiment. One treatment offers the opportunity to free-ride, the other two are placebo treatments. We compare results within subjects. In the free-riding treatment, subjects who contribute little in the lab are less productive. This effect is quantitatively as important as the effect of ability. The correlation between lab and field disappears in the two placebo treatments. We conclude that we can use lab experiments to learn about behavior in situations that share the game form but not necessarily the frame.

JEL-Code: C910, C930, D010, D640.

Keywords: field and lab experiments, external validity, public goods, team production.

Florian Englmaier
University of Konstanz
Department of Economics
Konstanz / Germany
florian.englmaier@uni-konstanz.de

Georg Gebhardt
Ulm University
Faculty of Mathematics and Economics
Ulm / Germany
georg.gebhardt@uni-ulm.de

October 10, 2011
We thank seminar participants at the Harvard University Behavioral and Experimental Economics Workshop, in the MELESSA Brownbag Lunch at Munich, the University of Cologne, the Toulouse School of Economics, the 2009 Workshop on “Natural Experiments and Controlled Field Studies”, the Econometric Society World Congress 2010 in Shanghai, the 2010 Workshop on “Learning with Field Experiments” at Wharton, the 2011 Workshop “Determinants and Implications of Prosocial Behaviour” in Southampton, and in particular Gary Charness, David Laibson, John List, Klaus Schmidt, Caspar Siegert, Robert Ulbricht, and Joachim Winter for helpful comments and suggestions. Christoph Lex and Johannes Kümmel provided excellent research assistance. The project was funded through SFB/TR-15.

1 Introduction

To explain behavior observed in laboratory experiments, economists have extended game theory to include heterogeneous, non-standard preferences.[1] One basic prediction of such models with stable preferences is that individuals behave consistently across similar game forms within and outside the lab. Consistent with this prediction, within subject correlation of behavior between lab and field settings has been documented.[2] In this paper, we present such evidence for the class of public good games that economists see as a metaphor for problems such as teamwork, free-riding, and contributions to a common good. We show that contributions in a lab public good game correlate with effort in a teamwork task in the field. For this task, we are able to generate quantitative estimates, allowing us to compare the impact of our measure of cooperativeness on productivity to the impact of ability.

An implicit but so far untested assumption in this line of work is that the correlation obtains because the lab and the field share the same strategic situation (game form); it could, however, be the case that lab experiments pick up other mechanisms such as norms about behavior in groups or broader fairness concerns that operate independent of the exact strategic situation. To test this alternative hypothesis, we add two placebo treatments that share the frame but not the game form with the original set-up. Averaging over the two placebo treatments, we document that the correlation disappears, suggesting that the correlation is indeed driven by the strategic situation.

As the first part of our study we run a field experiment and hire temporary employees (henceforth: *clerks*) to register books in a university library. In doing so, we conduct a natural field experiment as participants do not know that they participate in an experiment and work in a natural environment.[3] In the end, we invite them back for a lab experiment in which they play a public good game.

In the field, the clerks register books into a computer system. The clerks work in groups of four for a fixed wage. In the main treatment (Group) they can free-ride on the other group members' efforts as they can leave as soon as the group has entered a pre-specified number of books. In this treatment, the underlying game form is similar to the public good experiment insofar as employees have an incentive to free ride — yet, the framing between lab and field is different. We correlate contributions in the lab with effort in the field ("within subject"). We measure effort as productivity (books per minute) corrected for ability. To do so we divide the books-per-minute score by the result of a typing speed test (letters per five minutes) that job candidates had to perform when they applied for

[1] Examples for such theories include but are not limited to reciprocity (Rabin 1993, Dufwenberg and Kirchsteiger 2004, and Falk and Fischbacher 2006), inequity aversion (Fehr and Schmidt 1999 and Bolton and Ockenfels 2000), or the preference formulation of Charness and Rabin (2002).

[2] See, e.g., Karlan (2005), Benz and Meier (2008), or Fehr and Leibbrandt (2011). However, recently the discussion whether results obtained in laboratory settings can be readily extrapolated to field settings has intensified. See e.g., the papers by Levitt (2007), List (2007), or Falk, Meier, and Zehnder (2011).

[3] See John List's classification at *http://www.fieldexperiments.com/*

the job. We document that in the Group treatment, those clerks provide more effort that also contribute more in the laboratory public good game.

If we impose some additional structure, we can show that the effects are economically meaningful: In the Group treatment, choosing a clerk who contributes one additional standard deviation in the lab, results in a 11% higher productivity (corresponding to a 0.54 standard deviation increase). Choosing a clerk with a one standard deviation better test score results in 13% higher productivity. I.e., selecting clerks on *ability* and *fairness* (lab contributions), has an impact of a similar order of magnitude. This seems to be broadly consistent with the priorities applied in real world hiring practices.[4]

The correlation between the contribution in the laboratory public good game and the effort in the field may be due to the same underlying game form in the lab and in the field or to some features in the framing that trigger similar behavioral responses. To distinguish between these two channels we run two additional placebo treatments where there is no scope to free-ride but that share the otherwise identical frame with the Group treatment. In the first placebo treatment (Single), each employe can leave as soon as he or she has entered a certain number of books, independent of the performance of the other group members. In the second placebo treatment (No Incentives), the clerks can leave after a fixed period of time, no matter how many books they or the other group members have entered. We find that in the two control treatments, taken together, there is no correlation between contributions in the lab and effort in the field. If we analyze the two control treatments separately we obtain a clear zero correlation in the No Incentives treatment and a weak positive correlation (p-value around 9%) in the Single treatment.

Analyzing the field results in more detail, we show that incentivizing clerks has ambiguous effects depending on group heterogeneity with regards to ability: If we compare productivity conditional on the testscore, we find no improvement going from the No Incentives to the Single treatment. But if we control, by treatment, for group heterogeneity (within group standard deviation of the testscore), we find a large positive effect of the Single treatment for groups that are more homogenous than the average; for heterogeneous groups, however, incentives backfire, as heterogeneity has a negative effect in the Single treatment, but not in the other two. While our experiment was not designed to analyze this particular effect, we hypothesize that this effect is driven by the fact that the Single treatment is the only treatment to rank the group members explicitly in terms of task completion. Thus, lagging group members realize how much slower they are than their coworkers and their motivation drops off. Barankay (2011) documents in a field experiment with a different context a similar strong negative effect of rankings.[5]

[4]See for example Rosse et al. (1991) or Dunn et al. (1995) for studies on the relative importance of different characteristics in hiring decisions.

[5]Meidinger et al. (2003) analyze the relation of team heterogeneity and productivity in a laboratory experiment. They show that workers are better able to cooperate when the team is homogeneous.

While we do not suggest that we can always extrapolate from game forms in the lab to "similar" game forms in the field, we interpret our results as encouraging news that this practice can be useful.[6] We uphold this interpretation even given the uneven results over the two placebo treatments: if the correlation was driven by the the frame, e.g. the fact that the subjects always work or play in groups of four, it should appear over all three treatments. In the No Incentive treatment, however, we clearly observe no correlation. The potential correlation between the public good experiment and effort in the Single treatment would be consistent with certain aspects of preferences (types) being correlated; e.g., altruistic types that forego free-riding possibilities could also be less competitive and, therefore, immune to being ranked.

