Incentives versus sorting in tournaments: Evidence from a field experiment*

Edwin Leuven† Hessel Oosterbeek‡ Joep Sonnemans§
Bas van der Klaauw¶

Abstract

A vast body of empirical studies lends support to the incentive effects of rank-order tournaments. Evidence comes from experiments in laboratories and non-experimental studies exploiting sports or firm data. Selection of competitors across tournaments may bias these non-experimental studies, whereas short task duration or lack of distracters may limit the external validity of results obtained in lab experiments or from sports data. To address these concerns we conducted a field experiment where students selected themselves into tournaments with different prizes. Within each tournament the best performing student on the final exam of a standard introductory microeconomics course could win a substantial financial reward. A standard non-experimental analysis exploiting across tournament variation in reward size and competitiveness confirms earlier findings. We find however no evidence for effects of tournament participation on study effort and exam results when we exploit our experimental design, indicating that the non-experimental results are completely due to sorting. Treatment only affects attendance of the first workgroup meeting following the announcement of treatment status, suggesting a difference between short-run and long-run decision making.

Keywords: Tournaments, Incentives, Sorting, Field Experiments
JEL Codes: J33, C93, M52

1 Introduction

Rank-order tournaments as analyzed in Lazear and Rosen (1981) are widely viewed as labor market contracts that describe and explain relevant features of compensa-

*We thank Monique de Haan, Sandra Maximiano, Erik Plug, Holger Sieg and seminar participants in Amsterdam, Århus, Berlin, Bonn, Leicester, Madrid, Paris, Southhampton and St-Gallen for fruitful discussion and comments.

†ENSAE-CREST.
‡University of Amsterdam and Tinbergen Institute.
§CREED - University of Amsterdam and Tinbergen Institute.
¶VU University Amsterdam and Tinbergen Institute.

tion and incentive structures in organizations. Tournaments provide incentives by evaluating individual performance relative to the performance of competitors. Unlike piece rates, tournaments induce optimal effort in circumstances where absolute productivity is difficult to observe but where relative performance measures are more readily available. A pay-off scheme based on relative productivity is also attractive in situations where common shocks affect the productivity of all individuals.

There is a small but expanding literature examining the incentive effects of tournaments. Observational studies use firm level data to test indirect predictions of tournament theory, or data from sports events to estimate the effect of prices on performance. A second strand of studies are based on laboratory experiments. As we will discuss in more detail in the next section, observational studies potentially suffer from selection bias when more able individuals get sorted in the more competitive and higher prize tournaments. Experimental studies have not analyzed sorting across tournaments, and have been criticized for having limited external validity.

The aim of the current paper is to test the incentive effects of tournaments and in particular the confounding effects of sorting when observational data is used to test tournament theory. We ran a field experiment in a naturally occurring setting. In our experiment students enrolled in a standard introductory microeconomics course could win a substantial prize for having the best score on the course's final (multiple choice) exam. Participants had to select themselves into a tournament with either a low prize (1000 Euros), a medium prize (3000 Euros) or a high prize (5000 Euros). Within each tournament participants were then randomly assigned to a treatment group and a control group. In each tournament the prize was won by the student in the treatment group who performed best at the exam. Students in the control groups could not win a prize. We conducted this experiment two years in a row (in the academic years 2004/5 and 2005/6). The exam consisted of 35 multiple-choice questions. The number of correct answers on the exam is our measure of productivity. During the course we registered participants' attendance of the course's workgroups, and at the end of the exam we asked them how much time they spent preparing for the exam. These are our measures of effort.

The time span between the announcement of this field experiment and the day of the final exam is 3 months. While this period is shorter than the "several years" involved in a tournament to become a firm's CEO, it is substantially longer than the "two or three hours" of a typical laboratory experiment or a sport event and will therefore contribute to the external validity of our results. To assess the confounding effect of sorting we contrast our experimental findings with a non-experimental empirical analysis that exploits across tournament variation in prize money and group

size which has been the common way of testing tournament theory using sports data (e.g. Ehrenberg and Bognanno, 1990a), or firm data (e.g. Eriksson, 1999).

We find only very little evidence that the prospect of winning a prize affected students' effort. The exception is that treated students are significantly more likely to attend the first workgroup meeting immediately after assignment to treatment and control groups was announced. But there is no effect on attendance of subsequent workgroup meetings or on the amount of time spent preparing for the exam. Consistent with this, we also find no effect on students' achievement; not on its mean level and also not when we focus on students in the top of the achievement and ability distributions. This is in stark contract with the non-experimental analysis of our data which leads to the (erroneous) conclusion that higher rewards generate higher productivity. Finally, we exploit variation in treatment intensity across workgroups to test for spillovers, but do not find support for that.

The remainder of this paper is organized as follows. The next section discusses the related literature and the contribution of our study. Section 3 describes the design of our field experiment. Section 4 presents and discusses the results. Section 5 summarizes and concludes.

2 Related literature

Empirical support for tournament theory comes from various sources. First, there are several studies that use data from firms or executive pay. As Becker and Huselid (1992) point out, tournament structures are more likely to exist where individuals' absolute effort or performance is difficult or costly to observe. Studies based on firm data therefore do not conduct direct tests of the incentive effects of tournaments but typically test some of the indirect implications of the tournament model. An example is Eriksson (1999), who uses data from Danish firms to test among other things whether pay differentials between job levels are consistent with relative compensation, whether pay dispersion is higher in more noisy environments, and whether pay dispersion is affected by the number of participants. His analyses support most of the theoretical predictions. Gibbs (1995) criticizes such indirect tests of the tournament model by arguing that (many of) the indirect implications of the tournament model are also consistent with a model in which firms require a threshold level for performance in order to be promoted to the next rank. Another potential complication arises from the sorting of workers across firms.

A second source for empirical support for tournament theory comes from studies where sports events are analyzed in terms of rank-order tournaments. This line

of research started with the papers of Ehrenberg and Bognanno (1990a,b) that analyzed data from professional golf tournaments. Both papers regress players' final score in a tournament on the total available prize money with control variables for difficulty of the course, weather conditions, players' ability and opponents' quality. They also regress final-round scores on proxies of players' marginal returns to effort. The results of these analyses typically provide support for tournament theory: the level and structure of prizes influence players' performance.[1] Following Ehrenberg and Bognanno's lead, regression models inspired on tournament theory have been estimated on data from other sports including professional bowling (Abrevaya, 2002; Bognanno, 1990), tennis (Sunde, 2003) and auto racing (Becker and Huselid, 1992). The results of all these studies are in support of tournament theory.

A third group of studies is based on laboratory experiments, starting with Bull et al. (1987). In this study over 200 undergraduate students volunteered to participate in one of the ten treatments of this study. Treatments vary features as the prize spread, asymmetry of costs and information conditions. The main finding of this study is that while behavior is on average in agreement with theoretical predictions there is a very large variance of behavior at the individual level. This is not the case in the piece rate treatment in which some of the subjects participated. Moreover, low ability (high cost) subjects tend to choose higher effort levels than predicted. The authors hypothesize that these deviating findings are due to the game nature of tournaments (in the laboratory). Van Dijk et al. (2001) conducted a laboratory experiment to contrast the performance of piece-rates schemes, team incentives and tournaments.[2] Their findings regarding tournaments are very similar to those obtained by Bull et al. (1987). They too, report a large variance in behavior under tournaments and low ability subjects exerting too much effort.[3]

The evidence from sports events and from laboratory experiments is often used as the basis for rather strong statements concerning the optimality of tournaments. For instance, Van Dijk et al. (2001) conclude that "[f]rom the perspective of an

[1]The results in Orszag (1994) cast some doubt on the validity of the findings reported in Ehrenberg and Bognanno (1990a). Orszag replicates their analysis using data from another tour, and finds no effect of total prize money on performance. Further analysis reveals that in the data used by Ehrenberg and Bognanno, the (self-rated) weather data are not orthogonal to the prize money variable. Orszag concludes that "perhaps golf is not the ideal example to study tournament theory, or perhaps tournament theory does not elicit the desired incentive results."

