Absolute vs. Relative Incentives under Peer Effects in Education

Kosmas Marinakis

ICEF & Faculty of Economics,
NRU - Higher School of Economics,
Moscow, Russia

July 13, 2018

Abstract. The paper contrasts an absolute and a relative performance evaluation scheme theoretically and with a field experiment. The experiment compares agents' performance under two alternative grading methods in two separate college classes. The baseline results indicate that the conventional absolute system encourages effort more effectively than a cardinal tournament. Moreover, the experiment surprisingly shows that high-power incentive schemes, through relative performance evaluation, will consistently result in progressively lower effort levels by the students. The proposed explanation for these unexpected observations in the field is the existence of strong peer effects, which render highly socialized environments qualitatively different than the generic market setting assumed by traditional agency theory. A theoretical model with agents averse to taking actions that harm other agents is developed. All theoretical results indicate that peer effects play an important role in the determination of the superior evaluation method in socialized environments.

Keywords: Relative performance evaluation; piece rates; tournaments; field experiment; item discrimination index; answer sharing;

JEL Codes: C93, D82, I21.

Acknowledgments: I thank Walter Thurman for his help on experimental design, modeling and editorial issues. I also thank Dan Kovenock, Charles Knoeber, Thomas Noe, Alexis Belianin, Theofanis Tsoulouhas, Andrej Krickovic and Martsella Davitaya for their useful suggestions.

1. Introduction and Literature Review

A well-known result in the principal-agent literature is that the principal is able to influence the agents' choice of effort by providing appropriate incentives. Numerous contract theory articles (for example, see Holmström (1982)) have shown that, under simple linear contracts of absolute or relative evaluation, agents' efforts can be adjusted by the principal to the level that serves the principal's interests. Undoubtedly, the instructor-student grading relationship can be seen as an instance of the classic principal-agent model. In the classic model, the optimal level of effort for the agent is the one that maximizes the utility gained from monetary compensation net of cost of effort. If effort is unobservable or non-contractible, contracts must be contingent on final output. Likewise, in the instructor-student regime, students choose the optimal level of studying in order to keep the disutility of effort low but also, to maximize the utility gained from grades assigned by the instructor, who can only observe their performance in tests, exams and assignments. The grading method used by the instructor is announced ex ante in the course syllabus and it essentially resembles a compensation contract. Given the parallel nature of these two settings and the fact that in the standard principal-agent model, effort is known to be elicited by the principal, a reasonable question to ask is *can the instructor induce students to study harder by adjusting the incentives in the grading method*? The answer to this question is not straightforward.

Despite the similarities, the instructor-student relationship exhibits a number of substantial differences compared to the standard principal-agent model. First, the instructor, unlike the principal, is not the residual claimant of the educational success of the students. Second, the students do not face the possibility of binding participation constraints similar to those in standard agency theory. Third, and most important, in the standard principal-agent model, agents do not have social concerns, while in the instructor-student regime, social factors may play an important role in students' decisions. The classroom, far different from the impersonal market considered in agency models, is a highly socialized environment, in which students form long lasting social relationships with their peers and thus in addition to self-interest, likeability, popularity and social status are important concerns when making choices.

The existence of social considerations is not limited to the classroom but it is indeed a prevalent characteristic in many agency situations, where agents form long standing relationships and their payoffs are interdependent. Peer effects are common in many business workplaces, especially when agents' actions affect their own utility as well as the payoffs of their peers. In such environments, agents may exhibit distaste for actions that inflict negative results to other agents. It is not uncommon for over-performers to not be particularly likable in school, in the workplace or in any competitive settings with intense social dynamics. For example, consider long time co-workers competing for a promotion. Even though strategies that can be considered extreme from a social perspective may maximize the likelihood of receiving the promotion, we would expect that most candidates will refrain from choosing those in fear of becoming social outcasts in the workplace. Instead, the contestants are socially conscious and this causes them to moderate their actions. Instances of the above scenario may include associate attorneys who compete to become partners in law-firms and faculty whose annual raises are based on performance relative to their peers.

The aspect from which we consider peer effects is qualitatively different than the predominant view of Fehr and Schmidt (1999). We introduce a form of inequity aversion that is based on the assumption that agents are averse to taking actions that actively harm other agents. This is particularly pertinent under the relative performance evaluation, where over-performance by an agent will cause a decrease in the compensation of lower performing agents and this will decrease the over-performer's utility. In our notion of inequity aversion the agents feel no envy, while their feeling of altruism does not refer to the comparison of payoffs per se, but rather to whether the agent has actively caused other agents to receive a lower payoff.

Empirical investigation can indicate the extent to which peer effects make a difference in the behavior of agents. Deviation of the empirical observations in the classroom from the theoretical expectations generated by standard agency models may indicate that peer effects are of importance for the theory and practice of incentives in highly socialized environments. This paper adopts an experimental approach to detect the effect of incentives in the instructor-student grading relationship. We present a field experiment that was conducted in real college classes for the sole purpose of investigating if the provision of high-power incentives to students through different grading methods can stimulate student performance. The experiment considers two alternative grading schemes for identical homework assignments in two separate sections of the same college course. The first section was evaluated according to the traditional *absolute evaluation* method, while the second one according to a *relative evaluation* method. The contrast of the two methods can yield valuable insights. Under the traditional absolute grading system, the instructor calculates each student's score taking into account only the student's own performance. In this scheme the power of incentives cannot be adjusted in a meaningful way by changing the amount of points per correct answer because this would cause only a nominal change in grading.[1] Under the relative grading system, each student is evaluated according to individual performance relative to the average performance of the class.[2] The relative method incorporates a *coefficient of incentive*

[1] The discussion refers to linear schemes. In a non-linear absolute grading scale the power of incentives may be adjusted meaningfully.

[2] A usual form of the relative method is the cardinal tournament. That is, grading is calculated according

power which the instructor can willingly adjust. That is, the instructor can set the rate at which points from those who performed below the average will be transferred to those who performed above it.

The classroom turns out to be an ideal field for experimenting with incentives. Contrary to a laboratory setting, the subjects are studied in their natural environment doing exactly what they are used to during their entire school life. Effort choices under each grading method irreversibly affect their social status in the classroom, and more importantly, their grades. The duration of the experiment -an entire semester- was long enough for every student to adapt and familiarize themselves with the effect of the experiment on their outcome in the course. Participants anticipated that the experiment would impact real aspects of their life and, therefore, their response to the alternative incentive schemes was of great interest.

The experimental approach in evaluating incentive methods is far from novel. Dechenaux, Kovenock and Sheremeta (2015) provide a detailed survey of experimental studies in the field of incentives in groups. Sheremeta (2016) surveys the benefits and disadvantages of tournaments as they have been identified in experimental (but also theoretical and empirical) literature over the past years. Several experimental works deal with the effectiveness of incentives specifically in education. Kremer, Miguel and Thornton (2009) estimate the impact of a merit scholarship program and find evidence for positive program impacts on academic performance of primary school female students in comparison to the control group. Brownback (2017) uses a field experiment to investigate how changes in the class size affect students of different abilities. He finds that under relative grading, a larger class size elicits lower effort from weaker students, while a smaller class size causes those students to exert more effort. On the other hand, high ability students fail to take advantage of the increased expected variance of performance in smaller classes. On the contrary, Figlio and Lucas (2004) conclude that along with high-achieving students, higher standards benefit low-achieving students the most. However, they consider students in a much earlier stage of their education: the elementary school, where social considerations of the subjects are somewhat different from those in high school and college. Betts and Grogger (2003) consider how grading standards affect academic outcomes over the entire spectrum of student ability. They find significant variation in the effectiveness of incentives with respect to the ranking of students. They show that provision of incentives is effective for top students but may discourage students near the bottom of performance distribution. Becker and Rosen (1992) investigate the effects of incentives on student behavior by comparing different standards

to $b + \beta(x_i - \overline{x})$, where b and β are positive parameters defined by the instructor and $x_i - \overline{x}$ is the difference of individual performance from the average performance. The constant b is analogous to a "signing bonus" and β is the power of incentives.

in testing and focusing attention on the subject's position in the distribution of student attainment. One of their baseline findings is that the composition of the group which each student is evaluated against matters in the effort choice of the student, with smaller groups generating increased effort levels. Angrist, Lang and Oreopoulos (2009) examine a field experiment where college students receive a combination of academic support and monetary incentives in order to improve course performance. The monetary incentives increase the power of incentives, while academic support decreases cost of effort. They observed mixed results with female students persistently improving, while male students were not affected.

Our paper belongs to the stream of experimental literature that compares relative and absolute compensation on effort and performance. We draw from Bandiera, Barankay, and Rasul (2005), who contrasted an absolute and a single relative evaluation method considering personnel data collected in the field. They find that, when the principal switched payment methods from a relative performance tournament to an absolute performance piece rate, productivity significantly increased. Wu and Roe (2005) use a laboratory experiment to test the difference of effort levels under a tournament and under a fixed performance standard contract. They conclude that effort is significantly higher under the fixed performance standard, which is an absolute evaluation method. More recently, Paredes (2017) investigates the effect of a transition from an absolute grading scheme to a relative one on the effort of students. Using a model with no peer effect considerations, she finds that, when uncertainty is sufficiently low, absolute grading encourages effort from stronger students but discourages effort from relatively weaker students. Czibor, Onderstal, Sloof and van Praag (2016) conduct a field experiment in university classes to compare absolute and relative performance evaluation methods with respect to their effect on student effort and performance. They find no significant impact on effort or performance. They attribute this result to low academic motivation of students in their sample.

