

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. Direct evidence comes from experiments in laboratories or from non-experimental sports events (golf, tennis). The short duration of the tasks at hand or the lack of distractors may, however, limit the external validity of the findings from these studies. Moreover, non-experimental results can be biased due to (self-)selection. To address these concerns we conducted a field experiment where 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 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 due to sorting. The only exception is that treatment has a significant impact on attendance of the first workgroup meeting immediately after the announcement of treatment status, suggesting a difference between short-run and long-run decision making.

1 Introduction

Rank-order tournaments as analyzed in Lazear and Rosen (1981) are widely accepted as labor market contracts that describe and explain relevant features of compensation and incentive structures in organizations. These tournaments are characterized

*We thank Monique de Haan, Sandra Maximiano, Erik Plug and seminar participants in Amsterdam, Leicester, Madrid, Paris and St-Gallen for fruitful discussion and comments. Leuven, Oosterbeek and Sonnemans are affiliated with the University of Amsterdam, and Van der Klaauw with the Free University Amsterdam. Leuven and Van der Klaauw are also affiliated with CEPR and IZA. All authors are affiliated with the Tinbergen Institute.

by the evaluation of individual performance relative to the performance of competitors. Opposed to 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.

Economists have used various sources to search for empirical support for tournament theory. First, there are a number of studies that use data from firms or on 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 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.

A second important 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) dealing with professional golf. 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. Both papers 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, models inspired on tournament theory have been applied to 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

¹The results in Orszag (1994) cast 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."

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 on average behavior is 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 desirability of tournaments. For instance, Van Dijk et al. (2001) conclude that "[f]rom the perspective of an 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 test environments are rather limited representations of the actual situations people have in mind when they talk about the incentive effects of tournaments (competing for a promotion in an organization, to become CEO, etc.). 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

²A 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.

³Other 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.

intense competition with relatively large amounts of time between such 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. Levitt and List (2007) argue that the psychology literature shows 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. Gneezy and List (2006), for example, consider gift exchange experiments that have been popular in laboratory experiments. When they run such experiments in a natural setting they find that employees' positive responses to employers' gifts consisting of high wages are only short-lived.

A second 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. Lazear et al. (2006) show that introducing sorting affects observed sharing behavior in dictator game experiments.

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.

We organized a field experiment in a naturally occurring setting to test the incentive effects of tournaments and the confounding effects of sorting. In our experiment students enrolled in a standard introductory microeconomics course could win a substantial prize for having the most correct answers on the course's final (multiple choice) exam. The setup of our experiment is such that participants had to select themselves into a tournament with 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 consist of a test of 35 multiple-choice questions. The number of correct answers on the test is our measure of productivity. During the course we registered participants' attendance of the course's workgroups and at the exam of the course we asked them how much time they spent preparing for the exam. These are our measures of effort.

The period 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. To assess the confounding effect of sorting we contrast our experimental finding with non-experimental empirical analyses that exploit across tournament variation in prize money and group size which has been the common way of testing tournament theory using sports data, where golf or tennis players choose in which tournaments they participate.

We find only very little evidence for effects of the prospect of winning a prize on 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. The non-experimental analysis of our data leads to the erroneous conclusion that higher rewards generate higher productivity, suggesting that sorting is an important factor in understanding tournaments.

The remainder of this paper is organized as follows. Section 2 describes the design of our field experiment. Section 3 presents and discusses the results. Section 4 relates our findings to previous empirical tests of rank-order tournaments and concludes.

2 Experimental Design and Data

The subject pool in our field experiment 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 entered in 2005/2006. During the first year of their three years bachelor program all students follow exactly

the same program of 14 courses for a total of 60 credits.⁴

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 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.⁵ The prize differed between tournaments, and was 1000, 3000 and 5000 euros respectively. It was made clear that students could participate in one tournament, and therefore had to choose the prize for which they wanted to compete. After having chosen their preferred tournament, one out of two students was randomly selected to actually participate in the tournament. Students that were randomized into the tournament would therefore compete with other students that had i) selected the same tournament and, ii) were also randomized into the treatment. The size and composition of each tournament is therefore endogenously determined. Since our 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.

