Incentives versus Sorting in Tournaments: Evidence from a Field Experiment

Edwin Leuven, *CREST (ENSAE)*

Hessel Oosterbeek, *University of Amsterdam and Tinbergen Institute*

Joep Sonnemans, *CREED—University of Amsterdam and Tinbergen Institute*

Bas van der Klaauw, *VU University Amsterdam and Tinbergen Institute*

Existing field evidence on rank-order tournaments typically does not allow disentangling incentive and sorting effects. We conduct a field experiment illustrating the confounding effect. Students in an introductory microeconomics course selected themselves into tournaments with low, medium, or high prizes for the best score at the final exam. Nonexperimental analysis of the results would suggest that higher rewards induce higher productivity, but a comparison between treatment and control groups reveals that there is no such effect. This stresses the importance of nonrandom sorting into tournaments.

I. Introduction

Explicit financial incentives may not only lead to higher performance through their impact on agents' effort but also raise performance through

We thank Monique de Haan, Sandra Maximiano, Erik Plug, Holger Sieg, and seminar participants in Amsterdam, Århus, Berlin, Bonn, Leicester, Madrid, Paris, Southampton, and St. Gallen for fruitful discussion and comments. Contact the corresponding author, Edwin Leuven, at edwin.leuven@ensae.fr.

[*Journal of Labor Economics,* 2011, vol. 29, no. 3]

0734-306X/2011/2903-0006$10.00

sorting. Lazear (2000) analyzes worker performance in a firm that changed from paying fixed salaries to paying piece rates. He finds that half of the resulting productivity increase is attributable to an increase in productivity of the existing workforce, while the other half is due to less productive workers leaving the firm and more productive workers entering the firm. In the absence of Lazear's analysis, one might have attributed the whole productivity gain to the incentive effect of piece rates compared to fixed salaries.

In this article, we disentangle incentive effects from potentially confounding sorting effects in the context of a rank-order tournament (cf. Lazear and Rosen 1981). 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 prizes on performance. Other studies are based on laboratory experiments. Observational studies potentially suffer from selection bias when more able individuals get sorted in the more competitive and higher-prize tournaments. Experimental studies have been criticized for their limited external validity.

We ran a field experiment in which students enrolled in an introductory microeconomics course could win a substantial prize for having the best score on the course's final exam. Participants had to select themselves into a tournament with a low prize (€1,000), a medium prize (€3,000), or a high prize (€5,000). 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. The number of correct answers on the exam is our measure of productivity. During the course, we registered participants' attendance at its workgroups, and at the exam we asked them how much time they had spent preparing. These are our measures of (self-reported) effort.

The time span between the announcement of the experiment and the day of the final exam is 3 months. While this period is shorter than the time it takes to become a firm's CEO, it is substantially longer than the 2 or 3 hours of a typical laboratory experiment or sports event. To assess the confounding effect of sorting, we contrast our experimental findings with a nonexperimental 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 1990) or firm data (e.g., Eriksson 1999).

We find only very little evidence that the prospect of winning a prize affected students' efforts. 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 reported in 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 the ability distributions. This is in sharp contrast with the nonexperimental analysis of our data, which leads to the erroneous conclusion that higher rewards generate higher productivity.

The remainder of this article is organized as follows. To put our contribution in perspective, the next section discusses the related literature. Section III describes the design of our field experiment and presents and discusses the results. Section IV summarizes and concludes.

II. Related Literature

Empirical evidence regarding tournament theory comes from various sources. First, there are several studies that use data from firms or executive pay. These studies do not conduct direct tests of the incentive effects of tournaments but typically test some of the theory's indirect implications. Consistent with the theory, Eriksson (1999) finds that the pay difference increases when one moves up in the hierarchy, that an increase in the noise as measured by the variation in firm sales increases the variation in pay, and that there is a modest increase in the winning prize if the number of contestants increases. Eriksson's (1999) results are, however, equally consistent with a model in which firms use a tournament structure to attract more productive workers.

A second source of empirical support for tournament theory comes from studies in which sports events are analyzed in terms of rank-order tournaments. This research started with Ehrenberg and Bognanno (1990), who analyzed data from professional golf tournaments. They 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 support tournament theory: the level and structure of prizes influence players' performance. Other studies have applied the same framework to other sports, including bowling (Abrevaya 2002), tennis (Sunde 2009), and auto racing (Becker and Huselid 1992). Their results also support tournament theory.

A third group of studies is based on laboratory experiments, starting with Bull, Schotter, and Weigelt (1987). In this study, over 200 undergraduate students volunteered to participate in one of 10 treatments. Treatments vary features such as the prize spread, asymmetry of costs, and information conditions. The main finding is that while behavior in tournament treatments 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 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.[1]

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, 208) conclude that "from the perspective of an employer, relative payment schemes would therefore [higher effort on average] be superior." Likewise, Becker and Huselid (1992, 348) 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" or that "tournament systems have considerable motivational properties."