Our experiment yields several interesting results. *First*, the data show that under equal power of incentives, that is, when the response of expected compensation to own performance is equal, relative performance evaluation is less effective than absolute performance evaluation with respect to student performance. This is in sharp contrast to many prominent theoretical results shown by Lazear and Rosen (1981), Green and Stokey (1983), Nalebuff and Stiglitz (1983) and others, which indicate that relative evaluation is superior to absolute evaluation under conditions where it is costless to observe and compare agents' performance, output includes a sufficient level of common noise, and agents are risk averse. Even in the case where agents are assumed to be risk neutral, Lazear and Rosen have shown that both evaluation methods in their native form are expected to be equally effective. This weak superiority of the relative method can be further supported by a fair amount of experimental works, where peer effects are not present. In one of the first experiments on this topic, Bull, Schotter and Weigelt (1987) do not consider any social interaction between subjects and find that the baseline results of relative evaluation theory are confirmed. Agranov and Tergiman (2013) use a controlled laboratory experiment to evaluate the performance of relative and absolute compensation methods. Their findings verify the theoretical prediction that, when peer effects or other social factors are not important, the relative method is superior to the absolute. However, when social considerations come into play the classic theoretical predictions may not be materialized. Our first result is indeed consistent with that of Bandiera et al. (2005), Wu and Roe (2005) and Duflo, Dupas, and Kremer (2011), who attribute this theoretically unexpected result to the presence of peer effects. The importance of social identification on individual behavior in relative performance evaluation is also pointed out in Mago, Samak and Sheremeta (2016), who introduce a behavioral, non-monetary utility of winning and relative payoff maximization and they find that by creating social familiarity between subjects, over-expenditure of effort is decreased. Moreover, Dubey and Geanakoplos (2010) explore how grading schemes affect the utility of students. Their main result is that when students care about their relative rank, they are better motivated by revealing their performance in less informative grading categories, rather than exact numerical scores. They also conclude that absolute performance evaluation over-performs relative performance evaluation with respect to excreted effort. As we will demonstrate later on in our model presentation and discuss extensively in our results section, this deviation between theoretical and empirical research can be explained by the fact that a relative evaluation bonus comes together with a social cost, which reduces the overall value of over-performance and makes it optimal for over-performers to moderate their effort. It is obvious that a model which takes into account social considerations is salient in an entire class of agency relationships, such as professional corporations, where high performance by one agent may alter the distribution of resources among the group.

A *second* finding is that relative evaluation is more effective, when the power of incentives in the relative method is set to a sufficiently lower level than the (fixed) power of incentives in the absolute method. This is a surprising result that to our knowledge has never been observed before. When the power of incentives was set to a value half of that in the absolute method, performance appeared to achieve its maximum. In accordance with the hypothesis of peer effects, lowering the power of incentives causes compensation to depend less to own-performance, essentially diminishing the importance of the incentive component in the agents' compensation. This limits the social cost from over-performance and enables agents to maintain high effort in order to obtain higher utility from compensation.

Our *third* experimental result is also surprising and novel. The experiment indicates that under the relative method, the power of incentives has a negative impact on students' performance. That is, an increase in the power of incentives leads students to progressively decrease effort in an attempt to lower the social cost arising from the likelihood that the student will over-perform. This provides evidence that, when the power of incentives keeps increasing, the importance of peer effects will eventually offset the compensation drive in the determination of the optimal level of effort. In a survey after the experiment, several participants explained that they felt “quite uncomfortable” knowing that they have acquired bonus points which would have been awarded to someone else if the traditional absolute method was in place.

The paper also investigates whether the use of a relative grading method prevents dishonest behavior in assignments. Under absolute grading, answer sharing is a common phenomenon, especially when the assignments consist of multiple choice questions. In fact, under absolute grading, answer sharing may be an optimal strategy because it eliminates the social cost without affecting the student's own grade. Conversely, when a relative evaluation method is in effect, a cost is inflicted on those who share their work because, by doing so, the class average increases and the expected benefit of those who gave their work away decreases. This cost turns out to be a strong disincentive for answer sharing. The experiment confirms that the relative method reduces dishonest behavior.

A special contribution of this paper is the introduction of the ‘item discrimination index’ for weighing the observations for the purpose of ensuring the uniform quality of assignment questions. The index quantifies the capacity of a question to discriminate between well prepared and less prepared students. By weighing the value of each item according to the item's discrimination index, the transformed data provide a more clear description of effort choices and normalize different assignments with respect to difficulty.

Section 2 provides a modeling framework to motivate the theoretical expectations from the experiment. Section 3 describes the experimental design and section 4 discusses the data treatment. The findings are presented in section 5. Section 7 discusses the implications and concludes the paper.

2. Theoretical Framework

Before we present the experiment it would be beneficial to develop a theoretical framework. First, as a benchmark, we will lay out the standard model of incentives without social considerations. Here, the aim is to establish the theoretical expectations, that is, what should one expect to observe in the experiment from the point of view of the standard principal-agent framework. Then, we will incorporate peer effects in the theoretical analysis. The inclusion of peer effects in the model is claimed to be the missing piece in the explanation of the experimental results presented in Section 5.

It should be noted that the peer effects model below *is not estimated in the experiment* but rather is intended to provide an idea about what sort of results one would expect to obtain by altering the standard incentives paradigm to allow social aspects to be taken into account.

2A. Theoretical Expectations

Consider a risk neutral principal and N risk neutral agents of heterogeneous ability.[3] Agents must individually undertake a task that involves production of output. Each agent, i, produces output according to the production function

$$x_i = a_i + e_i + \varepsilon_i + \eta, \tag{1}$$

where a_i is the i th agent's ability, drawn iid from an interval $[0, \overline{a}]$ according to a distribution function F which is common knowledge[4]; e_i is the agent's effort; ε_i is an idiosyncratic random shock; and η is a random shock common to all agents. Ability is private information to the agent ex ante and effort is unobservable. Both shocks follow independent normal distributions with zero means and finite variances $var(\eta) = \sigma_\eta^2$ and $var(\varepsilon_i) = \sigma_\varepsilon^2$, $\forall\ i$. Moreover the idiosyncratic shocks are independent. Upon completion of the task, the principal compensates the agents according to a pre-announced compensation scheme. If the principal uses an absolute evaluation method (a piece rate contract) the compensation to the ith agent will be

$$w_i = \gamma x_i, \tag{2}$$

where γ is the coefficient of the incentive power. If the principal offers a relative evaluation method (a cardinal tournament) with coefficients b and β, the compensation to the ith agent will be

$$w_i = b + \beta(x_i - \overline{x}_{-i}), \tag{3}$$

where $\overline{x}_{-i} \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j\neq i}}^{N} x_j$, is the average production excluding agent i.[5] Agents are exogenously assumed to have accepted the terms of the agreement and thus, there is no participation issue. However, even if an agent performs the task exerting zero effort, his expected production will still be positive because of the assumption about a_i. That is, an individual rationality constraint with reservation value of 0 would not alter the setting, since it would

[3] For the effect of the heterogeneity of agents in relative performance evaluation see Tsoulouhas and Marinakis (2007).

[4] See Moldovanu and Sela (2006) for a more detailed treatment on the assumptions for the distribution of ability.

[5] The exclusion of i from the average is done for mathematical simplicity. All results hold if a regular average is assumed (see Marinakis and Tsoulouhas (2013), for instance).

be non-binding, and thus the agent would never refuse participation. The principal is not assumed to be the residual claimant of the difference between total production and compensation to the agents. However, the principal's satisfaction is increasing in total output, $\sum_{i=1}^{N} x_i$. Agent i's payoff is

$$U_i = w_i - \frac{e_i^2}{2a_i}. \tag{4}$$

Under *absolute evaluation* the agent solves

$$\max_{e_i} EU_i = \max_{e_i} \left[\gamma(a_i + e_i) - \frac{e_i^2}{2a_i} \right], \tag{5}$$

for which the solution satisfies

$$e_A^* = \gamma a_i. \tag{6}$$

Under *relative evaluation* the agent's problem is

$$\max_{e_i} EU_i = \max_{e_i} \left[b + \beta \left[(a_i - \overline{a}_{-i}) + (e_i - \overline{e}_{-i}) + E(\varepsilon_i - \overline{\varepsilon}_{-i}) \right] - \frac{e_i^2}{2a_i} \right] \tag{7}$$

and the solution satisfies

$$e_R^* = \beta a_i. \tag{8}$$

The optimal amounts of effort derived under absolute evaluation in (6) and under relative evaluation in (8) verify the long known fact for the provision of linear incentives that the optimal effort exerted by the agents is directly analogous to the product of ability multiplied by the power of incentives.[6] Observe that (6) and (8) are almost identical with their only difference being the symbol denoting the power of incentives under each method.

As it was pointed out, first in Lazear and Rosen (1981), and analyzed more extensively in Marinakis and Tsoulouhas (2011), the assumption of risk neutrality for the agents makes the principal indifferent between the two evaluation methods. That is, with risk neutral agents both schemes ensure the same expected payoff for the principal. Marinakis and Tsoulouhas (2013) have shown that in the not unusual case, where agents are considered to exhibit some positive degree of risk aversion, the presence of the common shock, η, will tilt the scale towards the relative evaluation method. The relative method is by construction able to filter out the common shock from the responsibility of the agent because the realization of η drops out of the calculation of w_i. The relative method allows for efficient risk shifting from the risk averse agent to the risk neutral principal. In a sense, in order to get rid of the common risk, the agent becomes more tolerant to higher-power incentives and participates

[6] See for example Lazear and Rosen (1981), Green and Stokey (1983) and Nalebuff and Stiglitz (1983).

in production even when β is higher than the maximum possible γ (the power of incentives he would have tolerated if an absolute method was in place). This sort of insurance offered by the relative method is a move closer to the first-best and thus, under risk aversion for the agent, a relative evaluation method is expected to yield a higher effort level by the agents.