[2]A novel feature of their experimental design is that subjects have to provide real effort (searching the maximum in a grid) rather than just choose a number from an effort-cost table.

[3]Other studies using laboratory experiments include Schotter and Weigelt (1992) who study the effects of affirmative action programs in a tournament setting, Harbring and Irlenbusch (2003) who focus on the effects of different tournament sizes and different prize structures, and Freeman and Gelber (2006) who examine the effects of inequality in rewards and different provision of information.

employer, relative payment schemes would therefore [higher effort on average] be superior." (p.208). Likewise, Becker and Huselid (1992) state that: "Employers want to encourage employees to take risks and to be entrepreneurial, but not to be careless in their actions. It would appear that tournament reward systems have the potential to achieve these goals" (p.348), or "[t]ournament systems have considerable motivational properties" (p.348).

It is unclear whether this kind of inference can be based on behavior observed in the laboratory or in sports events. These environments are arguably rather limited representations of the economic environments that are considered to be suitable for tournaments such as the competition for a promotion in an organization, or becoming CEO. A first important difference is the lack of potentially distracting factors which may side track people. A second important difference is the duration of the task at hand. Subjects in laboratory experiments spend at most two or three hours on their task, and also sports events are characterized by short periods of intense competition with relatively large amounts of time between events. In contrast, the interaction between employees at a particular hierarchical level who compete for promotion to the next level, can easily take some years. In their discussion of the external validity of laboratory experiments Levitt and List (2007) emphasize findings from the psychology literature which show that there are important differences between short-run (hot) and long-run (cold) decision making. In the hot phase, emotions can be very important, while these can be suppressed in the cold phase. This behavior is illustrated by the findings of Gneezy and List (2006) who consider gift exchange experiments which have been popular in laboratory experiments. When they run gift exchange experiments in a natural setting they find that employees' positive responses to employers' gifts consisting of high wages are only short-lived.

A further element that limits the external validity of experimental results is related to sorting. Inspired by the finding that some experiments conducted in the laboratory report high variance in behavior, Eriksson et al. (2006) show that this is an artifact of the designs implemented in these laboratory experiments. They argue that in reality participants self-select into tournaments. Consequently, in their laboratory experiment they let their subjects choose between payment schemes (tournament versus piece-rates) and find that the variance in behavior is reduced in comparison to a situation in which choice is not possible.[4] A related study is Cadsby et al. (2007) who compare performance under a piece rate scheme and a fixed wage contract in a laboratory experiment. An interesting aspect of this study is that

[4]Lazear et al. (2006) show that introducing sorting affects observed sharing behavior in dictator game experiments.

in some rounds subjects can sort into the payment scheme of their choice. Based on performance in other rounds where subjects could not choose the compensation scheme. Cadsby et al. estimate that the sorting effect is twice the size of the incentive effect. Similar results are reported by Dohmen and Falk (2006) who also find that sorting decisions vary by gender, risk attitudes, overconfidence and other personality traits.

Where the *absence* of sorting in laboratory experiments limits their external validity, the *presence* of sorting poses econometric challenges to the analysis of naturally occurring data. Lazear and Rosen (1981) observed that "[i]n the real world, where there is population heterogeneity, market participants are sorted into different contests. There players (and horses, for that matter) who are known to be of higher quality ex ante may play in games with higher stakes" (fn.5). Studies using field data typically ignore such selectivity and at most assume that all selection is on observables, which may severely bias the conclusions. Davis and Stoian (2005) analyze the performance of top runners in road-race competitions. They estimate a performance equation similar to earlier studies, but also introduce a sorting equation. They conclude that runners respond to prizes both in their performance and their sorting decisions. One important limitation of their analysis is that the sorting equation is assumed to be independent from their performance equation which implies that selection is essentially assumed to be on observables.

3 Experimental Design and Data

In our field experiment participants had to sort themselves into tournaments with different prizes. The experimental design allows us to estimate the impact of the tournament on effort and performance and also investigate the importance of the confounding effect of sorting. The subject pool is drawn from two cohorts of first-year students in economics and business at the University of Amsterdam. The first cohort entered in the academic year 2004/2005, the second in 2005/2006. During the first year of their three years bachelor program all students follow exactly the same program of 14 compulsory courses for a total of 60 credits.[5]

Students follow a standard introductory microeconomics course in the second term of the first year. The course is worth 7 credits, which implies a nominal study-load of 196 hours. The course was taught over a period of seven weeks in November and December. The exam was held at the end of January and consisted of 35 multiple choice questions. During each of the seven course weeks, there was

[5]There are no difference in the program between the two years.

a two-hours central lecture for all students together on Monday, and there were two two-hours workgroup meetings on Tuesday/Thursday or Wednesday/Friday. Attendance of the central lecture and the workgroup meetings is not compulsory.

We invited the students to participate in the field experiment during the first central lecture of the course, which was held in a lecture hall with almost all students present. We explained that we would be organizing three separate tournaments. Within each tournament the student who answered most multiple-choice questions correctly at the exam would be declared the winner and would receive a prize.[6] The prize differed between tournaments, and was 1000, 3000 and 5000 euros respectively. It was made clear that students could participate in only one tournament, and therefore had to choose the prize for which they wanted to compete and that after having chosen their preferred tournament, one out of two students was randomly selected to actually participate in the tournament. We explained that those randomized into the tournament would compete with others that i) selected the same tournament and, ii) were also randomized into the treatment.

Students could participate by filling out a form that asked their name, age, gender, math score in secondary school, the prize they wanted to participate for, their subjective evaluation of how well they expected to do on the exam relative to others, and their consent to link information from the experiment to information from the students' administration.[7] Application forms were distributed during the break of the first lecture and, for students not attending this lecture, also during the workgroups in the first week. Forms had to be handed in no later than 5pm of the Friday of the first course week.

The result of the randomization of students to the treatment and control group was announced at the start of the second week, during the central lecture in the lecture hall on Monday and on the teaching website on the intranet. The announcement also explicitly communicated the number of competitors in each tournament to the participants.

The participants in this field experiment are recruited from the same population as the subjects usually participating in laboratory experiments. Hence, the differences that we will report between the results emerging from laboratory experiments and the results emerging from our study cannot be explained by systematic

[6]If the highest score in a tournament was shared by more than one student, then the prize was divided among these students. This happened in one of the six tournaments we organized.

[7]The question asking about students' subjective rank reads: "Assume that this reward experiment would not take place. Out of 100 randomly chosen first-year economics and business students in this university, how many do you expect to perform better on the microeconomics exam than you?" For the purpose of the analysis we reversed the ordering of this measure so that a higher score means a better subjective rank.