Related Literature: Many variations of the classic public good experiment, Kim and Walker (1984) and Isaac et al. (1985), have been conducted to check the stability of the outcomes. For a survey of early variations, see the survey in the *Handbook of Experimental Economics* by Ledyard (1995) or the survey by Fehr and Gaechter (2000).

Recently, a few papers relate behavior in the lab and in the field. Karlan (2005) document that behavior in a laboratory investment game predicts repayment of microfinance loans one year later. Fehr and Leibbrandt (2011) combine laboratory and field observations of the same individuals, for example, of fishermen who face a common pool resource dilemma and Benz and Meier (2008) match laboratory observations to donations data from the University of Zurich. These papers also find a positive correlation between field and lab behavior, but they do not have control and treatment groups in the field. Palacios-Huerta and Volij (2008) document that soccer players play mixing strategies both in penalty shootouts and experimental matching pennies games, and Palacios-Huerta and Volij (2009) document that chess grand masters play the Nash equilibrium in centipede games. These papers find similar strategic behavior in lab and field tasks, but there is selection into the field task (professional athletes and chess players). Moreover, Levitt et al. (2010) are not able to replicate the findings of these studies.

The remainder of the paper is structured as follows. The next section presents the experimental set-up for the field and the lab study. Section 3 presents the results and Section 4 discusses the results and concludes.

[6] One of the reasons why public good experiments have received a lot attention in the lab experimental is that the simple public good game serves as a powerful metaphor for many economic situations which involve free riding, as classic public goods, intellectual property rights for informational goods, or team production in organizations.

2 Experimental Set Up

2.1 General

As the first part of our study we hire temporary workers to register books in a university library. As the clerks work in a natural work environment and do not know that they participate in an experiment, it is a natural field experiment. At the end of the field experiment we pay the subjects their wage and invite them to come to our lab and participate in a lab experiment roughly two weeks later. We still do not reveal to the subjects that they participated in a field experiment. In the lab, subjects play a public good game and answer a questionnaire.

2.2 Field Experiment

The field experiment took place between October 1 and October 21 of 2008 at the Economics Department of the University of Munich. We recruit 103 subjects to register books in an institute library. These clerks have to enter books' information into an Excel-based computer database. We advertise the positions as one-time half-day jobs via posters all over the university campus, in student dorms, and on internet job portals. We promise to pay a salary of €55 and give the job duration as "up to 5 hours". Applicants have to apply via a recruiting website where they enter demographic information, contact details, and complete a five minute typing test. The test mirrors their actual task: they are displayed pictures with book information which they must type into an online form.

A research assistant (RA) invites the applicants to come to our library at a specific date and time. Every day we have subjects come in two shifts, the morning shift starting at 8:30am and the afternoon shift starting at 1:30pm. Each shift consist of a group of four individuals. We allocate subjects to shifts on a first-come-first-serve base subject to the constraint that we avoid putting participants with the same last name, the same address or the same major into the same shift. We instruct our RAs to look for signs of prior relationship (eg. joint arrival, etc.), but do not observe any.

The RA welcomes every shift and explains the task to the group. He shows them the shelves with the books they have to enter, before leading each one of them to one of four adjacent but separate offices. In their office, the participants find a laptop computer with an opened Excel-based entry mask. All entries subjects make in the Excel mask are saved on a central log file so we can track their performance over time. In principle, it is possible for the subjects to talk and interact, but we observe no signs of more than casual interaction. Our RA brings the books to the offices of the four subjects and gives them feedback on their performance every 30 minutes. See Figures 4, 5, and 6 in Appendix B for pictures of the physical set up.

All subjects in one shift are allocated to one of three treatments.

Group: All subjects can leave once they have registered 680 books as a group.[7] The feedback every 30 minutes is given with respect to group progress ("In the last 30 minutes the group entered x books. So in total the group has entered y out of a total of 680 books."). In this treatment "private return from effort $<$ social return" holds.

Single: Each subject can leave once it has registered 170 books individually. The feedback every 30 minutes is given with respect to individual progress ("In the last 30 minutes you entered x books. So in total you have entered y out of a total of 170 books."). In this treatment "private return from effort = social return" holds.

No Incentives: All clerks must work 3.5 hours, irrespective of the number of books entered. The feedback every 30 minutes is given with respect to individual progress ("In the last 30 minutes you entered x books. So in total you have entered y books."). In this treatment "private return from effort = 0" holds.

In the Group treatment, purely egoistic sujects should free ride, as the social benefit from effort is larger than the private benefit. In contrast, in Single and No Incentives treatments there is no scope for free riding as there is no difference between private and social benefits from effort. We keep the 30 minute feedback structure constant across all treatments to avoid confounding the true treatment variation with perceived variations in monitoring.

Only at the end of the experiment one of us introduces himself (truthfully) as academic staff responsible for the library organization. He pays the participants, tells them about the opportunity to participate in paid laboratory experiments at *MELESSA (Munich Experimental Laboratory for Economic and Social Sciences)*, and invites them to experimental sessions roughly two weeks after the field experiment.[8] Subjects can either directly sign up for a session or leave their email address to be contacted a few days before the sessions take place.

2.3 Laboratory Study

Between October 28 and 30 — about 10 days after the end of the field experiment — we run five sessions of laboratory experiments at MELESSA in Munich. About half of our subjects from the field (49) come to the lab. Exactly one half of the participants in the Single treatment show up, as well as 48 % in the No Incentives and 46 % in the Group treatment; we do not expect non-random selection of our field subjects into the lab and a Kruskal-Wallis test does not reject the null hypothesis that the participation probability

[7] In the first sessions we had one session with 120 books and 2 sessions with 150 books per person. People were substantially faster than we expected so we increased the number of books to 170 per person.

[8] We deliberately chose to become visible only at that stage, to avoid that subjects made the link between the laboratory study and the prior employment.

is the same over the three groups (p-value: 0.93). We fill up the sessions with students from the MELESSA subject pool. The subjects were invited via ORSEE, Greiner (2004), and the experiment was implemented with zTree, Fischbacher (2007). In total we have 84 subjects (4*16 and 1*20) in the lab.

Each session lasts roughly 90 minutes and the average earnings are 23.85€. In each session we run six rounds of experiments. Rounds 4-6 are fully anonymous, but rounds 1-3 are not. In these non fully anonymous rounds, we randomize group composition for each of the rounds. Before the round starts, we lead all subjects out of the room, to lead them back in as a group of four telling them that they play together. Thus subjects know who is in their group, but do not know which group member does what in the experiment. We designed this treatment to resemble the field situation where clerks are welcomed and introduced as a group. In the fully anonymous treatments, we reshuffle groups randomly in each round and subjects can neither see the members of their group nor know who does what within the group.

In rounds 1 and 4 we run two standard public good games.[9] In this linear public good experiment subjects play in groups of four, get an initial endowment of 20 tokens, and must decide how many tokens they put into a common pool. Tokens in the pool are multiplied by 1.6 and distributed among the group members. Tokens outside the pool belong to the subject. This game is repeated ten times with the same groups (partner treatment). Therefore, purely egoistic players should contribute nothing while it is efficient to contribute 20. At the end of the experiment, subjects answer a short questionnaire.

