Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN ECONOMIA

Ciclo XXIV

Settore scientifico disciplinare di afferenza: SECS-P01

Essays in Behavioral Personnel Economics

Presentata da:

Tommaso Reggiani

Coordinatore Dottorato

Giacomo Calzolari

Relatore

Andrea Ichino

Esame finale anno 2012

Contents

Chapter 1

Earning Honor or Money? Self-Selection of Motivated Workers

ABSTRACT

In this paper we argue that paying higher wages does not necessarily attract the right workers in work environments where where job-specific intrinsic motivation matters. Specifically, we hypothesize that mission-oriented organizations should pay wages below the market wage: Hereby, the organization will attract only those applicants that genuinely care about the organization and not so much about the wage. We test this selection mechanism in a laboratory experiment in which subjects face a choice between two jobs. This choice represents a trade-off between personal monetary payoff and a contribution to a prominent mission-oriented organization. We find that a lowering the wage for the mission-oriented job leads to a smaller, but significantly more intrinsically motivated applicant pool for this job. However, effort that workers exert remains unchanged. We conclude that mission-oriented organizations profit from underpaying relative to the market wage.

Keywords: Intrinsic Motivation, Labor Economics, Selection, Lab Experiment.

JEL code: C91, J31, L31, M52.

1.1 Introduction

Economists have hypothesized that firms and organizations can attract harder-working or more qualified workers by paying wages higher than a worker's marginal productivity. Models of efficiency wages are based on the notion that higher wages reduce shirking or the need of monitoring (Shapiro and Stiglitz 1984), attract workers with better outside options (Akerlof 1970, Greenwald 1986), or induce workers' reciprocity (Akerlof 1984, Yellen 1984). In this paper, we examine an important mechanism inducing some organizations to offer wages *below* the market rate in order to attract the "right" workers that share the employers' goals. In particular our experiment allows to shed light on two main questions: First, how do wage gaps affect the self-selection of heterogeneous workers into jobs that they potentially care about? Do low wages sort those workers who have the highest intrinsic motivation for working towards the organization's goals?
Second, how much effort do those workers provide who sacrifice a higher wage in order to work towards a goal that they care about?

Recently, economists have hypothesized that identity (Akerlof and Kranton 2010) as well as workers' motivation to work for a particular firm or a particular cause are important determinants of self-selection in the labor market (Rebitzer and Taylor 2011, Besley and Ghatak 2005, Heyes 2005, Delfgaauw and Dur 2007, Brekke and Nyborg 2010, Barigozzi and Turati forthcoming). This line of research stresses that the utility derived from a job goes beyond wages and other perks, and also encompasses idealistic benefits such as the comfort to know that one works for a worthy cause. Casual observations and introspection suggest that the latter can indeed be an important motive in shaping vocational choices. When workers are heterogeneous in their motivation to work for a particular firm or organization, paying higher wages may lead to an applicant pool that is on average less motivated or identifies less with the potential employer. Conversely, by paying lower wages, an organization will attract applicants that genuinely care about the organization and its goals. If wages are high, the organization will also attract applicants that do not identify much with the organization but apply merely because of the high wage.

A number of reasons suggest that some organizations might attract the "right" workers by paying lower and not higher wages. There is empirical evidence indicating that such a selection mechanism may exist in real labor markets. For instance, some authors have documented a wage gap between workers in the non-profit and the for-profit sector although the evidence is somewhat mixed (see, e.g., Hansmann 1987, Leete 2001, Benz 2005). Insofar as non-profit organizations aim to attract workers with cause-specific motivation (more so than for-profits), this can be seen as evidence consistent with a selection mechanism such as the one described in the previous paragraph[1]. However, this evidence is far from conclusive in evaluating the validity of a selection mechanism based on job-specific motivation. Moreover, it is hard to test such a mechanism empirically because of a number of confounding factors. Stern (2004) finds that researchers in biology are willing to sacrifice a part of their wage in order to work in a job that allows them to do research. However, more recent work by Sauermann and Roach (2010) highlights the role of heterogeneity in ability and a "taste for research" in order to understand the pattern of wages and the types of jobs potential researchers choose. This heterogeneity highlights the role of sorting in labor markets — which is at the core of the mechanism we describe — and suggests that a controlled experiment is an important tool to understand the causal effects of wages in the presence of heterogeneity of worker's preferences.

The mechanism we have described so far was an analysis of worker behavior in response to different wages; however, the outcomes observed empirically by an econometrician are determined by a labor market equilibrium that depends on both workers' and firms' or organizations' optimizing behavior. Moreover, observational data usually lacks exogenous variation in the wage gap between different types of jobs. In addition, it is hard to obtain measures of workers' attitudes toward different potential employers before choosing jobs. All in all, these factors obstruct a clean identification of the effect of wage differentials, between, e.g., jobs in the non-profit and for-profit sector, on workers' self-selection.

[1] From a different perspective **?** claim that such organizations could provide lower wages aiming to signal their genuine non-profit nature to the society.

A laboratory experiment is well-suited to overcome the obstacles faced by empirical research using observational data since it provides a controlled environment that allows identification of the causal effect of wage differentials on workers' self-selection. Falk and Heckman (2009a) argue that laboratory experiments provide an important tool for causal inference in the social sciences: previous work (e.g., Charness and Kuhn 2011, Falk and Gächter 2008a, Falk and Ichino 2006, Falk and Fehr 2003a) demonstrates that laboratory experiments are an especially valuable source of knowledge about the functioning of labor markets. The key advantage of implementing a laboratory experiment in our case is the ability to exogenously vary the wage gap between different types of jobs and to obtain clean measures of subjects' ability and attitudes toward different causes before choosing a job.

The precise work environment we consider in our paper is individuals' support for protecting the environment. First, we elicit subjects' environmental attitudes and identification with a well-known environmental organization, Greenpeace Germany. This is conducted as part of a broader survey so that subjects are not too focused on the questions relating to environmental issues. Before subjects start the main part of the experiment that requires them to work on a monotonous real effort task (counting the frequency of specific numbers in a table), we familiarize subjects with this task and collect a skill measure. Next, subjects face a choice between two jobs: a "green" job and a "standard" job. In our three treatments, we exogenously vary the upfront wage in the "green" job across subjects, it is always weakly lower than the upfront wage in the "standard" job. Except for the donation to Greenpeace associated with working on the "green" job and potential differences in upfront wages, both jobs are identical: subjects engage in the in the real effort task for 40 minutes without the option to quit. Beyond the varied upfront wage, both jobs pay a small remuneration for each table counted correctly that is identical in both jobs. The donation that is generated in the "green" job consists of a piece rate that it paid for each correctly solved table in addition to the private piece rate. The donation is paid by the experimenters on behalf of the subjects. This design allows an assessment of how different wage gaps affect the composition of workers choosing the "green" job in terms of their environmental attitudes and identification with the

organization towards which their work contributes.

Our setup overcomes the econometric identification issues typically afflicting observational wage data in several ways as described above. Most important, we vary the wages in the "green" job across subjects and thus create the exogenous variation needed — but typically lacking in observational data — to identify which jobs individuals choose when faced with different wage gaps. Moreover, the experiment provides a clean measure of subject's skill and identification with their potential employer before job selection takes place. The laboratory environment is also appealing since it allows us to leave the two jobs completely identical — except for the donation generated in the "green" job. Thus, variations in behavior caused by wage changes identify the value of being able to generate donations to Greenpeace and are not affected by other differences between jobs that could generate other hedonic wage differentials in non-laboratory settings.

We find that offering lower wages is indeed an effective mechanism to screen motivated workers. In all treatments, we find that subjects who identify with Greenpeace are more likely to choose the "green" job. In the treatment with equal fixed wages of 8 euro for the "standard" and the "green" job, more than 94 % of subjects choose the "green" job. When imposing wage penalties of 1 €or 3 €for choosing the "green" job, the fraction choosing the "green" job shrinks to 74 % and 21 %, respectively. As hypothesized in the literature, the pool of subjects choosing the "green" job in the treatments with lower wages identifies more strongly with Greenpeace than the subjects choosing the "green" job in the treatment with equal wages; the difference is statistically significant. Comparing the subjects choosing the "green" job in the treatments with a wage penalty, reveals that increasing the wage gap from 1 € to 3 € leads to a "greener" pool of subjects choosing the "green" job, however, the difference is not statistically significant. This suggests that the compositional effect we find is driven by the existence of a wage penalty — with the magnitude of the wage penalty being less important than the presence of a wage penalty per se. Even small wage decreases relative to the outside option can be sufficient to screen workers that identify with Greenpeace. With respect to the output produced on the job we find that paying lower wages does not re-

duce output produced by workers choosing the "green" job. In all treatments, the difference in output between workers in the "standard" and the "green" job is not significant. Overall, our experiment suggests that underpaying workers relative to their outside option — even by a small amount — can be an effective mechanism for organizations to select the "right" workers.

Our work provides empirical support for theories positing an effect of worker motivation or identity on an individual's choice of job or employer (Besley and Ghatak 2005, Heyes 2005, Delfgaauw and Dur 2007, Brekke and Nyborg 2010, Sauermann and Roach 2010). To the best of our knowledge, there is no previous experimental evidence shedding light on how wages affect self-selection of heterogeneous workers in environments where identification with an employer matters. Our results are in line with the findings of Lazear, Malmendier, and Weber (2012) who investigate the role of sorting in experiments. In their experiment, individuals can avoid or opt into situations where they can share money with others. Lowering the cost of sharing leads to a larger fraction of subjects opting into the sharing situation — however, these subjects are the ones who share least. Relatedly, we find that reducing the wage penalty of the "green" job leads to a worker pool that identifies less with Greenpeace. Our findings are relevant for the literature on the importance of intrinsic motivation for the provision of effort (e.g., Ryan and Deci 2000, Kreps 1997, Besley and Ghatak 2005, Francois and Vlassopoulos 2008, Ariely, Bracha, and Meier 2009, Benabou and Tirole 2006). Our experiment can also inform the debate on why jobs in the non-profit sector typically pay lower wages than similar jobs in the for-profit sector (Hansmann 1987, Leete 2001, Benz 2005). Our results suggest that a mission-oriented organization can attract the "right" workers by underpaying relative to a worker's outside option. Further work needs to address how similar mission-oriented organizations, e.g., Greenpeace and the Green Belt Movement, compete for workers.

The rest of the paper is structured as follows: In Section 1.2 we develop a theoretical model that serves as the road map for our study. Section 1.3 describes the experimental design. Section 1.4 is devoted to the analysis of the data, Section 1.5 concludes.

1.2 Model

We present a conceptual framework to analyze output decisions and self selection in light of heterogeneity of workers' ability and job-specific intrinsic motivation. We assume that workers' preferences can be represented by the utility function

$$U(y, j|w, \alpha, \gamma) = w_j(y) - \alpha \cdot c(y) + \gamma \cdot I_j \cdot g(y), \tag{1.1}$$

with

$$\begin{aligned}
&\alpha > 0,\ \gamma \geq 0,\\
&j \in \{G, S\},\ \text{ with }\ I_S = 0,\ I_G = 1,\\
&c'(y) > 0,\ c''(y) > 0,\\
&g(y) \geq 0,\ g'(y) \geq 0,\ g''(y) \leq 0.
\end{aligned}$$

In this setup, $y \geq 0$ denotes the output produced by a worker and $w_j(y)$ is the wage function for job j.[2] The costs of producing output are captured by the convex function $\alpha \cdot c(y)$. Workers are heterogeneous in their ability which is captured by varying cost parameters α. The term $\gamma \cdot I_j \cdot g(y)$ captures a "non-standard" utility component, i.e., the component that is not captured by the monetary incentives and effort costs. There are two possible types of jobs $j \in \{G, S\}$, with $I_S = 0$ and $I_G = 1$. On the "green" job G, the worker experiences some non-monetary utility.

The model is flexible enough to account for two different types of preferences: Non-monetary utility could be derived from generating a positive externality (e.g., by contributing to a public good) while working on the job, but it could also be that the worker is just enjoying the job without caring about the output. The relative weight of this utility part is measured by the "identification" parameter γ. I can be interpreted as measuring the overlap between a worker's identity and the goals of a firm. The function $g(\cdot)$ is identical across workers, but the parameter γ varies

[2] We are reluctant to interpret y as "effort", because we assume heterogeneity in workers's ability, meaning that workers experience different effort costs when producing the same amount of output. We think that the concept of "effort" is closer related to the costs of producing some fixed amount of output than to the amount of output itself.

between workers. On the "standard" job S, the worker does not experience any non-monetary utility.

The functional form of $g(\cdot)$ reflects the property of the worker's social preference. If $g'(y) > 0$, the worker gets additional positive non-monetary marginal utility from producing output. This type of social preference can represent pure (or outcome-oriented) altruism as well as warm-glow (or action-oriented) altruism (compare Andreoni 1989, Andreoni 1990, and Tonin and Vlassopoulos 2010). In the special case of $g'(y) = 0$ and $g(0) > 0$, the worker receives utility from working at a job G as opposed to S, but he does not receive extra (non-monetary) marginal utility by producing more output. This social preference could be labeled as "participation utility".

Suppose that the worker's payment is a fixed wage plus a piece rate, i.e., $w_j(y) = w_j + py$. While fixed component varies between the jobs, the piece rate is assumed to identical.[3] In this case, the optimal output level y_j^* on job j is chosen satisfying the FOC[4]

$$\left.\frac{\partial U(y, j|w, \alpha, \gamma)}{\partial y}\right|_{y_j^*} = p - \alpha \cdot c'(y_j^*) + \gamma \cdot I_j \cdot g'(y_j^*) = 0. \tag{1.2}$$

Now suppose that workers can choose among the "green" job G and the "standard" job S. While the piece rate is equal for both jobs, the fixed payments w_G and w_S may differ. Having chosen some job $j \in \{S, G\}$, the optimal output level y_j^* produced by the worker is the one satisfying the FOC (1.2). By applying the implicit function theorem, we see that the optimal output level decreases with the cost parameter and (weakly) increases with the identification parameter:

$$\frac{\partial y_j^*}{\partial \alpha} = \frac{c'(y_j^*)}{\gamma_G \cdot I_j \cdot g''(y_j^*) - \alpha \cdot c''(y_j^*)} < 0 \qquad \text{and} \tag{1.3}$$

$$\frac{\partial y_j^*}{\partial \gamma} = \frac{I_j \cdot g'(y_j^*)}{\alpha \cdot c''(y_j^*) - \gamma \cdot I_j \cdot g''(y_j^*)} \geq 0, \tag{1.4}$$

where the second expression holds with equality if $I_j \cdot g'(y_j^*) = 0$, i.e., if the

[3] When piece rates are not identical, the results qualitatively change.

[4] In the following we assume that the functions $c(\cdot)$ and $g(\cdot)$ are continuously differentiable up to the degree needed and that internal solutions exist.

marginal non-monetary utility is zero.

Prediction 1. *Workers with a low cost parameter produce more output.*

Prediction 2. *On the "green" job, output increases with the motivation γ iff the marginal non-monetary utility is positive.*

One of the key question we want to address is what job a worker will select into when faced with different alternatives. When choosing among two alternative jobs associated with fixed wage components w_G and w_S, the worker compares the indirect utility associated with each job and chooses the green job over the standard job iff

$$U(y_G^*, G|w_G, \alpha, \gamma) \geq U(y_S^*, S|w_S, \alpha, \gamma). \tag{1.5}$$

Clearly, a worker with $\gamma > 0$ will choose the "green" job if the wage gap is negative, i.e. $w_S - w_G < 0$. However, we are interested in cases of a positive wage gap, i.e. $w_G < w_S$. This case is empirically more prevalent, as the observed wage gap between non-profit and for-profit organizations suggests. When w_G is lowered (increasing the wage gap $w_S - w_G$), inequality (1.5) holds for fewer subjects. The first sorting implications of the wage gap are straightforward:

Prediction 3. *Raising the wage gap $w_S - w_G$ results in a smaller pool of subjects opting for the "green" job.*

Prediction 4. *For a given wage gap, applicants for the "green" job have a higher degree of intrinsic motivation than applicants for the "standard" job.*

Beyond that, our goal is to analyze how the distribution of intrinsic motivation and cost parameters of those choosing the "green" job is influenced by the wage gap. We do this by analyzing how the marginal worker, i.e., the one who is indifferent between the standard and the "green" job, is affected by changes of w_G holding w_S fixed. If the characteristic (i.e., skill or identification) of the marginal worker rises, then the averages of this characteristic in applicant pools for both jobs move in the same direction. We can analyze the marginal type by applying the

implicit function theorem to the indifference condition

$$F(y_G^*, y_S^*, w_G, w_S, \alpha, \gamma) = U(y_G^*, G|w_G, \alpha, \gamma) - U(y_S^*, S|w_S, \alpha, \gamma) = 0. \quad (1.6)$$

Holding ability constant, this leads to a negative effect of w_G on the marginal worker's intrinsic motivation $\tilde{\gamma}$ and, using (1.4), to a weakly negative effect on his output $\tilde{e_G}$:

$$\frac{\partial \tilde{\gamma}}{\partial w_G} = \frac{-1}{g(1, y_G^*)} < 0 \qquad \text{and} \quad (1.7)$$

$$\frac{\partial \tilde{y}_G}{\partial w_G} = \frac{\partial \tilde{y}_G}{\partial \tilde{\gamma}_G} \frac{\partial \tilde{\gamma}}{\partial w_G} \leq 0. \quad (1.8)$$

Prediction 5. *Raising the wage gap increases the average intrinsic motivation of applicants for the "green" job.*

In a next step, we analyze how the ability of the marginal type $\tilde{\alpha}$ is affected by changes in w_G when holding motivation constant. It follows from (1.6) that the cost parameter of the marginal type $\tilde{\alpha}$ increases in w_G and output $\tilde{e}_G$ decreases:

$$\frac{\partial \tilde{\alpha}}{\partial w_G} = \frac{1}{c(y_G^*) - c(y_S^*)} \geq 0 \quad \text{and} \quad (1.9)$$

$$\frac{\partial \tilde{y}_G}{\partial w_G} = \frac{\partial \tilde{y}_G}{\partial \alpha} \frac{\partial \tilde{\alpha}}{\partial w_G} \leq 0. \quad (1.10)$$

Both equations hold with equality if $g'(y) = 0$, because this implies $y_G^* = y_S^*$.

Prediction 6. *Raising the wage gap lowers the average cost parameter of applicants for the "green" job iff the marginal non-monetary utility is positive.*

Both (1.8) and (1.10) provide the same and last implication:

Prediction 7. *Raising the wage gap increases the average output produced on the "green" job iff the marginal non-monetary utility is positive.*

Predictions 5 to 7 where derived using a ceteris paribus analysis, i.e., we assumed that the cost parameter and motivation respectively do not change. In fact,

the predictions also hold true under the assumption that the cost parameter and motivation are distributed independently. If they are not distributed independently, the joint distribution can cause non-monotonic sorting (compare Barigozzi and Turati forthcoming).