Table 1. Numbers of participants [test-taking participants]

	2004		2005	
	Treated	Controls	Treated	Controls
1000 prize	25 [23]	25 [24]	32 [32]	32 [29]
3000 prize	59 [51]	58 [51]	58 [50]	58 [55]
5000 prize	56 [48]	58 [48]	69 [64]	69 [67]
All	140 [122]	141 [123]	159 [146]	159 [151]

differences in subject pool. Moreover, with a subject pool of students in economics and business, we give the predictions of tournament theory a fair chance since these fields are likely to attract students who are more competitive and more sensitive to financial incentives than the average person in the population (Carter and Irons, 1991; Frank et al., 2000).

In our design both the size and composition of each tournament is endogenously determined. Since the randomization scheme randomizes half of the students out of the treatment we have a control group for each tournament. This means that we can estimate the incentive effect for each separate tournament, but across tournament comparison will be confounded by the self-selection of students. This setup allows us to contrast experimental and non-experimental results.

Table 1 shows how the participants in the experiment sorted themselves over the three tournaments, separately for both years. It also shows how many students in each group took the exam. In both years around 20 percent of the participants opted for the tournament with the low (1000 Euros) prize. The remaining 80 percent split about equally over the other two tournaments. The sizes of the tournaments are relevant as they indicate the number of competitors for those who were assigned to the treatment groups.

From the administrative record of the university we obtained, for each participant in the experiment, the results on the exams they took in the first term. These first term scores and the mathematics grade students achieved at the centralized matriculation exam for secondary school serve as our measures of ability. The score on the microeconomics exam is our measure of productivity. The subjective rank reported by the student in the form at the start of the experiment can also be a measure of ability, but it may also capture elements of motivation. Furthermore, the instructors of all workgroups kept track of students' attendance of the workgroups. Finally, we added an extra question at the end of the exam that asked students to report how many hours they spent preparing for the exam. Workgroup attendance and exam preparation are our measures of effort.

Table 2. Balancing of treated and controls by tournament - pre-treatment variables

	1000		3000		5000	
	Treated	Controls	Treated	Controls	Treated	Controls
Age	19.23	19.36	19.44	19.59	19.48	19.63
Male	0.61	0.68	0.71	0.72	0.74	0.73
Ability	7.12	7.08	7.25	7.27	7.31	7.44
Credits	6.42	6.42	7.10	6.34	7.09	6.95
GPA	5.49	5.40	5.76	5.47	5.88	5.91
Subjective Rank	60.95	60.23	62.50	66.53	65.07	65.33
Prior Attendance	1.63	1.53	1.54	1.47	1.51	1.48

Having not only data on productivity but also on effort make the data quite unique and particularly suitable for a empirical investigation of tournament theory. Data from sports tournaments are often restricted to either a measure for effort or a measure for productivity, while in laboratory experiments productivity is equal to effort plus some noise term.

For the assignment of students to the treatment and control group, we use stratified randomization based on high school math score. We constructed subsamples of students with similar math scores and divided within each subsample the students equally over the treatment and control group. The main reason for using stratified randomization is to reduce the risk of ending up with an unequal distribution of ability between treatment and control groups. Table 2 presents sample means of pre-treatment variables for treatment and control groups separately for each tournament. With the exception of subjective rank in the 3000 tournament which is significant at the 10 percent level, none of the differences between treated and controls is significantly different at the 10%-level or better based on chi-square tests for gender and on t-tests for the other variables. The main message from this table is that within each tournament, treated and controls are balanced in terms of their observed characteristics. This is to be expected given the random assignment to treatment and control groups.

Table 2 also indicates that students with better math scores and who performed better in the first term (more credits collected and a higher GPA) are more likely to choose a tournament with a higher prize. This provides some evidence that higher ability students are more likely to self-select into the higher prize tournament. This is also seen in students' subjective position, students who rank themselves higher seem to be more likely to choose a higher prize tournament.

Table 3 compares the different tournament for the two years to investigate the

Table 3. Differences between tournaments, Pooled Sample

				Rank-sum Test			
	1000	3000	5000	1vs3	1vs5	3vs5	J-T Test
A. Pre-Treatment Characteristics, Treated & Controls							
- Age	19.3	19.5	19.6	0.077	0.212	0.487	0.466
- Male	64.9	71.6	73.3	0.209	0.103	0.666	0.144
- Ability	7.1	7.3	7.4	0.273	0.028	0.147	0.020
- Credits	6.4	6.7	7.0	0.556	0.212	0.404	0.190
- GPA	5.4	5.6	5.9	0.353	0.015	0.076	0.010
- Subjective Rank	60.6	64.5	65.2	0.065	0.007	0.290	0.011
- Prior Attendance	1.6	1.5	1.5	0.426	0.321	0.830	0.388
B. Exam Outcomes, Controls							
- Test Taker (%)	93.0	91.4	90.6	0.717	0.589	0.823	0.606
- Score	18.7	18.9	21.2	0.850	0.006	0.001	0.001
- Pass (%)	35.1	32.8	46.5	0.761	0.151	0.030	0.048
- Attendance	6.2	6.3	6.8	0.866	0.339	0.397	0.299
- Preparation (hours)	26.6	22.4	23.6	0.110	0.228	0.545	0.475

Note: Sample averages per tournament. The column 1vs3 (1vs5, 3vs5) reports p-values from Wilcoxon rank-sum tests comparing the 1000 and 3000 tournaments (1000 and 5000, 3000 and 5000 resp.) based on the pooled sample. The last column reports the p-values from a two-sided Jonckheere-Terpstra (J-T) test for ordered alternatives.

sorting of students into the different tournaments.[8] Panel A shows shows the average individual characteristics and the different ability measures for both the treated and controls. The main result is that more able students tend to select themselves into the tournaments with the higher prizes. Students in the higher prize tournaments have on average better high school math scores and scored higher grades in the first term. Students who opted for a tournament with a higher prize also tend to report a higher perceived position in the exam score distribution. Students that rank themselves higher also tend to choose higher prize tournaments.

Finally, one can conclude from panel A that there are no substantial differences in student populations between the two years in which the experiment was conducted in terms of gender and age. Recently studies have documented gender differences in competitive behavior (e.g. Gneezy et al., 2003; Gneezy and Rustichini, 2004) and suggested that they may explain gender gaps in the labor market. Niederle and Vesterlund (2007) let participants in a laboratory experiment choose between a piece rate and a tournament. They find that men are twice more likely to choose the tournament than women. Interestingly we do not find important differences

[8]Table A3 reports these data separately for 2004/5 and 2005/6.