3 Results

3.1 Overview of Lab and Field Results

Before we analyze the relationship between the lab and the field experiment, we summarize the results of each stage separately. In both, the non-anonymous (PG1) and the fully anonymous (PG2), **lab public good games** we replicate standard results: Subjects differ substantially in their contributions, but even initially they contribute less than 20 on average before they reduce their contributions over time. In Figure 1 we plot the average contribution levels over the ten periods for PG1 on the left and PG2 on the right. In PG2, subjects initially contribute more (restart effect), but then they reduce their contributions again. We obtain a relatively small restart effect, but it is confounded by a treatment variation between the first and the second round: In the first round subjects knew with which other subjects they were in one group; in the second round, they did not. As we know from Fehr and Gaechter (1999) subjects contribute more if they meet the other people in

[9] In rounds 2, 3, 5, and 6 we run a different experiment; its design is described in Appendix C. We do not use the results of these experiments in our analysis.

their group before the experiment. This may explain why subjects contributed less in the second round, a behavior that looks like a smaller restart effect. In Figure 2 we present the average contributions combining PG1 and PG2. We use the average contribution over all 20 periods as our main measure of lab contributions. As discussed below, our results are robust to using different measures.

Contributions are not affected by participation or treatment assignment in the field. On average over PG1 and PG2, subjects contribute 7.9. The subjects from the MELESSA subject pool contribute slightly more (8.4) than the participants in the field experiment (7.6); this difference, however, is not significant (p-value Mann-Whitney U-test: 0.28). Participants in the Single treatment contribute slightly more (7.9) than participants in the No Incentives or Group treatments, who contribute almost exactly the same (7.4); a Kruskal-Wallis test does not reject the null hypothesis that contributions are the same over the three groups (p-value: 0.99).

Table 1 shows the descriptive statistics for the **field experiment**. We assigned slightly more subjects to the main Group treatment. We obtain an odd number of subjects in the No Incentives treatment because one session comprised only three subjects as one clerk failed to show up. We get less than 170 in the average number of books entered in the Group treatment because we ran the first three sessions with 120 or 150 books. We get less than 170 in the Single treatment because one subject had not finished his share of books after five hours (the advertised maximum working time) and we allowed him to leave. Subjects entered the highest number of books in the No Incentives treatment where also the variance in performance was highest. We have about two thirds of women in all our treatments.

Table 1: Descriptive Statistics Field

	Total	Group	Single	No Incentives
Number of Subjects	103	44	32	27
Number of Females	73	32	20	21
Avg. Speed (books/minute)	0.901	0.904	0.913	0.882
Std.Dev. Speed	0.199	0.182	0.217	0.211
Avg. # of Books entered	171.05	165.18	169.00	183.04
Std.Dev. Books entered	32.89	34.85	7.29	44.28

In Figure 3 we plot our measure of adjusted speed over time. We calculate adjusted speed as a measure of effort by dividing speed (books per minute) by the testscore (characters per two minutes). We plot the typing speed for the first two hours because even the fastest participants worked for at least two hours. In later hours, the fastest participants in the Single treatment were already finished so that we cannot compare average speeds. We calculate average speeds for four intervals of half an hour each because every half an hour

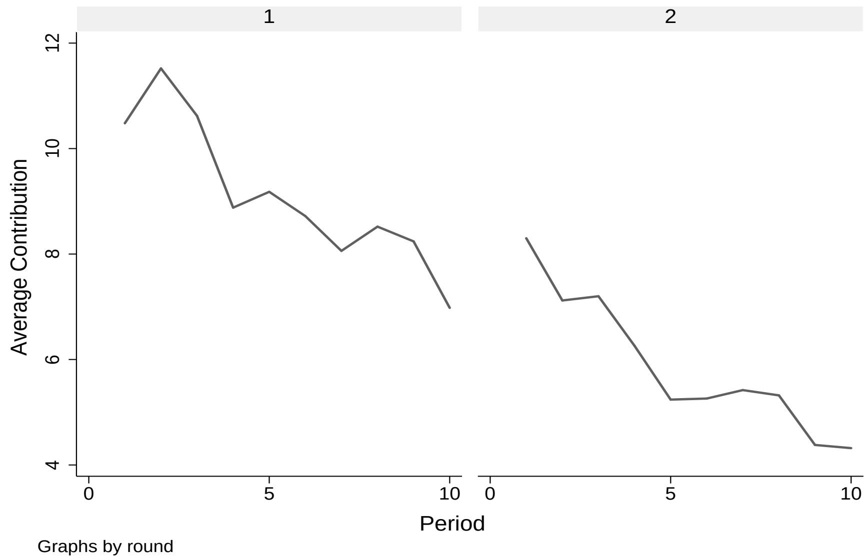

Figure 1: Avg. Public Good Contribution per Period
Left: PG1 (not fully anonymous); Right: PG2 (fully anonymous)

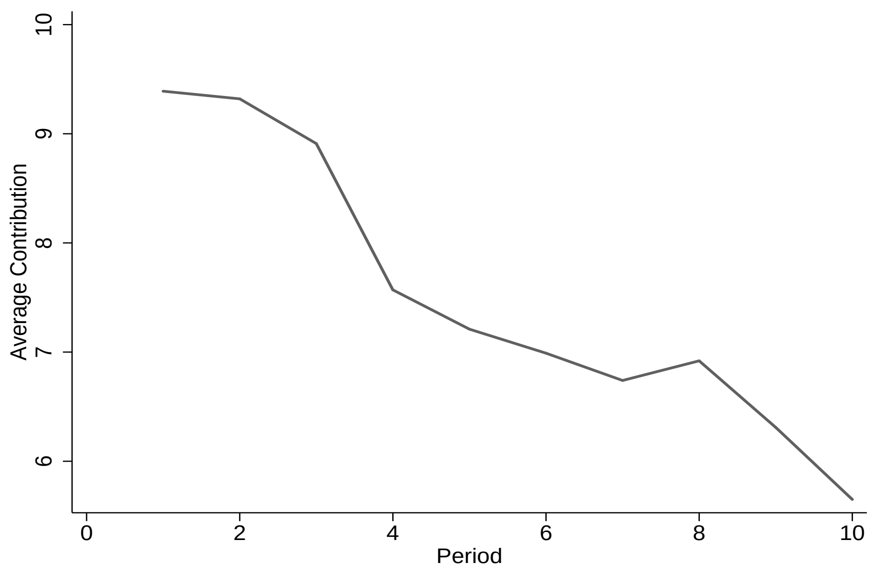

Figure 2: Avg. Public Good Contribution per Period (Sum PG1 + PG1)

the participants received feedback on their and the group's progress.

We find that adjusted speed changes little over time. In all three treatments subjects seem to improve their performance a little initially so that they work faster in the second and third 30 minute period. In the Single and No Incentives treatment, the subjects continue to work as fast or even faster in the last period, while we find a slight drop off in the last period in the Group treatment. Overall, however, we find little dynamic over time. Therefore, for the rest of this paper, we abstract from dynamic effects and present all results in terms of average (adjusted) speed over the whole working time of a subject.