This kind of inference on the basis of laboratory data or sports events seems rather strong. The environments are arguably limited representations of the those that are considered to be suitable for tournaments, such as the competition for promotion in an organization or becoming CEO. A first difference is the lack of potentially distracting factors that may sidetrack people. A second difference is the duration of the task at hand. Subjects in laboratory experiments spend, at most, a number of 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 take years. In their discussion of the external validity of laboratory experiments, Levitt and List (2007) emphasize findings from the psychology literature that 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 that 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 short lived.

The external validity of experimental results is further limited by sorting. Eriksson, Teyssier, and Villeval (2009) show that the high variance in behavior in laboratory tournaments is an artifact of the designs implemented in these experiments. They argue that, in reality, participants

[1] Other studies using laboratory experiments to analyze tournaments include Van Dijk, Sonnemans, and Van Winden (2001), who introduce real effort; Schotter and Weigelt (1992), who study the effects of affirmative action programs in a tournament setting; and Harbring and Irlenbusch (2003), who focus on the effects of different tournament sizes and different prize structures.

self-select into tournaments. Consequently, in their laboratory experiment, they let their subjects choose between payment schemes (tournament vs. piece rates) and find that the variance in behavior is reduced in comparison to a situation in which choice is not possible.[2]

While the absence of sorting in laboratory experiments limits their external validity, the presence of sorting poses econometric challenges to the analysis of real life data. Lazear and Rosen (1981) observed that "in 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" (n. 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 (e.g., Davies and Stoian 2006).

III. Experimental Design and Data

In our field experiment, participants had to sort themselves into tournaments with different prizes. The design allows us to estimate the impact of the tournament on performance and (self-reported) effort 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 2004–5, the second in 2005–6. In the first year of their 3-year bachelor program, all students follow the same 14 compulsory courses for a total of 60 credits. There are no differences in the program between the 2 years.

Students follow an introductory microeconomics course in the second term of the first year. The course is worth seven credits, which implies a nominal study load of 196 hours. The course was taught over a period of 7 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 7 course weeks, there was a 2-hour lecture for all students together on Monday, and there were two 2-hour workgroup meetings on Tuesday/Thursday or Wednesday/Friday. Attendance of the lecture and the workgroup meetings was not compulsory. The workgroup meetings discussed problem sets that students were expected to prepare in advance. No formal feedback on individual performance was given during the course.

We invited the students to participate in the field experiment during the first 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

[2] Lazear, Malmendier, and Weber (2006) show that, by introducing sorting affects, they observed sharing behavior in dictator game experiments.

the most multiple-choice questions correctly at the exam would be declared the winner and would receive a prize.

The prize differed between tournaments and was €1,000, €3,000, and €5,000, 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; after having chosen their preferred tournament, half of the students were randomly selected to actually participate in the tournament. We explained that those randomized into the tournament would compete with others who (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 approval to link information from the experiment to information from the students' administration.[3] 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 5:00 p.m. on Friday of the first course week. After handing in their form, students could not change tournaments. Upon recruiting participants, we also distributed a form specifying the exact rules governing this experiment.[4]

The result of the randomization of students to the treatment and the control groups was announced at the start of the second week, during the lecture on Monday and on the teaching Web site on the Intranet. The announcement also explicitly communicated the number of competitors in each tournament to the participants.

It is not common in the Netherlands to grade exams on a curve. This was also not the case in the exam performance we study in which the student's final grade is a step function of the number of correct answers. Moreover, students never receive formal information (neither in our experiment nor during the rest of their studies) about their rank but only about their own grades. This implies that there are no explicit incentives based on rank for the controls.

The participants in this field experiment were recruited from the same population as the subjects usually participating in laboratory experiments. Hence, differences between our results and results from laboratory experiments cannot be explained by differences in the subject pool. More-

[3] 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 introductory 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.

[4] These instructions are available on request.

Table 1
Numbers of Participants (Test-Taking Participants)

	2004		2005	
	Treated	Controls	Treated	Controls
€1,000 prize	25 (23)	25 (24)	32 (32)	32 (29)
€3,000 prize	59 (51)	58 (51)	58 (50)	58 (55)
€5,000 prize	56 (48)	58 (48)	69 (64)	69 (67)
All	140 (122)	141 (123)	159 (146)	159 (151)

over, with students in economics and business as participants, 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).

In our design, both the size and the composition of each tournament is endogenous. Since half of the students are randomized 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 nonexperimental results.