1.3 Experimental Design

Our experiment sheds light on self-selection in a context where identification with a potential employer matters (Besley and Ghatak 2005, Heyes 2005, Delfgaauw and Dur 2007, Brekke and Nyborg 2010). It is closely related to our model in order to test the prediction put forward in the last section. Therefore, the experiment offers subjects the choice between a "green" job that pays an equal or lower wage and supports a well-known environmental organization (Greenpeace), and a "standard" job that pays a higher wage but does not generate any positive externality. In both potential jobs, subjects engage in a tedious real-effort task for 40 minutes. This design captures the key difference between jobs at non-profit and for-profit organizations and allows us to exogenously vary the wage gap between the two jobs. We can directly observe job choice and subsequent output. Prior to the main phase of the experiment, we measure subjects' environmental attitudes and identification with Greenpeace in a survey.

The experiment was programmed using the "Bonn Experiment System" software BoXS by Seithe (2010), and participants were recruited through ORSEE (Greiner 2004). It is partitioned into four different phases. Subjects receive the instructions for the current phase at the beginning of each phase on the computer screen or on paper.

1.3.1 Survey Phase: Elicitation of Intrinsic Motivation

The experiment starts with a survey containing questions typically also included in socioeconomic surveys. The questions ask about important demographic characteristics like gender, age, and field of study and also measures preferences as well as personality traits using the "Big Five" inventory. The survey also asks how much

subjects identify themselves with the mission pursued by Greenpeace, and how much they identify with some other quite well-known, albeit non-environmental, non-profit organizations.[5] We will refer to this variable as "identification" in the following. The survey also contains questions asking subjects more general questions with regards to their attitude towards the environment and other causes.

From a methodological point of view, it is important to conduct the survey before the choice phase of the experiment because this allows a clear identification of subjects' attitudes toward Greenpeace that is not affected by treatment assignment. Conducting the survey after the main phases of the experiment would have been problematic because subjects' job choice may influence their survey answers. Our design circumvents this type of problem and provides us with a clean measure of subjects' attitudes that is orthogonal to treatment assignment.[6]

1.3.2 Training Phase

In a next step, subjects enter a training phase of 5 minutes that familiarizes them with the task in the main phase. The task consists in counting the number of ones in a table of 120 randomly ordered zeros and ones (compare Abeler, Falk, Götte, and Huffman 2011)[7]. Subjects have to count the number of ones in the table and enter it into the computer. For each correctly solved table they get a piece-rate of 0.10 €. After entering the number, "correct" or "false" is displayed for 2 seconds, then the next table appears (see Figure 1.3.1).

This phase provides an exogenous skill measure (i.e. independent of Treatment assignment) because it is part of all treatments and prior to Treatment assignment. We interpret the number of correctly counted tables in the training phase as an inverse measure of the cost parameter α from above. We will refer to this variable

[5] An analogous method has been used by Ariely, Bracha, and Meier (2009) and Fehrler and Kosfeld (2010). Our exact question was: "*How much do you identify yourself with the goals of Greenpeace (on an 11-point-scale, reaching from 0 to 10)?*"

[6] One could still object that job choice could partly be driven by the intention to behave consistent with the survey answer. However, only the job choice has payoff- and donation-relevant consequences. Thus, having the survey before the job choice reduces consistency issues.

[7] "*This task does not require any prior knowledge and performance is easily measurable; at the same time, the task is boring and pointless and we can thus be confident that the task entailed a positive cost of effort for all subjects. The task was also clearly artificial, and output was of no intrinsic value to the experimenter.*" (Abeler, Falk, Götte, and Huffman 2011)

as "skill" in the following. Before the training phase started, we made sure that subjects understand the task by displaying one table and having them count the numbers. The training only starts after a correct number has been entered.

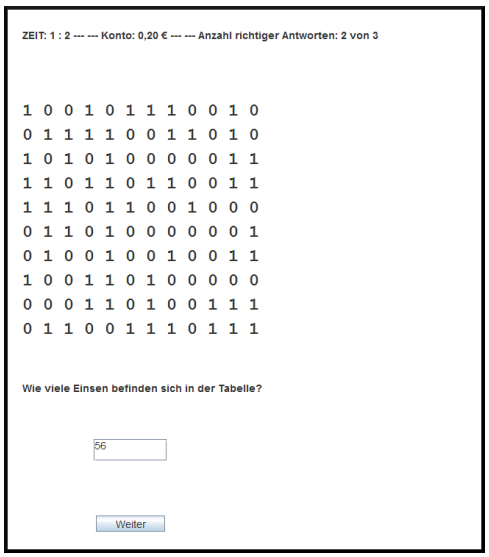


Figure 1.3.1: Real-Effort Task (Training Phase)

1.3.3 Treatment Phase

After the training phase, subjects are randomly assigned to one of the three treatments. In our between-subject design, each subject participated only in one of the treatments and was not informed about the other treatments. In each treatment, they have 40 minutes to work on the real effort task described before and get a (private) piece-rate of 0.10€ for each correctly solved table independent of their job choice. At the beginning of the treatment phase, subjects have the opportunity to choose between two alternative job arrangements: In all three treatments, subjects choosing the "standard" job earned a fixed wage component of $w_S = 8$€. The only parameter that was exogenously varied between the treatments is the fixed wage component for the green job. In our Treatment T0 it was 8 €, (wage gap$= 0$€), in Treatment T1 it was 7 €, (wage gap$= 1$€) and it was 5€ in Treatment T3 (wage gap$= 3$€). In the "green" job, subjects additionally generate a donation of 0.10€ for each correctly solved table that is donated by the experimenters on behalf of the subjects to Greenpeace Germany. An overview over the different treatments

can also be found in Table 1.6.1 in the Appendix. The jobs are explicitly labeled as "jobs" to make the job-market aspect somewhat salient. The exact labels in the experiment are "Job A" and "Job B (Greenpeace)".

To avoid possible peer-effect due to subjects leaving subjects are not allowed to quit the job (e.g. Falk and Ichino 2006, Linardi and McConnell 2011), but may stop counting whenever they want even though they have to wait until the 40 minutes are over[8]. While the job decision is our fist main outcome variable, the number of correctly counted tables in the treatment phase is our second main outcome variable and will be referred to as "output" in the following. We make sure that the subjects understand the payment schemes of both jobs and make an informed decision[9].

1.3.4 Final Phase: Robustness Check

After the main phase, all subjects are asked whether they plan to contribute money to Greenpeace in the near future. If they answer affirmatively, they are given the opportunity to donate any amount between zero and the total amount they earned for themselves in the experiment. This allows us to rule out a potential alternative strategy that subjects may choose because of potential "efficiency concerns": Subjects might choose the "standard" job associated with the higher upfront wage — even though this does not generate donations to Greenpeace on the job — in order to donate part of the personal payoff afterwards. If this was the case, our offer to donate part of their payoff is the transaction-cost-minimizing possibility to donate. The experiment ends with a short final survey section eliciting general personal characteristics and the attitude towards six general environmental issues[10].

[8] Subjects are allowed to bring readings and other items into the cubicles in order to use them during potential waiting times.

[9] Subjects have to answer three control questions of the form: *Imagine you managed to solve 60 tables in the following 40 minutes. How much would you earn, and how much would be donated to Greenpeace in "Job A" and in "Job B (Greenpeace)"?* Subjects cannot proceed until they found the correct solutions.

[10] For the environmental questions see Appendix 1.6.2.

1.4 Results

We conducted six sessions with a total of 144 subjects, 48 in each of the three treatments[11]. All sessions were conducted at the BonnEconLab using the laboratory's subject pool. 95% of the subjects were students of various disciplines, 56% of them were female.

The two key variables that we elicit in the first two phases will be called "identification" and "skill" in the following. "Identification" captures the degree of identification with the goals of Greenpeace (on a 11-point scale) and thus corresponds to the motivation parameter γ in the model. "Skill" measures the number of correctly solved tables during the training phase (5 minutes); it corresponds to the inverse of the cost parameter α from the model. The distribution of the variables "identification" and "skill" for the whole sample are displayed in Figure 1.4.

The subjects' average private payoff was 14.27 €. Among those choosing to work for Greenpeace the average contribution to Greenpeace (excluding the voluntary donation at the end of the experiment) was 6.29 €. Summary statistics by treatment and job choice can be found in Table 1.6.1 (Appendix).

Columns (1), (3) and (5) of Table 1.4.2 provide the results from a regression of output (i.e., number of correctly counted tables in the treatment phase) on skill, the identification with Greenpeace and treatment dummies. Columns (1) and (3) show the results for workers in the "standard" job and in the "green" job, respectively, while column (5) shows the results for all workers, adding a dummy variable indicating the choice of the "green" job as a control.

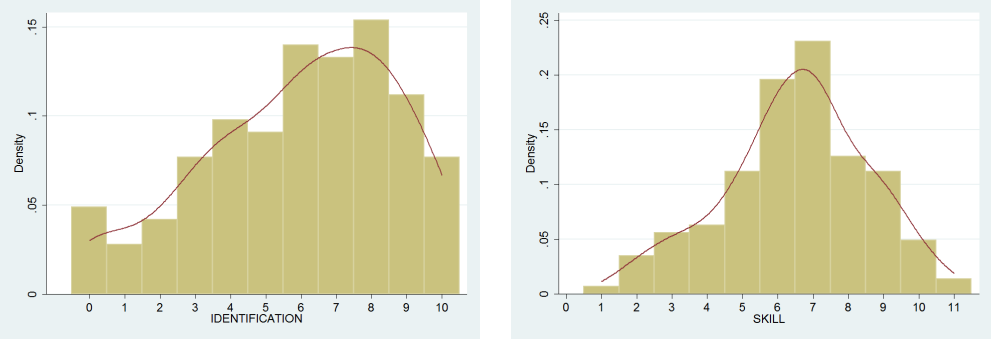
Finding 1. *High-skill workers produce more output.*

In fact, skill is the only variable having a significant effect on output, and it is sizable and highly significant in all specifications.[12]

[11] Due to technical problems, one subject dropped out before the treatment phase.

[12] Note that while the regression coefficient on skill is sizable, it is still surprisingly low: The time available in the treatment phase is eight times longer compared to the training phase, and subjects in fact managed to solve approximately ten times more tables in the treatment phase. The low coefficient (around 4, which is half of the expected effect) is probably due to a very specific "measurement error": Since we elicit "skill" in a phase of 5 minutes, our measure includes an error because (i) output has a random component (there is some probability of entering a wrong number), and (ii) we only measure an integer, i.e., the number of *completed* tables. The interpretation as an "errors in regressors" effect

Figure 1.4.1: Distribution of individual characteristics.



The two panels show the distributions (histogram and Gaussian kernel density) of the identification measure (degree of identification with the goals of Greenpeace, 11-point scale) and the skill measure (the number of correctly counted tables in the 5-minute Training Phase).

As stated earlier, we interpret skill as an inverse measure of the cost parameter, high skills correspond to low costs. By supporting Prediction 1, this finding indicates that there is in fact heterogeneity in the workers' ability which we are measuring with the variable skill.

Finding 2. *Output does not increase with identification in the "green" job, but it does not decrease either.*

As expected, identification has virtually no effect on output in the "standard" job (column 1). In light of Prediction 2, this suggests that workers do not receive positive marginal utility from generating the donation for Greenpeace. The non-monetary part of the utility function thus seems to exhibit the property that is described as "participation utility" in Section 1.2. This explanation is in line with Fehrler and Kosfeld (2010) who suggest that "the view that motivation by, or identification with, an employer's mission could work as a substitute for piece rate payments might be inadequate." It is also in line with recent evidence against pure (or output-oriented) altruism (Tonin and Vlassopoulos 2010). Though our finding of "participation utility" is not identical to the warm-glow altruism described in latter paper, there are similarities: In both cases the agent does not derive utility from his effective impact, but rather from being involved personally. When facing the

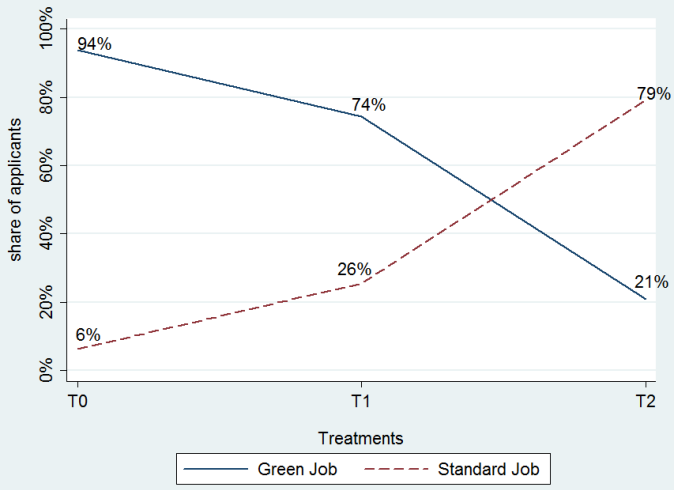
is also supported by the fact that the "inverse" regression of skill on output in the Treatment Phase gives a coefficient of about 0.07, which is approximately the inverse of 14.

hypothetical decision to match oneself and another agent to one "green" and one "standard" job, an agent with either "participation utility" or warm-glow altruism strictly prefers to match the "green" job to himself, while a purely altruistic agent is indifferent.[13]

Finding 3. *Raising the wage gap results in a smaller pool of workers opting for the "green" job.*

In Treatment T0, 45 out of 48 subjects choose the "green" job, in Treatment T1, 35 out of 47 subjects choose to work for the "green job", whereas only 10 out of 48 subjects choose the "green" job in Treatment T2. Figure 1.4.2 visualizes this result.The effect of the wage gap on sorting behavior can also be seen in the probit regressions of job choice on identification, wage gap or treatment dummies and skill, reported in columns (1) and (2) of Table 1.4.1. An increase of the wage gap by 1 € reduces the probability of choosing the "green" job by 34% on average.

Figure 1.4.2: Shares of job choice, by Treatment



The graph shows the shares of applicants for the "standard" job (dotted line) and the "green" job (solid line) for the three Treatments T0, T1, T2 (wage gap of 0 €, 1 € and 3 €).

[13]This illustrative example neglects effort decisions and just assumes that both agents perform identically and independently of the job match.

Table 1.4.1: Determinants of the Job Choice

Probit Regression (Marginal Effects) Dependent Variable: 1 if "green" job is chosen			
	(1)	(2)	(3)
Identification	0.0979***	0.0979***	
	(0.0215)	(0.0214)	
"green" factor			0.0795***
			(0.0205)
Wage Gap	-0.336***		
	(0.0482)		
T1		-0.435***	-0.434***
		(0.136)	(0.130)
T2		-0.874***	-0.827***
		(0.0594)	(0.0660)
Skill	0.00672	0.00691	-0.00183
	(0.0223)	(0.0220)	(0.0228)
Observations	143	143	143

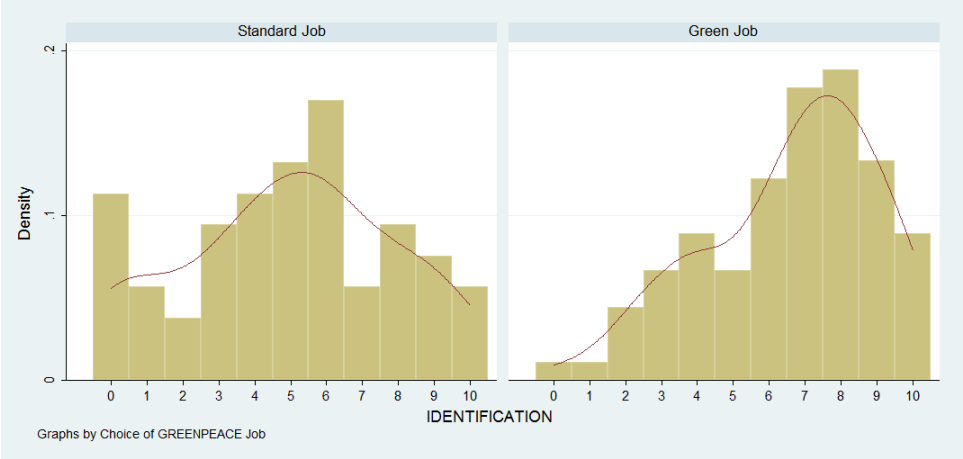
Standard errors in parentheses. (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)
The dependent variable is a dummy variable identical to one if a subject chooses the "green" job. This table presents the marginal effects at the mean of a probit regression. The explanatory variables are the "identification" (with the goals of Greenpeace, measured on an 11-point-scale) in the first two columns and the "green" factor constructed from the six questions about the attitude towards environmental issues (see Appendix 1.6.2). Column (1) uses the wage gap (treatment variable) as a linear argument, Column (2) uses the treatment dummies T1 and T2 (wage gap of 1 € and 3 €). Skill (the number of correctly counted tables in the 5-minute Training Phase) is always used as a control variable.

Finding 4. *The average identification with Greenpeace is significantly higher in the "green" job than in the "standard" job.*

Figure 1.4.3 depicts the strong difference in the distribution of subjects identification between the "green" and the "standard" job (this figure aggregates all treatments). The average identification with Greenpeace of workers in the "standard" job is 4.94, while it is 6.57 in the "green" job. A Wilcoxon rank-sum test reject the equality of the distribution (p-value < 0.01). This notion of favorable selection is also supported by the fact that the degree of identification with Greenpeace positively influences the probability of choosing the "green" job: The probit regression in column (2) of Table 1.4.1 shows that an increase in the identification measure by one point increases the likelihood of choosing the "green" job by 10%.

This effect is highly significant and sizable.

Figure 1.4.3: Distribution of "identification" by job choice.



The two panels contrast the distribution (histogram and Gaussian kernel density) of the identification measure (identification with the goals of Greenpeace, measured on an 11-point-scale) of those subjects choosing the "standard" job (left panel) and those choosing the "green" job (right panel). These panels pool the data from all three treatments.

Finding 5. *Raising the wage gap increases the average identification of workers in the "green" job.*

Figure 1.4.4 shows how the distribution of identification varies across treatments and job choice. The left diagram depicts the distribution of the identification measure of all subjects across the three treatments. Since the three distribution are virtually identical, this confirms that our randomized assignment to the different treatments produced comparable groups. The right diagram shows how the identification of those choosing the "green" job changes when moving from a wage gap of zero to a positive wage: The higher the wage gap, the more probability mass is at the upper identification levels, so that the distributions can be ranked in terms of statistical dominance. The Wilcoxon rank-sum test confirms that this difference between the two treatments with positive wage gaps (T1 and T3) and the treatment with zero wage gap (T0) is statistically significant (p-value of 0.0267). The differences of the average identification level across these two treatment groups is 1.22.

Figure 1.4.4: Self-Selection: Comparison of "identification" across treatments.

These two panels visualize the self-selection of workers based on their identification with Greenpeace across treatments. The three lines are the Gaussian kernel densities of the identification with Greenpeace. The solid line represents the distribution of subjects in Treatment T0 (zero wage gap), the dashed line of those in Treatment T1 (wage gap 1 €) and the dotted line of those in Treatment T2 (wage gap 3 €). The left diagrams displays the distribution for all subjects in the experiment, the right one only of those choosing the "green" job. While the left diagram shows that the populations are comparable across treatments, the right diagram shows that a an increasing wage gap leads monotonely to a "better" composition of applicants to the "green" job as compared to a zero wage gap.