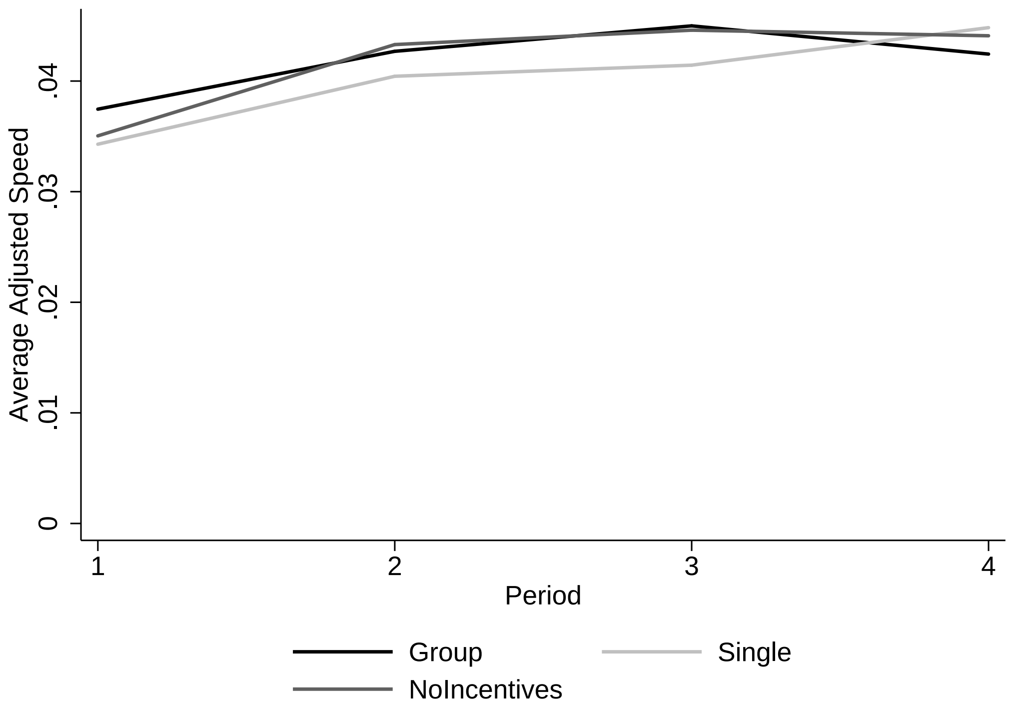

Figure 3: Skill Adj. Speed for the First Two Hours in the Field ($\frac{Books/Minute}{Testscore}$), by treatment
One period equals 30 Minutes, total time two hours

A casual inspection of Figure 3 seems to suggest that there is little difference between the treatments in the field experiment. While this is true for the averages, we will demonstrate in the remainder of the paper that distinct countervailing effects combine to produce these averages.

3.2 Correlation Lab and Field

In this section, we demonstrate qualitatively that subjects' behavior in the GROUP treatment, which allows free riding, is correlated with their lab behavior. At the same time the subject's behavior in the other two treatments, which do not allow free riding, is not significantly correlated with their lab behavior.

Various game theoretic models with stable (fairness) preferences make clear predictions for our set-up: Subjects with a preference for fairness contribute more in a public good game in the lab than subjects who have purely egoistic preferences. The same prediction holds for the GROUP treatment, in which subjects can free ride, i.e. work slowly, so that the other group members have to enter more of the books assigned to the group. No such prediction holds for the other two treatments (SINGLE and NO INCENTIVES), as in both cases the work load of the other group members is independent of how fast another member works.

As the theoretic prediction for the lab-field-correlation is the same for the Single and the No Incentive treatments, we initially pool the results from the two treatments into the category No Group. Because we obtain a larger sample size for the No Group treatment compared to the two single treatments, this makes it more difficult for us to reject the null hypothesis that there is no correlation in the No Group treatments.

Table 2: Average Lab Contribution - Skill adjusted Speed $\left(\frac{Books/Minute}{Testscore}\right)$

	Full Sample	Group	No Group
Kendall	0.1650	0.3789	0.0419
	(0.0961)	(0.0212)*	(0.7640)
Spearman	0.2351	0.5654	0.0389
	(0.1039)	(0.0094)**	(0.8411)
N. obs.	49	20	29

Notes: An observation is one subject present in the field and in the lab. Kendall and Spearman rank correlation coefficients between average public good contribution over 20 periods (PG1 + PG2) and *(Books/Minute)/Testscore*; *Testscore*: Letters per 5 Minutes; In parenthesis $Prob > |t|$ and $Prob > |z|$ respectively
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

In Table 2 we present the Kendall and Spearman correlation coefficients between the average contribution over all twenty periods in the two rounds in the lab experiment and the average adjusted speed in the field. For the full sample (column (1)) we find a slight correlation that is, however, not significant. If we split the treatments into Group and No Group, we find that there is almost no correlation in the No Group treatment. In the Group treatment however, we find a very strong correlation at least at the 5% level for both tests.

To put our results in perspective, we can compare them to findings from social psychology: Ross and Nisbett (1991), for example, argue that it is very hard to "predict" behavior based on information about personality traits; they report correlations between behavior and traits not exceeding 0.3. If we interpret the lab contribution as a measure of some personality trait, we find much stronger correlations (between 0.4 and 0.6 in the relevant comparisons), suggesting a much better predictability of behavior.

In Table 3, we present the results for a three way split, i.e., we separately list the correlation coefficients for the Single and the No Incentives treatment. In the No Incentives treatment we obtain a clear zero correlation and in the Single treatment a weak positive correlation with a p-value around 9%. Still, if the correlation was driven by the the frame as opposed to the game form it should appear over all three treatments. We will discuss later on the role of heterogeneity and competition for the results of the field experiment. Though our experiment was not designed to differentiate along this dimension, one potential explanation for the weak correlation between the public good experiment and effort in the

Table 3: Average Lab Contribution – Skill adjusted Speed $\left(\frac{Books/Minute}{Testscore}\right)$

	Group	Single	No Incentives
Kendall	0.3789	0.3250	−0.0256
	(0.0212)*	(0.0868)	(0.9512)
Spearman	0.5654	0.4356	−0.0854
	(0.0094)**	(0.0917)	(0.7815)
N. obs.	20	16	13

Notes: An observation is one subject present in the field and in the lab. Kendall and Spearman rank correlation coefficients between average public good contribution over 20 periods (PG1 + PG2) and *(Books/Minute)/Testscore*; *Testscore*: Letters per 5 Minutes; In parenthesis $Prob > |t|$ and $Prob > |z|$ respectively
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

Table 4: Lab Participation – Skill adjusted Speed $\left(\frac{Books/Minute}{Testscore}\right)$

	Full Sample	Group	Single	No Incentives
Kendall	−0.0198	0.0782	−0.1331	−0.0399
	(0.7338)	(0.3896)	(0.2206)	(0.7524)
Spearman	−0.0340	0.1330	−0.2234	−0.0666
	(0.7331)	(0.3894)	(0.2191)	(0.7413)
N. obs.	103	44	32	27

Notes: An observation is one subject present in the field. Kendall and Spearman rank correlation coefficients between decision to participate in the lab experiment and *(Books/Minute)/Testscore*; *Testscore*: Letters per 5 Minutes; In parenthesis $Prob > |t|$ and $Prob > |z|$ respectively
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

Single treatment would be a correlation of certain aspects of preferences (types), e.g., altruism with a lack of competitiveness.