From equations (6) and (8) it becomes clear that the theoretical expectation from the standard incentive model is that, in the worst case scenario where students are risk neutral, the two grading methods *should be expected to perform equally*, while the existence of some risk aversion would render the relative method superior to the absolute method. Another straightforward expectation from the model is that effort per agent (and thus performance) is expected to be strictly increasing in the power of incentives independently of the evaluation method used by the principal. None of these expectations materialized when tested in the field.

2B. A Model with Peer Effects

The payoff function in (4) assumes that the agents care to maximize their benefit from their individual compensation net the cost of effort ($e_i^2/2a_i$). In other words, the standard model ignores the fact that the agents may have social considerations when they choose effort. In the peer effects model, we still consider a risk neutral principal and N risk neutral agents of heterogeneous ability. The production function for the agents is still given by (1). All the assumptions for the arguments in (1) remain in place. The alternative evaluation schemes are still given by (2) and (3) and agents are still exogenously forced to participate. The principal is not assumed to be the claimant for residual output, however, the principal's satisfaction is increasing in total output.

During the production process the N agents form M groups, in which participants develop social relationships. We shall refer to these groups as "social blocks" throughout the paper. The social blocks are not necessarily of equal size but an agent is limited to belong to only one block. Agent i's payoff is now given by

$$U_i = w_i - \frac{e_i^2}{2a_i} - \theta\left(\beta\right)\max\{w_i - \widetilde{w}_{-i,m}, 0\}, \tag{9}$$

where m indexes the block in which i is participating and $\widetilde{w}_{-i,m}$ is the average compensation of the agents participating in block m excluding agent i.[7] According to (9), the agent still

[7] Again, the exclusion of agent i from the calculation of the mean is used for simplification purposes and it does not alter any of the model's results. To see this, consider that if i was included in the block mean, then the deviation of w_i from the mth block mean would be $w_i - \widetilde{w}_m = \frac{M_i-1}{M_i}w_i - \frac{M_i-1}{M_i}\frac{1}{M_i-1}\sum_{\substack{j=1\\ j\neq i}}^{M_i} w_j = \frac{M_i-1}{M_i}\left(w_i - \widetilde{w}_{-i,m}\right)$, where M_i is the size of the block to which agent i belongs. That is, the exclusion of i simply scales the deviation down uniformly by $\frac{M_i-1}{M_i}$. However, in the experiment, the standard definition

derives satisfaction from the difference of individual compensation minus the cost of effort. Yet, now, he also derives a disutility, at a rate $\theta(\beta)$, from not wanting to "appropriate" compensation from individuals in his social block. We assume that $\theta(\beta)$ is strictly increasing in β for $\beta > 0$ and $\theta(0) = 0$. The rate at which the agents dislike over-performance, $\theta(\beta)$, depends on the power of incentives, β.[8] This makes sense intuitively because β directly affects the amount of compensation transferred from those below the average to those above it. Since this transfer is the source of social disutility, it is reasonable for β to enter the rate at which peer effects influence the agent's payoff. Notice that according to the assumption that $\theta(0) = 0$, inequity aversion will matter only in the case where a relative method is applied.

Our form of inequity aversion is qualitatively different than that introduced by Fehr and Schmidt in 1999 and was followed among others by Grund and Sliwka (2002) and Demougin and Fluet (2003). Inequity aversion according to Fehr and Schmidt is directly imposed on agents' preferences. As such, it can influence any kind of compensation scheme -absolute or relative. In the Fehr and Schmidt setting, an agent faces an "altruism" cost if he performs above others; and an "envy" cost if he performs below others. This means that an over-performing agent experiences a decrease in utility solely from the fact that his performance is above others, regardless if his actions have affected other agents' payoffs. Likewise, under-performing agents suffer a utility loss simply because they envy higher-performing agents, even though their compensations may not be affected by those above them. In our model, inequity aversion originates from the assumption that agents are specifically averse to taking actions that actively harm other agents. The over-performer's utility is reduced only when his effort choice directly reduces the payoffs of other agents. Moreover, under-performing agents, depending on the compensation scheme, may experience a decrease in their compensations from the actions of those ranked above them but they do not forgo utility due to envy per se. Therefore, the distinctive feature of our model is that, contrary to the fairness literature, inequity aversion arises less from direct preference assumptions and more from the payoff structure.

The motivation behind this choice of modeling is that, when an absolute system is in place, the agent may still feel sorry if his social block peers do worse but knows that this cannot be prevented by holding back on own effort. That is, under the absolute system

of the mean was used for the relative evaluation method.

[8] To see why β can be an argument in the utility function, through θ, when an absolute method is used, consider the nested version of the two schemes, $w_i = b + \gamma' x_i - \beta' \overline{x}_{-i}$. Here, $b = \beta' = 0$ yields the absolute method and $\beta' = \gamma' > 0$ yields the relative method. In this setting, one can realize that the so-called 'power of incentives', β, can be decomposed into the 'power of absolute incentives', γ', and the 'power of relative incentives', β'. So, precisely speaking, function's θ argument is β' rather than β. This makes the absolute scheme a special case of the relative scheme, where $\beta' = 0$.

the agent cannot feel responsible for worsening the position of others. On the contrary, under a relative system, the over-performers carry some responsibility for the decrease in the marginal payoff of those behind them. One can think of several real world examples of such behavior. For instance, recall the general dissatisfaction voiced in a classroom when the professor announces that "even though the test was overly hard, some students did ace it and thus there is no need for a curve".[9] The feeling of guilt for a situation in which the subject feels indirect responsibility can be found in several instances in real life. A juror may feel guilt for years after rendering a (justified) verdict sentencing a defendant to death because of their active role in this decision. Similarly, it is documented in psychology that operators of machines or vehicles can go through major psychological trauma when they have an active role in a serious accident, even though they did not exhibit any sort of negligence in their duties and the accident could only theoretically be prevented had they taken rather unusual or exceptional measures.[10]

It would be useful to provide some connection of our modeling assumptions with the experiment. It is not unnatural to expect that social blocks are formed in every class during or long before the semester. The participants of such social blocks tend to connect their individual utility with the deviation of their individual performance from the average performance in the block, when their actions directly affect the outcome of their peers.[11] Each member of a block experiences a social cost when this person's over-performance becomes the reason that his or her friends receive a lower score. Part of this cost is that the over-performers' peers may consider that the over-performers took advantage of the relative grading system to seize points from those who ranked lower. Under relative evaluation, a constant total amount of compensation points is available (that is, $\sum_{i=1}^{N} w_i = Nb$) for every assignment. Hence, the only way for someone to improve is to beat the average and, in some sense, to "steal" points from those who ended up under the average. The last term of (9) states that, for social reasons, agents do not like to outperform their social block's average. In our model, agent's compensation falls when individual performance drops. When individual performance increases, however, agent's compensation improves but this comes at a social

[9] By 'curve' here we mean the additional fixed bonus to every student, so that the point average in a test will meet a prior standard. This adjustment is usual in college classes for tests, exams and other assignments.

[10] For example, a subway operator may go though trauma because he or she happened to be at the controls when someone decided to jump in front of the train. Quoting an article entitled "Subway Deaths Haunt Those at Trains' Controls" (New York Times, January 4, 2013):

> "Many workers involved in fatal hits can take months to return [to their duties. . .]. Some never return to their old jobs at all [. . .] or even retire if they have already worked many years."

[11] Our framework differs from that of Gill and Stone (2010) in that our agents adopt a more socially sophisticated conception of fairness, in the sense that they care only for a specific subset of the agents they compete with and only if their actions affect the payoffs of the members of this subset.

cost of $\theta\left(\beta\right)\left(w_i-\widetilde{w}_{-i,m}\right)$, if $\theta\left(\beta\right)>0$ and $w_i>\widetilde{w}_{-i,m}$.[12]

According to the previous analysis, under the absolute evaluation method the payoff function (9) reduces to (4) and thus, the optimal effort is still given by (6). That is

$$e_A^*=\gamma a_i. \tag{10}$$

Under relative evaluation now, if we substitute (3) and (1) into (9), the expected payoff to the agent, who expects to be over the social block average, can be written as

$$EU_i=\left(1-\frac{N}{N-1}\theta\left(\beta\right)\right)\beta e_i-\frac{e_i^2}{2a_i}+\Theta+\Lambda. \tag{11}$$

where

$$\Theta\equiv\theta\left(\beta\right)\frac{1}{N_m-1}\sum_{\substack{j=1\\j\neq i}}^{N_m}\left[b+\beta x_j\right]+\frac{1}{\left(N_m-1\right)\left(N-1\right)}\theta\left(\beta\right)\beta\sum_{\substack{j=1\\j\neq i}}^{N_m}\sum_{\substack{k=1\\k\neq j,i}}^{N}x_k,$$

$$\Lambda\equiv\left(1-\frac{N}{N-1}\theta\left(\beta\right)\right)\beta\left(a_i+\varepsilon_i+\eta\right)-\frac{1}{N-1}\theta\left(\beta\right)\beta x_i$$

and neither Θ nor Λ depend on e_i. Agent i solves

$$\max_{e_i}\left[\left(1-\frac{N}{N-1}\theta\left(\beta\right)\right)\beta e_i-\frac{e_i^2}{2a_i}\right]+\Theta+\Lambda,$$

which satisfies

$$e_R^*=\left[1-\frac{N}{N-1}\theta\left(\beta\right)\right]\beta a_i. \tag{12}$$

That is, the agent who believes that he is over the block average, will tend to decrease effort, compared to (8), discounting it by $1-\frac{N}{N-1}\theta\left(\beta\right)$. Two facts are worth noticing at this point. First, the usual structure of the optimal effort (power of incentives multiplied by individual ability) remains unaffected in (12), however, it is now adjusted for peer effects. Second, even though the agent dislikes over-performance at a rate $\theta\left(\beta\right)$, he should not be expected to select the effort level that equates compensation to the social block average because satisfaction is still affected by personal interest (individual compensation net of the cost of effort).