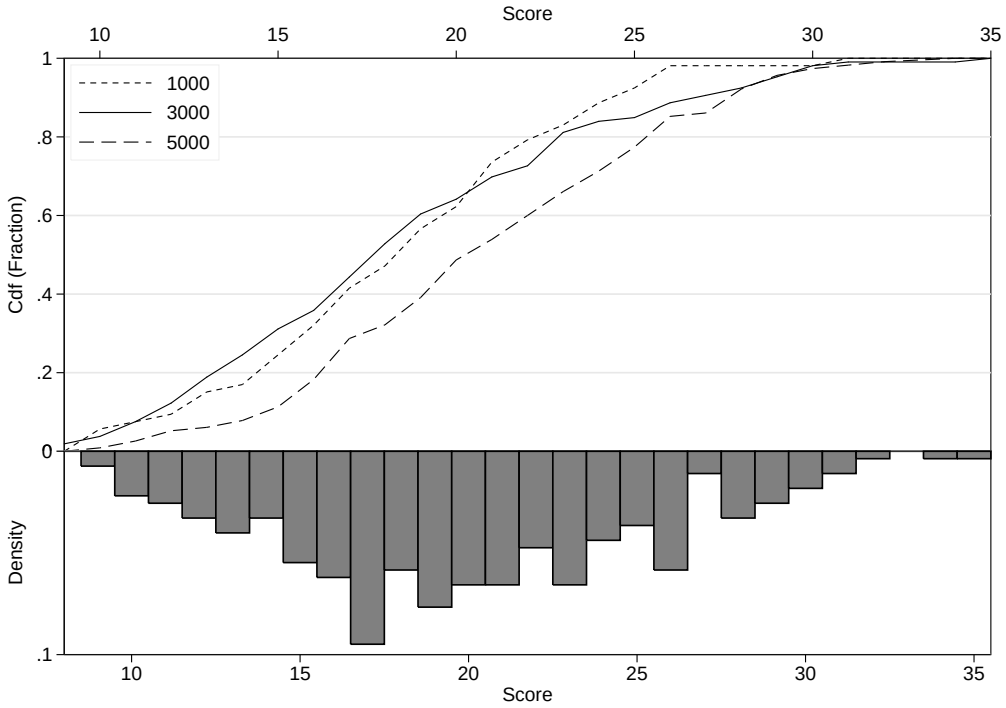

Figure 1. Exam score distributions in the control groups

between men and women with respect to sorting between tournaments.

Further evidence for the sorting of students comes from investigating the students in the control group. Assuming that the students in the control group are not affected by the tournament, differences in effort or productivity between students who selected into different tournaments are due to sorting. Panel B in Table 3 shows the results for outcome measures by tournament choice. The most important result is that students in the 5000 euro tournament score significantly better at the exam than other students. The productivity of these students is therefore higher independently of the financial incentives of the tournament scheme. There are however no significant differences in average effort between the students in the different tournaments.

The lower panel of Figure 1 shows the exam score distribution in the population (of controls) and the upper panel provides additional evidence with respect to sorting by plotting the cumulative distributions of the exam score distributions of the participants assigned to the control groups for each tournament separately. The distribution of the 5000 league clearly stochastically dominates those of the 1000 and 3000 leagues. The ordering of the 1000 and 3000 distributions is not monotonic in prize money, although above the mean (20) the cumulative distribution of the 3000 league dominates the 1000 league. Even though the correlation between the

means of the exam score distributions of the controls and the size of the reward across the six tournaments is nearly perfect, Figure 1 shows that the distributions overlap substantially.[9]

4 Results

4.1 Non-experimental analysis

Empirical papers using data from sports or companies only have information on those individuals who actually play for the prize. In the spirit of these non-experimental analyses we start out by restricting the analysis to participants who were assigned to the treatment group and thus could win a prize. We follow Ehrenberg and Bognanno (1990a) and estimate regressions of the following form

$$y_i = \delta p_i + \alpha n_i + \beta x_i + \varepsilon_i \tag{1}$$

where y_i is student's i performance on the exam, p_i is the size of the reward in the tournament the student is participating in, n_i the number of competitors the student faces, and finally x_i which is ability, or ability relative to the competitors. Table 4 shows estimates of (1) using different sets of covariates.

The first regression, which does not include any ability and tournament control variables, shows a positive effect of reward size on productivity but this effect is small and lacks significance. Recall however from Table 1 that the tournament with larger prices tend to have more competitors which suggests that the coefficient in column (1) is likely to be downward biased. This is confirmed by the second column where we control for the number of competitors in the tournament. The effect of the reward size increases and differs significantly from zero. The point estimate suggests that productivity goes up by one correct response (around 0.2 of a standard deviation in the score) for each 1000 euros increase in the prize. At the same time about ten additional competitors decrease the number of correct responses by one.

One might still be concerned by remaining omitted variable bias. Especially ability bias arising from sorting where higher prize tournaments attract more able participants is a concern. We can however control for participants' ability (math score) since we collected information on math grades students obtained at high school matriculation. Column (3) shows the results from this specification. The R-squared increases substantially confirming that the math score is a good measure of ability while at the same time the impact estimate on prize money remains similar.

[9]For some additional results on sorting and heterogeneity see Appendix B.

Table 4. Regressions of score on reward size for treated

	(1)	(2)	(3)	(4)
Prize Money/1000	0.318	1.118	1.023	1.094
	(0.223)	(0.445)**	(0.415)**	(0.415)***
# Competitors		−0.103	−0.107	−0.105
		(0.047)**	(0.043)**	(0.043)**
Ability			1.861	
			(0.251)***	
Ability − $\overline{\text{Ability}}_{\text{Competitors}}$				1.801
				(0.245)***
R-squared	0.01	0.03	0.21	0.20

Note: Robust standard errors in parentheses. Number of observations equals 260. **) significant at 5 percent level, ***) significant at the 1 percent level.

Finally, one might argue that what matters is not so much ability, but rather ability relative to the competitors in the tournament. In the final column the math score is therefore replaced by the difference between participants' own math score and the average math score of their competitors in the tournament.[10] We again find that average productivity goes up by one correct response for each 1000 euros increase in prize money.

The results in Table 4 support the predictions of the tournament model: An increase in reward size and a decrease in the number of competitors enhance productivity. These conclusions are also confirmed when we estimate the regressions in Table 4 on the sample of participants who were assigned to the control groups and were therefore not exposed to the tournament incentives. Using these observations the "effects" of rewards are even larger, and the point estimate in the first column is already significantly different from zero (see Table A1 in the Appendix). In the remainder of this section we will show that these conclusions are not confirmed by analyses that use the control groups for inference, and are in fact an artifact of participants' self-selection into tournaments.

[10]This last specification is very similar to the specifications in the papers of Ehrenberg and Bognanno. The difference is that they include own ability and the mean of competitors' ability as separate regressors. In our data these variables suffer from multicollinearity since the correlation between the mean math score of competitors and the size of the reward equals 0.83. Including the difference in math scores circumvents this multicollinearity problem and at the same time captures the spirit of the specifications in Ehrenberg and Bognanno (1990a,b).

4.2 Experimental results

We now turn to the experimental results of the paper. Using the control groups we can estimate the causal effect of participating in the tournaments on effort and productivity. Our measures of effort are attendance per workgroup meeting, aggregate workgroup attendance and time spent (in hours) preparing for the exam.

Before turning to the main results with respect to effort we start out by investigating the relation between effort and performance. Using the sample of controls we regressed exam performance on both workgroup attendance and a spline of preparation hours while controlling for ability, gender and cohort. The left panel of Figure 2 shows the results for preparation hours. The solid line is the best fitting spline with a single kink (at sixteen hours) in an OLS regression, and the dashed line is based on a more flexible local linear regression. For each level of preparation the circles indicate actual average performance and their size is proportional to the number of observations. As can be seen from the graph, the relation between performance and effort as measured by self-reported preparation hours is highly non-linear. In the beginning preparation time is positively and significantly correlated with achievement, and three hours of preparation time are associated with approximately one correct exam response. We see this up to sixteen hours – which covers about forty percent of the population – whereas after this point the positive relation disappears and even becomes slightly negative. The right panel proceeds in the same way for workgroup attendance. Here we find a strong monotonic and linear relationship between workgroup attendance and performance. Over the whole range three workgroup meetings imply one more correct exam answer.

The full regression results, reported in Table A2, show that the relationship between workgroup attendance and performance is highly statistically significant, as is preparation time up to sixteen hours. Although one must be careful with giving these results a causal interpretation, they suggest that in particular workgroup attendance, and also preparation time over a more limited range are both measures of student effort which significantly affect performance.