Students could apply to participate by filling out a form that asked them 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.⁶ Application forms were distributed

⁴There are no difference in the program between the two years.

⁵If the highest score in a tournament was shared by more than one student, then the prize was divided among these students.

⁶The 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

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

	2004		2005	
	Treatment	Control	Treatment	Control
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]

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 17h00 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.⁷

An interesting feature is that 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 results emerging from laboratory experiments and the results emerging from our study cannot be explained by systematic 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.

Table 1 shows how the participants in the experiment sorted themselves over the three tournaments, separately for both years. It 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 (which were communicated to the participants) 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 received for each participant in the experiment their results on exams they took in the first term. The score of the microeconomics exam is our measure of productivity, while the scores on exams of courses taught in the first term and the grade for mathematics in secondary school serve as measures of ability. The subjective rank reported by the student in the

score means a better subjective rank.

⁷Ideally participants in an experiment are unaware of the experiment going on. In the context of the current field experiment this was, however, not feasible as students in treatment and control groups interact.

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 to 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. Having not only data on productivity but also on effort make the data quite unique for an empirical investigation of tournament theory. For example, 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. Gender and age are used as control variables.

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 in the higher prize tournament. This is also expressed in students' subjective position, students who rank themselves higher seem to be more likely to choose a higher prize tournament.

To investigate the sorting of students into the different tournaments we regressed individual characteristics and the different ability measures on the students' tournament choices. In Table 3 we show the results from the OLS regressions where the data are pooled for the two years. 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

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

	1000		3000		5000	
	Treatment	Control	Treatment	Control	Treatment	Control
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
Math	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	62.5	66.53	65.07	65.33
Attendance pre	1.63	1.53	1.54	1.47	1.51	1.48

Table 3: OLS estimation results for differences between tournaments; Treated and Controls

	Age	Male	Math score	Credits	GPA	Rank
3000 prize	0.22 (0.17)	0.067 (0.054)	0.16 (0.14)	0.29 (0.53)	0.16 (0.18)	4.05 (2.08)*
5000 prize	0.26 (0.17)	0.084 (0.053)	0.27 (0.14)*	0.60 (0.52)	0.45 (0.18)**	4.64 (2.06)**
Year 2005	0.04 (0.12)	0.002 (0.037)	-0.04 (0.10)	-0.24 (0.38)	-0.30 (0.13)***	2.57 (1.50)*
Intercept	19.27 (0.16)***	0.648 (0.049)***	7.13 (0.13)***	6.55 (0.48)***	5.61 (0.16)***	59.14 (1.90)***

Note: Robust standard errors for dummy dependent variables. *) significant at the 10 percent level. **) significant at the 5 percent level.

Table 4: OLS estimation results for differences between tournaments; Controls only

	Took exam	Exam score	Pass rate	Attendance	Preparation hours
3000 prize	-0.011 (0.043)	0.06 (0.84)	-0.035 (0.077)	0.14 (0.65)	-3.99 (3.27)
5000 prize	-0.023 (0.043)	2.56 (0.82)***	0.110 (0.077)	0.62 (0.61)	-3.16 (3.27)
Year 2005	0.077 (0.033)**	-3.34 (0.60)***	-0.192 (0.055)***	0.63 (0.46)	5.63 (2.26)**
Intercept	0.886 (0.037)***	20.56 (0.76)***	0.458 (0.072)***	5.82 (0.58)***	23.46 (2.78)***

Note: Robust standard errors for dummy outcome variables. **) significant at the 5 percent level. ***) significant at the 1 percent level.

prize also tend to report a higher perceived position in the exam score distribution. For GPA these differences are also significant between the 1000 and the 5000 euro tournament. Finally, one can conclude from the table 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. Students in 2005 scored lower grades in the first term than students from the 2004 cohort, suggesting that the former are on average less able than the latter.