Our results hold up if we use different measures of lab or field performance. We obtain essentially identical results if we measure performance not as all books entered but as only correctly entered books. We report the main results for this approach in Table A-1 in Appendix A. Again, we get identical results with slightly more noise if we measure lab contributions by only one of the two public good games (see Table A-2 in Appendix A)

As a robustness check (Table 4), we present the Kendall and Spearman correlation coefficients between participation in the lab experiment and the average adjusted speed in the field. We find no correlation between performance in the field and lab participation, neither for the whole sample nor for any of the treatments. Hence we do not suspect any selection into lab participation along this dimension either.

3.3 Regression Analysis of Lab and Field Relationship

In this section, we impose more structure and run a regression analysis to quantify the impact of the differences in fairness preferences as measured by the lab contributions; we find the impact to be large. We estimate two specifications with OLS, where each observation, indexed by i, is one subject that we have observed in the field and in the lab. In the main specification, we distinguish Group and No Group treatments and estimate

$$speed_i = \beta_0 + \beta_1 I_{Group} + \beta_2 score_i + \beta_3 contr_i + \beta_4 contr_i I_{Group} + \varepsilon_i,$$

where $speed_i$ is the log of the average speed (in books per minute), I_{Group} is an indicator variable that takes on the value 1 if the subject participated in the Group treatment in the field and 0 otherwise, $score_i$ is the log of the testscore (in characters in five minutes), and $contr_i$ is the log of the average contribution to the public good over the two rounds. Note, that we no longer use our skill adjusted speed measure but control for ability directly in the regression. All standard errors are robust and clustered by field sessions.

In addition, we use a specification with a three way split:

$$speed_i = \beta_0 + \beta_1 I_{Group} + \beta_2 I_{Single} + \beta_3 score_i + \beta_4 contr_i + \beta_5 contr_i I_{Group} + \beta_6 contr_i I_{Single} + \varepsilon_i,$$

where I_{Single} is an indicator variable that is 1 if the subject participated in the Single treatment in the field and 0 otherwise.

Qualitatively, we get the same results from the regression (Table 5) as from the correlation coefficients: Average speed is significantly correlated with lab contributions in the Group treatment, but neither in the No Group treatments taken together, nor in the Single or No Incentive treatments separately, i.e. we cannot reject the Null hypothesis that contributions do not predict speed in the Single treatment at the 5% level, even if we impose the additional parametric structure of the regression, which increases the power of the test. In addition, we can demonstrate that our typing speed test captures a meaningful part of ability as the testscore is highly correlated with average speed.

We use the coefficient estimates of the main specification to obtain a measure of the economic significance of the effects: A one standard deviation increase in the log average contribution (67.0%) increases average speed by 10.8% in the Group treatment. We can compare this with the effect of the testscore: A one standard deviation increase in the log testscore (28.0%) increases speed by 12.9%, i.e., if subjects have the possibility to free ride in the field, selecting them on the basis of their lab contributions yields basically the same productivity advantage as selecting them on the basis of ability, i.e., the testscore.

Table 5: OLS Estimates of the Average Speed in the Field Experiment

	Books per Minute (Log)	
	Two Way Split	Three Way Split
GROUP (Dummy)	−0.231	−0.296
	(0.151)	(0.164)
SINGLE (Dummy)		−0.269
		(0.138)
Testscore (Log)	0.461	0.467
	(0.108)***	(0.114)***
Avg. Contribution (Log)	−0.002	−0.029
	(0.026)	(0.021)
Avg. Contribution (Log) × GROUP	0.161	0.191
	(0.074)*	(0.074)*
Avg. Contribution (Log) × SINGLE		0.129
		(0.073)
Constant	−3.144	−3.117
	(0.700)***	(0.729)***
N. obs.	49	49
R^2	0.451	0.486

Notes: An observation in the regression is one subject present in the field and in the lab. The dependent variable is log of *Books/Minute*. Robust standard errors (clustered by field session) in parenthesis
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

4 Discussion and Conclusion

4.1 What type of subjects drive our results?

We can shed some more light on the motivations of the people who contribute in the lab and in the field. Results from various lab studies suggest that there is a (small) fraction of subjects who are genuinely willing to contribute, often called *altruists*. These subjects will typically contribute even in the last round. There are also other subjects who contribute as long as they believe that others contribute. These may be either purely egoistic subjects, *strategic egoists*, who contribute in the early rounds for strategic reasons to increase the contributions of others or genuine *conditional cooperators* who are willing to cooperate as long as others cooperate. In general, both of these latter types are observationally almost equivalent and in equilibrium contribute initially but reduce their contributions towards the end, leading to the downward-sloping pattern of average contributions.

To analyze which of the two groups drives the correlation between the lab and the field, we correlate productivity in the field with the contribution levels in the first periods of the two rounds of the lab game and the last periods of the two rounds. If conditional

Table 6: Correlation: Lab Contribution – Skill adj. Speed $\left(\frac{Books/Minute}{Testscore}\right)$

	First Periods		Last Periods	
	Group	No Group	Group	No Group
Kendall	0.3579	0.0320	0.2316	0.0493
	(0.0292)*	(0.8215)	(0.1600)	(0.7200)
Spearman	0.5435	0.0237	0.3335	0.0547
	(0.0132)*	(0.9029)	(0.1508)	(0.7781)
N. obs.	20	29	20	29

Notes: An observation is one subject present in the field and in the lab. Kendall and Spearman rank correlation coefficients between sum (PG1+PG2) of the public good contribution in the first periods (1) respectively last periods (10) and *(Books/Minute)/Testscore*; *Testscore*: Letters per 5 Minutes; In parenthesis $Prob > |t|$ and $Prob > |z|$ respectively
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

cooperators and strategic egoists drive the correlation, we should find the correlation for the first round but not for the last round. But if genuine altruists are behind the correlation, we should observe a significant correlation for both periods. In Table 6 we present the correlation coefficients for the the first and last periods. In neither specification, we find a significant correlation in the No Group treatments. In the Group treatment, however, we find a correlation between first period contributions and adjusted speed in the field but no significant correlation with the last period contributions. This is consistent with the hypothesis that strategic egoists and conditional cooperators drive the lab-field-correlations.

4.2 Regression Analysis Field: Group Heterogeneity

In this section, we return to the question why incentives have seemingly so little impact on productivity in the field experiment. We show that incentivizing clerks by allowing them to leave early if they work faster, has ambiguous effects on productivity depending on group heterogeneity with regards to ability: Incentives may even backfire in heterogeneous groups.

To investigate this hypothesis in a regression analysis, we need a measure of group heterogeneity. We cannot use the standard deviation of actual performances for obvious endogeneity problems. Instead, we use the standard deviation of testscores. We include for each subject the standard deviation of testscores in his or her group together with the (log)level of the testscore of the individual subject. We interact the standard deviation of the testscore with the treatment dummy, as we suspect that heterogeneity plays an especially important role in the Single treatment as it is the only treatment to rank the group members in terms of task completion.