Table 1 shows how the participants in the experiment sorted themselves over the three tournaments. In both cohorts, around 20% of the participants opted for the tournament with the low prize. The remaining 80% split about equally over the other two tournaments. The sizes of the tournaments indicate the number of competitors for those who were assigned to the treatment groups. The table also shows how many students in each group took the exam. There are no systematic differences in exam taking between treated and controls.[5]

Given the numbers of competitors in the tournaments and the prizes, the expected value of exposure to treatment, assuming an equal probability to win, ranges from €31–€89. A relevant comparison is the wage rate of €7.5 per hour that freshmen at the University of Amsterdam earn in side jobs (Leuven, Oosterbeek, and Van der Klaauw 2010a). Hence, the expected reward is worth 4–12 hours of work, which is equivalent to attending two to six workgroup meetings.

Information on exam results before the introductory microeconomics course was obtained from the administrative record of the university. These results, together with students' mathematics grades at the centralized matriculation exam for secondary school, are our ability measures.

[5] It is common in the Netherlands that students do not take an exam. An important reason is that there are retake exams during the year. Furthermore, dropout rates from university for freshmen are substantial.

Table 2
Balancing of Treated and Controls by Tournament—Pretreatment Variables

	€1,000		€3,000		€5,000	
	Treated	Controls	Treated	Controls	Treated	Controls
Age (years)	19.2	19.4	19.4	19.6	19.5	19.6
Male (%)	61	68	71	72	74	73
Math (1–10)	7.1	7.1	7.3	7.3	7.3	7.4
Credits (0–12)	6.4	6.4	7.1	6.3	7.1	7.0
GPA (1–10)	5.5	5.4	5.8	5.5	5.9	5.9
Subjective rank (0–100)	61.0	60.2	62.5	66.5	65.1	65.3
Prior attendance (0–2)	1.6	1.5	1.5	1.5	1.5	1.5

The score on the introductory 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 capture elements of motivation. Furthermore, the instructors 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 it. Workgroup attendance and self-reported exam preparation are our measures of effort.

We used the control group to estimate a regression of exam performance on workgroup attendance and a spline in self-reported preparation hours and control for ability, gender, and cohort. We find a coefficient of 0.293 (SE = 0.073) for workgroup attendance, a coefficient of 0.284 (SE = 0.073) for preparation up through 16 hours, and a coefficient of −0.033 (SE = 0.019) for preparation hours when they exceed 16. While we cannot give these coefficients a causal interpretation, they suggest that, in particular, workgroup attendance is a meaningful measure of performance-enhancing effort.

For assignment of students to treatment and control groups, we used stratified randomization on the basis of high school math score. We constructed a small number of subsamples of students with similar math scores and divided the students equally over the treatment and the control group within each subsample. Stratified randomization reduces the risk of ending up with an unequal distribution of ability between treatment and control groups.

Table 2 presents sample means of pretreatment variables for treatment and control groups by tournament. It shows that within each tournament, treated and controls are balanced in terms of observed characteristics. Only the subjective rank in the €3,000 tournament differs significantly at the 10% level between treated and controls.

Table 3 compares subjects' characteristics across tournaments to investigate sorting into different tournaments. More able students are more likely to select into the tournaments with the higher prizes. Students in the higher-prize tournaments have, on average, better high school math

Table 3
Differences between Tournaments, Pooled Sample

	€1,000	€3,000	€5,000	Rank-Sum Test* 1 vs. 3	1 vs. 5	3 vs. 5	J-T Test†
Pretreatment characteristics, treated and controls:							
Age	19.3	19.5	19.6	.077	.212	.487	.466
Male	64.9	71.6	73.3	.209	.103	.666	.144
Ability	7.1	7.3	7.4	.273	.028	.147	.020
Credits	6.4	6.7	7.0	.556	.212	.404	.190
GPA	5.4	5.6	5.9	.353	.015	.076	.010
Subjective rank	60.6	64.5	65.2	.065	.007	.290	.011
Prior attendance	1.6	1.5	1.5	.426	.321	.830	.388
Exam outcomes, controls:							
Test taker (%)	93.0	91.4	90.6	.717	.589	.823	.606
Score	18.7	18.9	21.2	.850	.006	.001	.001
Pass (%)	35.1	32.8	46.5	.761	.151	.030	.048
Attendance	6.2	6.3	6.8	.866	.339	.397	.299
Preparation (hours)	26.6	22.4	23.6	.110	.228	.545	.475

NOTE.—Sample averages per tournament.
* p-values from Wilcoxon rank-sum tests comparing the €1,000 and €3,000 tournaments (and €1,000 vs. €5,000 and €3,000 vs. €5,000, respectively) based on the pooled sample.
† p-values from a two-sided Jonckheere-Terpstra (J-T) test for ordered alternatives.

scores and scored higher grades in the first term. Students that rank themselves higher also tend to choose higher-prize tournaments.