[12] An alternative way to model inequity aversion, where the aversion is symmetrical for deviations from above and below the average is by using the function $U_i=w_i-\frac{e_i^2}{2a_i}-\lambda\delta_i(w_i-\widetilde{w}_{-i,m})^2$, where m indexes the block in which i is participating, $\widetilde{w}_{-i,m}$ is the average compensation of the agents participating in block m excluding agent i, $\delta_i\in[0,1]$ is a parameter indicating the importance of peer effects and λ is a binary variable which is 0 when compensation is absolute and 1 when compensation is relative. The optimal effort in the relative method, then, may also be consistent with the experimental findings in section 5.

3. The Experiment

The aim of the experiment was to study how the method of grading and the power of incentives could affect the choices of college students.[13] Two separate sections of a 'Principles of Economics' course were taught by the same instructor. The first section consisted of 42 students and the second of 40 students. Both sections were in session two times a week and were held consequently with a 15 minute break in-between. The content and the style of the lectures were identical for both sections. Attendance was monitored in both sections and there were no cases of students attending a section other than the one they were registered for. From the very first lecture students became aware that an unusual grading method for homework assignments would be tested in the section they were enrolled. A short orientation for the method of grading was given during the first lecture, making the students aware of the grading process for the entire semester. A few students who were absent from the orientation, sat through a make-up orientation session within the first week of classes, before any assignment was available or due. No student dropped or added the course after the first day of classes.[14]

All students were required to turn in 8 homework assignments during the semester.[15] Homework was announced to count for 40% towards the final score in the course, while the two midterm tests combined with the final examination had a total weight of 60%. The homework weight was set higher than the usual 20 - 30% in order to increase the importance of homework in the determination of the final grade and thus, to add to the robustness of the experiment's results. Students were well aware that homework points were valuable for a positive result in the course. Every two weeks the instructor would post a new assignment on the course website and all students would be notified that the new assignment was available. Assignments were common for both sections. Students had to log in using their personal university credentials to download the assignment document. This document consisted of three parts: (i) a cover page with information about the deadline and instructions for the completion and submission of the assignment, (ii) a part containing the details about the grading scheme applied in each section and (iii) the part with the questions. Each assignment contained exactly 25 multiple choice items. From those, the two first questions were always

[13] This experimental study was done in an attempt to investigate ways to improve the quality of education provided through a novel method of incentives of more intense competition among peers. As such, the study did not legally require a permit other than the approval of the head of the department of economics of the university, which was gladly granted. This paper simply documents the observations and the methods used along with the analysis of the data collected in the field.

[14] Principles courses are of high demand by students. Most of these sections are completely full from the very beginning of the registration and students rarely drop out after the first day of classes.

[15] The number of homework assignments was set to 8 because according to the setting, it was necessary to have an even number of assignments and there was not enough time for 10 assignments.

about the grading method. This ensured that before exerting effort, every student was well informed about the grading method. Questions 3 - 25 were related to course material. Every question offered 4 alternative responses, out of which only one was considered correct. Every homework question was created by the instructor, was relevant to the lecture and had never been used before in order to prevent leakages. Students had ten days to submit their answers from the moment the assignment became available. They had to log in again and use an electronic form to submit a string of letters (A, B, C or D) that had to be typed into designated fields. Each student was allowed a unique submission but before they submitted, they had the chance to review all answers and verify that they were indeed ready to submit. After the deadline expired the system did not accept any late submissions.

The assignments and the deadlines were common for both sections but the grading methods were always different (see Table 1). For the first 4 assignments, section 1 was evaluated according to a relative method with progressively increasing power of incentives, while section 2 served as a control section and was evaluated according to the traditional absolute method. For assignments 5 through 8 this regime was inverted. That is, section 2 was evaluated according to the same progressive relative method and section 1 became the control section. In the absolute method, each correct answer was worth 4 points, so that the maximum score would be the usual 100 points. Since students are well aware that their real result is the ratio of their score to the maximum possible score, the power of incentives under the absolute method cannot be adjusted meaningfully. Namely, it will make no difference if one gives 5 points per correct question, making the maximum score 125, since the ratio will remain unaffected. On the other hand, when a relative method is used, the power of incentives can be meaningfully altered by changing β, the coefficient of relative performance $(x_i - \overline{x}_{-i})$ and, thus, own performance, x_i. This is because β adjusts the marginal penalty/reward for deviating an additional unit from the average performance. In this experiment, a progression of such coefficients was used. In the first treatment the coefficient was 2 (making the power of incentives half of that of the absolute method), in the second treatment the coefficient became 4 (equal to that of the absolute method), in the third treatment it was set to 6 (one and a half times the absolute one) and in the fourth treatment its value was 8 (carrying double the incentive power of the absolute method). The base compensation for the relative method was set to 75 and remained unchanged throughout the duration of the experiment. This value was selected because the average score in similar assignments for a large number of past sections of the same course was 75 out of 100 with a notably small variance. Table 1 sums up the grading methods for each section.

After the deadline for the submission of the last assignment, all students were required to fill out a questionnaire about their experience with the experiment. The questionnaire

	Section 1	Section 2
hw #1	$75+2(x_i-\overline{x})$	$4x_i$
hw #2	$75+4(x_i-\overline{x})$	$4x_i$
hw #3	$75+6(x_i-\overline{x})$	$4x_i$
hw #4	$75+8(x_i-\overline{x})$	$4x_i$
hw #5	$4x_i$	$75+2(x_i-\overline{x})$
hw #6	$4x_i$	$75+4(x_i-\overline{x})$
hw #7	$4x_i$	$75+6(x_i-\overline{x})$
hw #8	$4x_i$	$75+8(x_i-\overline{x})$

x_i : number of correct answers for the ith student

$\overline{x}$: section average

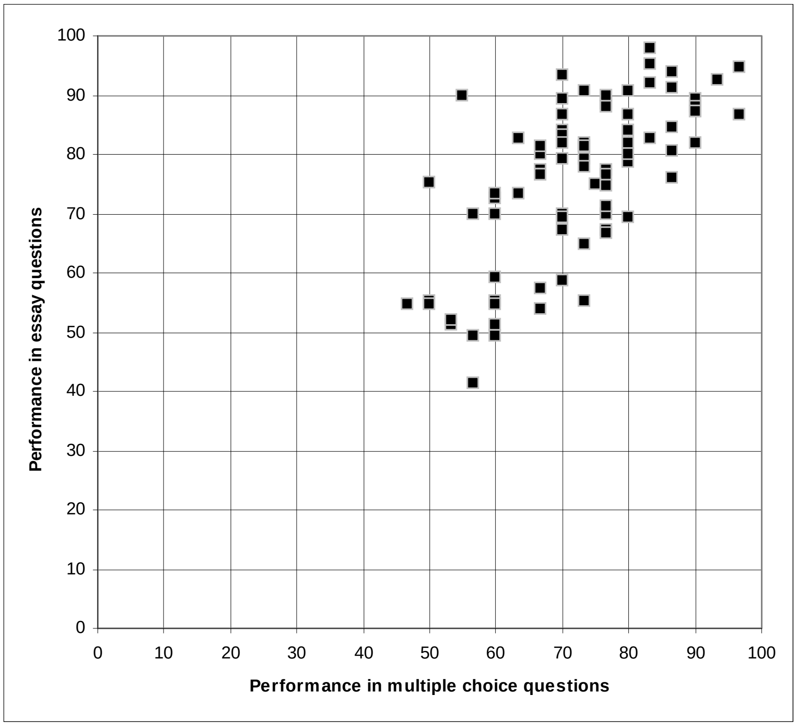
Table 1: The grading methods used throughout the experiment in the two sections

asked specific questions, the answers of which could be tabulated and used as binary variables if needed. It also included some open-ended questions where the students were asked to share their opinion or share comments they thought would be helpful for assessing the experiment. Everyone was required to identify themselves on the questionnaire. However, after the completion of the process, a volunteer student from each section collected the questionnaires, sealed them in an envelope and carried them to a university official, who had pledged to not release the documents until the grades for the class had been finalized. Moreover, before the process began, the students received a clear assurance that their statements on this document could not be used against them or against anyone else under the code of student affairs of the university.[16]

The goal of the experimental study was to investigate if the real world incentives live up to the theoretical expectations of the principal-agent framework. The primary concern was the effect of incentives on effort and thus performance. In the experiment, the provision of incentives was embedded in each alternative grading scheme for the assignments. Since effort was unobservable, the students' performance was used as a proxy for effort. Performance in multiple choice questions is affected by effort but also ability and luck. This makes it hard to identify the individual effects of effort and ability on performance in the experiment. However, student ability is not time variant and thus, changes in performance can most likely be attributed to changes in effort. Moreover, we are able to use the results of everyone's performance in tests and exams to obtain an estimate for individual ability. Even though the performance observed in such exams is also affected by both effort and ability, the homework performance, relative to exam performance, is a good measure of 'pure effort'.

[16] According to the constitution, a statement produced under a binding promise of privacy creates 'reasonable expectation of privacy' and is not admissible as direct proof of guilt in a court of law including the court of student affairs.

Figure 1: The performance of each student in the essay question portion versus the multiple choice portion for each exam.



By the term 'pure effort' we mean the actions a student can take in order to increase performance in homework which, however, they cannot take during a classroom test or exam, such as studying from notes or textbooks when working on a specific question, online search, attending office hours or paying increased attention during the lecture.