Table 5 shows results from equations in which the various effort measures are regressed on a single treatment dummy. The set of control variables consists of age, a dummy for gender, math score, subjective rank and dummies for reward size and cohort. We do not find any impact on total preparation time for the exam and on total workgroup attendance. However, when we split up workgroup attendance and consider the impact on separate meetings, we see that the treatment effect for the first meeting is significantly different from zero (at the 5%-level): treated participants were 7 percentage points (at a base of 74 percent) more likely to attend

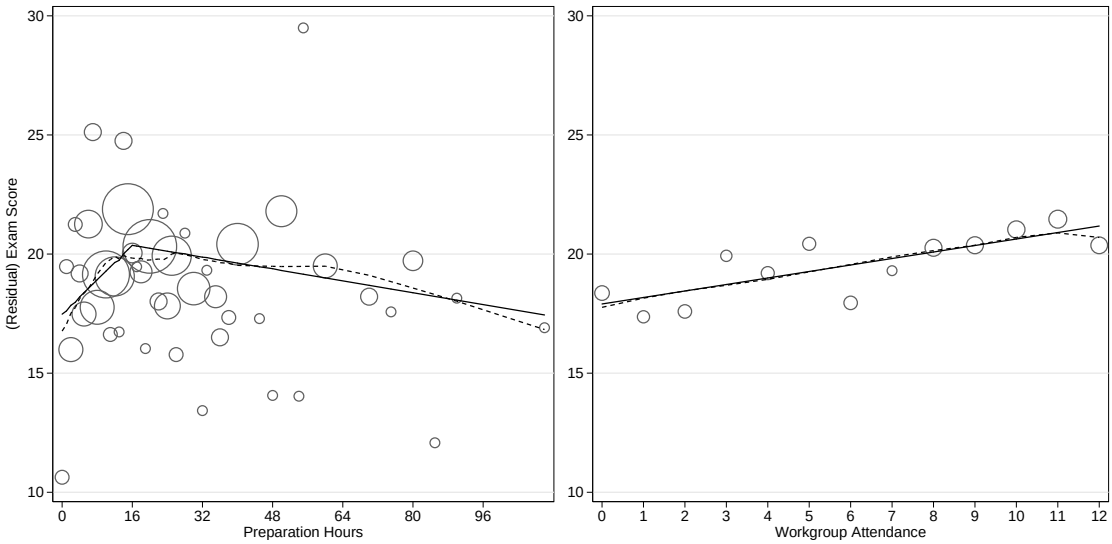

Figure 2. Effort and performance, Controls only

the first meeting than the controls. This suggests different behavior in the short-run (hot) and in the long-run (cold).

Given that there appear to be no (lasting) effects of treatment on effort we should not expect any impact on productivity, unless our measures of effort do not pick up all relevant dimensions. This could be the case if the tournament incentives do not change the extensive margin but rather the intensive margin of study time, i.e. the efficiency of time spent studying.

Table 6 reports the effects of treatment on two measures of productivity: a binary indicator for taking the exam and the actual exam score, which is the number of correct responses on the 35 multiple choice questions. The first column shows some minor differences in test taking between the treated and control which are never statistically significant. The second shows how the treatment affected average performance in the different leagues. Results are mixed, although the treated score on average one more correct answer in the 3000 prize group, we also find a negative point estimate for the 5000 tournament. In the final column we included the student who did not take the test and assigned zero correct answers as their exam score. For the 1000 and 5000 tournaments this does not change our findings, and the positive effect for the 3000 league disappears. It seems therefore that the treated exam takers in the 3000 tournament are on average somewhat better than the control students in that league who end up taking the test. We interpret these results as being consistent with the findings for effort: we find no indication of any impact of the treatment on mean productivity.

Incentives in the experiment are given by rewarding the highest exam score in

Table 5. Effect of tournament incentives on effort

Effort Measure	Estimate (s.e.)
Preparation hours	−0.302 (1.461)
Total attendance	0.075 (0.271)
Attendance by Meeting	
- 1st meeting	0.067 (0.034)**
- 2nd meeting	−0.008 (0.037)
- 3rd meeting	−0.014 (0.036)
- 4th meeting	−0.001 (0.037)
- 5th meeting	−0.041 (0.038)
- 6th meeting	0.047 (0.037)
- 7th meeting	−0.029 (0.039)
- 8th meeting	0.042 (0.038)
- 9th meeting	−0.019 (0.039)
- 10th meeting	0.027 (0.039)
- 11th meeting	−0.023 (0.040)
- 12th meeting	0.027 (0.038)

Note: Each estimate comes from a separate linear probability regression. The specification controls for age, gender, math and subjective rank and dummies for reward size, year and attendance during the workgroups before randomization. Robust standard errors are in parentheses. Number of observations equals 574, except in first row where only 512 test-takers are included. **) significant at 5 percent level.

Table 6. Mean effect estimates on productivity

	Test taking (1)	Exam Score (Test takers) (2)	Exam Score (All) (3)
1000 prize	0.032 (0.044)	0.974 (0.877)	1.164 (1.166)
3000 prize	-0.057 (0.043)	1.184 (0.617)*	0.005 (0.975)
5000 prize	-0.026 (0.037)	-0.629 (0.644)	-0.847 (0.969)
Pooled	-0.027 (0.024)	0.383 (0.403)	-0.153 (0.599)

Note: Each cell comes from a separate regression. Controls are the same as in Table 5. Standard errors are in parentheses.

Table 7. Ranksum tests based on top of each tournament/cohort

		Ranking by exam score			Ranking by math score			Ranking by subjective rank		
		Ranksum			Ranksum			Ranksum		
		Controls	Treated	p-value	Controls	Treated	p-value	Controls	Treated	p-value
Year	League	Top 3			Top 3			Top 3		
2004	1000	9.5	11.5	0.66	9.5	11.5	0.66	12	9	0.51
	3000	9.5	11.5	0.66	12	16 (4)	1.00	17.5 (5)	18.5	0.13
	5000	8	13	0.27	14 (4)	14	0.48	21 (5)	15	0.65
2005	1000	8.5	12.5	0.37	8	13	0.26	38.5 (6)	27.5 (5)	0.65
	3000	15	6	0.03	12	9	0.49	12.5	15.5 (4)	0.86
	5000	8.5	12.5	0.38	11 (4)	17	0.08	19.5 (4)	8.5	0.21
Pooled		313.5	352.5	0.53	354.5	425.5	0.20	665	463	0.38
		Top 5			Top 5			Top 5		
2004	1000	21.5	33.5	0.20	24.5	30.5	0.53	37.5	53.5 (8)	0.71
	3000	27	28	0.91	30.5	24.5	0.53	22	33	0.25
	5000	20	35	0.11	53.5 (8)	37.5	0.72	31.5	34.5 (6)	0.78
2005	1000	25	30	0.59	21.5	33.5	0.19	38.5 (6)	27.5	0.65
	3000	38	17	0.02	29	26	0.75	40	65 (9)	0.74
	5000	25.5	29.5	0.67	23.5	31.5	0.40	30.5	24.5	0.52
Pooled		872.5	957.5	0.53	988	1028	0.35	1124.5	1290.5	0.63
		Top 10			Top 10			Top 10		
2004	1000	81	129	0.07	89	121	0.22	112	98	0.60
	3000	107	103	0.88	110	121 (11)	1.00	206.5 (15)	118.5	0.52
	5000	87.5	122.5	0.18	106.5	103.5	0.91	257 (18)	149	0.85
2005	1000	105.5	104.5	0.97	111.5	141.5 (12)	0.82	171.5 (15)	179.5 (11)	0.11
	3000	113	97	0.54	130	123 (12)	0.32	119.5	111.5 (11)	0.50
	5000	111	99	0.64	91	119	0.29	117.5	92.5	0.34
Pooled		3527.5	3732.5	0.59	3621	4254	0.43	5705	4165	0.39

Note: In parentheses number of observations if different from 3, 5 or 10 because of ties.