Further evidence for sorting comes from the outcomes of students in the control group. Assuming that the students in the control group are not affected by the tournament, differences in (self-reported) effort or productivity between students who selected into different tournaments are due to sorting. Table 3 shows the results for outcome measures by tournament choice. The most important result is that students in the €5,000 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 (self-reported) effort between the students in the different tournaments.

The bottom of figure 1 shows the exam score distribution in the population (of controls), and the top 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 €5,000 league stochastically dominates those of the €1,000 and €3,000 leagues. The ordering of the €1,000 and €3,000 distributions is not monotonic in prize money, although above the mean (20) the cumulative distribution of the €3,000 league dominates the €1,000 league. Even though the correlation between the means of the exam score distributions of the controls and the size of the

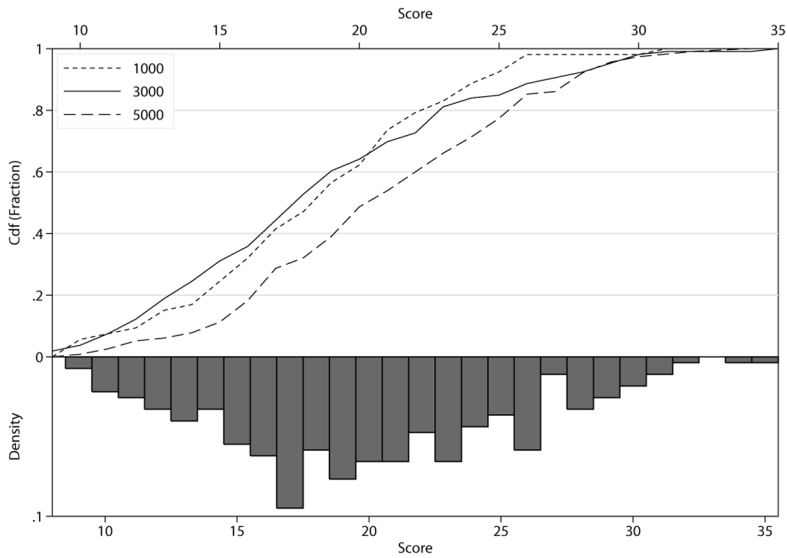

FIG. 1.—Exam score distributions in the control groups. Cdf = cumulative distribution function.

reward across the six tournaments is nearly perfect, figure 1 shows that the distributions overlap substantially.[6]

Leuven, Oosterbeek, and Van der Klaauw (2010b) analyze a simple model with two types of individuals (high ability and low ability) and two tournaments (high prize and low prize). We show that even in such a simple model, many different equilibriums may arise. In particular, there exist mixed-strategy equilibriums in which both types mix. In that case, a fraction of the high-ability individuals chooses the low-prize tournament, while a fraction of the low-ability individuals chooses the high-prize tournament. The probabilities of choosing the tournaments depend on the sizes of the prizes and the difference in ability between the two types. The overall sorting pattern that we observe in our data is consistent with such a mixed-strategy equilibrium and, therefore, consistent with ex ante rational choices. To assess the ex ante rationality of students' sorting into different tournaments at the individual level, we would need information about their utility function, their true ability (including its unobserved component), their cost of effort, and their beliefs about other students' types and strategies.

[6] For some additional results on sorting and heterogeneity, see appendix A.

Table 4
Regressions of Score on Reward Size for Treated—Nonexperimental Analysis

	(1)	(2)	(3)	(4)
Prize money/€1,000	.318	1.118	.973	
	(.223)	(.445)**	(.409)**	
I (prize money = €3,000)				8.759
				(2.251)***
I (prize money = €5,000)				10.450
				(2.663)***
No. of competitors		−.103	−.104	−.285
		(.047)**	(.044)**	(.072)***
R^2	.01	.03	.24	.28
Controls[a]	No	No	Yes	Yes

NOTE.—Robust standard errors in parentheses. Number of observations equals 260. I(X) is an indicator variable that equals one if X is true and is zero otherwise.
[a] Includes age, gender, math, subjective rank, and workgroup attendance before randomization.
** Significant at 5% level.
*** Significant at 1% level.

A. Nonexperimental Analysis

Studies using data from sports or firms only have information on individuals who actually play for the prize. In the spirit of these nonexperimental analyses, we start out by restricting the analysis to participants who were assigned to the treatment group and thus could win a prize. We thus intentionally ignore the information from observations in the control group. We follow Ehrenberg and Bognanno (1990) and estimate regressions of the following form:

$$y_i = \beta_0 + \beta_1 p_i + \beta_2 n_i + \beta_3 x_i + \varepsilon_i, \quad (1)$$

where y_i is student i's performance on the exam, p_i is the size of the reward in the tournament the student is participating in, n_i is the number of competitors the student faces, and x_i is a set of control variables including ability. Table 4 shows estimates of equation (1) for different specifications.