In the experiment, tests were identical for both sections, they took place in class, with closed book and closed notes and they were proctored by the author. The papers from both sections were put together, shuffled and marked blindly by the author according to the absolute method. No student had to be given a makeup test. All tests were 50% multiple choice and 50% essay questions. This format was chosen as an a-priori safe choice, in order to identify potential problematic situations. For example, it is not uncommon that a hard working student under-performs in tests using multiple choice items. However, this precaution turned out to be unnecessary because for every student the performance in multiple choice questions was not significantly different than the performance in essay questions. This can be seen in Figure 1.

4. Data Treatment

Upon the completion of the experiment, data on the total number of correct answers for each student and for each of the eight assignments were gathered. The data produced a panel with the assignment number as the time series dimension and the student ID number as the cross-sectional dimension. The raw panel was used to test various statistical hypotheses about the effect of the incentive structure of the grading method on student performance. However, most of those tests were inconclusive. In contrast to basic economic intuition, incentives seemed to have an unclear effect on the students. Even though the fixed effects model controls for individual and assignment effects, the raw dataset evidently contained a significant amount of noise because the observations were not properly weighted. Regression Panel 1 illustrates the impact of incentives on effort according to the raw data. The binary variables *hw2* - *hw8* control for assignment effects relative to *hw1,* which is the omitted variable. A similar structure of 81 binary variables was used to control for individual (i.e. for each student) effects. The regression coefficients for these individual binary variables are not shown in Regression Panel 1 in order to economize on space. The binary variables indicated by *'beta* $= 2$*', 'beta* $= 4$*', 'beta* $= 6$*'* and *'beta* $= 8$*'* take the value 1 if the incentive power coefficient was 2, 4, 6 or 8 respectively and 0 otherwise. The regression coefficients of these variables measure the effect of each incentive power coefficient on performance, relative to the effect of absolute grading, which is the omitted variable. From this regression it is unclear how incentives affect performance because, only when the tournament coefficient is 6, does there seem to be a significant relationship between incentives and performance. In this case, performance decreases by 7.3 points relative to the performance under the traditional piece rate. Regression Panel 1 suggests that raw performance is not useful in explaining effort choices.

The chief suspect for the noise in the dataset was the possible variability in the quality of questions. This became clear from survey testimonials, after the completion of the experiment. Several students indicated that the quality of multiple choice questions varied from question to question and from assignment to assignment. In general, multiple choice questions are considered of good quality when only those who master the material can come up with the correct response. On the contrary, low quality questions allow students who do not prepare well to get the answer right or they confuse good students and induce them to get the answer wrong. In other words, a high quality multiple choice item must have the ability to *discriminate* between students who exerted the appropriate effort from those who did not. An additional issue on the quality of the questions used in the experiment was that the difficulty of some items was solely due to the fact that they required high observational skill. An anonymous student aptly pointed out this issue in the semester's teaching evaluation:

> *"[. . .] Some multiple choice questions are confusing and/or misleading and they do not simply test the knowledge of the material but they are made to test how skilled we are spotting the tricky part of the question. [. . .]"*

Regression Panel 1: Incentives on raw performance

Coefficient	value	t-stat	p-val
intercept	101.8693	15.1860	0.0000
...	...	...	...
hw2	-4.6226	-1.3608	0.1741
hw3	0.4702	0.1384	0.8900
hw4	-6.4202	-1.8900	0.0593
hw5	-12.7644	-4.5228	0.0000
hw6	-18.7078	-5.4146	0.0000
hw7	-4.8631	-1.4075	0.1598
hw8	-13.4578	-3.8951	0.0001
beta = 2	4.2882	1.5194	0.1292
beta = 4	-0.2957	-0.1048	0.9166
beta = 6	-7.3333	-2.5984	0.0096
beta = 8	-3.2481	-1.1509	0.2503

n	d.o.f.	R square	Var	F-stat	p-val
640	549	0.3688	317.8053	3.5642	0.0000

Model: "perf = id-dums hw-dums rel-dums", No logs are used; zero submissions are not dropped unless more than 2. All sections, all questions, all classes.

Such questions appear difficult not because they test difficult material but because they require high observational skill or the ability to use logic in generic situations. We can conclude that, since success in such questions is not necessarily correlated with high effort, these questions are expected to add noise to the dataset if they are weighted the same with high quality items.

The ability of a question to discriminate high from low effort students can be illustrated in the following example of an actual question from one of the experiment's assignments.

> If household income increases and, ceteris paribus, the supply of a good increases then the good is:
>
> A. A luxury.
>
> B. An inferior good.
>
> C. A normal good.
>
> D. None of the above is correct.

Unquestionably, household income and the supply of a good are not related in principle and the average economics student is expected to be familiar with both notions, so that this task should not be a problem. However, this question proved to be a "mind game" because the majority of students wrongly answered C. The trick is that this particular item mentally conditions the students to focus on the direction of the effect rather than to think if income and supply are in fact related. That is, in the way this question is phrased, the typical student becomes confused and tends to think that the question examines the definitions of normal and inferior goods rather than whether there exists a causal relationship between income and supply. Indeed, most students were distracted from noticing that the question referred to a supply increase rather than to a demand increase. In such questions, success depends more on observational skill rather than effort. It is evident that, when the same question was used in a test for a later class and was slightly altered such that the alternative D was: "*Household income and supply are not related*", the majority of students answered correctly D, while only a small minority chose C.

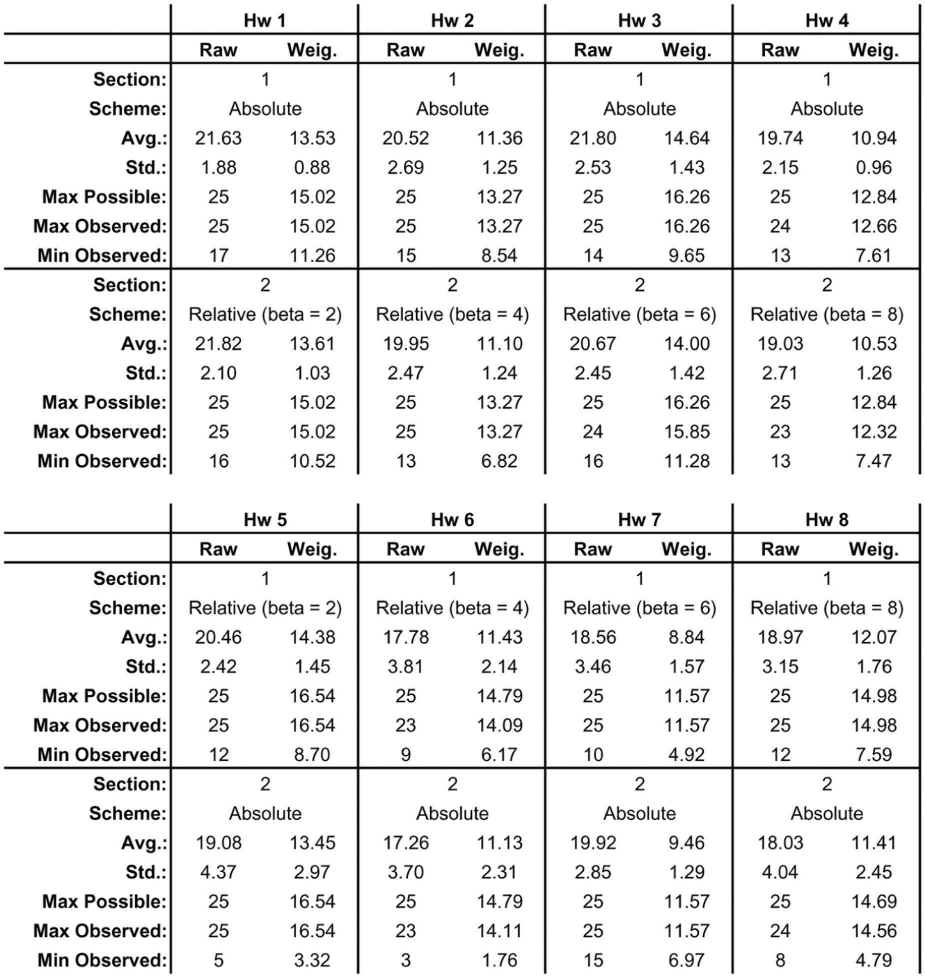
Performance will be a better proxy for effort if we weigh the items according to their individual discriminating ability. For this purpose the "item discrimination index" is used to define the weight of each question.[17] The item discrimination index is a measurement of the correlation between the item response and the overall performance in the assignment. As a correlation measurement, the discrimination index appears in popular grading software packages, which accompany scantron machines and online class management applications (Aplia, Blackboard, WebAssign etc.). The experimental data in this paper was transformed according to one of the most common and simple discrimination indices used in practice. That is,

$$d_q = \frac{\overline{X}_{c,q} - \overline{X}}{S_X} \sqrt{\frac{P_{c,q}}{1 - P_{c,q}}}, \tag{13}$$

where d_q is the discrimination index for the qth question, $\overline{X}_{c,q}$ is the mean score of those who answered question q correctly, $\overline{X}$ is the mean score on the assignment, S_X is the standard deviation of scores on the assignment and $P_{c,q}$ is the proportion of those who answered question q correctly. Obviously, the index is constructed to take values between 0 and 1. After this index was used for weighting each question in the dataset, the transformed dataset clearly showed statistically significant patterns for the effects of incentives on effort.[18] These patterns are discussed in the next section. The discrimination index evidently cleared the noise in the dataset. For instance, not surprisingly, the discrimination index weighted the question mentioned above nearly 25% less than the average item. Notably, when the question was rephrased and reused in a later class, the discrimination index assigned a weight not

[17] See Pyrczak (1973) for an extensive investigation of the item discrimination index.

[18] The weight for question q is d_q. Then, the weighted performance is d_q if the student's answer was correct and 0 otherwise.