Finding 6. *Raising the wage gap does not affect the average skill of workers in the "green" job.*

We find hardly any difference in the skill level of workers in the "green" job across treatments (compare Figure 1.6.1). In terms of the model prediction this again indicates that the "green" job does not seem to provide additional marginal non-monetary incentives. However, it is important to notice that no negative selection occurs either. This finding indicates that there is no self-selection of workers based on skills. This is confirmed by the analysis of the selection decision: The probit regression displayed in Table 1.4.1 shows no effect of skills on the probability of choosing the "green" job.

Finding 7. *Raising the wage gap does not influence the average output of workers in the "green" job.*

The regression of output in Table 1.4.2 shows that the wage gap does not have any effect on the output produced in the "green" job. This finding is in line with Findings 4 and 6. Again, the effect is neither positive nor negative. The latter point deserves some attention. Though our model predicted a weakly positive effect, one

could also argue that when workers exert negative reciprocity or similar preferences towards the experimenter, a lower wage in the "green" job might lead to lower productivity. However, this is not the case. Although paying less to the workers, the virtual "green employer" does not suffer any negative consequence in our experiment, neither in terms of sorting nor in terms of subsequent effort decisions.[14]
In the model section we mentioned the relationship between output in the Treatment Phase, skills and effort. So far, we only made statements about the former two. To assess effort, we now analyze the "skill-normalized output", i.e., the ratio between output in the Treatment Phase and skill. This ratio measures how hard a subject works in the treatment as compared to the training phase and can thus be interpreted as an effort measure when assuming that all subjects worked equally hard (according to their skills) in the training phase. Figure 1.6.2 (Appendix) shows the averages across treatments and jobs. None of the differences are statistically significant. It appears that neither job choice nor treatment variations can induce subjects to work harder.

The dependent variable output is the number of tables correctly counted in the Treatment Phase. The first two columns contain the sample of subjects choosing the "standard" job, the two columns in the middle the sample choosing the "green" job and the last two columns contain the whole sample. The explanatory variables are the "identification with the goals of Greenpeace" (measured on an 11-point-scale) in columns (1), (3) and (5), and the "green" factor constructed from the six questions about the attitude towards environmental issues (see Appendix 1.6.2) in columns (2), (4) and (6). The treatment dummies T1 and T2 (wage gap of 1 € and 3 €) and skill (the number of correctly counted tables in the 5-minute Training Phase) are used for all specifications. In the last two columns (whole sample), a dummy variable identical to one if the "green" job is chosen is included.

[14] Of course, the lower number of applicants might be a downside in case the employer has to fill many vacancies (compare Delfgaauw and Dur 2007).

Table 1.4.2: Determinants of Output in the Treatment Phase (OLS)

Dependent Variable	Output in the Treatment Phase					
Sample	"standard" job		"green" job		all	
	(1)	(2)	(3)	(4)	(5)	(6)
Identification	-0.161		0.357		0.194	
	(0.691)		(0.682)		(0.485)	
"Green" factor		0.231		0.746		0.486
		(0.636)		(0.705)		(0.478)
Skill	3.650***	3.650***	4.508***	4.486***	4.161***	4.131***
	(0.916)	(0.913)	(0.784)	(0.781)	(0.594)	(0.592)
T1	-9.254	-9.575	0.777	0.117	-0.738	-1.217
	(8.794)	(8.782)	(3.416)	(3.464)	(3.030)	(3.057)
T2	-2.387	-3.173	-0.821	-1.663	1.455	1.018
	(8.370)	(8.161)	(5.366)	(5.388)	(3.958)	(3.873)
Choice of "green" job					2.523	1.857
					(3.527)	(3.434)
Constant	40.87***	39.56***	30.08***	28.01***	31.29***	30.47***
	(9.822)	(9.763)	(7.100)	(7.037)	(5.667)	(5.611)
Observations	53	53	90	90	143	143
R-squared	0.261	0.262	0.288	0.295	0.276	0.281

Standard errors in parentheses. (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

We replicated some of our key results using a factor analysis to extract a single factor of individual "greenness" instead of the 11-point scale identification with Greenpeace. For the factor analysis, we use subjects' attitude towards six environmental issues (e.g. climate change, pollution, ...) as well as the original 11-point scale identification with Greenpeace. Responses to the different items are highly correlated, which is consistent with the different questions capturing the same factor. The factor analysis indicates that actually only one factor should be kept, we call it the "green" factor. Our results are basically unchanged when we use the "green" factor instead of the identification with Greenpeace for the analysis (compare Tables 1.4.2 and 1.4.1). Further details can be found in Appendix 1.6.2.

At the end of the experiment, only 15 subjects said that they planned to donate money to Greenpeace in the near future, and only 7 subjects actually made a

positive donation. Only one of them had not chosen the "green" job before. The average donation amongst those who made a positive donation was 1.16 €. This alleviates the potential concern that subjects might choose the "standard" job in order to earn more money and donate part of it afterwards.

1.5 Conclusion

We study how wages and identification with an organization's goals influences self-selection into jobs. To the best of our knowledge, we are the first to experimentally assess the sorting behavior in a context where individuals' identity regarding the job and skill matters. While "standard" (non-behavioral) theory tells us that an employer can attract better workers by paying higher wages, we conjecture that the opposite may be true in some important contexts. Specifically, we show that paying low wages might be beneficial for mission-oriented organizations which aim to attract employees that identify with the organization's goals: Hereby, the organization will attract those applicants that genuinely care about the organization, but less of those applicants that apply because of the high wage. Furthermore, the organization safes money by paying lower wages without incurring any negative effects in terms of workers' output. We describe a model of this selection mechanism and test its predictions in a laboratory experiment in which subjects can choose between two alternative real-effort jobs: a "green" job which pays a lower wage and in which effort provision generates donations to an environmental organization as well as a "standard" job which pays a higher wage but generates no extra benefits to said organization. Using a between-subject design, we vary the wage in the "green" job to analyze how the size of the wage gap between the two jobs affects the type and behavior of the individuals choosing the "green" job. We observe that increasing the wage gap — or, put differently, lowering the wage in the "green" job — reduces the number of subjects choosing the "green" job. We find that the individuals opting for the "green" job in the low-wage treatment have higher degrees of identification with the environmental organization than the ones who do so when there is no wage gap. Moreover, the fact that output does not vary with the

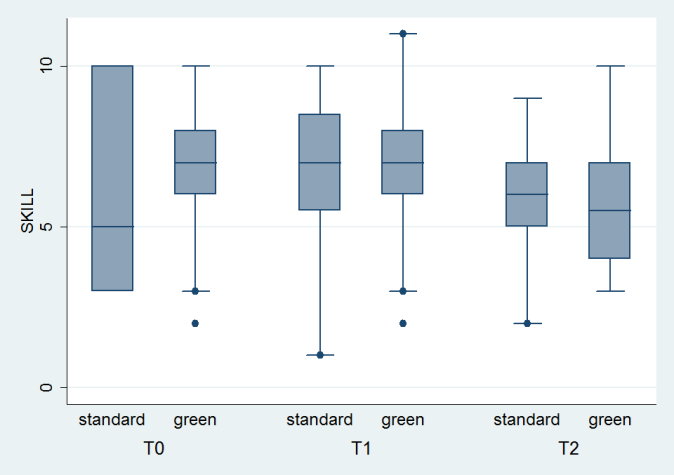
wage gap implies that it can pay off for mission-oriented organizations to underpay their employees relative to the relevant market wage. Ability does not influence the job decision. Interestingly, highly motivated agents do not exert more effort when choosing the "green" job. This is broadly consistent with empirical evidence against pure (or output-oriented) altruism as put forth by Tonin and Vlassopoulos (2010).

One interesting avenue for future research is the relationship between identity and image concerns. In our setting, individuals act isolated from any community and their actions are not visible to any outsiders. Evidence from social psychology and economics (Benabou and Tirole 2003, Benabou and Tirole 2006, Kosfeld and Neckermann 2011, Ariely, Bracha, and Meier 2009 and Linardi and McConnell 2011) suggests that social recognition is an important factor in understanding pro-social behavior. A further extension of our current design could shed some light on the interplay of intrinsic motivation, monetary incentives and social recognition on the selection and the effort provision of intrinsically motivated workers.

1.6 Appendix

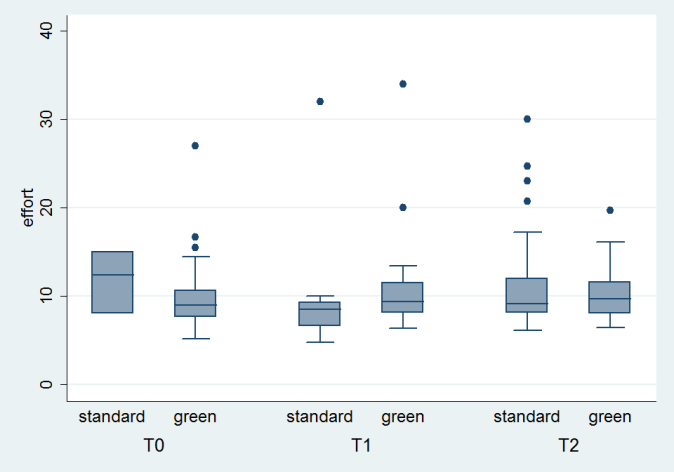
1.6.1 Tables and Figures

Figure 1.6.1: Distribution of workers' skills: comparison by job choice and treatments



This figure shows box plots of workers' skill (i.e. the number of correctly counted tables in the training phase) for each combination of job choice and treatments. We see that the distributions are independent of job choice: For each treatment, the distribution of skills is very similar in the "green" and in the "standard" job. The only variation is that the skill distribution in T2 seems to be shifted down a little bit. Since assignment to treatments occurred after the skill measure is taken, this cannot be a causal effect and thus has to be interpreted as randomization effect of finite numbers.

Figure 1.6.2: Workers "effort": comparison by job choice and treatments



This figure shows box plots of workers' "effort" for each combination of job choice and treatments. "Effort" is defined as "skill-normalized output", it is the ratio of the number of correctly counted tables in the Treatment Phase (40 minutes) and number of correctly counted tables in the Training Phase (5 minutes). There does not seems to be any difference across the six categories.

Table 1.6.1: Treatment overview and Descriptive Statistics

	Treatment 0		**Treatment 1**		**Treatment 2**	
	"standard" job	**"green" job**	**"standard" job**	**"green" job**	**"standard" job**	**"green" job**
Fixed wage	8 €	8 €	8 €	7 €	8 €	5 €
Private Piece-rate	0.10 €	0.10 €	0.10 €	0.10 €	0.10 €	0.10 €
Piece-rate for Greenpeace	0 €	0.10 €	0 €	0.10 €	0 €	0.10 €
Applicants (share)	6%	94%	26%	74%	79%	21%
Applicants (absolute numbers)	3	45	12	35	38	10
Average GP identification	2.66 (2.51)	5.95 (2.61)	3.41 (2.53)	7.02 (2.03)	5.60 (2.80)	7.70 (1.80)
Average skill	6.00 (3.60)	6.91 (1.81)	6.83 (2.59)	6.71 (2.20)	5.92 (1.81)	5.90 (2.23)
Average output (treatment phase)	62.3 (17.5)	63.4 (18.0)	56.0 (19.4)	63.6 (15.8)	59.2 (13.8)	58.6 (18.7)
Average normalized effort[a]	11.80 (3.53)	9.72 (3.72)	9.85 (7.14)	10.59 (5.16)	11.16 (5.39)	10.73 (4.18)

Standard Deviations in parentheses.

[a]Ratio of the numbers of tables counted correctly in the treatment phase (40 minutes) and in the training phase (5 minutes, "skill")

1.6.2 Factor Analysis

We use the answers to the following questions - as well as the question about identification - to construct a measure of "greenness":

"On a scale from 0 to 10, how worried are you about:

- *nuclear power stations and nuclear waste?*
- *air pollution?*
- *damage of the ozone layer?*
- *pollutants in food?*
- *climate change?*
- *species extinction?"*

The answers to these questions are all highly correlated with each other and with the original measure of identification with Greenpeace, as Table 1.6.2 shows.

Table 1.6.2: Inter-item correlation of environmental questions

Variable	identification	nuclear	air	ozone	food	climate	species
identification	1.000						
nuclear	0.602	1.000					
air pollution	0.553	0.806	1.000				
ozone	0.534	0.765	0.843	1.000			
food pollution	0.279	0.391	0.405	0.401	1.000		
climate change	0.600	0.657	0.684	0.758	0.473	1.000	
species	0.620	0.488	0.546	0.533	0.420	0.602	1.000

This table presents the inter-item correlations of the identification with Greenpeace and the answers to the six questions about environmental concerns.

We conducted a factor analysis that suggests that one factor should be retained (eigenvalue of 4.43) which we interpret as "greenness". See the Table 1.6.3 for details for the factor analysis.

This factor is highly correlated with the original measure for identification with Greenpeace. For the rest of the appendix, we standardize the factor to have the same mean and variance as the original 11-point identification measure we use. We conduct the same analyses as in the results section of the paper yielding essentially

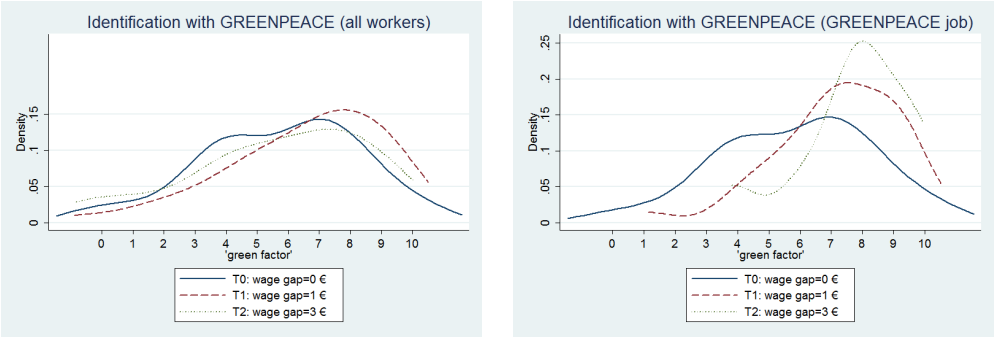
Table 1.6.3: Factor Analysis of the environmental questions

Variable	Factor1	Factor2	Factor3	Factor4	Uniqueness
gp	0.6986	0.2446	-0.208	-0.0046	0.4089
nuclear	0.8378	-0.1731	-0.1115	0.0195	0.2553
air pollution	0.877	-0.2254	-0.0265	0.0118	0.1793
ozone	0.8783	-0.2137	0.0589	-0.024	0.1789
food pollution	0.4945	0.1176	0.2295	0.0229	0.6884
climate change	0.8306	0.0972	0.1167	-0.0198	0.2867
species	0.6872	0.3211	-0.0004	0.0039	0.4247

the same results. The statistical tests we conducted to test the hypothesis whether the mean identification of the workers choosing the "green" job is the same in the wage gap and the no-wage gap treatments now reject the null hypothesis at much lower p-values (p=0.0017). Next, consider the analysis of job choice. As can be seen in Table 1.4.1, there is virtually no difference if we replace our identification measure in the probit regression with the "green" factor from the factor analysis. The different point estimates stem from the normalization of the factor variable.

Figure 1.6.3 shows the distribution of the factor in the whole sample and the sample choosing the "green" job (split by treatment). We replicate the result we obtained using the 11-point scale: The analysis using the "greenness" factor shows the robustness of the original measure of identification with Greenpeace that we use.

Figure 1.6.3: Self-Selection: Comparison of the “green” factor across treatments.



This diagrams are equivalent to Figure 1.4.4, but use the “green” factor instead of the standard measure of identification with Greenpeace. These two diagrams visualize the self-selection of workers based on their “greenness” across treatments. The three lines are the Gaussian kernel densities of the “green” factor constructed from the six questions about the attitude towards environmental issues. The solid line represents the distribution of subjects in Treatment T0 (zero wage gap), the dashed line of those in Treatment T1 (wage gap 1 €) and the dotted line of those in Treatment T2 (wage gap 3 €). The left diagrams displays the distribution for all subjects of the experiment, the right one only of those choosing the “green” job. While the left diagram shows that the populations are comparable across treatments, the right diagram shows that a an increasing wage gap leads monotonely to a “greener” composition of applicants to the “green” job as compared to a zero wage gap.

1.6.3 Translation of the Instructions

After Phase 2 (Training), subjects were given printed instructions. This is the translated version of the treatment T2 in which the fixed wage in the "green" job was 5 €.

You have the choice between two jobs, "Job A" and "Job B". Both jobs are identical to the task of the "counting ones" that you are already familiar with; the payment however will be different. The working time will be 40 minutes for both jobs. After this time there will be a short survey (about 5 minutes).
While "Job A" offers a higher personal payoff than "Job B", the latter one gives you the opportunity to generate a donation to GREENPACE Germany.

Job A:
You get a fixed payment of 8 € regardless of the number of correct answers. In addition, you get 10 cents for each correct answer. GREENPEACE does NOT receive any money when you choose this job.
The following examples illustrate the payment scheme for "Job A":

- If you correctly count 100 tables during the 40 minutes, you receive a payoff of 8 € + 100*0.10 € = 18 €.
- If you correctly count 40 tables during the 40 minutes, you receive a payoff of 8 € + 40*0.10 € = 12 €.
- If you correctly count 70 tables during the 40 minutes, you receive a payoff of 8 € + 70*0.10 € = 15 €.

Job B:
You get a fixed payment of 5 € regardless of the number of correct answers. In addition, you get 10 cents for each correct answer. Furthermore, the experimenters also pay 10 cents for each correct answer as a donation to Greenpeace Germany. Please note that the donation which you generate in this way is not deducted from your personal payoff, but is paid in addition. After the experiment, you will get a

receipt for the donation, which enables you to verify that the money will actually be donated to Greenpeace.

The following examples illustrate the payment scheme for “Job B”:

- If you correctly count 100 tables during the 40 minutes, you receive a payoff of 5 € + 100*0.1 € = 15 €.

  In addition, GREENPEACE receives a donation of 100*0.10 €= 10 €.

- If you correctly count 40 tables during the 40 minutes, you receive a payoff of 5 € + 40*0.1 € = 9 €.

  In addition, GREENPEACE receives a donation of 40*0.10 €= 4 €.

- If you correctly count 70 tables during the 40 minutes, you receive a payoff of 5 € + 70*0.1 € = 12 €.

  In addition, GREENPEACE receives a donation of 70*0.10 €= 7 €.