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. Tournaments with larger prizes have more competitors, however, which suggests that the coefficient in table 4, column 1, may be downward biased. In the second column, 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 SD in the score) for each €1,000 increase in the prize. Ten additional competitors decrease the number of correct responses by one.

One might still be concerned by omitted variable bias. Ability bias arising from sorting in which higher-prize tournaments attract more able participants is a concern, especially. Table 4, column 3, shows results from a specification including math score and other controls. The R^2 increases

Table 5
Effect of Tournament Incentives on Effort

Effort Measure	Estimate (SE)
Self-reported preparation hours	−.448 (1.463)
Total attendance	.073 (.271)
Attendance by meeting:	
First meeting	.068 (.034)**
Second meeting	−.008 (.037)
Third meeting	−.017 (.035)
Fourth meeting	−.000 (.037)
Fifth meeting	−.043 (.038)
Sixth meeting	.045 (.037)
Seventh meeting	−.026 (.039)
Eighth meeting	.040 (.038)
Ninth meeting	−.016 (.039)
Tenth meeting	.025 (.039)
Eleventh meeting	−.025 (.040)
Twelfth meeting	.031 (.038)

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

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.[7] In the final column, we relax the assumption that prize money has a linear impact on the score, by including dummies for the medium prize and the high prize. The results reject linearity ($p = .004$) and suggest that the impact is concave in the prize money ($\partial y/\partial p > 0$, and $\partial^2 y/\partial p^2 < 0$).

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. 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.

B. Experimental Results

We now turn to the experimental results of this article. Table 5 shows results from equations in which workgroup attendance and self-reported preparation hours are regressed on the treatment dummy and control variables for age, gender, math score, subjective rank, reward size, and cohort. We do not find any impact of treatment on self-reported preparation time for the exam and on total workgroup attendance. However, when we consider the impact on attendance of separate meetings, we see

[7] Results are virtually identical when we replace math score by the difference between participants' own math score and the average math score of their competitors in the tournament.

Table 6
Mean Effect Estimates on Productivity

	Test Taking (1)		Exam Score: Test Takers (2)		Exam Score: All (3)	
€1,000 prize	.035	(.049)	.895	(.822)	1.120	(1.228)
€3,000 prize	−.056	(.044)	1.246	(.616)**	.072	(.988)
€5,000 prize	−.028	(.036)	−.629	(.636)	−.869	(.958)
Pooled	−.028	(.024)	.447	(.400)	−.115	(.596)

NOTE.—Each cell comes from a separate regression. Controls are the same as in table 5. Robust standard errors are in parentheses. Number of observations equals 574, except in col. 2, where only 512 test takers are included.
** Significant at 5% level.

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%) more likely to attend the first meeting than the controls. We do not find such differences for any of the subsequent workgroup meetings. This suggests that there is a response of treatment on workgroup attendance but that this is short lived.[8]

Given that there appears to be no lasting effect of treatment on (self-reported) 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, that is, 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. Column 1 shows some minor differences in test taking between the treated and controls, which are never statistically significant. Column 2 shows how the treatment affected average performance in the different leagues. Results are mixed: the treated score, on average, one more correct answer in the €3,000 prize group, but the point estimate for the €5,000 tournament is negative. In column 3, we included the students who did not take the test and assigned zero correct answers as their exam score. For the €1,000 and €5,000 tournaments, this does not change our findings, and the positive effect for the €3,000 tournament disappears. It seems, therefore, that

[8] Another explanation is that students find out at the first workgroup meeting that they did not get lucky in terms of all the smart people ending up in another tournament. This seems less likely since most competitors come from other workgroups (there are nine or 10 of these). The central lecture will be a better occasion to learn about competitors' identities. Moreover, as we will show below, predicting the winners is not easy. Although the winners come from the upper part of the ability distributions, they are (with one exception) not the students with the highest ability scores.

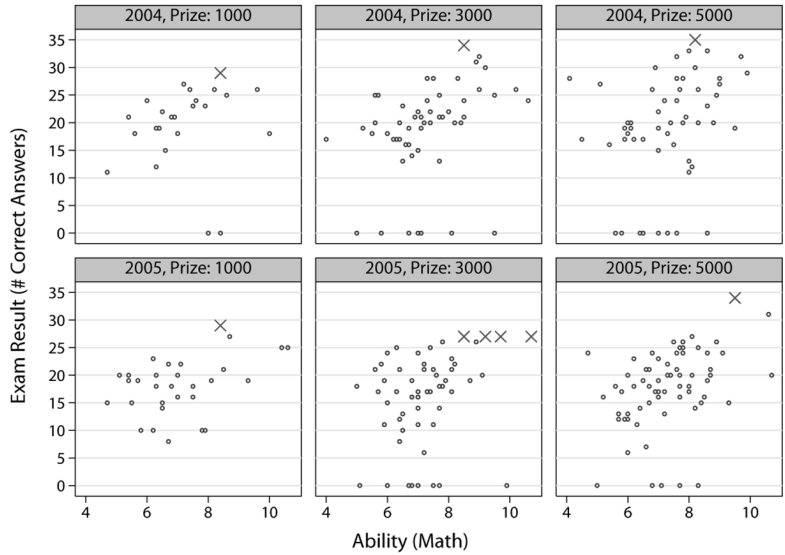

FIG. 2.—Tournament performance and math score

the treated exam takers in the €3,000 tournament are, on average, somewhat better than the control students in that league who end up taking the test. We interpret the results as being consistent with the findings for (self-reported) effort: we find no indication that the treatment had an impact on mean productivity.