	Hw 1		Hw 2		Hw 3		Hw 4	
	Raw	Weig.	Raw	Weig.	Raw	Weig.	Raw	Weig.
Section:	1		1		1		1	
Scheme:	Absolute		Absolute		Absolute		Absolute	
Avg.:	21.63	13.53	20.52	11.36	21.80	14.64	19.74	10.94
Std.:	1.88	0.88	2.69	1.25	2.53	1.43	2.15	0.96
Max Possible:	25	15.02	25	13.27	25	16.26	25	12.84
Max Observed:	25	15.02	25	13.27	25	16.26	24	12.66
Min Observed:	17	11.26	15	8.54	14	9.65	13	7.61
Section:	2		2		2		2	
Scheme:	Relative (beta = 2)		Relative (beta = 4)		Relative (beta = 6)		Relative (beta = 8)	
Avg.:	21.82	13.61	19.95	11.10	20.67	14.00	19.03	10.53
Std.:	2.10	1.03	2.47	1.24	2.45	1.42	2.71	1.26
Max Possible:	25	15.02	25	13.27	25	16.26	25	12.84
Max Observed:	25	15.02	25	13.27	24	15.85	23	12.32
Min Observed:	16	10.52	13	6.82	16	11.28	13	7.47

	Hw 5		Hw 6		Hw 7		Hw 8	
	Raw	Weig.	Raw	Weig.	Raw	Weig.	Raw	Weig.
Section:	1		1		1		1	
Scheme:	Relative (beta = 2)		Relative (beta = 4)		Relative (beta = 6)		Relative (beta = 8)	
Avg.:	20.46	14.38	17.78	11.43	18.56	8.84	18.97	12.07
Std.:	2.42	1.45	3.81	2.14	3.46	1.57	3.15	1.76
Max Possible:	25	16.54	25	14.79	25	11.57	25	14.98
Max Observed:	25	16.54	23	14.09	25	11.57	25	14.98
Min Observed:	12	8.70	9	6.17	10	4.92	12	7.59
Section:	2		2		2		2	
Scheme:	Absolute		Absolute		Absolute		Absolute	
Avg.:	19.08	13.45	17.26	11.13	19.92	9.46	18.03	11.41
Std.:	4.37	2.97	3.70	2.31	2.85	1.29	4.04	2.45
Max Possible:	25	16.54	25	14.79	25	11.57	25	14.69
Max Observed:	25	16.54	23	14.11	25	11.57	24	14.56
Min Observed:	5	3.32	3	1.76	15	6.97	8	4.79

Table 2: The descriptive statistics for the raw and weighted performance in each section for each assignment.

significantly different than to the other questions. Table 2 illustrates the descriptive statistics for the raw and weighted performance in each section for each assignment.

Another issue concerning the data was how to handle missing observations. From the 656 possible submissions a total of 38 submissions were missing. A missing observation could be either because a student had a personal reason not allowing submission in time or because the student decided not to exert any effort for the assignment. The two reasons are qualitatively different. The first reason does not indicate intention for exerting zero effort while the second reason clearly does so. The way missing observations were handled did not materially affect the results. Therefore, it was considered reasonable to handle the missing observations in the following manner: All missing observations were interpreted as zero performance unless a student submitted less than 6 out of the 8 assignments. In the latter case, the student was entirely dropped from the panel. According to the selected method, only 2 students, both from the second section, were dropped. The rationale behind this choice was that students who failed to submit more than two assignments were considered to be uninterested in the course and the grading method, so they should be left out of the experiment.

5. Experiment Findings

The weighed data collected from the experiment were analyzed according to the fixed effects model in order to capture the effects of the grading method and the power of incentives on performance. Groups of binary variables were used to control for various effects. Specifically, the group of 7 binary variables *hw_ dums* was used to control for difficulty among the assignments. Each binary variable, hw_h in this group, where $h \in \{2 : 8\}$, takes the value 1 for the hth assignment and the value 0 otherwise. The regression coefficients of the hw_h variables measure the difficulty of each assignment relative to the difficulty of the first assignment because the omitted variable was hw_1. The group of 81 binary variables *id_ dums* controlled for idiosyncratic effects among subjects. Each id_i variable of this group, where $i \in \{2 : 81\}$, indicates that the assignment was submitted by the ith student. The omitted binary variable was id_1. The variable id_1 represents the student with the highest performance in tests and exam (homework assignment scores were not used for ranking the students), so the regression coefficients on the id_i will always appear to be negative. The group of binary variables *rel_ dums* is used to capture the effects of the relative performance evaluation coefficients on performance. The notation of the binary variables *rel_ dums* is explained in the following table (Table 3).

Variable	Grading method
'beta = 2'	$75 + 2(x_i - \overline{x})$
'beta = 4'	$75 + 4(x_i - \overline{x})$
'beta = 6'	$75 + 6(x_i - \overline{x})$
'beta = 8'	$75 + 8(x_i - \overline{x})$

Table 3: Binary variables indicating the power of incentives under relative evaluation

The regression coefficients of the *rel_ dums* variables indicate the effects of the incentive power on performance relative to the effect of the 4-point piece rate. The binary variable *tourbot* becomes 1 when any relative method is used, and additionally, when the student is ranked in the bottom 20 students of his or her section according to performance in tests and exams, and 0 otherwise. The binary variable *tourtop* takes the value 1 when any relative

Regression Panel 2: The effect of incentives on weighted performance.

Coefficient	value	t-stat	p-val
intercept	13.9484	12.6311	0.0000
…	…	…	…
beta = 2	1.9013	4.9002	0.0000
beta = 4	-0.8599	-2.2163	0.0271
beta = 6	-0.9358	-2.4118	0.0162
beta = 8	-1.0845	-2.7951	0.0054

n	d.o.f.	R square	Var	F-stat	p-val
640	556	0.2770	9.6352	2.5671	0.0000

Model: "discr = id-dums rel-dums", No logs are used; zero submissions are not dropped unless more than 2. All sections, all questions, all classes.

method is used and the student is ranked in the top 20 of students of his or her section according to performance in tests and exams, and 0 otherwise. The variables *tourtop* and *tourbot* are useful to define the variable $tt \equiv tourtop + tourbot$. The variable *tt* is used later on to test the hypothesis of equality of the regression coefficients of *tourtop* and *tourbot*.

5A. Performance under the Two Methods

Regression Panel 2 illustrates the effect of incentives under the two compensation methods when we control for idiosyncratic effects. The dependent variable in the regression model for Regression Panel 2 is 'weighted performance' and the regressors are the groups of variables *id_ dums* and *rel_ dums*. When we use the discrimination index to weigh the data, there is no need to use the *hw_ dums* in the regression to control for assignment effects. This is because the weighing of the data using the discrimination index has already normalized the dataset for assignment difficulty effects.[19] Recall that the omitted variable from *rel_ dums* is the binary variable that becomes 1 when the absolute method is used and 0 when any relative method is used. Therefore, the coefficients of the *id_ dums* represent the total weighted score of the four relative methods relative to the absolute method. As one can see in Table 2, the maximum total weighted score for each assignment varies from 11.57 (hw 7) to 16.54 (hw 5). The average maximum total weighted score is 14.41 and one can use this average to interpret the magnitude of our results.

According to the Regression Panel 2, when $\beta = \gamma = 4$ (equal incentive power for both schemes), performance, and thus effort, seem to be lower under the relative than they are

[19] When the set *hw_ dums* was included in this regression, the coefficients of these variables were not significant at a 0.1 level.

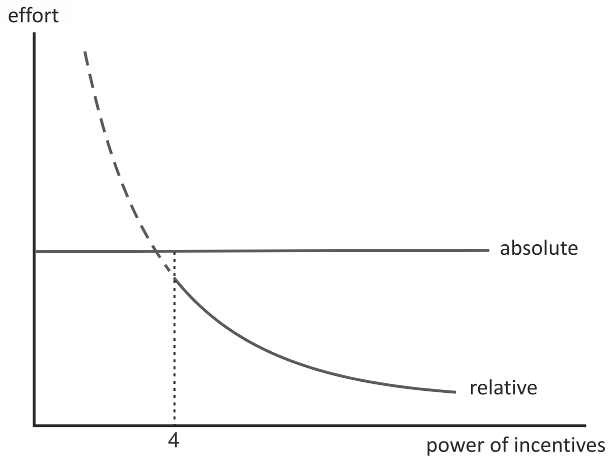
under the absolute evaluation method. This means that when the relative method was used, performance dropped by 0.86 units (that is a 5.96% decrease in comparison to the average performance of 14.41). The difference appears to be quite significant with a p-value equal to 0.0271. This is directly opposite to the expectations formed by a standard model of incentives as we illustrated in section 2A, where the relative method is expected to be weakly dominant to the absolute method, depending on the degree of risk aversion. Nevertheless, in the field, when the power of incentives is held constant, students turn out to perform better under an absolute method than under a more competitive relative method. This observation is in line with prominent empirical papers such as Bandiera et al. (2004). A possible explanation for such -incompatible with the elementary theory- result is that interactions within highly socialized environments, such as the classroom or the workplace, might expose the agents to some additional kind of cost other than the disutility of effort or the risk aversion assumed in the benchmark model in section 2A. Social interactions in the classroom are of primary concern for the students. As shown in section 2B, under a relative method, the choice of effort may inflict a social cost on the student. This is because, individual effort is inversely related with compensation to other students for whom the student at hand might care. This effect is captured by our peer effects model in section 2B from equations (10) and (12) provided that $\theta(\beta) < 1$, when $\beta = \gamma$.