Chapter 2

Severity vs Leniency Errors in Performance Appraisal

ABSTRACT

Supervisors can fail performance appraisal mainly in two ways: with *leniency* errors that assign predominantly high evaluations thus rewarding even undeserving agents that have exerted low effort or with *severity* errors that assign predominantly low evaluations and thus neglecting rewards to deserving agents that have exerted high effort. The basic principal-agent model with moral hazard predicts both errors to be equally detrimental to effort provision. We then show this prediction fails in the lab. In fact, failing to reward deserving agents is significantly more detrimental than rewarding undeserving agents. We discuss our result in the light of different economic theories of behavior. Our result may have interesting implications for strategic human resource management and personnel economics and may also contribute to the debate about incentives and organizational performance.

Keywords: Agency theory, Type-I and Type-II errors, Real effort, Rater error, Leniency errors, Severity errors, Performance appraisal.

JEL code: C91, M50, J50.

2.1 Introduction

Supervisors manipulate rewards for performance as a means of incentivizing effort. High yet achievable goals, remunerated with non-negligible rewards, should induce rational agents to exert effort. However agent's actual effort and supervisor's (we will use the synonyms *supervisor*, *rater*, and *principal* interchangeability) performance appraisal can be misaligned. This may happen both i) because effort is only stochastically related with performance and ii) because rater's evaluation is subjective. In both cases imperfect performance appraisal can lead to two types of error:

- A supervisor (she) may assess low performance when in fact the agent (he) is duly exerting high effort and thus she does not reward a deserving agent. This is a Type-I error[1].
- A supervisor may observe high performance when in fact the agent is not exerting high effort. Therefore she may reward the undeserving agent. This is a Type-II error.

Whenever actual agent's effort and supervisor's performance assessment are misaligned, any reward system necessarily produces a certain balance of Type-I and Type-II errors. The rater can affect this balance by – for instance – setting different performance goals that the agent must match in order to obtain the reward. A *lenient* goal implies a high probability that an agent exerting low effort is nevertheless rewarded (Type-II error). This clearly demotivates the agent from exerting effort. On the other hand, a *severe* goal implies a high probability that an agent willing to exert effort is nevertheless not rewarded (Type-I error). Both errors are thus detrimental to effort provision.

The challenge for the supervisor is to tune the reward scheme so that the trade-off between the two error types is optimized. This depends – *inter alia* – upon the

[1] In an ideal contract with perfect monitoring the agent should receive a high remuneration whenever he exerts high effort. The agent's compliance with the prescribed behavior thus can be interpreted as the null hypothesis, so that the rater can both incorrectly reject the null and not reward a deserving agent (a Type-I error) and incorrectly accept the null and reward an undeserving agent (Type-II error).

elasticities to each type of error of individuals' willingness to exert effort. Some examples may illustrate the pervasiveness of the problem.

A **quality control manager** samples the production of a group of workers. The number of defective products depends on i) machinery random errors and ii) workers' effort. Workers are paid a premium if the number of defective products sampled for each of them is below a certain threshold. The manager's problem is where to set such threshold: if the threshold is too low (*lenient*) it produces many Type-II errors so that undeserving agents will get the reward. If it is too high (*severe*) it produces many Type-I errors so that deserving agents will not get the reward. In both cases workers' willingness to exert high effort are weakened. Are workers more demotivated by severe or lenient production targets?

A **board of directors** sets objective goals (revenues, profits, share prices etc.) for the firm's CEO for bonus compensation. Firm's performance depends both on CEO's own effort and on external factors such as the business cycle and regulation. CEO's ability and firm's performance are therefore only weakly related: in a given year the firm may produce disappointing performance notwithstanding CEO's effort (Type-I error) or may produce good performance in spite of CEO's lazy conduct (Type-II error) . The board must find the correct balance between setting *lenient* goals that do not challenge the CEO enough and *severe* goals that discourage him.

A **firm** sets up a subjective performance appraisal system for its employees. In training two managers to become the firm's rater it discovers that both are affected by systematic biases that skew the distribution of rating or grading. In particular one manager tends to assign predominantly high ratings committing thus leniency errors while the other has the tendency to deliver low ratings to the same individuals. Both raters demotivate the employees: *lenient* appraisals induce employees to lower their effort while **severe** targets discourage. The firm must decide which manager to put in charge of the system. Is it more beneficial to put in charge the lenient or the severe one?

Evaluation errors are a problem in other areas as well.

A **school teacher** wants to motivate her students to study hard for the final exam by showing them some final assessment tests. She knows however that if she shows them a *lenient* test, students will underestimate the challenge and think they can pass the exam with little effort while if she shows them a *severe* test they might be discouraged to exert high effort fearing that it might not be enough.

The **social planner** must deter crimes by setting up a good criminal procedure. He knows that if she sets the burden of evidence too high, then the procedure is *lenient* as judges will not be able to prove easily guilt of culpable individuals. Therefore individuals may find crime convenient. On the other hand, if she sets the burden of evidence too low, then the procedure is *severe* as judges may easily prove guilt also for innocents. In this case individuals may find that honesty does not pay.

A **parent** wants his child to be good at running. In order to motivate him she promises a nice Christmas present for achieving certain goals. She knows that if she sets a *lenient* goal the child will not train hard because the goal can be easily reached. On the other hand if the goal is too *severe* the child will be discouraged to train.

All these situations point to a common problem which is the research object of the paper: supervisor's activity is prone to both Type-I and Type-II errors and both are detrimental to agent's performance. In all cases she controls the error's trade-off and therefore it becomes crucial to assess how bad one error is compared to the other.

Knowing the actual marginal costs of each error in terms of agent's effort provision, we in turn have a sense of the optimal error trade-off.

To provide informed recommendations on the topic, in this paper we formalized a simplified principal-agent model where both Type-I and Type-II in performance appraisal are considered. The main theoretical prediction delivered by the model states that both error types should be treated equally as they both jeopardize agent's effort performance by the same token. To test the main theoretical prediction, we devised a real effort laboratory experiment to study how agents behave facing the different type of evaluation errors. Our main finding shows that there is a substantial gulf between the theoretical predictions and the empirical laboratory evidence. In

particular failing to reward a deserving agent (Type-I error) is significantly more detrimental to effort provision than rewarding an undeserving agent (Type-II error).

To the best of our knowledge, this is the first experimental exercise where the effects of both Type-I and Type-II errors on agents' effort provision are studied. We believe that this could shed more light on the incentive-effort-performance schema, and contribute both theoretically and to managerial practice.

The paper is organized as following: Section two provides a review of the the related literature. Section three introduces a simplified principal-agent model where both Type-I and Type-II errors are considered. Section four describes the both the experimental design and the procedures adopted to test the theoretical predictions delivered by the model. Data analysis is discussed in section five. In section six, we discuss the experimental findings in light of some behavioral theories. Section seven concludes.

2.2 Literature Review

Several streams of literature, including personnel economics (Lazear, 1999), agency theory (Hölmstrom, 1979a; Aron and Olivella, 1994; Prendergast, 1999) human resources management and organizational studies (Steers, Mowday, and Shapiro, 2004) deal with errors in performance appraisal. In a principal-agent relation, when only an objective evaluation of performance is feasible (output is observable), evaluation errors arise because i) agent's effort is non observable and ii) agent's effort provision and observable performance are stochastically related.

However, organizations where performance is directly observable are rare (Gibbons, 1998; Prendergast, 1999) and the "fascination with an 'objective' criterion, [where] individuals seek to establish simple, quantifiable standards against which to measure and reward performance" leads many pay-for-performance schemes to establish severely distorting incentives (Kerr, 1975). The problem of "rewarding A while hoping for B" (Kerr, 1975; Baker, Gibbons, and Murphy, 1994) can only be partly mitigated by adding more indirect information on agent's effort (Hölmstrom, 1979b) as very often the problem lies in defining exactly what the principal's

objective is (Baker, 1992; Jensen and Meckling, 1976).

Most organizations then rely on subjective performance appraisal in order to motivate their employees (Prendergast and Topel, 1993; Prendergast, 1999; MacLeod, 2003; Kambe, 2006). Compared to firm owners, managers usually have private information concerning overall performance of their subordinates (MacLeod, 2003; Thiele, 2011). Subjective performance measures can be used alone (Bull, 1987; MacLeod and Malcomson, 1989; Levin, 2003) or in combination with objective measures (Schmidt and Schnitzer, 1995; Pearce and Stacchetti, 1998).

In any case, subjective performance evaluation is also prone to errors.

The two most important rater's errors are classified by the literature with *leniency errors* and *severity errors*[2]

Leniency errors occur when the rater assigns predominantly high ratings on a given scale. Thus in our terminology leniency errors are situations where type-II errors are relatively abundant.

Severity errors, that is to say situations where type-I errors are relatively abundant, occur when the supervisor assigns predominantly low ratings.

These errors reduce the scope of appraisal because they restrict the range of useful measures of performance, and thus weaken the incentive (MacLeod, 2003).

Kane (1994) distinguishes between i) *nonvolitional* systematic errors and (ii) *volitional* systematic errors. Along the same lines Prendergast (2002) distinguishes between i) biases based on personal feelings and ii) biases based on personal returns. Among the first group there are errors arising from unconscious cognitive and behavioral biased in observing, elaborating, recalling of ratee performance information or in the process of generating the appraisal rating. Also feelings such as empathy and affection play an important role (Cardy and Dobbins, 1986; Varma, Denisi, and Peters, 1996) and the rater's assessment can also be manipulated by the

[2]Other rater's errors are: (i)*central tendency* error derives from a propensity to avoid assigning extreme values ; (ii) *halo effect* refers to raters judgment on one scale influencing ratings on other scales; (iii) *contamination errors* affect the construct validity of ratings by relying on irrelevant information; (iv) *similar-to-me error* occurs when ratings are influenced because the ratee has affinity with the rater; (v) *recency errors* happens when recent performance is given too much weight as opposed to early performance within a given time interval and on the opposite (vi) *first impression error* when early performance is given too much weight as opposed to more recent performance within a given time interval (See Thomas and Meeke 2010 on classification of rater's errors. See Rabin and Schrag 1999 specifically on first impression bias).

ratee (Higgins, Judge, and Ferris, 2003). Among the second group there are intentional distortions of appraisals done in order to serve rater's goals. For instance, if the principal is the residual claimant on the agents' production and the agents' pay is based on principal's subjective appraisal, she may under report the performance of her subordinates in order to save costs. This would amount to a volitional systematic severity error. On the other hand many raters are no residual claimants but themselves part of hierarchies and therefore their utility functions may deviate from the principal's objectives. In particular a supervisor may find it convenient to provide lenient evaluations because she colludes with the agent (See Tirole, 1986; Prendergast and Topel, 1996; Strausz, 1997; Vafaï, 2010; Thiele, 2011) or because of more complex motivation (Judge and Ferris, 1993; Grund and Przemeck, 2012; Giebe and Guertler, 2012). This paper complements this vein of literature that focuses on supervisors' behaviour as we focus our attention on agents' behaviour when exposed to *leniency* environments(setups with high levels of Type-II errors) or *severity* situations (where instead Type-I errors abound).

2.3 Theoretical Setup

In the following paragraphs we model a simple relation between a supervisor and an agent where the supervisor can not contract the agent's effort e_i and only the final binary output of the project p_i (failure / success) can be monitored.

The supervisor thus can evaluate performance by observing the outcome achieved by the agent and rewarding him accordingly.

Agent's Disutility of Effort. Let e be a measure of effort. The agent's choice is binary: either he invests a low level of effort (e_L) or a high level (e_H) such that the set of possible actions is $(e_L, e_H) \in A$. Effort implies disutility for the agent. Following the literature, we define this disutility as a generic function of the level of effort $c(e)$ such that $c(e_H) > c(e_L)$, meaning that higher effort creates more disutility. Each agent has his own disutility of effort for both the low and high level and therefore it is associated with a particular level of $c(e_H)$ and $c(e_L)$. In this model, effort can take up only two values, therefore we can safely assume

that $c(e_L) = e_L$ and $c(e_H) = e_H$ (see also Mas-Colell, Whinston, and Green, 1995; Bolton and Dewatripont, 2005). We assume that each agent suffers the same disutility of effort e_L and e_H. Although this is a simplifying assumption, it does not affect the results of the model[3].

Agent's utility. The agent is an expected utility maximizer with a utility function $u_i(w, e) = v_i(w) - e$ where w is the wage and can take the following two values (w_0, w_r) (w_0 is the baseline wage and w_r is the rewarding wage) and where $e \in (e_L, e_H)$. The utility function is separable in monetary utility and disutility of effort following the usual assumptions of concavity for the former and convexity for the latter. Note that $v_i(w_r) - v_i(w_0)$ is the net reward. Each agent is uniquely associated with a value $\Delta_i = v_i(w_r) - v_i(w_0)$ that simply represents his monetary utility of the net reward.

Supervisor's profit. Let p_i denote the agent's performance in terms of project realization as observed by the supervisor. The supervisor is risk-neutral and gains the project's output less any wage payment made to the agent: $p_i - w$ with $w \in (w_0, w_r)$ and with $p_i - w_r > 0$ so that the supervisor's profits are always positive.

2.3.1 Performance and Error Trade-off

The agent's effort e is not observable, hence it is only stochastically related to p_i. The supervisor decides the project target $\overline{p}$ which triggers the rewarding wage w_r. Notice that the supervisor may commit the following evaluation errors:

- Type-I error: with probability α, the agent that has spent a high level of effort does not meet the project target and thus he is not rewarded;
- Type-II error: with probability β, the agent that has invested a low level of effort nevertheless meets the project target and thus he is undeservedly rewarded.

Every agent has a different utility of the monetary reward Δ_i and therefore, for some levels of $\bar{p}$, e_L,e_H, α and β, there will be only a certain proportion of agents

[3]Heterogeneity among agents is preserved in the utility of income. As the relevant variable for us is given by the difference in the utility of income and the disutility of effort, we can simplify e_H and e_L to be the same for all agents.

willing to exert high effort in return for the utility of the monetary reward. It is reasonable to assume that $1 - \alpha > \beta$, that is to say that the probability of correctly rewarding the performance of the high-effort agent is larger than the probability of wrongfully rewarding the performance of the low-effort agent. If this were not the case then the evaluation procedure would be equivalent to or worse than tossing a coin. The derivation of the probabilities of errors from the definition of p and e_L, e_H is outside the scope of the present work. However, it is intuitive to say that the sum of errors $(\alpha + \beta)$ is minimized for some intermediate levels of project target $\bar{p}$. This is because when the performance target is set very low it is very easy to meet the target both with high and with low effort. Therefore with low $\bar{p}$ we have no Type-I errors (i.e. not meeting the target when exerting high effort) and many Type-II errors (i.e. meeting the target when exerting low effort). Therefore there is little incentive for the agent to invest high effort. The more the performance target increases, the smaller the probability of Type-II error becomes and thus switching to high effort becomes convenient. At some intermediate level of $\bar{p}$ we have a few Type-I errors and a few Type-II errors. Finally, when the project target becomes extremely high, the probability of Type-II errors (i.e. meeting the target when exerting low effort) becomes virtually nil but at the same time the probability of Type-I error (i.e. not meeting the target when exerting high effort) is very high and therefore there is little incentive to exert high effort.

2.3.2 Agent's Choice

The payoff for the agent (utility of income less the disutility of effort) of complying with the supervisor's request to exert high effort are thus the following:

$$\alpha \left[(v_i(w_0) - e_H \right] + (1 - \alpha) \left[v_i(w_r) - e_H \right] \tag{2.1}$$

while if low effort is exerted the utility is the following

$$\beta \left[(v_i(w_r) - e_L \right] + (1 - \beta) \left[v_i(w_0) - e_L \right] \tag{2.2}$$

The agent i will comply with the prescribed behavior and choose to invest high

effort if the expected payoff of high effort (Equation 2.1) is higher than the payoff of choosing to exert low effort (Equation 2.2). The equation simplifies as follows: $(v_i(w_r) - v_i(w_0))\,(1 - \alpha - \beta) > e_H - e_L$ which implies the following compliance condition:

$$\bar{\Delta} = \frac{e_H - e_L}{1 - \alpha - \beta} \tag{2.3}$$

Equation 2.3 suggests that for each agent there exists a value $\bar{\Delta}$ for which the agent is indifferent between exerting high effort or low effort. Given that the disutility of effort and the probabilities of effort are exogenously determined, the choice of exerting high effort thus depends on whether each individual $\Delta_i \overset{?}{\gtrless} \bar{\Delta}$.

Note that on the right hand of Equation 2.3 we have the net disutility of effort for the agent i discounted by both Type-I and Type-II errors. Note also that on one hand the larger the probability of β (being rewarded undeservingly), the larger are the returns from not exerting effort. On the other hand, however, the larger the probability of α (not being rewarded when deserving it), the larger are the returns of exerting effort. More formally, given an increase in the probability of Type-II error β, compensated by an equal decrease in the probability of Type-I error α leaves the individual indifferent in choosing whether to exert high effort or not. The same is true when Type-I errors probability (α) increase and Type-II errors probability (β) decrease. We may define the sum of errors $\alpha + \beta$ as the *accuracy* of the evaluation process. Accuracy can be kept constant with very different error trade-offs as long as $\alpha_{low} + \beta_{high} = \alpha_{high} + \beta_{low}$. According to this analysis both error types should be treated equally as they both jeopardize agent's effort performance by the same token.

2.3.3 Experimental Treatments

To test the behavioural implications of the model, we exploit the *accuracy* property to devise three experimental treatments described in Table 2.3.1.

Table 2.3.1: Table of treatment parameters

	α	β	$1-\alpha-\beta$
$T0$ **- Just**	0	0	1
$T1$ **- Severe**	$4/5$	0	$1/5$
$T2$ **- Lenient**	0	$4/5$	$1/5$

Just Treatment ($T0$). In T0 there are no evaluation errors ($\alpha, \beta = 0$) and thus the expected returns are $v(w_r)$ and $v(w_0)$ for exerting high (e_H) and low effort (e_L) respectively. The expected returns of exerting high effort are thus $(v(w_r) - v(w_0))$. In this treatment the choice of exerting high effort is always rewarded while the choice of exerting low effort never is. This treatment is "just" in the sense that the agent gets what he deserves.

Severe Treatment ($T1$). In T1 there are no Type-II errors ($\beta = 0$) but the probability of Type-I error is significant ($\alpha = 0.8$).[4] Given the high number of errors, the net returns from exerting high effort are smaller but still positive $\frac{(v(w_r)-v(w_0))}{5}$. In this treatment the high-effort choice is seldom rewarded while the choice of low effort is never rewarded. This treatment is "severe" in the sense that the deserving agent very often does not get what he deserves.

Lenient Treatment ($T2$). In T2 there is a significant probability of Type-II error ($\beta = 0.8$) but there are no Type-I errors ($\alpha = 0$). Given the high number of errors, the net returns from exerting high effort are smaller but still positive $\frac{(v(w_r)-v(w_0))}{5}$. In this treatment the high-effort choice is always rewarded while the choice of low effort also is very often rewarded. This treatment is "lenient" in the sense that the undeserving agent very often gets what he does not deserve.

[4]The choice of $\alpha = 0.8$ was made upon considering this probability high enough to be salient and clearly low enough to leave space to the realization of the complementary state-of-the-word.