In Table 7 we present the results of this regression. We find that the point estimates of

Table 7: OLS Estimates of the Average Speed in the Field Experiment

	Books per Minute (Log)
GROUP (Dummy)	0.381
	(0.380)
SINGLE (Dummy)	0.918
	(0.400)*
Testscore (Log)	0.455
	(0.084)***
Std. Testscore (Log)	−0.012
	(0.061)
Std. Testscore (Log) × GROUP	−0.075
	(0.083)
Std. Testscore (Log) × SINGLE	−0.189
	(0.089)*
Constant	−3.006
	(0.621)***
N. obs.	103
R^2	0.360

Notes: An observation in the regression is one subject present in the field. The dependent variable is log of *Books/Minute*. Robust standard errors (clustered by field session) in parenthesis
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

the coefficient on the standard deviation of testscores is negative in all three treatments: heterogeneous groups perform worse. But the size of the effect varies over treatments. It is very small for the NO INCENTIVE treatment, somewhat larger for the GROUP treatment, and by far largest for the SINGLE treatment. It is statistically significant (at the 5% level) only in the SINGLE treatment.

If we control for group heterogeneity, we get different estimates of the treatment dummies. These dummies now capture the treatment effect for a group that has an average degree of homogeneity. We find that the point estimates are now positive for the GROUP treatment (which provides some incentives) and even more positive for the SINGLE treatment (which provides the strongest incentives). Only the coefficient for the SINGLE treatment is significant (at the 5% level). The point estimates are large, in particular for the effect of SINGLE treatment relative to the NO INCENTIVES treatment, implying a strong increase in speed if we go from NO INCENTIVES to the SINGLE treatment.

These results suggest that heterogeneity and incentives interact to produce the at first glance counterintuitive impact of our treatments. Because we do not select or group the participants in the field experiment on the basis of their testscores, some of the groups are very heterogeneous with regards to ability. This heterogeneity does not translate into differences in payment or working time in the Group and No Incentive treatments; everybody is paid the same and everybody comes and leaves together. In the Single treatment, however, some subjects have to work much longer than others: In some groups the fastest clerk can leave after two hours, while his or her slower coworkers have to work for up to five hours. Even if the subjects cannot directly see each other working, they notice when one of their coworkers leaves and their motivation drops off. As this effect can only happen in the Single treatment, we suspect that this effect may drive the seeming ineffectiveness of incentives in our field experiment.

We did not design our experimental set-up to explicitly answer the reasons behind the strong effect of heterogeneity. It could be that the slow members of a group feel unfairly treated because they must work much longer for the same amount of money than the fast members. Therefore, they may experience a decrease of motivation. Maybe, the Single treatment introduces a notion of competition that motivates group members that work at a similar speed, but discourages laggards. Whatever the reason for the interaction effects of treatment and heterogeneous ability, the effects are large. Whenever we introduce incentives, we reward not only effort but also ability, and we know little how this is perceived by workers, as most lab experiments abstract from heterogeneous ability. Our results suggest that this interaction should be added to the growing list of potential pitfalls associated with the provision of explicit incentives.

4.3 Summary

We run a public good experiment in the field and in the lab with (partly) the same subjects. The field experiment is a true natural field experiment as subjects do not know that they are exposed to an experimental variation. Our study offers several contributions. We can establish that lab behavior is externally valid, as behavior in the lab and in the field correlates, even within subjects. We can go beyond the existing literature, as we have treatment and control groups in the field. Using the placebo treatments, we can show that the correlation is only present in the public good treatment (Group) but not in the other treatments (Single, No Incentives). I.e., subjects behave similarly under "similar" incentive structures. We take this as indication for the external validity of laboratory experiments. This might also be indicative for the existence of stable preference types. Furthermore, the simple game theoretic structure of the public good game seems to capture important aspects of real life situations. Moreover, we can show that the effect of "fairness" preferences on performance is economically relevant, in our setting comparable in size to the effect of ability. Finally we document the (detrimental) effect of heterogeneity w.r.t. ability on performance under explicit incentive structures.

References

Barankay, I. (2011) "Rank Incentives: Evidence from a Field Experiment," mimeo Wharton School of Business, University of Pennsylvania

Benz, M. and S. Meier (2008) "Do people behave in experiments as in the field-evidence from donations," *Experimental Economics*, Vol. 11(3), pp. 268-281

Bolton, G.E. and A. Ockenfels (2000) "ERC – A theory of equity, reciprocity and competition," *American Economic Review*, Vol. 90(1), pp. 166-193

Charness, G. and M. Rabin (2002) "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, Vol. 117(3), pp. 817-869

Dufwenberg, M. and G. Kirchsteiger (2004) "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, Vol. 47, pp. 268-98

Dunn, W., M.K. Mount, M.R. Barrick, and D.S. Ones (1995) "The Big Five personality dimensions, general mental ability and perceptions of employment suitability," *Journal of Applied Psychology*, Vol. 80, pp. 500-509

Falk, A. and U. Fischbacher (2006) "A Theory of Reciprocity," *Games and Economic Behavior*, Vol. 54(2), pp. 293-315

Falk, A., S. Meier, and C. Zehnder (2011) "Do Lab Experiments Misrepresent Social Preferences? The case of self-selected student samples," *Journal of the European Economic Association*, forthcoming

Fehr, E. and S. Gaechter (1999) "Collective action as a social exchange," *Journal of Economic Behavior & Organization*, Vol. 39(4), pp. 341-369

Fehr, E. and S. Gaechter (2000) "Fairness and Retaliation: The Economics of Reciprocity," *Journal of Economic Perspectives*, Vol. 14, pp. 159-181

Fehr, E. and A. Leibbrandt (2011) "A Field Study on Cooperativeness and Impatience in the Tragedy of the Commons," *Journal of Public Economics*, forthcoming

Fehr, E. and K.M. Schmidt (1999) "A theory of fairness, competition and cooperation," *Quarterly Journal of Economics*, Vol. 114(3), pp. 817-868.

Fischbacher, U. (2007) "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, Vol. 10(2), pp. 171-178.

Greiner, B. (2004) "An Online Recruitment System for Economic Experiments," GWDG Bericht 63, Göttingen : Ges. für Wiss. Datenverarbeitung, pp. 79-93

Isaac, R.M., K. McCue, and C.R. Plott (1985) "Public Goods Provision in an Experimental Environment," *Journal of Public Economics*, Vol. 26, pp. 51-74

Karlan, D. (2005) "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions," *American Economic Review*, Vol. 95(5), pp. 1688-1699

Kim, O. and M. Walker (1984) "The Free Rider Problem: Experimental Evidence," *Public Choice*, Vol. 43, pp. 3-24

Ledyard, J. (1995) "Public Goods: A Survey of Experimental Research," in: *The Handbook of Experimental Economics*, Eds. A. Roth and J. Kagel, Princeton University Press, Princeton/NJ

Levitt, S.D. (2007) "Viewpoint: On the Generalizability of Lab Behavior in the Field," *Canadian Journal of Economics*, Vol. 40(2), pp. 347-370

Levitt, S.D., J.A. List, and D.H. Reiley (2010) "What Happens in the Field Stays in the Field: Exploring Whether Professionals Play Minimax in Laboratory Experiments," *Econometrica*, Vol. 78(4), pp. 1413-1434