5B. Incentive Power and Performance under the Relative Method

According to Regression Panel 2, again, performance seems to be *negatively affected* by β, the coefficient of the power of incentives under the relative method. That is, when the incentive power is low (*beta = 2*), performance is even higher than the performance under the absolute method (13.18% higher). For β equal to or higher than γ (*beta = 4, beta = 6* and *beta = 8*), the relative method causes progressively lower performance (reductions of 5.96%, 6.52% and 7.53%, respectively). According to the p-values for the t-statistics for the regression coefficients, these results are highly significant. Moreover, the null hypothesis that the coefficients for *beta = 2, beta = 4, beta = 6* and *beta = 8* are all equal, is rejected with an F-statistic of 17.07, for which the p-value (for 3 and 556 degrees of freedom) is practically zero. The observations concerning the effect of incentive power on effort, from Regression Panel 2 are abstracted in Figure 2.

Under the relative method, when the power of incentives is lower than 4, incentives are in fact diluted in the sense that the marginal benefit of a correct item is smaller than the value of the item itself as a proportion of the maximum score. That is, with a power of incentives equal to 2, an additional correct item yields 2 points, while it actually represents 4% of the total output. This "weakening" of incentives under the relative method seems to encourage effort more than the traditional absolute method, where the power of incentives

Figure 2: Effort vs. power of incentives under both methods.



is fixed and equal to the proportional value of each item (4 points). The latter, combined with the decreasing effect of the incentive power on performance (see Regression Panel 2 and figure 2), strongly indicate that students have some sort of "incentive power aversion". Clearly, our observations in the field do not fulfill the theoretical expectations established in the benchmark model in section 2A and also in seminal works in the agency literature, where effort is shown to be directly analogous to the power of incentives. This striking result has never been observed before to our knowledge and it indicates that in highly socialized environments incentive effects may be reversed. One can state that agents seem to care a lot about the results of their actions on others. This is not a novel observation in behavioral sciences.

Connecting this finding to our peer effects model in Section 2B, the factor $\left[1 - \frac{N}{N-1}\theta\left(\beta\right)\right]$ in (12) can explain the decrease of effort as β increases. This adjustment for peer effects seems to be what is missing from the conventional model. As β increases, $\theta\left(\beta\right)$ also increases making the agent to exert lower effort as long as $\beta d\theta\left(\beta\right)/d\beta + \theta\left(\beta\right) > 1$. That is, for sufficiently high values of β, the increase in social cost offsets the increase in individual benefit inducing the agent to moderate effort.

Even though, this result was unexpected according to conventional theory, the participants of the experiment seemed to have a quite accurate feeling for the reversal of the function of incentives. In the survey questionnaire they filled out immediately after the completion of the experiment, the most common statement was that the relative performance evaluation makes them uncomfortable when choosing the amount of effort they will exert. Students mentioned that their decision on their effort was becoming more awkward as the incentive scheme was enforcing higher-power incentives and that the grading system created tension in the relationships between classmates. Participants were conscious that over-performance would provide them with a personal benefit but would make their social peers worse off. Students were well aware that, under this system, a constant total amount of points was available for every assignment and the only way for someone to improve was to "steal" points from someone else. Relative evaluation transformed the learning process in the classroom to a zero-sum-game. This reversed the function of incentives in the sense that competition was turning classmates against each-other instead of enabling study groups to improve the overall level of the class. Several participating students stated in the survey that, even though they do not care about the performance of the rest of the class as a whole, they feel uncomfortable experiencing situations in which their performance was an outlier relative to the performance of their closed group of friends. This indicates that social blocks, such as those assumed in section 2B, are formed among students.

All results in sections 5A and 5B can be fully explained by the model of peer effects provided in section 2B. Normalizing the power of incentives so that each unit of output is worth 1 point (thus maximum score will be 25) and assuming a rate of over-performance aversion, $\theta(\beta)$, such that $\theta(0) = 0$ (that is, when absolute evaluation is used), $\theta(.5) \in (-1, 0)$ and $\theta(\beta) \in (0, 1)\ \forall\ \beta \in \{1,\ 1.5,\ 2\}$, equations (10) and (12) can fully explain our results of incentives on effort, namely, (i) absolute performance is superior to relative performance when $\beta = \gamma$, (ii) relative performance is superior when the (normalized) power of incentives is less than 1 and, (iii) under the relative scheme, effort is decreasing in β.[20]

5C. Answer Sharing

Another interesting topic addressed in the experiment was the possibility of preventing dishonest behavior in the assignments. In this course, as in many others worldwide, students were encouraged to form study groups. Study groups allow students to put their individual strengths together producing a positive learning outcome for all participants. Study groups -when they work right- promote trading of knowledge between students. However, as some students pointed out in the survey, students often developed a "more general" kind of trading in study groups. Since the answers in multiple choice questions are simply a 25-character string of letters, answers can easily be shared at a low risk of detection. Thus, it is possible, under the traditional absolute grading system, for some strong students, at zero cost to them, to trade their answers for non-educational benefits. A participant to the experiment aptly stated: "*[...] the homework key can easily become the... key to the exclusive [...] Saturday party*". After a strong student leaks their answers, they become a public good and may be distributed freely among weaker students. Those who share gain valuable popularity, while

[20] An example of such a function can be $\theta(\beta) = \begin{cases} 0 & \text{if} \quad \beta = 0 \\ \frac{\beta-\vartheta}{\beta+\vartheta} & \text{if} \quad \beta > 0 \end{cases}$, where $\vartheta \in (0, 1)$.

Regression Panel 3: How are Top 20 and Bottom 20 students affected by the introduction of the tournament?

Coefficient	value	t-stat	p-val
intercept	11.5700	61.2357	0.0000
tourbot	-0.8360	-2.5962	0.0096
tourtop	0.4088	1.2278	0.2200

n	d.o.f.	R square	Var	F-stat	p-val
640	637	0.0180	11.4236	5.8316	0.0031

Model: "discr = tourbot tourtop", No logs are used; zero submissions are not dropped unless more than 2. All sections, all questions, all classes.

weaker students may appear to perform well. Conversely, when a relative evaluation method is in effect, a cost is imposed on strong students who share because answer sharing increases the class average and decreases the expected score of those who shared.

Regression Panel 3 illustrates our findings on answer sharing. The groups of variables *tourbot* and *tourtop* were regressed on the weighted performance. Recall that *tourbot* is 1 when any relative method is used, and additionally, the student is ranked in the bottom 20 students in his or her section and 0 otherwise. The variable *tourtop* is 1 when any relative method is used and the student is ranked in the top 20 of students in his or her section and 0 otherwise. The ranking of students was done with respect to their results in tests exclusively. Regression Panel 3 examines the difference in performance between each section's top 20 students and bottom 20 students. The analysis assumes that students who perform well in tests are less likely to turn in answers copied from others. The model in Regression Panel 3 contrasts the performance response of top and bottom students when the grading method was switched from a piece rate to a tournament. Top students seem to have a slight (non significant) positive performance response (approximately 5.8%) to the switch, while bottom students' performance drops significantly by almost 8.67%.[21] Regression Panel 4 presents the regression of the variable *tt* (where $tt \equiv tourtop + tourbot$) and *tourtop* on weighted performance. This model is run with the purpose of testing the significance of the difference in performance of top and bottom students through the p-value of the coefficient of the variable *tourtop*. As can be seen in Regression Panel 4, the difference is also highly significant with a p-value of 0.0011. This is sufficient evidence that some answer sharing took place while the absolute method was used and it was discontinued after the introduction of the relative method.

[21] Notice that the regression coefficients in Panel 3 contrast the absolute method with all four relative methods used in the experiment combined together.

Regression Panel 4: Test: When a tournament is in effect, do Bottom 20 students drop more than Top 20 students?

Coefficient	value	t-stat	p-val
intercept	11.5700	61.2357	0.0000
tt	-0.8360	-2.5962	0.0096
tourtop	1.2448	3.2900	0.0011

n	d.o.f.	R square	Var	F-stat	p-val
640	637	0.0180	11.4236	5.8316	0.0031

Model: “discr = (tourtop + tourbot) tourtop”, No logs are used; zero submissions are not dropped unless more than 2. All sections, all questions, all classes.

We have to make clear that the results presented in sections 5A and 5B are robust despite the prospect of answer sharing. Cheating would not cause a qualitative difference in the incentive effects because of two reasons. *First*, the overall extent of cheating on performance was significant but relatively small. The 8.67% decrease reported above reflected the performance of the lower ranks of the class for half of the assignments that sharing was not penalized by the scheme. *Second* and most importantly, all results in Regression Panel 2 are driven by the middle and upper mass of the class. This makes sense, since students at the bottom, who would benefit from answer sharing, are not the ones who will choose to moderate effort to avoid over-performance.

6. Discussion and Conclusions

Highly socialized environments may not comply with the predictions of the standard principal-agent model. The main reason for this deviation is that in the standard model agents are assumed to consider exclusively their narrow self-interest, while social considerations are largely ignored. In the instructor-student relationship, popularity, likability and "coolness" are the most common words used in the classroom to express status. The social reputation they develop during their college years may follow many for their entire lives. Thus, student behavior is more likely to be subject to social considerations that distort the effects of incentives as we know them from theory.

The conventional linear absolute method of grading does not allow for a material adjustment of the power of incentives. A linear relative method, on the other hand, is able to provide adjustable incentive power through the coefficient of the deviation of individual performance from the average. When a relative method is used and the incentive power is set to a high level, larger amounts of effort yield higher individual compensation. This, however, may lead to a reduction in the agent’s overall utility because, under a relative scheme, over-performance effectively decreases the payoffs of the group, which also concerns the agent. By design, the relative scheme rewards those who perform above the average and proportionally penalizes those who perform below it. Since the total compensation in the group is not affected by the overall performance of the participants, the reward an agent receives for beating the average comes from the penalty charged to those who ranked behind this agent. In this context, keeping effort down to a moderate level may decrease individual payoff but it also decreases the agent's disutility from harming others. In some cases, the latter turns out to be highly desirable, to the point that it may offset the loss in individual payoff. Hence, the answer to the question if the provision of incentives is effective in the classroom, involves the comparison of two factors in tension with each other: the *compensation effect*, which is positively affected by effort; and the *social interest*, which might be negatively affected by effort when effort is sufficiently high. The interaction of these two factors suggests that, when agent behavior is sufficiently affected by peer effects, a relative evaluation method can discourage effort because the social interest may prevail over the compensation effect.