2.3.4 Testable Predictions

For the purpose of the experimental test we focus in particular on Equation 2.3 which describes the condition under which each agent may switch from low to high effort provision.

Claim 1. Neglecting due rewards (α) decreases agents' effort provision.
Equation 2.3 shows that Type-I errors are detrimental to effort provision w.r.t. an error free scenario. By increasing the probability of Type-I error, the model predicts that the share of agents exerting high effort should diminish. This prediction can be tested by contrasting the share of agents exerting high effort in $T0$ and $T1$.

Claim 2. Rewarding undeserving agents (β) decreases agents' effort provision.
Equation 2.3 shows that Type-II errors also are detrimental to effort provision compared to an error free scenario. By increasing the probability of Type-II error, the model predicts that the share of agents exerting high effort should decrease. This prediction can be tested by contrasting at the shares of agents exerting high effort in $T0$ and $T2$.

Claim 3. Type-I and Type-II errors are equally detrimental to agents' effort provision.
A given increase in the probability of Type-II error β, compensated by an equal decrease in the probability of Type-I error α and viceversa (*accuracy* kept constant), leaves the individual indifferent in choosing whether to exert high effort or not. In order to test this prediction we compare the share of agents exerting high effort in $T1$ and $T2$.

2.4 The Experiment

The use of a lab experiment to test our theoretical predictions provides several important advantages (Falk and Heckman, 2009b; Charness and Kuhn, 2010) in comparison with observational datasets that are typically used in labor/personnel

economics or managerial case studies; above all the opportunity to control for all the crucial variables of the economic environment and the possibility to vary *ad hoc* the precise variables of interest (Falk and Fehr, 2003b; Falk and Gächter, 2008b). On the other hand the external validity of lab findings can be questioned (Gneezy and List, 2006). However, the research question of the present work deals with a variable - evaluation errors - that is basically impossible to observe in the field because of the unobservability of effort and the stochastic relation between performance and effort. In the lab instead we can superimpose an exogenous probability of error in evaluating performance and at the same time we can perfectly observe effort. This ideally allows us to identify precisely the impact of errors on effort provision and thus on performance.

2.4.1 Experimental Design

The experimental design is made up of three phases: the preliminary Phase I is used to elicit individuals' risk attitudes via a standard incentivized choice of lotteries (Holt and Laury 2002) . This is followed by Phase II, where individual productivity in the default task is measured, and then there is Phase III, where individuals have to carry out the task facing the different evaluation errors. This phase is our actual main treatment phase. Exploiting a within-subject design, effort actions are elicited under all three treatments : $T0, T1, T2$ (see Table 2.3.1) then just one of the tree scenario and related agent's action are actually implemented and considered to determine the payments.

The three main treatments are featured by two different configurations. This is to check whether different levels of initial endowment could play a role in determine systematic different perceptions of the evaluation errors. In the first endowment configuration, labeled as *low* endowment configuration, the agent has no initial endowment. The reward amount, linked to the eventual evaluation errors, represents the main portion of her final revenue. In the second endowment configuration, labeled as *high* endowment configuration, the agent receives by default an initial endowment. In this second case, the reward amount linked to the eventual

evaluation errors still represents an important portion of the agent's final pay-off, nevertheless under this configuration he can rely on a quite satisfactory minimum outcome.

The experiment adopts a within-subject design (which in its actual implementation is very close to a strategy-method elicitation mechanism) given the two alternative treatments configurations (*high* vs. *low*). To control for any possible ordering effect, the order in which subjects are asked to make their choices under treatments T1 and T2 are randomized across the different experimental sessions[5]. Feedback information, on the outcomes of the lotteries in Phase I and on whether the supervisor-automaton makes an evaluation error in the implemented scenario, are provided at the end of the experimental session. This is to assure full independence of the different treatment phases, free of historical contagion and therefore statistically independent.

Between each of the phases, subjects have the opportunity to rest. Common instructions for the subsequent phase are read and described aloud while instructions concerning each single treatment are delivered on screen. Control questions for each of the different phases and treatments are administered through the computer.

The experiment was programmed and conducted with z-Tree (Fischbacher, 2007a) and sessions took place at the Einaudi Institute of Economics and Finance in Rome on April 6, April 8, April 14 and May 2, 2011. We ran a total of four sessions with 84 participants. Subjects were recruited online with ORSEE (Greiner, 2004). Mann-Whitney tests indicate that there are no significant differences in the socio-demographic characteristics of the subjects across sessions: mainly undergraduate students with very different backgrounds (humanities, medicine, hard sciences, social sciences). Average age was 22.47 (s.d. 2.16), females 40%, males 60%. Via the strategy method we elicited 84 observations for each treatment. Average payoff was about €10.21.

Phase I - Risk attitude elicitation.

Following Holt and Laury (2002) subjects are asked to carry out a standard series

[5]Treatment T0 is always submitted first as it represents the benchmark case.

of lotteries (see Table 2.8.1 in the Appendix) to measure individual risk-aversion. Outcomes of the lotteries are communicated to subjects only at the end of the experiment.

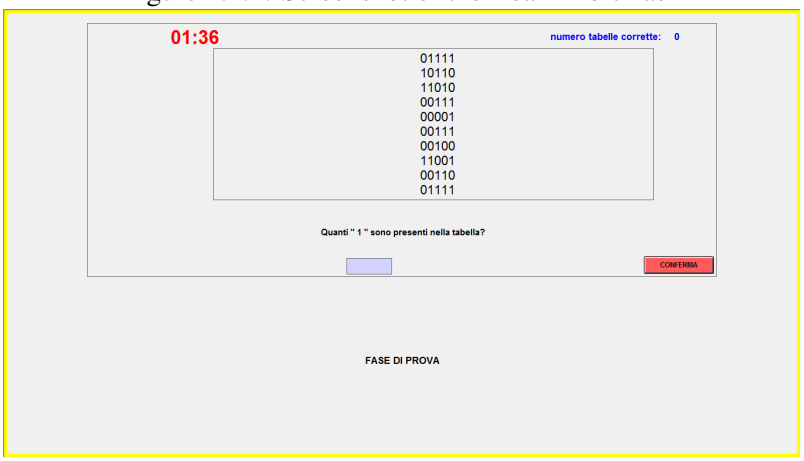
Phase II - High effort productivity elicitation

Following Abeler, Falk, Götte, and Huffman (2011), the work of the real effort task consists in counting the number of occurrences of the digit *1* in as many tables as possible, where each table is composed of 50 digits and among these the number of *1*s is randomly generated (see Figure 2.4.1). This task has several advantages: it does not require any prior specific knowledge; performance is objective and easily measurable; and there is little room for learning effects. At the same time, the task is boring and pointless at least for most of the subjects and thus it can be claimed that the task entails a positive cost of effort. The task is also clearly artificial, and output does not provide intrinsic or extrinsic value to the experimenter. This should rule out any tendency for subjects to use effort provision during the experiment as a way to reciprocate for incentives provided by the experimenter or the possibility that subjects carry out the task for some intrinsic motivation. To elicit the individual productivity, subjects are offered a pure piece-rate compensation scheme. They receive a €0.03 payment for each table correctly processed and 0 for each table incorrectly processed. Furthermore both a countdown timer and a counter reporting the number of tables processed are provided. After 10 minutes subjects receive a summary statistical report concerning the number of tables correctly processed $(q_{phase-II})$, the number of tables incorrectly processed and the total amount of money generated in this phase . The average score is 45.9 (s.d. 11.5) correctly counted tables[6]. The number of tables counted in this stage, featured by a pure piece-rate regime, represents a reliable measure of the individual specific productivity/ability. In the following session, and under different treatment conditions (T0, T1 and T2) this measure contributes to define the individual target. By scaling

[6]Despite the fact the piece-rate had a low magnitude, agents worked intensively. The small piece rate aims to provide a clear incentive to exert effort in this phase,at the same time – considering the flow of the whole experimental protocol – it is important to do not provide high payments in this preliminary phase to prevent distortions in the two different treatments configurations adopted in phase III.

the individual specific target on individual ability, we roughly normalize the individual cost of effort for the task across subjects. In other words, individuals with different abilities count different numbers of tables in Phase II. However, by setting individual targets of Phase III proportional on these numbers, the costs of effort for reaching the respective targets should be roughly equivalent.

Figure 2.4.1: Screenshot of the Real Effort Task



Phase III - Experimental treatments.

Since the perception of the project feasibility/success could be highly subjective and vary significantly between subjects – for instance due to differences in overconfidence – we exogenously impose an objective failure/success probability in term of the supervisor's error probability of correctly observing the actual project outcome.

Phase III is 40 minutes long, four times the length of Phase II. The task consists in correctly processing 90% of four times the $q_{phase-II}$ measured for each subject in Phase II ($\bar{q} = q_{phase-II} \times 4 \times 0.9$). The 10% discount is justified by the higher fatigue created by the longer task and at the same time it signals that the task is feasible by exerting a high but not extraordinarily high level of effort. In addition, this allows to get rid of any uncertainty concern related to actual feasibility of the target[7]. We implemented the 3 treatments (T0, T1 and T2) featured by two differ-

[7]This is confirmed by the data. All subjects except one that decide to exert high effort choosing

ent endowment configurations (*low* and *high*): $T0_{low}$, $T1_{low}$, $T0_{low}$ and $T0_{high}$, $T1_{high}$, $T2_{high}$.

Absent any evaluation error (e.g. in T0) the accomplishment of the task is rewarded with €6.60 [8] on top of the initial endowment that characterized the two alternative treatment configurations: €0 in the *low* endowment configuration, €5.28[9] in the *high* endowment configuration.

Table 2.4.1: Experimental Sessions

Session	Conf.	Date	Treatment Order	Endowment €	Reward €	Subjects
A	Low	April 4	T0,T1,T2	0	6.60	23
B	Low	April 6	T0,T2,T1	0	6.60	17
C	$High$	April 8	T0,T1,T2	5.28	11.88	23
D	$High$	May 2	T0,T2,T1	5.28	11.88	21

Subjects were presented with the following text common to all treatments.

> In Phase III you have to take a decision. You can decide (i) to undertake the task and meet the target of counting, according to your capacity, a feasible number of tables in 40 minutes time ; (ii) to skip the task and proceed immediately to the payment phase and leave the lab or to undertake the task and fail to purse the target. You may be rewarded with a payment of €6.60 for undertaking and accomplish the task. The assignment of the reward is subject to errors. Under different situations you might be subject to, the supervisor might provide you the payment of €6.60 when you do not undertake or do not accomplish the task properly , conversely it might deny the payment of

to carry out the duty eventually match the performance target

[8] This amount is proportional to a hourly wage of €10.

[9] The unrounded amount of €5.28 aims simply to do not provide any particular salient signal or reference.

€6.60 when you actually duly accomplish the task.

Then we present the three treatments randomized across sessions avoid to any ordering effect. The text is tailored to each treatment *(T0) [T1] {T2}* and configuration $\underline{Low}$ / $\overline{High}$ as follows:

`(i - treatment configuration / initial endowment)`

In this phase you receive a basic compensation of €$\underline{0}$ / €$\overline{5.28}$.

Your task consists in processing <*number* $\bar{q}$> tables in 40 minutes and you are monitored through an automaton-supervisor. The supervisor *(will not) [may] {may}* commit an evaluation error:

`(ii - high effort action)`

If you undertake and accomplish the task meeting the target, you will *(certainly receive) [not receive with a 80% probability] {certainly receive}* the €6.60 reward payment.

`(iii - low effort action)`

If you do not undertake it or if you undertake it but you do not meet the target, you will *(certainly not receive) [certainly not receive] {receive with a 80% probability}* the €6.60 reward payment.

`(iv - selection of the action)`

Please, make your choice:

[A – I will perform the task] / [B – I will skip the task]

Each subject is asked to state her choice (*A* or *B*) for each of the three possible scenarios characterized by the treatments. However, after the three choices, only one scenario is randomly selected[10] and its parameters applied.

[10]The design is thus a within-subject as we are able to observe the variations of subjects' effort choice across the three treatments and it implements the strategy method as the choices are elicited before one is randomly chosen and implemented. This procedure avoids income effects and also rules out any potential order effect of subjects' choice being influenced by previous decisions.

To ensure a truthful revelation, the subject is informed about which scenario is actually randomly implemented only after she states her decisions for all 3 scenarios. The design ensures that the subject makes truthful choices for the three scenarios. This is because the choices imply real consequences: if the subject chooses *A* then she must spend 40 minutes in the lab anyway before progressing to the questionnaire and payment phase and if she chooses *B* then the real effort task is skipped entirely. Therefore the subject has no reason to misrepresent her true preferences.

If a subject decides not to perform the task in a given treatment, and this treatment is then randomly implemented, she can proceed immediately to the next step - filling the questionnaire in - and then she is paid, viceversa, she has 40 minutes available to carry out the task.

2.5 Results

In order to test the theoretical predictions and the magnitude of the detrimental effect entailed by the different evaluation errors, we proceed contrasting the shares Z_{Ti} of complying agents when exposed to the different treatments T. According to the theoretical model, we define as complying agents the ones who accepted to exert high effort deciding to carry out the task and accomplish the project meeting the target. In this respect, we only focus on the "extensive margin" of subjects willing to exert high effort. Tuning the target to the 90% of the maximal individual capacity, it assures the feasibility of the goal and allows to get rid of uncertainty concerns related to the actual feasibility of the duty. For this main reason – in this setting – the analysis of the effort "intensive margin" results to be trivial by construction: all subjects except one that decided to exert high effort choosing to carry out the duty eventually matched the target.

Result 1. Neglecting due rewards (α) decrease agents' effort provision

In order to test whether neglecting due rewards decreases agents' effort provision, we contrast the share of complying agents (define as Z) in $T0$ (just treatment with

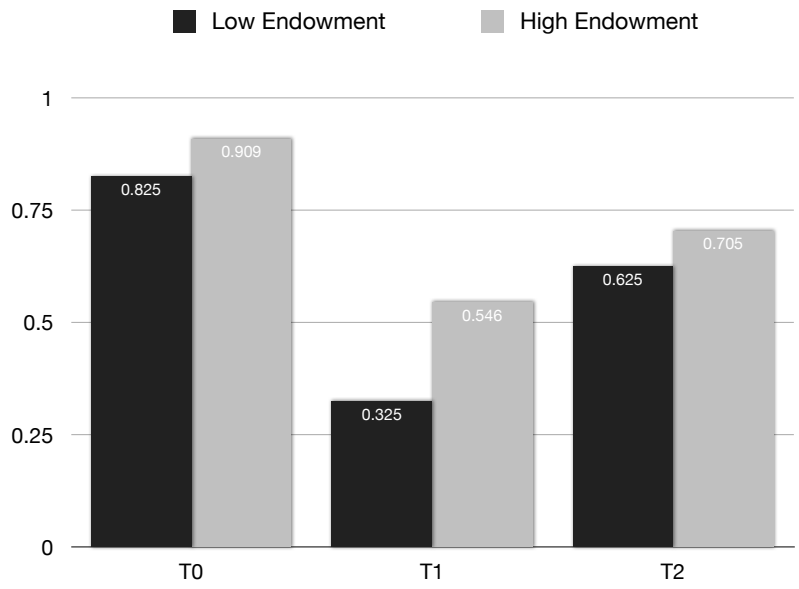


Figure 2.5.1: Percentage of population exerting effort in Sessions A & B and in Sessions C & D

$\alpha = 0, \beta = 0$) and $T1$ (the severe treatment with $\alpha = 0.8, \beta = 0$). We test the following:

$$H_0 : Z_{T0} > Z_{T1} \text{ - vs - } H_1 : Z_{T0} = Z_{T1}$$

Low configuration (A & B): Under perfect monitoring (T0), a share of population equal to 82.5% chooses to exert high effort to accomplish the task while the same share falls to only 32.5% when subjects are exposed to Type-I error (T1). This negative effect is strongly statistically significant at the 0.01% level, on the basis of a two-sided null hypothesis and 40 independent paired observations (McNemar's[11] $\chi^2 = 20, p - value < 0.0001$).

High configuration (C & D): We repeat the same test for the sessions with the high endowment configuration. Under perfect monitoring (T0), a share of popula-

[11]Our within-subject design enables us to observe the choices of subjects under all the different treatment conditions. The McNemar test fits particularly well with our experimental setting since paired-sample tests are used to assess the differences in the population shares of agents exerting high effort under the different treatments. See (Fehr, Falk, and Fischbacher, 2003; Enderer and Manso, 2009; Caplan, Aadland, and Macharia, 2010) Analogous qualitative results on statistical significance in mean differences are replicated adopting a proportions test for differences in proportions.

tion equal to 90.9% chooses to exert effort. When exposed to the Type-I error (T1) only 54.5% of agents exert high effort. This negative effect is strongly statistically significant at the 0.01% level, on the basis of a two-sided null hypothesis and 44 independent paired observations (McNemar's $\chi^2 = 16$, $p - value < 0.0001$).

Result 2. Rewarding undeserving agents (β) decrease agents' effort provision
As before, in order to test whether rewarding undeserving agents decreases agents' effort provision, we compare the share of complying agents in $T0$ (just treatment with $\alpha = 0, \beta = 0$) and $T2$ (lenient treatment with $\beta = 0.8, \alpha = 0$). We test the following:

$$H_0 : Z_{T0} > Z_{T2} \text{ - vs - } H_1 : \ Z_{T0} = Z_{T2}$$

Low configuration (A & B): Facing perfect monitoring (T0), 82.5% of agents are willing to exert high effort achieving the goal while the share decreases to 62.5% when subjects are exposed to a Type-II errors (T2) the scenario. This negative effect is statistically significant at the 5% level, on the basis of a two-sided null hypothesis and 40 independent paired observations (McNemar's $\chi^2 = 5.33$, $p - value = 0.022$).

High configuration (C & D): In the configuration with high initial endowment, the percentage of subjects exerting high effort drops from 90.9% in T0 to 70.5% in T2. This negative effect is statistically significant at the 0.05 % level, on the basis of a two-sided null hypothesis and 44 independent paired observations (McNemar's $\chi^2 = 9$, $p - value = 0.0027$).

Result 3. Neglecting due rewards (Type-I) is more detrimental to agents' effort provision than rewarding undeserving agents (Type-II)
Claim 3 of the model predicts that neglected rewards to complying agents (Type-I error , α) and undeserved rewards to non-complying agents (Type-II error , β) are identically detrimental for effort provision. In order to test this hypothesis we compare the share of complying agents in $T1$ and $T2$. We test the following:

$$H_0 : Z_{T1} = Z_{T2} \text{ - vs - } H_1 : \ Z_{T1} \neq Z_{T2}$$

Low configuration (A & B): Facing a substantial probability of Type-I error (T1), only a share of population equal to 32.5% is willing to exert high effort, while the share increases to 62.5% when subjects are exposed to a scenario characterized by Type-II errors (T2). The positive effect is statistically significant at the 0.05% level, on the basis of a two-sided null hypothesis and 40 independent paired observations (McNemar's $\chi^2 = 8$, $p - value = 0.0047$).