List, J.A. (2007) "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy*, Vol. 114(1), pp. 1-37

Meidinger, C., J.-L. Rulliere, and M.-C. Villeval (2003) "Does Team-based Compensation Give Rise to Problems When Agents Vary in Their Ability?," *Experimental Economics*, Vol. 6(3), pp. 253-272

Palacios-Huerta, I. and O. Volij (2008) "Experientia Docet: Professionals Play Minimax in Laboratory Experiments," *Econometrica*, Vol. 76(1), pp. 71-115

Palacios-Huerta, I. and O. Volij (2009) "Field Centipedes," *American Economic Review*, Vol. 99(4), pp. 1619-35

Rabin, M. (1993) "Incorporating fairness into game theory and economics," *American Economic Review*, Vol. 83(5), pp. 1281-1302

Ross, L. and R.E. Nisbett (1991) "The Person and the Situation - Essential Contributions of Social Psychology," McGraw-Hill, New York/NY

Rosse, J.G., H.E. Miller and L.K. Barnes (1991) "Combining personality and cognitive ability predictors for hiring service-oriented employees," *Journal of Business and Psychology*, Vol. 5(4), pp. 431-445

Appendix A Tables

Table A-1: Correlation: Avg. Lab Contribution - Skill adj. Speed $\left(\frac{CorrectBooks/Minute}{Testscore}\right)$

	Full Sample	GROUP	NO GROUP
Kendall	0.1480	0.3368	0.0271
	(0.1358)	(0.0410)*	(0.8511)
Spearman	0.2190	0.5368	0.0318
	(0.1307)	(0.0147)*	(0.8700)
N. obs.	49	20	29

Notes: An observation is one subject present in the field and in the lab. Kendall and Spearman rank correlation coefficients between average public good contribution over 20 periods (PG1 + PG2) and *(Correct Books/Minute)/Testscore*; *Correct Books*: Core Information without typing mistakes; *Testscore*: Letters per 5 Minutes; In parenthesis $Prob > |t|$ and $Prob > |z|$ respectively
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"

Table A-2: Correlation: Lab Contribution – Skill adj. Speed $\left(\frac{Books/Minute}{Testscore}\right)$

	First Public Good Game (PG1)		Second Public Good Game (PG2)	
	GROUP	NO GROUP	GROUP	NO GROUP
Kendall	0.3158	0.1133	0.3421	−0.1207
	(0.0553)	(0.3984)	(0.0378)*	(0.3677)
Spearman	0.4891	0.1311	0.4761	−0.1562
	(0.0286)*	(0.4980)	(0.0338)*	(0.4184)
N. obs.	20	29	20	29

Notes: An observation is one subject present in the field and in the lab. Kendall and Spearman rank correlation coefficients between average public good contribution over 10 periods in PG1 or PG2 respectively and *(Books/Minute)/Testscore*; *Testscore*: Letters per 5 Minutes; In parenthesis $Prob > |t|$ and $Prob > |z|$ respectively
* = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%"
We present Kendall and Spearman Rank Correlation Coefficients between a subject's average contribution and her/his skill adjusted speed measured as *(Books/Minute)/Testscore* where *Testscore* is the result of our ability test upon hiring separate for contributions in PG1 and PG2. In parenthesis we report $Prob > |t|$ and $Prob > |z|$ respectively. Stars denote significance: * = "significant at 5%"; ** = "significant at 1%"; *** = "significant at 0.1%".

Appendix B Set-Up of the Field Experiment

Figure 4: Workspace

Figure 5: Office location

Figure 6: Books that had to be entered

Appendix C Instructions: Laboratory Experiment

INSTRUCTIONS (translated from German)

Welcome to the experiment and thanks a lot for your participation! From now on, please do not speak to other participants of this experiment any more!

General remarks on the procedure

This experiment serves to analyze economic decision making. You can earn money by taking part, which will be paid in cash after the experiment. During the experiment, you and other participants are asked to make decisions. Both your own decisions as well as those of other participants will determine your pay-off according to the following rules. The whole experiment lasts about 90 minutes. You will receive detailed instructions at the beginning. If you have questions regarding the instructions, please raise your hand. One of the conductors of the experiment will then come to you and answer your question personally. Each participant receives a fixed ID number by which he or she is identified during the experiment. During the experiment, we do not speak in terms of Euro, but Experiment-Points (EP). Your earnings will be calculated in EP in the course of the experiment. At the end of the experiment, all EPs you have earned will be converted into Euro using the following exchange rate:

1 Experiment-Point = 2 Euro-Cent

At the End: The experiment consists of 6 parts. It ends after the 6th part. The results from the individual rounds will be added up and converted into Euro. Your earnings will be paid out after the 6th part.

Anonymity: In each of the 6 parts, you will play in a new group. In each part your group will be re-matched randomly. Sometimes you know who is in your group, sometimes you do not. The other participants do not know during nor after the experiment which role you have taken or how much you have earned. We will evaluate the data from the experiment anonymously and delete all your personal and person-related information as soon as we have matched the data. At the end of the experiment you have to sign a receipt of your earnings which serves for accounting purposes only.

The Experiment

Part 1

Groups: At the beginning of part 1 you will be randomly divided into groups of four. In each later part your group will be re-matched randomly. You can see the other members of your group. But you cannot attribute decisions to other group members. Thus all group members stay anonymous. The 1st part of the experiment consists of 10 rounds in total. The composition of the group stays the same over the whole 1st part.

Budget and alternatives in each round: Each participant receives 20 Points in the beginning of each round. You can freely divide the points into two alternatives, X and Y:

1. You can put 0 to 20 points into pot X. The sum of the points in pot X will be multiplied by 1.6 and distributed equally among the group members. That is, for each point in pot X you will get 0.4 (=1.6/4) points. For example, if the sum of the points in pot X in your group is 60, each group member will get 60*0.4=24 points from pot X. If all group members together contribute 10 points to pot X, you and all other group members will get 10*0.4=4 EP from pot X.

2. You can put 0 to 20 points into pot Y. This amount goes then one-to-one into your earnings. So if you, for instance, contribute 6 points to pot Y, you get exactly 6 points. Your earnings per round is then the sum of your earnings from pot X and that from pot Y.

In mathematical terms, the result is:
Result (for member i) = $(20 - x_i) + (S * 1.6)/4$
x_i is the contribution of member i to pot X
S is the sum of the contributions of all group members to pot X
You will be asked on the screen how many points you want to contribute to pot X. The rest of the 20 points automatically goes to pot Y. Therefore, it is not possible to save points. You can enter only an integer number between 0 and 20 (i.e. 0; 1; 2; ...; 19; 20). After each round you will get to know the contributions of your group members to pot X and your total earning in this round in EPs. In the course of part 1 you will make 10 decisions according to the instructions above. Please remember: You will receive 20 points in each round, which you have to distribute to pot X and Y.
You will determine your deposit with the slider (See Figure 7). A click on the right-arrow increases your deposit by 1 point, whereas a click on the left-arrow reduces your deposit by 1 point.
You will get the instructions for the 2nd part of the experiment after end of the 1st part.