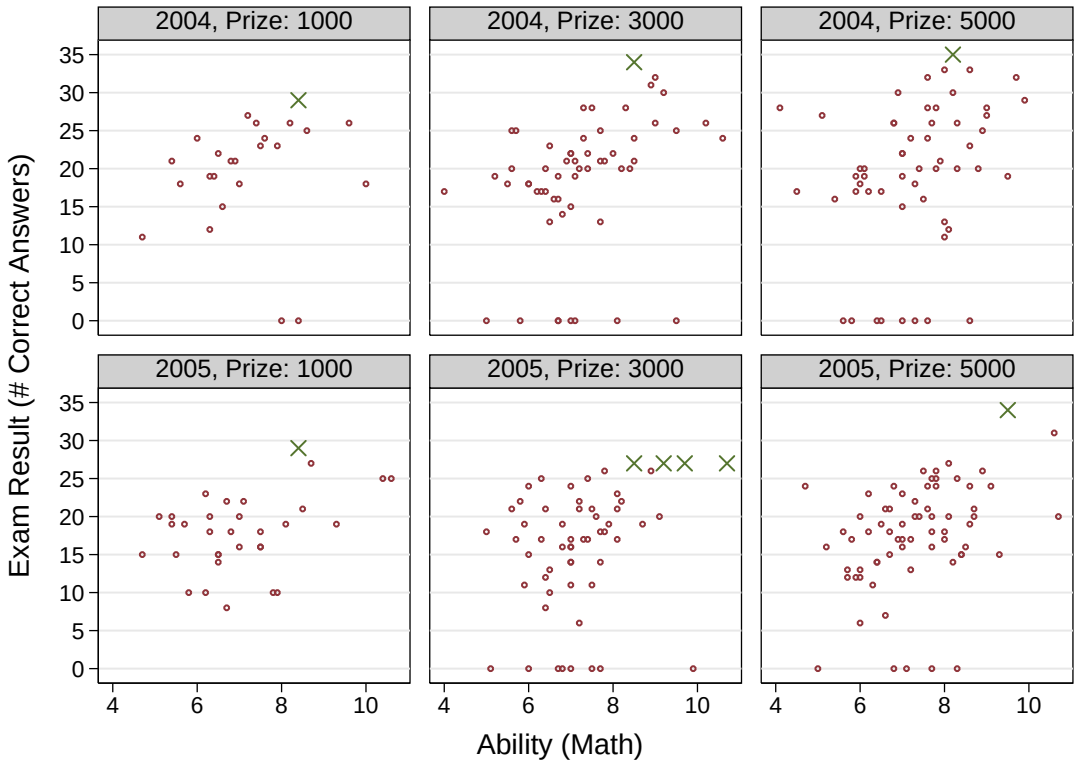

Figure 3. Tournament performance and ability

each tournament. One might therefore expect that especially exam scores at the top of the distributions are affected by exposure to treatment. This is illustrated by Figure 3 which plots for all six tournaments the performance of the participating students as a function of their ability as measured by their high school math score. The winners are marked by a cross, the losers by a circle.

It is immediately apparent from the graph that all winners come from the top of the ability distribution. With one exception the winners belonged to the top 10 in the ability distribution and are, again with one (other) exception, never the student with the highest ability as measured by the math score.

In the six tournaments we ran, only one student achieved a perfect score (35) which rules out ceiling effects. Finally, students in the 1000 and 3000 Euro tournaments performed less well than participants in the 5000 Euro tournaments. This suggests that there is scope for higher performance by increasing effort.

To examine incentive effects in the top, we ranked within each tournament all exam scores from the highest to the lowest, separately for control and treatment groups. The left part of Table 7 reports the sums of the ranks of the top 3, the top 5 and the top 10 for the controls and the treated in each tournament. The p-values of ranksum tests are reported in the final columns in these panels. Each separate test is

based on only 6, 10 or 20 observations and has therefore limited power. By pooling data from the separate tournaments we achieve more power. The p-values of the tests on the sum of ranksums are in the bottom right cell of each panel in the table. These p-values indicate that also at the top ends of the exam score distributions no treatment effects are found.

The right part of the table repeats the same analysis but now participants have been ordered on the basis of their ability (math score). Here too, we find no significant differences between the students in the tournament and those in the control group. This conclusion does not change when we order students on their subjective rank.

We also investigated gender differences in performance (not reported here). Just like we did not find differences in sorting behavior between men and women, we do not observe differences between the genders in performance either. Women do not perform worse than men in the tournaments, and women who compete (treated) do not perform worse than women who do not compete (controls). Again, this contrasts with the findings of Gneezy et al. (2003); Gneezy and Rustichini (2004); Niederle and Vesterlund (2007) who all find that compared to men women underperform when they find themselves in a competitive environment.

4.3 Spillovers

An alternative explanation for the absence of effects of the tournaments on achievement are spillovers to the controls. The most likely mechanism seems to be the one where the increased effort of treated students improves the outcomes of students in the control groups. If individuals work together then increased class attendance of treated students may also lead to higher attendance among control students. In a similar vein, if treated students become more active during class this may benefit all attending students, not only the treated ones.

Although spillovers are notoriously difficult to identify and are therefore difficult to rule out, we implemented a tentative test of this hypothesis. To do so we exploit the fact that, with the exception of central lectures, students are taught in small workgroups (as explained above). We consider it likely and assume here that if spillovers exist, they will operate at this level.

There were 9 workgroups in 2004, and 10 in 2005. Because these workgroups are relatively small (they are all around 40 students), the number of treated students will vary across workgroups because of small sample variation. The number of treated students per workgroup varied from 10 to 21 (16 on average with a standard deviation of 3). Our test consists of relating exam performance and workgroup

Table 8. Spillovers by tournament

		$n_g^T(1-t_i)$		$n_g^T t_i$	
A. Exam Score					
2004	1000	0.208	(0.256)	-0.570	(0.215)**
	3000	-0.017	(0.193)	0.321	(0.137)**
	5000	0.211	(0.188)	-0.007	(0.291)
2005	1000	0.285	(0.229)	-0.004	(0.424)
	3000	-0.067	(0.196)	-0.368	(0.284)
	5000	0.268	(0.217)	-0.119	(0.201)
Pooled		0.130	(0.095)	-0.084	(0.104)
B. Workgroup Attendance					
2004	1000	-0.276	(0.464)	-0.426	(0.303)
	3000	0.124	(0.149)	-0.146	(0.193)
	5000	-0.078	(0.204)	-0.367	(0.193)*
2005	1000	-0.090	(0.294)	0.370	(0.219)
	3000	0.281	(0.139)*	0.023	(0.100)
	5000	0.002	(0.133)	-0.092	(0.179)
Pooled		0.058	(0.084)	-0.107	(0.097)

Note: Pooled regressions include separate indicator variables for each tournament. Standard errors (between parentheses) are heteroscedasticity robust and are clustered at the workgroup level.

attendance of students to the number of treated students in their workgroup. We estimated the following regression separately for each group of students that applied to the same tournament (the treated and their controls):

$$y_{ig} = \beta n_g^T(1-t_i) + \gamma n_g^T t_i + \alpha t_i + \delta n_g + x_{ig}'\eta + \varepsilon_{ig} \tag{2}$$

Here y_{ig} is either the exam score or attendance of individual i in workgroup g, t_i is an indicator of the treatment status, and n_g^T are the number of treated students in workgroup g, and n_g the number of students in the workgroup that participated in the experiment. We control (in x_{ig}) for age, gender, ability, workgroup attendance prior to treatment assignment, subjective rank and grade point average in the first term. We allow for arbitrary heteroscedasticity in ε_{ig} and clustering at the workgroup level.