High configuration (C & D): In the configuration with high initial endowment the share of subjects that are willing to exert effort is equal to 54.5% in T1 and increases to 70.5% in T2. The positive effect is statistically significant at the 1% level, on the basis of a two-sided null hypothesis and 44 independent paired observations (McNemar's $\chi^2 = 7$, $p - value = 0.0082$).

Contrary to the predictions of the standard theory, this analysis provides evidence that Type-I and Type-II errors do not generate symmetric effects: the detrimental effect of Type-I errors is substantially greater than the negative effect entailed by Type-II errors.

Treatment's configurations analysis

The two configuration have been implemented to stylized the two following real world situations. Low endowment configuration (no endowment) represents the case in which the evaluation error affects the assignment of the whole wage payment. High endowment configuration (€5.28 endowment) reproduces the case in which evaluation errors affect the assignment of an additional bonus up to the baseline wage that is granted.

To assess how the perception of the evaluation errors changes according to these two different situations, we run a between-subjects analysis contrasting treatments outcomes by configurations.

An examination of Figure 2.5.1 suggests the percentage of population exerting high effort in each of three treatments qualitatively increases in the high endowment

Table 2.5.1: Summary of results

Z_{Ti}	**T0**	**T1**	**T2**
Share of population exerting high effort in *Low conf.* (A & B)	0.825	0.325	0.625
Share of population exerting high effort in *High conf.* (C & D)	0.9	0.545	0.705

$Switchers$	**T0-T1**	**T0-T2**	**T2-T1**
Share of switchers between treatments in *Low conf.* (A & B)	$0.5^{(***)}$	$0.2^{(**)}$	$0.3^{(***)}$
Share of switchers between treatments in *High conf.* (C & D)	$0.355^{(***)}$	$0.204^{(***)}$	$0.16^{(**)}$

configuration with respect to the low endowment setting.

Under T0 the share of agent exerting high effort in the low endowment configuration is equal to 82.5% while the correspondent share in the high endowment configuration increases up to 90.9%. The difference between configurations results to be not statically significant at any conventional level.

Under T1 the share of agent exerting high effort in the low endowment configuration is equal to 32.5% while the correspondent share in the high endowment configuration results to be 54.5%. The difference between configurations results to be significant at 5% level.

Under T2 the share of agent exerting high effort in the low endowment configuration is equal to 65.2% while the correspondent share in the high endowment configuration is equal 70.5%. The difference between configurations results to be not statically significant at any conventional level.

From a qualitative point of view, this analysis suggests that both the Type-I and Type-II errors appear to be more detrimental when they affect the whole wage

assignment (low configuration) than the provision of an additional bonus (high configuration). The negative effect is more pronounced in case of Type-II error.

Please note that from a rational point of view – in all the three treatments – the agent's marginal utility of the reward payment (w_r) should by higher in the low endowment configuration that in the high endowment one.

Gift-exchange theory (Akerlof 1982) could represent a plausible candidate explanation for this behavioural outcome. As far as agents get a fair basic wage (high endowment configuration) they are willing to comply more frequently pursuing the task. In particular, this consideration is corroborated by the fact that more agents exert high effort also in the error free scenario T0 when featured by the high endowment configuration.

2.6 Discussion

The asymmetry in behavior of subjects under the $severe$ (T1) and $lenient$ (T2) treatment is the interesting puzzle that emerges by the experimental test. In the following paragraphs we rule out some potential explanations for the asymmetry and we discuss the main result at the light of different economic theories of behavior.

2.6.1 Why it can not be risk aversion

The experiment has been designed in order to rule out potential differences between T1 and T2 in terms of risky choices. To see why, consider the following table of standard generic concave utility functions with separable costs of effort. The subject decides whether to exert high effort whenever the difference in utility (line 3) is positive. Note that the difference in expected utility between T1 and T2 is exactly the same.

Whether the attempt to rule out risk aversion by construction can be considered successful depends crucially on the acceptance of the separability of the utility functions in monetary utility and effort (see Laffont and Martimort 2002 - pg. 149) and whether we focus on standard risk aversion derived by the decreasing marginal utility of money.

Moreover, in order to control for risk aversion in the data, we have also run an incentivized Holt and Laury (2002) lottery test in Phase I. Correlations between the individual measure of risk aversion[12] and the choice of exerting effort both for treatment T1 and T2 is very weak and statistically not significant (Spearman's $\rho - correlation = 0.032$, $p - value = 0.77$ in T1 and $\rho - correlation = 0.027$, $p - value = 0.57$ in T2 respectively).

It is however well known that modest-scale risk aversion also can be explained by behavioral biases such as loss aversion and myopic loss aversion (Rabin, 2000). We are however skeptical on whether loss aversion can explain our asymmetric result. This is because the net monetary returns of exerting high effort under all three treatments are positive (€6.60 with certainty and €1.32 and €1.32 in expected terms respectively for T0, T1 and T2)

Table 2.6.1: Risk aversion

	$T0$ **- Just**	$T1$ **- Severe**	$T2$ **- Lenient**
Utility with high effort	$v(w_r) - c$	$\frac{v(w_r)+4v(w_0)}{5} - c$	$v(w_r) - c$
Utility with low effort	$v(w_0)$	$v(w_0)$	$\frac{4v(w_r)+v(w_0)}{5} - c$
Difference in utility	$v(w_r) - v(w_0) - c$	$\frac{1}{5}v(w_r) - \frac{1}{5}v(w_0) - c$	$\frac{1}{5}v(w_r) - \frac{1}{5}v(w_0) - c$

2.6.2 Reciprocity, Fairness and Inequity

The asymmetry can be modeled by introducing some psychological costs of Type-I errors and Type-II errors.

Consider first the subject that exerts high effort. In a *just* scenario with no errors (T0) she would deserve the high wage $v(w_r)$ and would thus get the net reward of $v(w_r) - v(w_0)$. If this is taken as reference, then given the probability of

[12] In terms of switching point from risky to safe option in Table 2.8.1

Type-I error, $\alpha\left(v(w_r) - v(w_0)\right)$ is the expected value of the reward that is withheld from the subject. Now consider the subject that exerts low effort. In an error-free scenario (T0) she would deserve the low wage $v(w_0)$ with no reward. If this is taken as a reference, then the occurrence of a Type-II error generates an amount of undeserved reward equal to $\beta\left(v(w_r) - v(w_0)\right)$ in expected value. There exists thus two types of departure from the reference *just* scenario with no errors. On one hand the agent exerting low effort gets undue rewards and on the other hand the agent exerting high effort does not get due rewards. We extend a model in such a way that both departures represent a cost for the subject although they are weighted differently by the parameters ϵ^+ and ϵ^- respectively with $0 \leq \epsilon^+ \leq \epsilon^- \leq 1$. Figure 2.6.1 illustrates the payoff structure of the model extension.

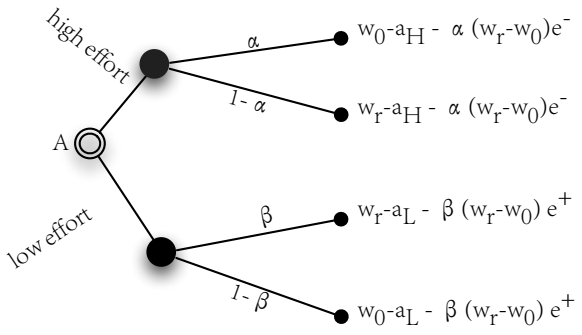


Figure 2.6.1: Choice of effort with fairness costs

The expected returns of exerting high effort are as in Equation 2.3 minus the disutility of departing from the reference point. The disutility is proportional to the net reward and to the magnitude of the error so that the expected returns are $\alpha\left[(v(w_0) - e_H\right] + (1-\alpha)\left[v(w_r) - e_H\right] - \alpha\left(v(w_r) - v(w_0)\right)\epsilon^-$. On the other hand the returns of exerting low effort are as in Equation 2.3 minus the disutility of departing from the reference point. Again, the disutility is proportional to the net reward and to the magnitude of the error so that the expected returns are $\beta\left[(v(w_r) - e_L\right] + (1-\beta)\left[v(w_0) - e_L\right] - \beta\left(v(w_r) - v(w_0)\right)\epsilon^+$. The new perfor-

	$T0$ **- Just**	$T1$ **- Severe**	$T2$ **- Lenient**
α	0	$4/5$	0
β	0	0	$4/5$
$\bar{\Delta}$	$e_H - e_L$	$(e_H - e_L)(5 + 4\epsilon^-)$	$(e_H - e_L)(5 - 4\epsilon^+)$

Table 2.6.2: Performance condition with fairness under the three treatments

mance condition is defined by the following equation:

$$\bar{\Delta} = \frac{1 + \epsilon^- \alpha - \epsilon^+ \beta}{1 - \alpha - \beta}(e_H - e_L) \tag{2.4}$$

In the following table, Equation 2.4 is computed with the parameters of our treatments.

Note that

$$\bar{\Delta}_{just} > \bar{\Delta}_{lenient} > \bar{\Delta}_{severe}$$

as

$$e_H - e_L > (e_H - e_L)(5 - 4\epsilon^+) > (e_H - e_L)(5 + 4\epsilon^-).$$

For a given monetary reward $\bar{\Delta}$ the individual is willing to exert a level of effort which is relatively high in T0, low in T2 and lower still in T1. Therefore the performance condition is higher in T0 and lower in T1 and takes an intermediate level in T2. This model is thus compatible with the results we obtain in the lab experiment.

In the next session, we proceed analyzing the main finding at the light of a set of behavioural economic theories.

2.6.3 Equity Theory and Inequity Aversion

Equity theory anticipates that not rewarding employees in accordance with their contributions undermines performance (Folger, 2001; Leventhal, 1980). Equity

theory (Adams, 1965) distinguishes between negative and positive inequity. Negative inequity happens in situations such as our T1: when individuals exerting high effort are not deservingly rewarded. In our model the costs of negative inequity are $\alpha\left(v(w_r) - v(w_0)\right)\epsilon^-$, therefore proportional to the magnitude of the Type-I error and to the net reward as well.

In T2 all subjects exerting high effort are rewarded. Higher effort provision (as compared to T1) can still be explained by equity theory in terms of positive inequity. Perception of unfairness persists when unfair distributions of outcomes are in favor of the employee. Though employees report being proud of their performance even when their success is the result of cheating, they also tend to feel guilty for their unfair behavior (Krehbiel and Cropanzano, 2000). In the model the costs of positive inequity is represented by $\beta\left(v(w_r) - v(w_0)\right)\epsilon^+$: they are proportional to the probability of Type-II error and to the magnitude of the net reward as well.

There is also a prolific stream of research in behavioral economics dealing with inequity aversion, described as a recurring preference for fairness and aversion to distributive inequality. Inequity aversion has been incorporated into several formal models of decision (Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). These models all encompass two common features: subjects dislike inequitable outcomes and suffer more from disadvantageous inequality than from advantageous inequality.

Inequity aversion can thus be fitted into our context, although in a looser sense. In fact in our context there is no interaction between agents and the supervisor is an automaton. Thus there can not be inequality in outcomes arising from the choices of the agent. However, inequity may arise with respect to the reference $just$ treatment and thus it resembles more closely an issue of organizational and procedural justice. If the subject measures the equity of the treatment she is subject to against the one she thinks ought to apply under the $just$ treatment, then inequity can take the form of both disadvantageous/negative inequity (as in the $severe$ treatment) and advantageous/positive inequity (as in the $lenient$ treatment). Following this literature, it is sound and compatible with our results to assume that $\epsilon^+ < \epsilon^-$, as favorable although unfair outcomes are preferred (less costly) to unfavorable and

unfair outcomes. However, note that the result $\bar{\Delta}_{just} > \bar{\Delta}_{lenient} > \bar{\Delta}_{severe}$ holds also when $\epsilon^+ = \epsilon^-$.

2.6.4 Gift-exchange and Reciprocity

Another stream of literature that is of use in interpreting the result is Akerlof's theory (1982) of gift-exchange. According to this theory the employer offers a gift to the employee in the form of an above-market-equilibrium salary in return for the worker's gift of high non-observable effort for the firm. Within the gift-exchange framework, our $lenient$ treatment (T2) represents a positive departure from the $just$ reference (T0) and therefore it could induce the individual to exert more effort . On the contrary, our $severe$ treatment (T1) is a negative departure from the $just$ reference (T0) that may induce him to exert no effort at all. A broader perspective on gift-exchange is given also by the reciprocity literature (Bruni, Gilli, and Pelligra, 2008; Falk and Fischbacher, 2006; Fehr and Schmidt, 2003). Reciprocating individuals typically respond in kind to others' actions: they are *severe* with the mean and *lenient* with the indulgent. The model above can also fit the reciprocity approach. In fact, the larger is the departure from the just reference, the larger is the willingness for the subject to sustain costs in order to either exchange gifts with the employer in the case of indulgent behavior ($\beta\left(v(w_r) - v(w_0)\right)\epsilon^+$) or negatively reciprocate in the case of severe behavior $\alpha\left(v(w_r) - v(w_0)\right)\epsilon^-$.

Our results seem to fit such behavior, as the propensity to exert high effort is ostensibly higher in the $lenient$ treatment than in the $severe$ one. The trouble in properly interpreting our result in terms of reciprocity and gift-exchange comes from the fact that in our experiment there is no real supervisor (our supervisor is represented by a passive automaton) to be indulgent or severe with. A well-established result in the literature is that, facing an automaton instead of a human subject playing as a proposer in an ultimatum game, respondents are less apt to reject unfair offers (Blount, 1995). There are two common and complementary interpretations of this result: a) a low proposal offered by an automaton can not be associated with the intention of an actual proposer that it is possible to harm

by refusing the offer (negative reciprocity) and b) a random draw is perceived as a fair procedure. In our experimental setting the supervisor is clearly an automaton and the evaluation error realization is random. Therefore subjects should have roughly the same inclination to exert high effort under both the *severe* and the *lenient* treatment. In other words, subjects do not have a real supervisor with which to negatively reciprocate in the case of severe treatment or with which to exchange gifts in the case of indulgent treatment. Even the alternative explanation that subjects are actually positively (negatively) reciprocating with the experimenter in the indulgent (severe) treatment seems to be weak. In fact it should be noticed that the effort task is clearly purposeless and therefore subjects can anticipate that the experimenter does not gain anything from having more effort exerted. Moreover they might also think that the experimenter is hurt by having to pay more. Therefore a good way to reciprocate positively (negatively) would be to exert low (high) effort in T2 (T1) so that the experimenter pays on average lower (higher) payoffs. These considerations might provide further evidence in support of the idea that a significant share of agents is intrinsically adverse to Type-I error per se.

2.7 Implications and Conclusion

The experiment finds strong support for the existence of an asymmetric impact of errors on agents' willingness to exert high effort. In particular an agent exposed to evaluation errors is more sensitive to Type-I errors. This result is not predicted by the model even when considering risk aversion within the expected utility framework. Further work is needed to explain the result, which seems to be robust against some preliminary treatments manipulations. From a theoretical perspective, the experiment sheds new light on the relation between reward systems and motivation that should inform agency theory, organizational behavior and personnel economics. From an organizational perspective, our result expands the notion of organizational justice: Departures from the just treatment can be both advantageous and disadvantageous even in absence of distributional implications with third parties as assumed by equity theories; subjects react differently when

they suffer (enjoy) disadvantageous (advantageous) injustice. Our results can also be interpreted as further evidence of gift-exchange behavior in the lenient treatment and negative reciprocity in the severe treatment.

Although the experimental method has limited external validity, this particular result may have direct practical implications in real-world contexts. Since intangibles are increasingly important in business organizations and knowledge-intensive jobs are difficult to assess, errors in evaluating employees' performance may well be a relevant phenomenon. Our research suggests that, when a perfect assessment of employees' effort provision is not viable, it may be wise for the supervisor to be cautious when neglecting rewards and - in general - have a pro-employee bias in conducting her assessment, as this may well be beneficial for employees' motivation and effort provision in the longer term.

2.8 Appendix

2.8.1 Instructions

We report here the instructions used for the T0 *high* treatments with baseline wage = €5.28 and the rewarding wage = €6.60. In *low* treatments instructions differ only in that the baseline wage = €0.