Incentives in the experiment are given by rewarding the highest exam score in 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 2, 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.

The graph shows 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. Also with one exception, the winner is never the student with the highest ability as measured by the math score. This suggests that it was not a priori obvious who was going to win the tournament. Figure 2 also shows that some students who lost their own tournament could have been the winner in another tournament in the same cohort (assuming everything else equal).

In the six tournaments we ran, only one student achieved a perfect score (35), which rules out ceiling effects. Also, students in the €1,000

and €3,000 tournaments performed less well than participants in the €5,000 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. Table 7 reports the sums of the ranks of the top three, the top five, and the top 10 for the controls and the treated in each tournament. The *p*-values of rank-sum tests are also reported. Each separate test is based on only six, 10, or 20 observations and, therefore, has limited power. By pooling data from the separate tournaments, we achieve more power. The *p*-values of the tests on the sum of rank sums indicate that also at the top ends of the exam score distributions, no treatment effects are found.

Table 7 repeats the same analysis for participants 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. The rank-sum tests are only informative about the order and may not be that informative about the size of the potential effect. The additional results presented in appendix B also do not show evidence of an effect on the results of top students. While these estimates are, of course, not very precise, the results for the top 10 in exam scores exclude an impact of more than one additional correct answer with 95% probability.

C. 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 in which 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 just the treated ones.

To test this explanation, we exploit that students are taught in small workgroups. We consider it likely and assume here that if spillovers exist, they will operate at this level. There were nine workgroups in 2004 and 10 in 2005. Because these workgroups are relatively small (around 32 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 SD = 3). Our test consists of relating exam performance and workgroup attendance to the number of treated students in their workgroup. We estimated the follow-

Table 7
Rank-Sum Tests Based on Top of Each Tournament/Cohort

	Ranking by Exam Score			Ranking by Math Score			Ranking by Subjective Rank		
	Rank Sum			Rank Sum			Rank Sum		
Cohort/League	Controls	Treated	*p*	Controls	Treated	*p*	Controls	Treated	*p*
Top 3:									
2004:									
1,000	9.5	11.5	.66	9.5	11.5	.66	12	9	.51
3,000	9.5	11.5	.66	12	16 (4)	1.00	17.5 (5)	18.5	.13
5,000	8	13	.27	14 (4)	14	.48	21 (5)	15	.65
2005:									
1,000	8.5	12.5	.37	8	13	.26	38.5 (6)	27.5 (5)	.65
3,000	15	6	.03	12	9	.49	12.5	15.5 (4)	.86
5,000	8.5	12.5	.38	11 (4)	17	.08	19.5 (4)	8.5	.21
Pooled	313.5	352.5	.53	354.5	425.5	.20	665	463	.38
Top 5:									
2004:									
1,000	21.5	33.5	.20	24.5	30.5	.53	37.5	53.5 (8)	.71
3,000	27	28	.91	30.5	24.5	.53	22	33	.25
5,000	20	35	.11	53.5 (8)	37.5	.72	31.5	34.5 (6)	.78
2005:									
1,000	25	30	.59	21.5	33.5	.19	38.5 (6)	27.5	.65
3,000	38	17	.02	29	26	.75	40	65 (9)	.74
5,000	25.5	29.5	.67	23.5	31.5	.40	30.5	24.5	.52
Pooled	872.5	957.5	.53	988	1,028	.35	1,124.5	1,290.5	.63
Top 10:									
2004:									
1,000	81	129	.07	89	121	.22	112	98	.60
3,000	107	103	.88	110	121 (11)	1.00	206.5 (15)	118.5	.52
5,000	87.5	122.5	.18	106.5	103.5	.91	257 (18)	149	.85
2005:									
1,000	105.5	104.5	.97	111.5	141.5 (12)	.82	171.5 (15)	179.5 (11)	.11
3,000	113	97	.54	130	123 (12)	.32	119.5	111.5 (11)	.50
5,000	111	99	.64	91	119	.29	117.5	92.5	.34
Pooled	3,527.5	3,732.5	.59	3,621	4,254	.43	5,705	4,165	.39

NOTE.—Number of observations in parentheses if different from 3, 5, or 10 because of ties.