The paper presents an experiment, which investigates how peer effects impact the role of incentives and tests whether a relative evaluation method is able to improve student effort. The paper introduces an innovative transformation for the panel data. The dataset is weighed by the item discrimination index in order to distinguish between questions requiring high amounts of effort and questions requiring other skills. The results were quite surprising.

Our first result claims that relative performance evaluation has a significant negative impact on students' performance compared to absolute evaluation. This is not what one would expect, having in mind the theoretical implications of the articles by Lazear and Rosen (1981), Green and Stokey (1983) and Nalebuff and Stiglitz (1983), where relative evaluation does not lead agents to exert lower effort than they would under an absolute method. This is also corroborated by empirical work such as Knoeber and Thurman (1994 and 1995) and Tsoulouhas and Vukina (2001), in which the two evaluation methods are contrasted in market environments with no significant peer effects (agents may not even know who they are competing against). When social considerations affect the interaction between principal and agents, relative evaluation inflicts a social cost on those who over-perform. In our classroom field, social concerns play a major role in the behavior of students. This places relative compensation methods at a disadvantage. It is evident that, in the survey administered after the experiment, several participants expressed their discomfort with the relative method by characterizing it as a "cut-throat method". According to our first result, students seriously considered the social cost imposed on them under the relative method and significantly reduced their effort. This result is in line with other empirical investigations, such as Bandiera et al. (2005), who include peer effects in their analysis.

Our second result suggests that relative evaluation is more effective when it delivers sufficiently lower-power incentives than the absolute evaluation. To our knowledge this fact has never been observed in the literature. The logic behind this finding is that, when the power of incentives was set below the real marginal value of output, the weight of the social interest factor diminishes and agents have the liberty to increase effort in order to maximize their individual compensation.

Our third result has also never appeared in the literature before. It claims that under relative evaluation, effort and power of incentives are inversely related. This is in sharp contrast with the elementary theory of incentives. Nevertheless, this result is intuitive if one considers the importance of peer effects. In relative evaluation, the power of incentives affects the social cost factor in a direct way. If peer effects are important for the participants, the rate at which the social cost increases offsets the rate at which individual compensation increases and, thus, the agent moderates effort. In the experiment, students respond to the increase in the power of incentives by decreasing their effort in an attempt to lower their social cost. This behavior shows that they dislike the transfer of points from those who performed lower than the average to those who performed above the average. Even when students do not care for every other student in the class, most of them do feel discomfort when they over-perform comparatively to their closed social circle of classmates, the so-called social block. Block participants connect their utility to the deviation of their individual performance from the block's average performance instead of the average of the entire class. When a relative evaluation method is used and the power of incentives increases, the social cost inflicted on those who outperform their social block rises unless they moderate their effort. Conversely, under an absolute grading method, the formation of social blocks is not expected to alter the effort choices for students. This is because individual results have no impact on the other participants of the block, and high effort is not accompanied by a social cost that can be handled by adjusting effort.

Another issue examined in the experiment is whether the relative evaluation method can prevent the occurrence of answer sharing. Answer sharing is very common in multiple choice questions. It can take the form of altruism, that is, someone can share his or her work with others for personal satisfaction; it can take the form of transaction, when answers are exchanged for some kind of other favor; finally, it can take the form of free riding in study groups. Under the absolute system, answer sharing is free of cost to the individual who shares. The relative system, however, will charge a penalty to those who give their work away. In the case of multiple choice assignments, for which the answers can easily be transmitted (and then re-transmitted to several recipients), this penalty might become severe. Answer sharing can be verified by conducting a simple "means test" for homework and test scores for the class. Typically, it is expected that students who benefit from answer sharing will exhibit significant differences between their homework and test performance. For this purpose students were split into two groups according to their average performance in tests only. The means test showed that answer sharing occurred less frequently under the relative method than under the absolute method.

All the experimental results are in line with the peer effects hypothesis. Social behavior is an important factor that should be taken into consideration when one examines the provision of incentives in the principal-agent framework. The implication of the paper is that higher-power relative incentives will not be effective when agents are subject to significant peer effects.

References

[1] Agranov M. and C. Tergiman, “Incentives and Compensation Schemes: An Experimental Study”, *International Journal of Industrial Organization* (2013), 31:3, 238-247.

[2] Angrist J., D. Lang, and P. Oreopoulos, “Incentives and Services for College Achievement: Evidence from a Randomized Trial” *American Economic Journal: Applied Economics* (2009), 1(1), 136-63.

[3] Bandiera, O., I. Barankay, and I. Rasul "Social Preferences and the Response to Incentives: Evidence from Personnel Data", *Quarterly Journal of Economics* (2005), 917-962.

[4] Becker W. and S. Rosen, “The Learning Effect of Assessment and Evaluation in High School”, *Economics of Education Review* (1992), 11(2), 107-118.

[5] Brownback A., “A Classroom Experiment on Effort Allocation under Relative Grading”, *Economics of Education Review* (2018), volume 62, 113-128.

[6] Bull C., A. Schotter and K. Weigelt, “Tournaments and Piece Rates: An Experimental Study,” *The Journal of Political Economy*, (1987), 1–33.

[7] Czibor E., S. Onderstal, R. Sloof, and M. van Praag, “Does Relative Grading Help Male Students? Evidence from a Field Experiment in the Classroom”. *Tinbergen Institute Discussion Paper* (2014) 14-116/V. 8/2014.

[8] Dechenaux E., D. Kovenock, R. Sheremeta, 2015 “A Survey of Experimental Research on Contests, All-pay Auctions and Tournaments” *Experimental Economics* 18 (4), 609-669.

[9] Demougin, D. and C. Fluet, (2003), “Inequity Aversion in Tournaments”, *Cahiers de recherche, CIRPEE.*

[10] Dubey P. and J. Geanakoplos, “Grading Exams: 100, 99, 98, . . . or A, B, C?”, *Games and Economic Behavior* (2010), 69(1): 72-94.

[11] Duflo, E., P. Dupas, and M. Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101(5): 1739-74.

[12] Fehr, E. and K.M. Schmidt "A Theory of Fairness, Competition and Cooperation", *Quarterly Journal of Economics* 114(3) (1999), 817-868.

[13] Figlio D. and M. E. Lucas “Do High Grading Standards Affect Student Performance?” *Journal of Public Economics* (2004), 88(9-10) 1815-1834.

[14] Gill, D., Stone, R., (2010), “Fairness and Desert in Tournaments”, *Games and Economic Behavior*, 69(2): 346-364

[15] Green, J. and N. Stokey, “A Comparison of Tournaments and Contracts," *Journal of Political Economy* 91 (1983), 349-364.

[16] Grund, C. and D. Sliwka, (2002), “Envy and Compassion in Tournaments”, *Bonn Econ Discussion Papers*, University of Bonn, Germany.

[17] Holmström, B., “Moral Hazard in Teams," *Bell Journal of Economics* 13 (1982), 324-40.

[18] Knoeber C. and W. Thurman, “Testing the Theory of Tournaments: An Empirical Analysis of Broiler Production,” *Journal of Labor Economics* 10(4) (1994), 357-379.

[19] Knoeber C. and W. Thurman, “ ’Don’t Count Your Chickens...’ Risk and Risk Shifting in the Broiler Industry," *American Journal of Agricultural Economics* 77 (1995), 486-96.

[20] Kremer M., E. Miguel and R. Thornton, “Incentives to Learn”, *Review of Economics and Statistics* (2009), 91(3): 437-456.

[21] Lazear, E., and S. Rosen, “Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy* 89 (1981), 841-64.

[22] Mago S., A. Samak and R. Sheremeta, “Facing Your Opponents: Social Identification and Information Feedback in Contests”, *Journal of Conflict Resolution* (2016), vol. 60, 459-481.

[23] Marinakis, K. and T. Tsoulouhas, "Are Tournaments Optimal over Piece Rates under Limited Liability for the Principal?", *International Journal of Industrial Organization* 31(3) (2013), 223-237.

[24] Marinakis, K. and T. Tsoulouhas, "A Comparison of Cardinal Tournaments and Piece Rate Contracts with Liquidity Constrained Agents", *Journal of Economics* 105(2) (2012), 161-190.

[25] Moldovanu B. and A. Sela, "Contest architecture", *Journal of Economic Theory* (2006), 126(1), 70-96.

[26] Nalebuff, B. and J. Stiglitz, “Prizes and Incentives: Towards a General Theory of Compensation and Competition,” *The Bell Journal of Economics* 14(1) (1983), 21-43.

[27] Paredes V., “Grading System and Student Effort”, *Education Finance and Policy* (2017) 12:1, 107-128.

[28] Pyrczak, F., “Validity of the Discrimination Index as a Measure of Item Quality,” *Journal of Educational Measurement* (1973)(3), 227-231.

[29] Sheremeta R., “The Pros and Cons of Workplace Tournaments” *IZA World of Labor* (2016).

[30] Tsoulouhas T. and K. Marinakis, "Tournaments with Ex Post Heterogeneous Agents", *Economics Bulletin*, Vol. 4 no. 41 pp. 1-9 (2007).

[31] Tsoulouhas, T. and T. Vukina, “Regulating Broiler Contracts: Tournaments versus Fixed Performance Standards,” *American Journal of Agricultural Economics* 83 (2001), 1062-73.

[32] Wu, S. and B. Roe, "Behavioral and Welfare Effects of Tournaments and Fixed Performance Contracts: Some Experimental Evidence," *American Journal of Agricultural Economics* (87) (2005), 130-146.