The first term on the right hand side captures spillovers to the control group, and the second term eventual spillovers or competition effects to the treated students. This is similar in spirit to Philipson (2000) and the way Miguel and Kremer (2004) estimate (across school) externalities in their analysis of spillovers of worm treatments in Kenia.

Estimates for the first two terms are reported in Table 8, the first column showing the spillover effects for students in the control group, and the second column those for students assigned to the treatment.

The results for exam scores are reported in panel A. While some of the estimates are significantly different from zero, their pattern is mixed. This is confirmed by the last row in the panel which reports results from a regression that pools all tournaments while adding tournament indicator variables. We cannot reject that there are no spillover effects. A similar conclusion arises from the estimates for workgroup attendance, the majority of which is insignificant and tend to vary in sign.

The overall picture emerging from Table 8 is that we fail to find support for spillovers. This suggests that explanations for the absence of a treatment impact need to be sought elsewhere.

5 Conclusion

We conducted a field experiment to test key predictions of rank-order tournaments in a natural setting where participants are potentially exposed to various natural distracters. To make the pool of competitors in tournaments more homogeneous and to mimic a realistic feature of real-world tournaments, we let participants sort themselves into tournaments with different reward sizes.

Our main findings are that:

1. Participants of higher quality select into tournaments with higher rewards;

2. Those who could win a prize were more likely to attend the first workgroup meeting immediately following the announcement of assignment to treatment and control groups. Exposure to treatment has, however, no lasting impact on workgroup attendance or exam preparation;

3. Exposure to treatment has no effect on students' productivity, nor on its mean level and neither when we focus on students in the top of the achievement and ability distributions;

4. A non-experimental analysis of our data falsely leads to the conclusion that higher rewards generate higher productivity. Instead the positive correlation between productivity and reward size is due to sorting.

Our findings contrast with results from previous studies that empirically test the predictions of tournament theory. These studies almost invariably find that participants in tournaments choose their effort and/or realize productivity in line with the

predictions of that theory. As the discussion in the introduction makes clear, the previous studies that measure effort and/or productivity use either data obtained in the laboratory or data gathered from sports events.

We fail to find evidence for spillovers that harmonize effort between the treated and controls. This supports our preferred explanation for the difference between our findings and the findings from laboratory experiments, namely that in laboratory experiments tasks are of short duration and participants can do nothing besides performing the task at hand, whereas the participants in our field experiment had many alternative uses of their time. Once subjects in laboratory experiments have entered the laboratory they know that they will be there for the next two or three hours. The only use of time available to them is to play the game.

The same explanation applies to the difference between our findings and the results from sport events. Professional golf players for example know that once they are on the green they have to play the nine or 18 holes. Again, there is no alternative to playing the game. In addition, the non-experimental results reported in this paper suggest that the empirical results based on data from sports events may be biased due to sorting.

The duration of our field experiment is three months from its announcement to the final exam. While a period of three months is clearly shorter than the perhaps years involved in tournaments in organizational life, it is closer to that context than laboratory experiments or sports events. Moreover, following an academic course, attending workgroups and preparing for an exam are naturally occurring events in the life of university students.

Like others before us (e.g. Gneezy and List 2006) we also find a difference between hot and cold decision-making. Participants in our experiment initially responded to assignment to treatment, by being more likely to attend the first workgroup. This supports the interpretation that the difference in duration between our field experiment and previous laboratory experiments/sports events is responsible for the difference in findings.

An alternative explanation for the absence of any lasting effect on effort or an effect on productivity is that the stakes in the field experiment are not sufficiently large. A first piece of evidence against this alternative explanation is again our finding that those assigned to treatment were more likely to attend the first workgroup. If the stakes were too small, there is no reason why they would do so. Second, in objective terms we argue the stakes are not small. Given the size of the tournaments and the prize money the expected value of exposure to treatment, assuming an equal probability to win, ranges from 31 to 89 euros. In an earlier study Leuven

et al. (2009) found that freshmen at the University of Amsterdam earn about 7.5 euros per hour in side jobs. At this wage, this means that (risk neutral) treated participants should have put in 4 to 12 additional hours into the microeconomics course. This translates into two to six extra workgroups attended. A response of that size would certainly have been identified in our data, but we observe nothing of the kind. Moreover, this is likely to be a conservative lower bound since if there are only a few potential (high ability) winners then the stakes are much higher.

The tournament model as formulated by Lazear and Rosen is an attractive model. With a simple mechanism it potentially explains a number of relevant features of internal labor markets. The evidence in this paper, however, suggests that the effort inducing effects of tournaments in rich and naturally occurring environments are less straightforward than previously thought. It may thus be that firms run tournaments not only because they provide incentives but also because they sort more productive workers into the firms that organize (higher prize) tournaments. This is comparable to the results reported by Lazear (2000), who finds that half of the productivity gain resulting from a switch from salaries to piece rates is due to sorting more productive workers. From the point of view of an individual firm it may not matter so much whether incentives are more important than sorting or vice versa. This issue has however potentially important implications for social welfare.

Finally it should be noted that the relative importance of sorting vs incentives is likely to vary with the nature of the task, the distribution of task related skills in the relevant population and the size of the incentive. Further exploration of these interactions will be an interesting area of research.

References

Abrevaya, J. (2002). Ladder tournaments and underdogs: lessons from professional bowling. *Journal of Economic Behavior and Organization*, 47(1):87–101.

Becker, B. and Huselid, M. (1992). The incentive effects of tournament compensation systems. *Administrative Science Quarterly*, 37(2):336–350.

Bognanno, M. (1990). *An Empirical Test of Tournament Theory*. PhD thesis, Cornell University.

Bull, C., Schotter, A., and Weigelt, K. (1987). Tournaments and piece rates: An experimental study. *Journal of Political Economy*, 95(1):1–33.

Cadsby, C., Song, F., and Tapon, F. (2007). Sorting and incentive effects of pay

for performance: An experimental investigation. *The Academy of Management Journal*, 50(2):387–405.

Carter, J. and Irons, M. (1991). Are economists different, and if so, why? *The Journal of Economic Perspectives*, 5(2):171–177.

Davis, T. and Stoian, A. (2005). Measuring the sorting and incentive effects of tournament prizes.

Dohmen, T. and Falk, A. (2006). Performance pay and multi-dimensional sorting: Productivity, preferences and gender. Discussion Paper 2001, IZA, Bonn.

Ehrenberg, R. and Bognanno, M. (1990a). Do Tournaments Have Incentive Effects? *Journal of Political Economy*, 98:1307–24.

Ehrenberg, R. and Bognanno, M. (1990b). The Incentive Effects of Tournaments Revisited: Evidence from the European PGA Tour. *Industrial and Labor Relations Review*, 43(3):74–88.