Table 8
Spillovers by Tournament

	Control Group: $n_g^T(1 - t_i)$		Treatment Group: $n_g^T t_i$	
Exam score:				
2004:				
1,000	.208	(.256)	−.570	(.215)**
3,000	−.017	(.193)	.321	(.137)**
5,000	.211	(.188)	−.007	(.291)
2005:				
1,000	.285	(.229)	−.004	(.424)
3,000	−.067	(.196)	−.368	(.284)
5,000	.268	(.217)	−.119	(.201)
Pooled	.130	(.095)	−.084	(.104)
Workgroup attendance:				
2004:				
1,000	−.276	(.464)	−.426	(.303)
3,000	.124	(.149)	−.146	(.193)
5,000	−.078	(.204)	−.367	(.193)*
2005:				
1,000	−.090	(.294)	.370	(.219)
3,000	.281	(.139)**	.023	(.100)
5,000	.002	(.133)	−.092	(.179)
Pooled	.058	(.084)	−.107	(.097)

NOTE.—Pooled regressions include separate indicator variables for each tournament. Standard errors (in parentheses) are heteroscedasticity robust and are clustered at the workgroup level.
* Significant at 10% level.
** Significant at 5% level.

ing regression separately for each group of students that applied to the same tournament (the treated and their controls):

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

where y_{ig} is exam score or attendance of individual i in workgroup g, t_i is an indicator of the treatment status, n_g^T is the number of treated students in workgroup g, and n_g is the number of students in the workgroup. We control for age, gender, ability, workgroup attendance before treatment assignment, subjective rank, and GPA 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 captures spillovers or competition effects to the treated students. Estimates for these terms are reported in table 8. The results for exam scores are reported, and while some of the estimates are significantly different from zero, the pattern is mixed. This is confirmed by the 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.

IV. Conclusion

We conducted a field experiment to disentangle the incentive and sorting effects in the context of a rank-order tournament in which participants

are potentially exposed to various natural distracting factors. We let participants sort themselves into tournaments with different prizes.

We find that participants of higher ability are more likely to select themselves into tournaments with higher rewards. Treatment induces higher attendance for the workgroup meeting immediately after the announcement of treatment assignment. There is, however, no lasting impact on workgroup attendance or self-reported exam preparation. Treatment also has no effect on students' productivity. The mean level is unaffected, and students in the top of the achievement and the ability distributions are not more productive when exposed to treatment. These findings contrast with results from a nonexperimental analysis of our data, which falsely lead 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 find strong incentive effects from rank-order tournaments. As the discussion in the introduction makes clear, the previous studies that measure effort 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 apart from performing the task at hand, whereas the participants in our field experiment had many alternative uses of their time. The same explanation applies to the difference between our findings and the results from sports events.

Like others before us (e.g., Gneezy and List 2006), we 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 meeting. 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 effects is that the stakes are too low. 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. Also, in objective terms the stakes are not small. The expected reward is equivalent to attending two to six extra workgroup meetings. A response of that size would certainly have been identified in our data.

The tournament model as formulated by Lazear and Rosen (1981) is an attractive model. With a simple mechanism, it potentially explains a number of relevant features of internal labor markets. The evidence in

this article, however, suggests that the effort-inducing effects of tournaments in rich and naturally occurring environments are not straightforward. It thus may 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 in line with the results reported by Lazear (2000).

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. However, this issue has potentially important implications for social welfare. If tournaments provide incentives, the firms that use them increase their output without imposing a negative externality on other firms. Instead, if tournaments merely sort productive workers to firms that use them, this sorting incurs a cost on other firms that see their most productive workers leave. The precise welfare implications depend on the extent to which other payment schemes provide incentives for optimal effort, workers' disutility of effort, and the degree of competition between firms.

We find that the difference in performance between students playing for different prize levels can be attributed entirely to sorting effects as opposed to incentive effects. This result may depend on the specific context of our experiment. While it shows that sorting effects are potentially important, there is no claim that these effects are generally more important than incentive effects. The relative importance of sorting versus 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.

Appendix A

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 A1 shows to what extent heterogeneity has been reduced. First, we show the standard deviation in math score. For every cohort, 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 population. 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.

Table A1
Within-Tournament Heterogeneity and Sorting—Standard Deviations

	Math Score—All		Exam Score—Controls			
	Actual Tournament		Actual Tournament		Assigned Tournament	
	2004	2005	2004	2005	2004	2005
€1,000 prize	1.25	1.31	4.47	4.77	4.42	4.16
€3,000 prize	1.29	1.19	5.33	5.51	4.08	4.94
€5,000 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

Appendix B

Incentive Effects at the Top

We selected from each tournament the best three (five or 10) students from the treatment group and the best three (five or 10) students from the control group. We use three different measures to define the best students: exam score, math score, and subjective rank.[9] Pooling the data from the six different tournaments—and due to some ties in the ranking—this gives us 38–47 observations for the top 3, 62–68 observations for the top 5, and 123–34 observations for the top 10.[10] We estimated the same specification as the one reported in table 6, column 3, using these subsamples. The results are reported in table B1.