SITUATION – A – ($T0_{High}$) # #

In Situation A you will receive a fixed payment of €<5.28> Your task (to count <goal> tables in 40 minutes) is supervised by an automatic supervisor. In this situation, the automatic supervisor does not commit any error of observation:

- If you accomplish the task (that is to correctly count the number of *1* in at least <goal> tables), the supervisor will certainly (probability 100%) commit no evaluation error and it will assign to you the payment of €<6.60> (tot. 11.88=5.38+6.60)
- If instead you do not accomplish the task (that is to correctly count the number of *1* in at least <goal> tables) the supervisor will certainly (probability 100%) commit no evaluation error and it will not assign you the payment of €<6.60> (tot. 5.38=5.38+0)

 # CONTROL QUESTIONS

- In this situation, if you do accomplish the task, you will receive the €<6.60> payment with a probability of [please, provide the answer -_____%-] (Correct answer is 100)
- In this situation, if you do not accomplish the task, you will receive zero payment with a probability of [please, provide the answer -_____%-] (Correct answer is 100)
- In this situation, you will receive a fixed payment of €<5.28> with a probability of [please, provide the answer -____% -] (Correct answer is 100)

 # EFFORT CHOICE

If Phase III of the experiment corresponds to SITUATION A as just described, will you perform the task or will you skip the task? Remember that:

- If you press the "A - I will perform the task" button and Phase III corresponds to SITUATION A, you will have to wait 40 minutes anyway before proceeding to the questionnaire phase.
- If you press the "B - I will skip the task" button and Phase III corresponds to SITUATION A, you will proceed directly to the questionnaire phase

[A – I will perform the task] / [B – I will skip the task]

 # # # SITUATION – B – ($T1_{High}$) # # #

In Situation B you will receive a guaranteed fixed payment of €<5.28>. Your task (to count <goal> tables in 40 minutes) is supervised by an automatic supervisor In this situation, the automatic supervisor might commit an error of observation:

- If you accomplish the task (that is to correctly count the number of *1* in at least <goal> tables)
  - the supervisor with a probability of 80% will commit an evaluation error and it will not assign to you the payment of €<6.60>
  - the supervisor with a probability of 20% will commit no evaluation error and it will assign to you the payment of €<6.60>
- If instead you do not accomplish the task (that is to correctly count the number of *1* in at least <goal> tables), the supervisor certainly (probability 100%) commit no evaluation error and it will not assign to you the payment of €<6.60>

 # CONTROL QUESTIONS

- In this situation, if you do accomplish the task, you will receive the €<6.60> payment with a probability of [please, provide the answer -_____%-] (Correct answer is 20)

- In this situation, if you do not accomplish the task, you will receive the €<5.28> payment with a probability of [please, provide the answer -_____%-] (Correct answer is 100)

- In this situation, you will receive a fixed payment of €<5.28> with a probability of [please, provide the answer -____% -] (Correct answer is 100)

 # EFFORT CHOICE

If Phase III of the experiment corresponds to SITUATION B as just described, will you perform the task or will you skip the task? Remember that:

- If you press the "A - I will perform the task" button and Phase III corresponds to SITUATION B, you will have to wait 40 minutes anyway before proceeding to the questionnaire phase

- If you press the "B - I will skip the task" button and Phase III corresponds to SITUATION B, you will proceed directly to the questionnaire phase

[A – I will perform the task] / [B – I will skip the task]

SITUATION – C – ($T2_{High}$)

In Situation C you will receive a guaranteed fixed payment of €<5.28>. Your task (to count <goal> tables in 40 minutes) is supervised by an automatic supervisor. In this situation, the automatic supervisor might commit an error of observation:

- If you accomplish the task (that is to correctly count the number of *1* in at least <goal> tables), the supervisor will certainly (probability 100%) commit no evaluation error and it will assign to you the payment of €<6.60>

- If instead you do not accomplish the task (that is to correctly count the number of *1* in at least <goal> tables)

  - the supervisor with a probability of 80% will commit an evaluation error and it will assign to you the payment of €<6.60>
  - the supervisor with a probability of 20% will commit no evaluation error and it will not assign to you the payment of €<6.60>

 # CONTROL QUESTIONS

- In this situation, if you do accomplish the task, you will receive the €<6.60> payment with a probability of [please, provide the answer -_____%-] (Correct answer is 100)
- In this situation, if you do not accomplish the task, you will receive the €<5.28> payment ii) with a probability of [please, provide the answer -_____%-] (Correct answer is 20)
- In this situation, you will receive a fixed payment of €<5.28> with a probability of [please, provide the answer -____% -] (Correct answer is 100)

 # EFFORT CHOICE

If Phase III of the experiment corresponds to SITUATION C as just described, will you perform the task or will you skip the task? Remember that:

- If you press the "I will perform the task" button and Phase III corresponds to SITUATION C, you will have to wait 40 minutes anyway before proceeding to the questionnaire phase
- If you press the "I will skip the task" button and Phase III corresponds to SITUATION C, you will proceed directly to the questionnaire phase

[A – I will perform the task] / [B – I will skip the task]

2.8.2 Screenshots

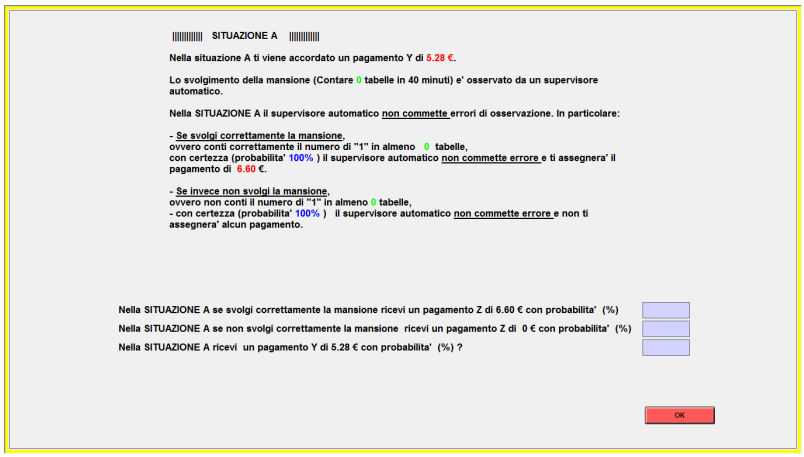


Figure 2.8.1: Screenshot of the Situation presentation and Control questions

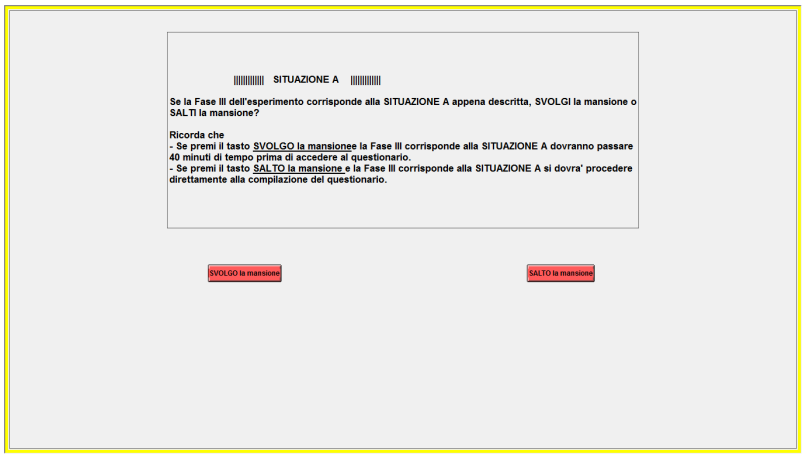


Figure 2.8.2: Screenshot of the Effort Choice Phase

Table 2.8.1: **Holt & Laury Table**

	Option A			**Option B**	
# Choice	Probability	Gain €		Probability	Gain €
#1	50% 50%	3.00 23.00	vs	100 %	7.00
#2	50% 50%	3.00 23.00	vs	100 %	8.00
#3	50% 50%	3.00 23.00	vs	100 %	9.00
#4	50% 50%	3.00 23.00	vs	100 %	10.00
#5	50% 50%	3.00 23.00	vs	100 %	11.00
#6	50% 50%	3.00 23.00	vs	100 %	12.00
#7	50% 50%	3.00 23.00	vs	100 %	13.00
#8	50% 50%	3.00 23.00	vs	100 %	14.00
#9	50% 50%	3.00 23.00	vs	100 %	15.00
#10	50% 50%	3.00 23.00	vs	100 %	16.00

Chapter 3

Teams or Tournaments? An Experiment on the Effectiveness of Alternative Grading Policies

ABSTRACT

We assess the effect on students' effort of two antithetic non-monetary incentive schemes based on grading rules using experimental data. We randomly assigned students to a tournament scheme that fosters competition between coupled students, a cooperative scheme that promotes information sharing and collaboration between students and a baseline treatment in which students can neither compete, nor cooperate. In line with theoretical predictions, we find that competition induces higher effort with respect to cooperation and cooperation does not increase effort with respect to the baseline treatment. However, we find strong gender effect since this result holds only for men while women do not react to this type of non-monetary incentive.

Keywords: Education, Gender Differences, Experimental Economics, Incentives, Competition, Cooperation.

JEL code:A22, C93, I20.

3.1 Introduction

In the last years a large debate has focused on the ways to improve schooling achievement at every level of education. The relevance of this goal is not disputable, since education contributes to the accumulation of human capital, the development of societies and it is considered as one of the main channels for the reduction of inequality. Recent studies have addressed this goal measuring the impact of monetary incentives both on input (e.g. subsidizing the purchase of learning supports) and on output (e.g. giving money based on grades, or conditional on passing the exam). In this paper we take a different approach exploring whether specific instructional choices and pedagogical practices can affect college students' learning. We do not use monetary rewards but provide grade-incentives to learn, throughout the semester. We study the effect of different grading rules on schooling achievement by assigning students to different incentive schemes: a tournament, a piece rate and a scheme that promotes cooperation. The analysis is performed on a sample of students enrolled in a undergraduate course in econometrics at the University of Bologna (Italy).

The design of the experiment is based on a theoretical model that contemplates three different incentive schemes. As a benchmark we consider the effect on effort of a piece rate reward. Then we analyze two alternatives: a tournament that fosters competition among matched students and a cooperative scheme in which they can share information and collaborate. The model suggests a weak ordering between the three: in the competitive environment the exerted effort should be higher than in the benchmark while the effort under the benchmark should be weakly higher than in the cooperative scheme. We also show that the detrimental effect of cooperative incentives on effort does not depend on the distribution of types in the population, while the magnitude – but not the sign – of the effects of a competitive incentive scheme depends on this distribution. To test these theoretical predictions, we randomly assign students to the treatments and we adopt a between-subjects design, i.e. each subject is exposed only to a single incentive scheme.

Data confirm the theoretical predictions in the full sample. Moreover, we show

that an important difference emerges between genders: promoting competition appears to have a strong positive effect on the exerted effort only for males. In contrast, promoting cooperation reduces effort with respect to the case where students can neither compete nor cooperate, but this effect is not statistically significant for both genders. These findings are in line with the literature on how competition affects behaviour depending on gender (see for example Gneezy, Niederle, and Rustichini 2003) and provide an interesting comparison with respect to the result of Angrist and Lavy (2009) who find that monetary incentives improve performance especially on girls.

We depart from this branch of the literature, complementing the results obtained through monetary incentive, by focusing on non-monetary ones since the latter represent a relatively cheap way to increase student's effort.[1] More specifically, we reward students with extra points for their course grade, up to 10% of the maximum mark: these additional points may result in a student passing the exam rather than failing, which has also consequences on the timing of the graduation.

The paper proceeds as follows. After a brief review of the related literature (Section 3.2) we describe and discuss in detail our experimental design (Section 3.3). In Section 3.4, we present a simple model, and derive the theoretical predictions which will serve as a reference for the analysis of the experimental data, presented in Section 3.5. Section 3.6 concludes, and presents possible extensions of this research.

3.2 Related Literature

Our study focuses on the impact of alternative pedagogical practices on students' performances. As such, it relates to the literature on economic education (see for example Becker 1997), which investigates how to assess and enhance the efficacy of teaching practices. Within this stream of literature, our work is particularly akin

[1] Studies on monetary incentives proved to be successful in improving students' performance but the cost of inducing higher effort is not negligible. In a study conducted in the New York City school system $600 have been awarded for each passing grade, the Baltimore City Public School District has paid up to $110 to improve scores on state graduation exams and similar programs in the US award up to $500 for each exam passed.

to the paper by Grove and Wasserman (2006), who study whether the adoption of plain individual-based grade incentives on problem sets improves students' exam performance, finding a positive effect only for freshmen. From a different perspective our study focuses on the effects, in terms of students' effort provision, generated by two alternative relative-based grading schemes applied to within-course tests.

On the general issue of how to foster students' effort and school achievement through explicit incentive schemes, several papers explore the role of financial incentives. Among those, Blimpo (2010) represents the closest study to our experiment. Analysing data from a field experiment in Benin with a pool of 100 secondary schools, he studies whether individual incentives or different kind of team incentives can lead to a higher students' school performance. He considers three treatments. In the first treatment, each student obtained an individual monetary reward if and only if his or her performance exceeded a minimal threshold at the final exam. In the second treatment, participants were randomly assigned to teams of four students and each team-member received a monetary reward depending on the average team performance, if and only if all the team-members achieved a target performance level. Finally, in the third treatment, participants were randomly assigned to teams of four students but in this case only the components of the three top-performer teams were awarded with a monetary prize. Blimpo (2010) finds that the individual based incentive scheme with cut-off target is most effective for students at an intermediate performance level: at the lower tail of the skills distribution, students reduce effort, probably because they perceive the target out of reach; at the higher tail of the distribution, students know that they are able to get the prize without any extra effort, thus the average impact of such incentives is smaller. When teams are evaluated according to the average performance of the group conditionally on the achievement of a minimal performance target (2^{nd} treatment), students across all levels of ability are positively affected: the effort exerted by the different team-mates is pushed toward the target. The tournament scheme (3^{rd} treatment) yields the most beneficial effects: it induces all the teams to work harder as students exposed to this treatment do not have any prior information about

the quality and the skills of their competitors in the other teams.

Recent papers consider tournaments at school with financial rewards. Kremer, Miguel, and Thornton (2009) focus their study on the evaluation of a merit scholarship programme dedicated only to female students in an elementary school in Kenya. They observe a substantial increase in the exams scores: in particular girls with low pre-test scores, who were unlikely to win a scholarship (and actually did not get it), reported positive and significant gains in terms of higher school performance. De Paola, Scoppa, and Nistico (2010) studied the effectiveness of monetary incentive schemes in enhancing students' performance using a randomized experiment involving undergraduates in an Italian University. Students participating in the experiment were assigned to three different groups: a high reward group, a low reward group and a control group. Rewards were assigned according to a ranking rule to the top performing students in each treated group. The authors report that financial rewards contributed to increase the students' performance: a very strong reaction emerged among high ability students who were likely to win the contest, while no significant effect was observed for low ability students that have fewer chances to win the tournament competition. Along the same lines, Leuven, Oosterbeek, and van der Klaauw B. (2010) present results of a randomized field experiment in which freshman students at the Amsterdam University had the opportunity to earn financial rewards for passing all first year requirements. Their findings provide evidence that high ability students perform significantly better when assigned to rewarded groups. On the contrary low ability students' outcome decreases if assigned to rewarded groups. The small aggregate average effect that they observe is therefore the sum of a positive effect for high ability students and a negative off-setting effect for low ability students. These previous results highlight the importance of controlling for students' ability and individual characteristics when assessing the impact of incentive schemes on their school performance.

A recent study, Fryer (2010), has addressed this goal measuring the impact of monetary incentives both on input (e.g. subsidizing the purchase of learning supports) and on output (e.g. giving money based on grades, or conditional on passing the exam). Results show that incentives can raise achievement among even

the poorest minority students in the lowest performing schools if the incentives are provided on "inputs". Incentives focused on "output" result to be much less effective.

Among the authors who studied the effects of financial incentives on 'output', some focused specifically on gender differences. Angrist and Lavy (2009) evaluate the effectiveness of financial rewards on the achievement of Israeli students using a randomized experiment providing monetary awards to students who obtain the university admission. The authors show how the program led to significant effects for girls but not for boys. Differences in gender-scheme interaction emerge also from the field experiment by Angrist, Lang, and Oreopoulos (2009). In this study, researchers randomly assigned a sample of students enrolled in a Canadian university to one of three different treatments: the first group was provided with a set of support services (e.g. tutoring); the second group was offered financial rewards for good academic scores; the third one was offered a combination of support services and monetary incentives according to the academic performance. The results of the experiment show that while males did not react to any of the treatments, females improved significantly their academic performance when monetary incentives were provided.

While females appear to react more than males to monetary incentives awarded for achieving an exogenously given target, incentive schemes based on competition may yield opposite effects. Gneezy, Niederle, and Rustichini (2003) find that males are more prone to engage in competition than females and in general males' performance increases more than the females' one when subjects are exposed to a competitive setting. Similarly, Niederle and Vesterlund (2007) find that, when given the opportunity to choose between a piece-rate payment scheme or a tournament, men select the tournament twice more frequently than women, suggesting that women tend to avoid competition when they have the chance to do so. Azmat and Iriberri (2010) find that, even when the incentive scheme is based solely on the subject's performance, providing information about the *relative* performance promotes higher levels of effort among men, but not among women. We explore the role of gender, and we find that males tend to respond to incentives as predicted by

the theory, while females do not.

From a theoretical standpoint, Bratti, Checchi, and Filippin (2011) propose a model of student cooperation/competition in learning activities, showing that free riding opportunities lead to an insufficient degree of cooperation between schoolmates, which in turn decreases the overall achievement of the group. According to their analysis, a cooperative learning approach may successfully emerge when the class is homogeneous in terms of students' ability. In our study we consider an experimental design and a theoretical model where the incentive scheme is exogenous but similarly to Bratti, Checchi, and Filippin (2011) we focus on student cooperation/competition in learning activities. Our theoretical model suggests that in a competitive environment individual performance should be higher than in the cooperative environment.

3.3 The Experimental Design

The experiment involved all the undergraduate students enrolled in the Introductory Econometrics course of the major in Management Studies at the University of Bologna, in year 2010.[2] The course lasted 10 weeks (a three-hour-lecture per week). Students participating to the experiment had to undertake 5 tests whose marks were translated into bonus points for the final exam. The bonus points for the final exam were equal to the average mark the student obtained in the five tests.[3]

Tests have been scheduled every two weeks and each test consisted of five multiple-choice questions to be answered in 50 minutes. Each test concerns all topics taught in the course until the last lecture before the test.

Tests were computerized, and were held in the computer laboratory of the School of Economics of the University of Bologna.[4] Desks were arranged so to exclude the possibility for students to talk during the exams (see Figure 3.7.1 in Appendix).

[2]The University of Bologna is considered the oldest University in Europe and counts on average nearly 8000 enrolled students each academic year.

[3]Marks in the final exam range from 0 to 30. The exam is passed with a mark equal or above 18. The bonus points ranged from 0 to nearly 4.

[4]The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007b).

The mark in each test consisted in an individual component, based on the number of correct answers in the test, and a number of extra points related to the treatment.

Our study included two treatment conditions – characterized by a competitive and by a cooperative incentive scheme, respectively – and a baseline treatment. In all treatments including the baseline, part of the incentive depended solely on individual effort. Treatments differed in how tests two, three and four were performed, while the first and the last test were identical across treatments. The first and the last tests were taken individually by each student. In contrast, in the second, third and fourth tests students in the two treatment conditions were randomly matched in couples at the beginning of each test, and had the opportunity of exchanging messages with their partner via a controlled chat program, running on their computer. In both treatment conditions, the total score in tests 2, 3, and 4 of the test depended not only on the student's individual effort (i.e. the net score), but also on the partner's performance. Table 3.3.1 summarizes the treatments, which are described in detail below.

Students were assigned to treatments between the first and the second test. Before starting tests 2, 3, and 4, students assigned to the two treatment conditions were asked whether they wanted to use the chat or not to communicate with the paired partner. This decision was taken simultaneously by all students. During the test, coupled students could use the chat program only if both of them declared to be willing to communicate, at the beginning of the test. If the two students chose to communicate, for each of the questions of the test they could send only one "signal" to indicate what the right answer was, and one short text message of up to 180 characters. Interactions were anonymous, as students could not know the identity of their partner. In the baseline treatment no interaction between students was allowed.[5]

In each test, the value p^q of correct answers to each question q ranged between

[5]Figure 3.7.4 presents a screen-shot of the graphical interface of the program used for the tests. On the left-hand side of the screen students could read the question, and the multiple-choice answers. On the top-right part of the screen they could send messages to their partner, while on the bottom-right part of the screen they could read the messages possibly sent to them by the partner.

0.3 and 1.2 points. Across all treatments, the number of points v_i^k a student could get by correctly answering the questions of test k was:

$$v_i^k = s_i^k \cdot \mathrm{I}\left(s_i^k \geq 1.5\right), \quad s_i^k = \sum_{q=1}^{5} p_i^{q,k}, \;\; k = 1, \ldots 5$$

In each test, the maximum number of points $\bar{v}$ was equal to 3. This is the individual part of the mark in the test, i.e. the component which is common across all treatments.

In the COMPETITIVE treatment, student i's mark in a test was increased by 2 extra points if her score resulted to be strictly higher than the partner's. The k-th test's mark $\hat{v}_i^k$ for student i under this incentive scheme is described in equation (3.1).

$$\hat{v}_i^k = v_i^k + 2 \cdot \mathrm{I}\left(s_i^k > s_j^k\right), \quad k = 2, 3, 4 \tag{3.1}$$

This provides an incentive for both matched students to compete.