Part 2

Groups: At the beginning of part 2 you will be randomly assigned to a new group of four participants. In each later part your group will be re-matched randomly. You can see the other members of your group. But you cannot attribute decisions to other group members. Thus all group members stay anonymous.
Budget and alternatives in each round: In this 2nd part, each group has to deposit a sum of 240 points into a pot. The game ends as soon as the group has raised these 240 points. In the beginning, each member receives a budget of 220 points. From these points you must deposit at least 10 per round into the group pot. You can increase your contribution in steps of 2.5 points up to 60 points. If the necessary 240 points have not been deposited after a certain round, the experiment goes to the next round and you can again choose your contribution. The experiment continues to move to further rounds until the number of deposited points in the group pot reaches 240. For each round played, expenses of 20 points will be subtracted from every member.
Consider the following case:
In the first and second round the participants have deposited altogether 180 points into the group pot. This means that at the beginning of the 3rd round there are still 60 points to be raised. In the third round, two participants contribute 10 points each and the other two pay in 15 points each. Then, at the beginning of the 4th round, the number of points still to be raised decreases to 10 points (=60 points - 2*10 points - 2*15 points). In addition, the budget of the participants who have contributed 15 points decreases by 35 points (15 points contribution and 20 points expenses per round). The budget of the participants who have contributed 10 points decreases by 30 (10 points contribution and 20 points expenses per round). If in a round the participants raise more points than the total necessary amount, the points and the expenses will be subtracted proportionally from each participant.
Consider the following case:
At the beginning of the third round, the participants have deposited altogether 180 points into

the group pot. So they still need to raise 60 points. If now in the third round each participant deposits 30 points, then the total contribution would be 120 instead of the 60 points needed. In this case, the budget of each participant decreases on a pro-rata basis only by 15 points (=60/120 * 30 contributed points) and the 10 (=60/120 * 20) points of expenses for this round.

In the end you will get as payoff the points that remain from your budget after the last round. You will be asked on the screen how many points you want to contribute to the group pot. You can enter a number between 10 and 60 in steps of 2.5 (i.e. 10; 12.5; 15; . . . ; 57.5; 60). After each round you get to know the sum of the contributions of your group members to the group pot and your remaining budget in experiment points.

You determine your deposit with the slider (See Figure 8). A click on the right-arrow increases your deposit by 2.5 points, whereas a click on the left-arrow reduces your deposit by 2.5 points.

You will get the instructions for the 3rd part of the experiment after the end of the 2nd part.

Part 3

Groups: At the beginning of part 3 you will be randomly assigned to a new group of four participants. In each later part your group will be re-matched randomly. You can see the other members of your group. But you cannot attribute decisions to other group members. Thus, all group members stay anonymous.

Budget and alternatives in each round: Same as in part 2.

Part 4

Groups: At the beginning of part 4 you will be randomly assigned to a new group of four participants. In each later part your group will be re-matched randomly. You will not come to know the other members of your group, neither during nor after the experiment. So your decisions stay anonymous. The 4th part of the experiment consists of 10 rounds in total. The composition of the group stays the same for the whole 4th part.

Budget and alternatives in each round: Same as in part 1.

Part 5

Groups: At the beginning of part 5 you will be randomly assigned to a new group of four participants. In each later part your group will be re-matched randomly. You will not come to know the other members of your group, neither during nor after the experiment. So your decisions stay anonymous.

Budget and alternatives in each round: Same as in part 2.

Part 6

Groups: At the beginning of part 6 you will be randomly assigned to a new group of four participants. You will not come to know the other members of your group, neither during nor after the experiment. So your decisions stay anonymous.

Budget and alternatives in each round: Same as in part 2.

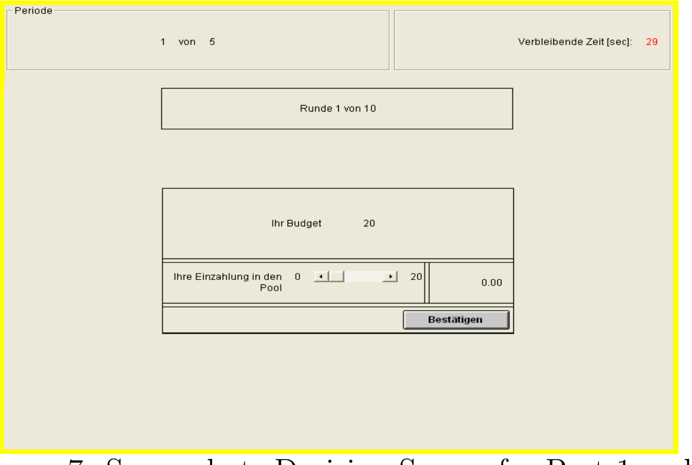

Figure 7: Screenshot: Decision Screen for Part 1 and 4

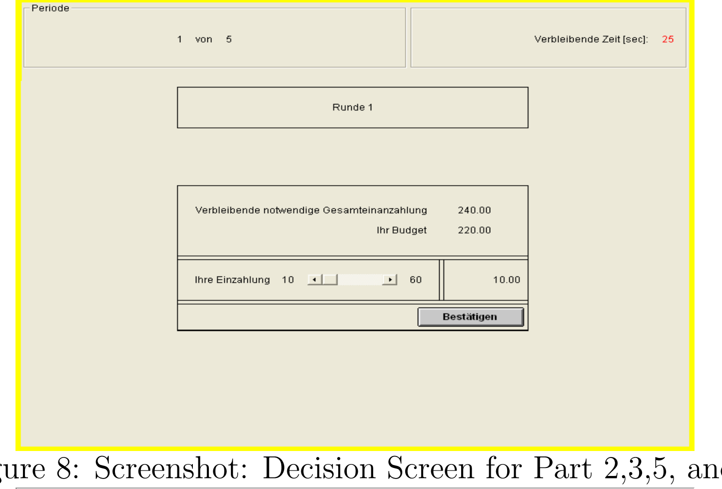

Figure 8: Screenshot: Decision Screen for Part 2,3,5, and 6

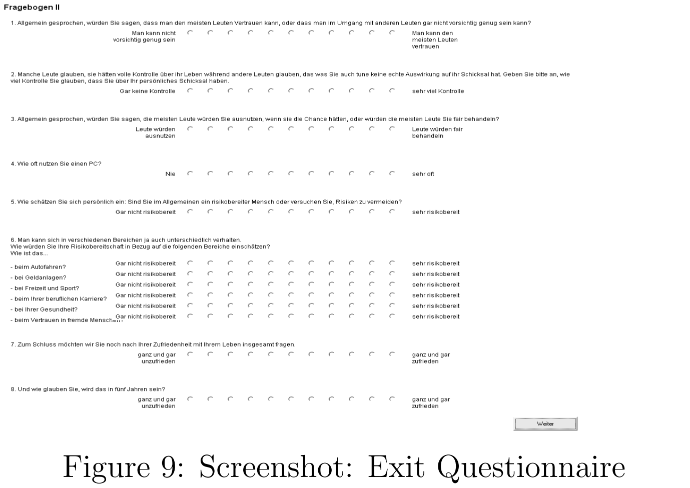

Figure 9: Screenshot: Exit Questionnaire