None of the reported coefficients is significantly different from zero, and with one exception all point estimates are either close to zero or even negative. This suggests that for students in the top of the ability distribution, exposure to tournament incentives has no impact on their exam score. Due to the small numbers of observations, the estimates in table B1 are not very precise. Because of this, we cannot in all cases exclude substantial positive effects of treatment. For the top three based on the

[9] Note that when we select the best students on exam score, sampling is endogenous, and estimates may be biased.

[10] Without ties, these numbers of observations would have been 36, 60, and 120, respectively.

math score, the estimate is very imprecise, and a positive impact of eight correct answers falls in the 95% confidence interval. In most of the other cases, the upper bound of the 95% confidence interval is, however, much smaller and usually below 2 and even close to 1.

Table B1
Effect Estimates on Exam Score for Top Students in Each Tournament

	Ranked by Exam Score		Ranked by Math Score		Ranked by Subjective Rank	
	Coefficient (SE)	*N*	Coefficient (SE)	*N*	Coefficient (SE)	*N*
Top 3	.488 (.697)	42	1.631 (3.120)	38	−2.223 (2.390)	47
Top 5	.557 (.561)	66	−.786 (2.153)	62	−1.430 (1.867)	68
Top 10	.263 (.381)	135	−.590 (1.238)	123	−1.228 (1.379)	134

NOTE.—Each cell comes from a separate regression. Controls are the same as in table 5. Robust standard errors are in parentheses.

References

Abrevaya, Jason. 2002. Ladder tournaments and underdogs: Lessons from professional bowling. *Journal of Economic Behavior and Organization* 47, no. 1:87–101.

Becker, Brian E., and Mark A. Huselid. 1992. The incentive effects of tournament compensation systems. *Administrative Science Quarterly* 37, no. 2:336–50.

Bull, Clive, Andrew Schotter, and Keith Weigelt. 1987. Tournaments and piece rates: An experimental study. *Journal of Political Economy* 95, no. 1:1–33.

Carter, John R., and Michel D. Irons. 1991. Are economists different, and if so, why? *Journal of Economic Perspectives* 5, no. 2:171–77.

Davies, Tom, and Adrian Stoian. 2006. Measuring the sorting and incentive effects of tournament prizes. Working Paper no. 06-08, Department of Economics, University of Arizona.

Ehrenberg, Richard G., and Michael L. Bognanno. 1990. Do tournaments have incentive effects? *Journal of Political Economy* 98, no. 6:1307–24.

Eriksson, Tor. 1999. Executive compensation and tournament theory: Empirical tests on Danish data. *Journal of Labor Economics* 17, no. 2:262–80.

Eriksson, Tor, Sabrina Teyssier, and Marie-Claire Villeval. 2009. Self-selection and the efficiency of tournaments. *Economic Inquiry* 47, no. 3: 530–48.

Gneezy, Uri, and John A. List. 2006. Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica* 74, no. 5:1365–84.

Harbring, Christine, and Bernd Irlenbusch. 2003. An experimental study on tournament design. *Labour Economics* 10, no. 4:443–64.

Lazear, Edward P. 2000. Performance pay and productivity. *American Economic Review* 90, no. 5:1346–61.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2006. Sorting in experiments with application to social preferences. NBER Working Paper no. 12041, National Bureau of Economic Research, Cambridge, MA.

Lazear, Edward P., and Sherwin Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89, no. 5: 841–64.

Leuven, Edwin, Hessel Oosterbeek, and Bas Van der Klaauw. 2010a. The effect of financial rewards on students' achievement: Evidence from a randomized experiment. *Journal of the European Economic Association* 8, no. 6:1243–65.

———. 2010b. Splitting tournaments. CEPR Discussion Paper no. 8016, Center for Economic Policy Research, London.

Levitt, Steven D., and John A. List. 2007. What do laboratory experiments measuring social preferences tell us about the real world? *Journal of Economic Perspectives* 21, no. 2:153–74.

Schotter, Andrew, and Keith Weigelt. 1992. Asymmetric tournaments, equal opportunity laws, and affirmative action: Some experimental results. *Quarterly Journal of Economics* 107, no. 2:511–39.

Sunde, Uwe. 2009. Heterogeneity and performance in tournaments: A test for incentive effects using professional tennis data. *Applied Economics* 41, no. 25:3199–3208.

Van Dijk, Frans, Joep Sonnemans, and Frans Van Winden. 2001. Incentive systems in a real effort experiment. *European Economic Review* 45, no. 2:187–214.