Eriksson, T. (1999). Executive Compensation and Tournament Theory: Empirical Tests on Danish Data. *Journal of Labor Economics*, 17(2):262–280.

Eriksson, T., Teyssier, S., and Villeval, M.-C. (2006). Self-selection and the Efficiency of Tournaments. Discussion Paper 1983, IZA.

Frank, R., Gilovich, T., and Regan, D. (2000). Does studying economics inhibit cooperation? In Connolly, T., Arkes, H. R., and Hammond, K. R., editors, *Judgment and Decision Making: An Interdisciplinary Reader*. Cambridge University Press.

Freeman, R. and Gelber, A. (2006). Optimal inequality/optimal incentives: Evidence from a tournament. Working Paper No. 12588, NBER.

Gibbs, M. (1995). Incentive compensation in a corporate hierarchy. *Journal of Accounting and Economics*, 19:247–277.

Gneezy, U. and List, J. (2006). Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments. *Econometrica*, 74(5):1365–1384.

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

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *The American Economic Review, Papers and Proceedings*, 94(2):377–381.

Harbring, C. and Irlenbusch, B. (2003). An Experimental Study on Tournament Design. *Labour Economics*, 10:443–464.

Lazear, E. (2000). Performance pay and productivity. *American Economic Review*, 90(5):1346–1361.

Lazear, E. and Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, 89(5):841–864.

Lazear, E. P., Malmendier, U., and Weber, R. A. (2006). Sorting in experiments with application to social preferences. Working Paper 12041, NBER.

Leuven, E., Oosterbeek, H., and van der Klaauw, B. (2009). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association*, forthcoming.

Levitt, S. D. and List, J. A. (2007). What do laboratory experiments measuring social preferences tell us about the real world? *Journal of Economic Perspectives*, 21(2):153–174.

Miguel, E. and Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. *Econometrica*, 72(1):159–217.

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

Orszag, J. (1994). A New Look at Incentive Effects and Golf Tournaments. *Economics Letters*, 46(1):77–88.

Philipson, T. J. (2000). External treatment effects and program implementation bias. Technical Working Paper Series T0250, NBER.

Schotter, A. and Weigelt, K. (1992). Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *Quarterly Journal of Economics*, 107(2):511–539.

Sunde, U. (2003). Potential, prizes and performance: Testing tournament theory with professional tennis data. Discussion Paper 947, IZA.

Van Dijk, F., Sonnemans, J., and Van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review*, 45:187–214.

A Appendix

Table A1. Regressions of score on reward size for controls

	(1)	(2)	(3)	(4)
Price Money/1000	0.690	2.296	2.165	2.338
	(0.201)***	(0.440)***	(0.382)***	(0.380)***
Number of competitors		−0.200	−0.196	−0.200
		(0.047)***	(0.042)***	(0.042)**
Math score			1.642	
			(0.275)***	
Ability − $\overline{\text{Ability}}_{\text{Competitors}}$				1.583
				(0.271)***
R-squared	0.04	0.10	0.24	0.24

Note: Robust standard errors in parentheses. Number of observations equals 265. ***) significant at the 1 percent level.

Table A2. Regression of effort measures on performance, Controls only

	Exam Score
Workgroup attendance	0.293
	(0.073)***
Preparation Hours (spline)	
0-16 Hours	0.284
	(0.073)***
16+ Hours	−0.033
	(0.019)*
Ability	1.654
	(0.226)***
Male	0.433
	(0.624)
Year = 2005	−3.468
	(0.560)***
Intercept	3.833
	(2.075)*
Adj. R-squared	0.325
N	259

Table A3. Differences between tournaments, Separately by year

		1000	3000	5000	1vs3	1vs5	3vs5	J-T Test
A. Pre-Treatment Characteristics, Full Sample								
- Age	2004	19.0	19.4	19.7	0.030	0.019	0.712	0.044
	2005	19.5	19.6	19.4	0.664	0.653	0.172	0.348
- Male	2004	70.0	69.0	73.5	0.895	0.650	0.455	0.523
	2005	60.9	74.1	73.2	0.066	0.080	0.864	0.166
- Ability	2004	7.1	7.3	7.4	0.656	0.169	0.223	0.119
	2005	7.1	7.3	7.4	0.282	0.088	0.418	0.089
- Credits	2004	6.5	6.6	7.4	0.777	0.174	0.166	0.110
	2005	6.4	6.8	6.7	0.588	0.628	0.907	0.733
- GPA	2004	5.6	5.7	6.1	0.774	0.035	0.023	0.012
	2005	5.3	5.6	5.7	0.374	0.192	0.754	0.264
- Subjective Rank	2004	41.2	37.6	35.3	0.206	0.026	0.139	0.017
	2005	38.0	33.5	34.4	0.108	0.098	0.957	0.192
- Prior Attendance	2004	1.6	1.5	1.5	0.207	0.412	0.584	0.673
	2005	1.5	1.5	1.5	0.910	0.541	0.405	0.434
B. Exam Outcomes, Control Sample								
- Test Taker (%)	2004	96.0	87.9	82.8	0.256	0.105	0.433	0.107
	2005	90.6	94.8	97.1	0.447	0.165	0.513	0.185
- Score	2004	19.6	20.5	23.7	0.648	0.001	0.001	0.000
	2005	18.0	17.4	19.4	0.429	0.228	0.015	0.061
- Pass (%)	2004	40.0	39.7	62.1	0.977	0.065	0.016	0.017
	2005	31.3	25.9	33.3	0.587	0.836	0.362	0.621
- Attendance	2004	6.3	6.1	6.2	0.815	0.897	0.936	0.932
	2005	6.1	6.5	7.3	0.641	0.151	0.361	0.157
- Preparation (hours)	2004	20.2	19.5	21.9	0.502	0.846	0.413	0.840
	2005	31.7	25.1	24.8	0.140	0.131	0.958	0.212

Note: See Table 3.

B Sorting and heterogeneity

Having participants sort themselves into tournaments may also reduce heterogeneity of participants within tournaments compared to the overall population. Theory predicts that a more homogeneous pool of competitors will, other things equal, induce more effort of participants. Table B1 shows to what extent heterogeneity has been reduced. First we show the standard deviation in math score. For every year we do this for the complete (pooled) population and for the separate tournaments. Within each tournament the standard deviation in math scores is not much lower than in the total population. For participants assigned to the control groups, the table shows the standard deviation of the exam score within each tournament and in the popu-

Table B1. Within tournament heterogeneity and sorting – Standard deviations

	Math Score - All		Exam Score - Controls			
	2004	2005	2004	2005	2004	2005
1000 prize	1.25	1.31	4.47	4.77	4.42	4.16
3000 prize	1.29	1.19	5.33	5.51	4.08	4.94
5000 prize	1.24	1.22	4.81	4.51	5.64	4.79
Pooled	1.26	1.23	5.22	5.00	5.22	5.00
Average	*1.26*	*1.22*	*4.96*	*4.93*	*4.80*	*4.73*
Tournament	Actual		Actual		Assigned	

lation. Participants in the low and high prize tournaments experience a reduction in heterogeneity relative to what they would experience in a randomly selected group, participants in the medium prize tournaments in contrast are confronted with a more heterogeneous group of competitors. Overall we observe, however, a reduction in heterogeneity. The final columns of the table report the standard deviations that would have been realized if participants would have been assigned to tournaments on the basis of their math scores (where the fraction of individuals in each tournament matches the actual distribution). Compared to the assignment on math score, participants' self-selection led to more homogeneous high prize tournaments and less homogeneous medium and low prize tournaments.