Further evidence for the sorting of students comes from investigating the students in the control group. Since 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. Table 4 shows the results from OLS regressions of outcome measures on 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

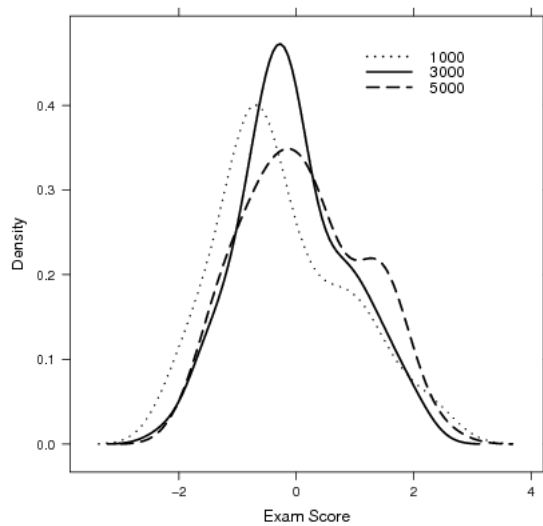


Figure 1: Exam score distributions in the control groups

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. In 2005 students prepared more for the exam and also a larger share of the students participated in the exam, but the pass rate and average exam score were significantly lower than in 2004.⁸

Figure 1 provides additional evidence by plotting the histograms of the standardized score distributions of the participants assigned to the control groups for each tournament separately. This shows that the controls in league 1000 and 3000 have more mass at the lower end of the exam scores than the controls in the 5000 league. Although 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,

⁸Note that this does not affect our results given the experimental setup of the data.

Table 5: Within tournament heterogeneity and sorting – Standard deviations

	Tournament Choice					
	Observed				Assigned	
	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
All	1.26	1.23	5.22	5.00	5.22	5.00
Weighted average			4.96	4.93	4.80	4.73

Figure 1 shows that the distributions overlap substantially.

Having participants sort themselves into tournaments may also reduce heterogeneity of participants within tournaments. Theory predicts that a more homogeneous pool of competitors will, other things equal, induce more effort of participants. Table 5 shows to what extent heterogeneity has been reduced. First we show the standard deviation in math score for the complete population in each year and for each tournament in each year. Within each group 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 overall. 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.

3 Results

We start with a non-experimental analysis of our data, meaning that we will ignore the fact that we have a control group and use across tournament variation in prize money and group size to estimate incentive effects. After that we analyze the effects

of exposure to tournament incentives on participants' effort and on their productivity exploiting the experimental design.

3.1 *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 type of non-experimental analyses we restrict 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

$$y_i = \delta p_i + \alpha n_i + x_i' \beta + \varepsilon_i \quad (1)$$

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 controls x_i , which may include ability, or ability relative to the competitors. Table 6 shows estimates of (1) using different sets of covariates.

The first specification, which does not include control variables, shows a positive effect of reward size on productivity but this effect is small and lacks significance. Recall 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 significantly differs 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 competitor decreases 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. 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

Table 6: Regressions of score on reward size for treated

	(1)	(2)	(3)	(4)
Prize Money/1000	0.318 (0.223)	1.118 (0.445)**	1.023 (0.415)**	1.094 (0.415)***
# Competitors		-0.103 (0.047)**	-0.107 (0.043)**	-0.105 (0.043)**
Ability			1.861 (0.251)***	
Ability - $\overline{\text{Ability}}$ 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.

average math score of their competitors in the tournament.⁹ We again find that average productivity goes up by one correct response for each 1000 euros increase in prize money.

The results in Table 6 support the predictions of the tournament model: we find that an increase in reward size and a decrease in the number of competitors enhance productivity. 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. These conclusions are also confirmed when we estimate the regressions in Table 6 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).

3.2 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 work-

⁹This 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).

Table 7: Effect of tournament incentives on effort

<i>Effort Measure</i>	<i>Estimate (s.e.)</i>
Preparation hours	-0.302 (1.461)
Total attendance	0.075 (0.271)
<i>Attendance by Meeting</i>	
- 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.

group attendance and the number of hour spent preparing for the exam. In Table 7 we show 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 both 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 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

Table 8: Mean effect estimates on productivity

	Test taking (1)	Score (2)
1000 prize	0.032 (0.044)	0.974 (0.877)
3000 prize	-0.057 (0.043)	1.184 (0.617)
5000 prize	-0.026 (0.037)	-0.629 (0.644)
Pooled	-0.027 (0.024)	0.383 (0.403)

Note: Each cell comes from a separate regression. Controls are the same as in Table 7. Standard errors are in parentheses.

all relevant dimensions of it. This could be the case if the tournament incentives do not change so much the time input but rather the intensity with which treated participants study.

Table 8 reports the effects of treatment on two measures of productivity: binary indicators for taking the exam and passing the exam, and the actual exam score, which is the number of correct responses to the 35 multiple choice questions. The results are 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 each tournament. We might therefore expect that especially exam scores at the top of the distributions are affected by exposure to treatment. To examine this, for each tournament we ranked all exam scores from the highest to the lowest separately for control and treatment groups. The left part of Table 9 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 students exposed to treatment and those in the control group.

Table 9: Ranksum tests based on top of each tournament/cohort

		Ranking by exam score			Ranking by math score		
		Ranksum		<i>p</i> -value	Ranksum		<i>p</i> -value
Year	League	Controls	Treated		Controls	Treated	
		Top 3			Top 3		
2004	1000	9.5	11.5	0.66	9.5	11.5	0.66
	3000	9.5	11.5	0.66	12	16 (4)	1.00
	5000	8	13	0.27	14 (4)	14	0.48
2005	1000	8.5	12.5	0.37	8	13	0.26
	3000	15	6	0.03	12	9	0.49
	5000	8.5	12.5	0.38	11 (4)	17	0.08
Pooled		313.5	352.5	0.53	354.5	425.5	0.20
		Top 5			Top 5		
2004	1000	21.5	33.5	0.20	24.5	30.5	0.53
	3000	27	28	0.91	30.5	24.5	0.53
	5000	20	35	0.11	53.5 (8)	37.5	0.40
2005	1000	25	30	0.59	21.5	33.5	0.19
	3000	38	17	0.02	29	26	0.75
	5000	25.5	29.5	0.67	23.5	31.5	0.40
Pooled		872.5	957.5	0.53	988	1028	0.35
		Top 10			Top 10		
2004	1000	81	129	0.07	89	121	0.22
	3000	107	103	0.88	104	106	0.94
	5000	87.5	122.5	0.18	110	121 (11)	1.00
2005	1000	105.5	104.5	0.97	111.5	141.5 (12)	0.82
	3000	113	97	0.54	130	123 (12)	0.32
	5000	111	99	0.64	91	119	0.29
Pooled		3527.5	3732.5	0.59	3621	4254	0.43

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

4 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 sharply 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.

Our preferred explanation for the difference between our findings and the findings from laboratory experiments is 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. Additionally, findings from sports events can be biased due to sorting, just as the non-experimental results reported in this paper.

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.

In accordance with the findings reported by 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 think that the stakes are not small. In the low prize tournaments 25 and 32 participants were competing for 1000 euros, in the medium prize tournaments 58 and 59 participants were competing for 3000 euros and in the high prize tournaments 56 and 69 participants were competing for 5000 euros. The expected value of exposure to treatment, assuming an equal probability to win, thus ranges from 31 to 89 euros. In an earlier study Leuven et al. (2003) 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, 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 because they provide powerful incentives but 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.

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).
- 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.
- 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. IZA Discussion Paper 1983.
- 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.

- 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. (2003). The effect of financial rewards on students' achievement: Evidence from a randomized experiment. CEPR Discussion Paper No. 3921.
- 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.
- Orszag, J. (1994). A New Look at Incentive Effects and Golf Tournaments. *Economics Letters*, 46(1):77–88.
- 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. IZA Discussion Paper 947.
- 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 (0.201)***	2.296 (0.440)***	2.165 (0.382)***	2.338 (0.380)***
Number of competitors		-0.200 (0.047)***	-0.196 (0.042)***	-0.200 (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.