Conversely in the COOPERATIVE treatment, student i's score in a test was increased by 1 extra point if the partner's score was sufficiently good. The k-th test's mark $\hat{v}_i^k$ for student i under this incentive scheme is presented in equation (3.2).

$$\hat{v}_i^k = v_i^k + \mathrm{I}\left(s_j^k \geq 1.5\right), \quad k = 2, 3, 4 \tag{3.2}$$

Finally, students in the BASELINE treatment received 1 extra point in tests 2, 3 and 4. [6]

Time-line of the experiment. The experiment started in February 2010, and ended in July of the same year. In the first lecture of the course, on February 25th, the full set of instructions was distributed to students and each student had two days to decide whether to take the tests or not. At this stage, students were not

[6] This is done so that the maximum number of bonus points per team is constant across treatments. For the design to be correctly balanced, incentives in the cooperative and competitive treatment should have the same size in expectation, i.e. holding θ constant (i.e., in each test the probability to get the bonus points under the competitive treatment should be half the probability to get the bonus points under the cooperative treatment). We tested this assumption in our data and it is never rejected at 5% level. More specifically, the probability to get the bonus points in test 2,3,4 respectively is: under the cooperative treatment 0.73, 0.95, 0.95; under the competitive treatment: 0.5, 0.39, 0.44.

Table 3.3.1: Summary of the treatments, in tests 2, 3 and 4

treatment	**extra points (rounds 2, 3, 4)**	**messages available**
BASELINE	1	no
COOPERATIVE	$\mathrm{I}(s_j^k \geq 1.5)$	yes
COMPETITIVE	$2 \cdot \mathrm{I}(s_i^k > s_j^k)$	yes

explicitly informed that they were taking part in an experiment and only at the very end of the course, participating students were asked to sign a consent form authorizing the treatment of data collected during the tests.[7] In this sense, our study is a "field experiment" under the terminology defined by Harrison and List 2004.

On March 1st, during a standard class, students were asked to fill in a questionnaire collecting data about some personal characteristics (age, gender, familiarity with computers, e-mail and chat programs, mother and father education). Questionnaire answers are used in the econometric analysis to baseline for individual-specific characteristics.[8]

On March 22nd students took the first test. Notice that at this stage students had not yet been assigned to treatments, so the grade in this first test can be used as a measure of their effort level before being exposed to the treatment. Students received information about what treatment they had been assigned to only three days later, on March 25th.[9] In the same day, students were informed about their own result in the first test, and about the distribution of the first test score among participants. In this way we tried to convey common knowledge of the distribution of competences and ability in the population. Section 3.4 will show how this is relevant from the theoretical point of view.

The remaining four tests were taken approximately every two weeks, in April and May 2010 with the exception of the fifth which was administered one week

[7] The experiment was authorized by the ethics committee of the the University of Bologna *(Comitato Bioetico per la Valutazione di Protocolli di Sperimentazione)*.

[8] An overview of the answers to the questionnaire is provided in Section 3.5, and a translation of the questions is reported in Table 3.7.2 in Appendix.

[9] Students taking part in our experiment were then randomly assigned to two groups of about 65 people each, because the computer lab can host only up to 80 students at a time. All students assigned to the competitive treatment and half of those assigned to the baseline treatment were in the first group, while all students in the cooperative treatment and the remaining students of the baseline treatment were in the second group.

after the fourth.[10] Student could benefit of the bonus points gained in the tests only if they took the final exam in June or July 2010. Before the experiment started, students were informed that the bonus points would expire after the summer.

3.4 The Model

This section describes the main features of the model we use to derive theoretical predictions and inform the experimental design. After briefly characterizing the general features of the model, we illustrate its implications in terms of expected effort under the different incentive schemes. We first describe what happens without competitive or cooperative incentives (BASELINE treatment). We then characterize the optimal effort under incentives to cooperation and to competition and finally we highlight the testable predictions of the model.

General features We assume that students' abilities are in the interval $\theta \in [0, 1]$ and are distributed according to a non-degenerate distribution function $F(\cdot)$. Students choose a level of effort $e_i \in [0, 1]$, which determines their score in the tests and the grade in the final exam. The dis-utility from effort is $c(e_i)$ where $c(\cdot)$ is the same across the population[11]. We further assume that $c(\cdot)$ is such that $c'(\cdot) > 0$ and $c''(\cdot) > 0$.

The expected score in test k is a function increasing in ability and effort and is given by the following expression:

$$s_i^k = e_i \cdot \theta_i \cdot \bar{v} \tag{3.3}$$

We thus assume that the productivity of effort is higher for higher-ability students, and that only students with $\theta = 1$ can get the maximum score ($\bar{v}$) if they exert the maximum level of effort ($e_i = 1$).

[10] This is made on purpose since the last test is taken by students individually and covers the last contents of the program as well as some of the previous ones. Hence, it will reflect the effort exerted in the previous stages.

[11] The dis-utility from effort can be thought both as the mental effort of being concentrated in studying a certain amount of hours and as the cost of spending these hours studying instead of meeting friends or doing some other activity.

The utility of each student is positively affected by the score and negatively affected by the effort. We assume that students choose their level of effort two times: the first time they choose $e_{i,0}$ when the course starts, before the first test and before the assignment to the treatments; later, after having been assigned to treatments they choose the level of effort e_i that determines their scores in tests 2 to 5 and in the final exam. At this point, their expected utility is given by 3 components: the bonus points obtained in the four remaining tests to be taken – which in the two treatment conditions is the outcome of the interaction with the matched agent – the individual mark in the final exam[12] and the cost of effort. Under the assumption of risk neutrality, the expected utility at the time in which e_i is chosen is:

$$E[U_i] = \frac{1}{5}\sum_{k=2}^{5}\int_0^1 \hat{v}_i^k(\theta_i, e_i, \theta_j, e_j)\cdot f(\theta_j)\mathrm{d}\theta_j + \bar{V}\cdot e_i\cdot\theta_i - c(e_i) \tag{3.4}$$

where $\bar{V}$ is the maximum mark in the final exam.

Baseline treatment A student assigned to the baseline treatment does not interact with any other student. As a consequence, considering the four tests and the final exam, the expected utility (3.4) simplifies in:

$$U_i^{BL} = \bar{V}\cdot e_i\cdot\theta_i + \frac{1}{5}\cdot(4\cdot e_i\cdot\theta_i\cdot\bar{v}) + \frac{3}{5} - c(e_i) \tag{3.5}$$

from this utility function we can derive the optimal effort exerted:

$$\frac{\partial U_i^{BL}}{\partial e_i} = (\bar{V} + \frac{4}{5}\cdot\bar{v})\cdot\theta_i - c'(e_i) \tag{3.6}$$

Normalizing the quantity $V + \frac{4}{5}\cdot\bar{v} = 1$, we get the baseline effort:

$$c'(e_i^{BL}) = \theta_i$$

[12]Remember that the bonus adds points on top of this mark.

that implies

$$\frac{\partial e_i^{BL}(\theta_i)}{\partial \theta_i} > 0$$

i.e., we expect more able individuals to exert more effort in the baseline treatment with respect to less able individuals and this reflects the different productivity of students. Given the higher return from studying, the more able individuals are more willing to trade off effort (time spent studying) for grades than less able individuals.

Competitive treatment To model student's behavior under the two treatments and to derive predictions, we look for the equilibrium in the Bayesian-Nash games where students have private information about their own type and a common knowledge on the distribution of ability in the population.

Under the competitive scheme, students get bonus points if their score is higher than the partner's. Equation (3.7) describes the expected utility in this case.

$$\begin{aligned} U_i^{comp} &= \bar{V} \cdot e_i \cdot \theta_i + \frac{1}{5}\left[4 \cdot e_i \cdot \theta_i \cdot \bar{v}\right] + \frac{3}{5} \cdot 2 \cdot Pr(e_i \cdot \theta_i > e_j \cdot \theta_j) - c(e_i) = \\ &= \bar{V} \cdot e_i \cdot \theta_i + \frac{1}{5}\left[4 \cdot e_i \cdot \theta_i \cdot \bar{v}\right] + \frac{6}{5} \cdot \int_0^{\theta_i \cdot \frac{e_i}{e_j}} f(\theta_j) d\theta_j - c(e_i) = \\ &= \bar{V} \cdot e_i \cdot \theta_i + \frac{1}{5}\left[4 \cdot e_i \cdot \theta_i \cdot \bar{v}\right] + \frac{6}{5} \cdot F\left(\theta_i \cdot \frac{e_i}{e_j}\right) - c(e_i) \end{aligned} \tag{3.7}$$

where $6/5 \cdot F(\theta_i \cdot e_i/e_j)$ is the expected number of additional points obtained in the second, third, and fourth test in case the student outperforms his partner. Hence, the expected utility can be expressed as:

$$e_i(\theta_i) \in \underset{e_i}{\text{argmax}} \Big\{ E[U_i] = \bar{V} e_i \theta_i + \frac{1}{5}\left[4 e_i \theta_i \bar{v}\right] + \frac{6}{5} \int_{\theta_j | \theta_j e_j < \theta_i e_i} f(\theta_j) d\theta_j - c(e_i) \Big\} \tag{3.8}$$

Under regularity assumption on the distribution of types in the population, it can be shown that the first order conditions are:

$$\theta_i - c'(e_i) + \frac{6}{5} f(\Phi_j(e_i)) \Phi'(e_i) = 0 \tag{3.9}$$

where Φ_k is the mapping from the effort to the type (individual ability).[13] Now, since $\Phi' = 1/e'$, we have the following solution for the optimal effort in the competitive treatment:

$$c'(e_i) = \theta_i + \frac{6}{5} f(\theta_i) \cdot \frac{1}{e'_i} \tag{3.10}$$

From this equation we see that the optimal effort exerted under this scheme is higher than the optimal level of effort e_i^{BL} in the baseline treatment.

Cooperative treatment Under this scheme, each student has a clear incentive to share her information (in tests 2, 3 and 4) and the mark depends also on the partner's effort.

In this case the expected utility becomes:

$$\begin{aligned} U_i^{coop} = & \bar{V} \cdot e_i \cdot \theta_i + \frac{1}{5} \cdot [e_i \cdot \theta_i \cdot \bar{v}] + \\ & + \frac{1}{5} \int_0^1 3 \cdot [\bar{v} \cdot (e_i \cdot \theta_i + e_j \cdot \theta_j - e_i \cdot \theta_i \cdot e_j \cdot \theta_j) + \\ & + \mathrm{I}(e_i \cdot \theta_i + e_j \cdot \theta_j - e_i \cdot \theta_i \cdot e_j \cdot \theta_j > 0.5)] \cdot f(\theta_j) \mathrm{d}\theta_j - c(e_i) \end{aligned} \tag{3.11}$$

The second term in equation (3.11) represents the points obtained form the fifth test, where no interactions among student was allowed, while the third term represents the bonus obtained in tests 2, 3 and 4. The assumption that information is shared by the students is crucial and implies that the probability of answering a question correctly is given by the probability that either one of the two students knows the solution, and the that optimal effort is given by:

$$c'(e_i^{coop}) = \theta_i - \frac{3}{5} \cdot \bar{v} \cdot \theta_i \int_0^1 \theta_j \cdot e_j \cdot f(\theta_j) \mathrm{d}\theta_j \tag{3.12}$$

The second term in the right-hand side of equation (3.12) is always non-positive, and its absolute value increases with θ_i. This shows that, since information is shared, each team member has an incentive to exploit the effort of the other low-

[13] In order to have a pure strategy Nash equilibria, the distribution function of types must be non-degenerate and the mapping from type to effort must be continuous and increasing. The requirement on the distribution of types is plausible given the heterogeneity in the population, while the two on the mapping between type and effort can be proven true in our case. In the non-heterogeneous case, that is when the distribution of types is degenerate, it can be easily shown that no pure-strategy equilibrium exists.

ering his own contribution. As a consequence, under the cooperative treatment, team members have an incentive to shrink their effort, and this detrimental effect of cooperation on effort is stronger for students with higher ability (θ_i).

Testable predictions To sum up, our theoretical model predicts that, given the ability θ_i, the effort exerted by student i in the three treatments is such that:

$$e_i^{coop} \leq e_i^{BL} < e_i^{comp}$$

i.e., we expect that on average students randomized into the COOPERATIVE treatment exert lower or equal effort than students randomized into the BASELINE treatment whereas students randomized into the COMPETITIVE treatment should exert more effort.[14] Conversely, at test 1, all students have the same individual incentives to increase effort and optimal effort depends only on their ability level, i.e. $e_{i,0}^{coop} = e_{i,0}^{BL} = e_{i,0}^{comp} = e_{i,0}$. Moreover, the model predicts that the detrimental effect of the cooperative scheme is stronger for high ability individuals while the same type of individuals should exert more effort with respect to the less able individuals in the baseline treatment. Note that our main testable predictions involve the differential changes in effort across treatments and ability levels. Our design allows to measure these changes, as discussed in more detail in section 3.5.1.

We also expect that students assigned to the cooperative treatment will use the chat more frequently and will use it to exchange information. Conversely, students assigned to the competitive treatment should use the chat less frequently and could potentially use it for acts of sabotage, i.e. to suggest the wrong answers. We collected data to check these aspects. Results of our inquiry are discussed in section 3.5.3.

3.5 Results

In this section we first discuss our choice of outcome measure, then present the data and discuss the results of the incentives to compete or to cooperate on information

[14] The ordering holds if the distribution of abilities is the same in the three treatments.

sharing and effort.

3.5.1 Measuring Effort

Our theoretical model predicts that for a given level of ability, there is a weak ordering in the effort exerted by each student i, namely $e_i^{coop} \leq e_i^{BL} < e_i^{comp}$. We thus expect that on average students randomized into the COOPERATIVE treatment exert lower or equal effort than students randomized into the CONTROL treatment whereas students randomized into the COMPETITIVE treatment should exert more effort.[15]

Equation (3.3) in our simple model describes the relationship between expected student's score at each test and effort, namely $s_i = \theta_i e_i \bar{v}$, where s_i is the net score of individual i, θ_i is a measure of individual ability, e_i is the effort exerted, and $\bar{v}$ is the maximum score.

Taking logs and allowing for noise in the way in which effort generates the score, we get

$$y_i = \zeta_i + \epsilon_i + log(\bar{v}) \tag{3.13}$$

where $y_i \equiv log(s_i)$ is the log of the net score of individual i, $\zeta_i \equiv log(e_i)$ is the log of the effort exerted, while $\epsilon_i = log(\theta_i) + \varepsilon_i$ and $E[\epsilon_i] = log(\theta_i)$, i.e. we assume that only the idiosyncratic component ε averages to 0 for any i, while the error ϵ_i has a possibly non-zero mean equal to an individual specific constant.

Our experimental design provides a way to measure effort under weak assumptions. Recall that we observe students' performance in similar tests both before the assignment to the treatments (test 1) and after the exposure to the treatments (test 5). Both these tests are taken individually under all treatments, cover similar topics, share the same structure, have the same number of multiple choice questions.[16] However, by construction, the score in the first test and the effort exerted to pass it cannot be affected by the treatments since both performance and effort are

[15]The ordering holds if the distribution of abilities is the same in the three treatments, which is guaranteed by randomization.

[16]The last test covers a larger set of arguments which includes also those covered by the first and is more closely spaced over time with respect to the other tests. Questions of each test are designed to keep the difficulty constant.

pre-determined with respect to the assignment to the different incentive schemes. Conversely, the score in the last test should reflect changes in effort induced by the treatment. Indeed, moving from equation (3.13) and contrasting the performance in test 5 and 1, we have $y_i - y_0 = \zeta_i - \zeta_{i0} + \varepsilon_i - \varepsilon_{i0}$.[17] It follows that $E[y_i - y_{i0}] = E[\zeta_i - \zeta_{i0}]$, i.e. by looking at the change in the logarithm of score between the first and last test, we measure the change of the logarithm of effort net of the direct effect of any fixed individual specific factor.

Recall that all our treatment conditions have a common individual incentive to increase effort but differ in the incentives to compete or cooperate and only in the baseline students can neither compete, nor cooperate. Following the theoretical predictions of our simple model, we expect an increase in effort in all treatments with respect to a set up where no individual incentives are granted. Our experiment is not designed to estimate this common effect – none of our groups has no individual incentives – but to capture the differential changes induced by the different treatments. The testable prediction of our model involves the differential increase in effort under the cooperative and competitive scheme with respect to the baseline. This weak ordering holds also if we consider $log(e)$, since the logarithm is a monotonic transformation.

To test the theoretical predictions, we first contrast the distribution of effort under the three schemes and check for heterogeneity in the treatment effect over the effort distribution. We then assess the effect on the average change in $log(e)$ and run the following regression

$$E[\zeta_i - \zeta_{i0}] = \beta_0 + \beta_1 Coop + \beta_2 Comp \tag{3.14}$$

where β_0 represents the average change in $log(e)$ under the baseline, β_1 is the average differential change in $log(e)$ under the cooperative scheme with respect to the baseline, and β_2 is the average differential change in $log(e)$ under the competitive scheme with respect to the baseline. The theory predicts $\beta_1 \leq 0$ and $\beta_2 > 0$. There is an additional prediction that $\beta_0 = 0$, i.e. no change in effort under the base-

[17] Note that we cannot use the results at test 2, 3 or 4 to compute our measure of effort: differently from test 5, that is taken individually, in test 2,3,4 paired students can communicate.

line. However, our model does not allow for learning which may occur in practice. Namely, after the first test the score of the students in the baseline improves because they are becoming more familiar with the types of tests and the way the tests are performed in the laboratory. Allowing for learning will not affect our theoretical predictions provided that learning is constant across treatments. If learning occurs in practice, $\beta_0 > 0$.

3.5.2 Data and Descriptive Statistics

Among the 145 students attending the course, 131 applied for participation into the experiment. Our elaborations are based only on the records of the *stayers*, i.e. 114 students who participated to all 5 tests.

We exclude from the elaborations the records of 17 students who missed at least one test: 10 students assigned to the BASELINE treatment, 2 students assigned to the COOPERATIVE treatment and 5 students assigned to the COMPETITIVE treatment (see table 3.7.1 in Appendix). We shall highlight that 6 of these students were late at the 3rd test and were thus excluded from that test. The experimental program is run in z-Tree Fischbacher (2007b): when the test (the experimental session) starts, additional subjects can participate only shutting down and restarting the entire session. Students were informed that not being on time for the test would result in being excluded from the test session. Out of these 17 students, 8 dropped out after the first test: all these students were assigned to the baseline treatment after test 1. When we compare stayers and dropouts in the full sample, we cannot reject the null that drop-outs had a worse performance in the first test.[18]

Once we limit the analysis to the students who participated at all tests, the samples are relatively balanced across treatments with respect to observed and pre-determined characteristics: we do not detect differences in the distribution of the score the first test (score 1) and the average score at previous exams (GPA) be-

[18]There are no significant differences between the subpopulation of excluded students and the stayers in observable and pre-determined characteristics among the students who where assigned to the COOPERATIVE treatment. We do not reject the null of equal means at 1% level -but we reject at 5%- in the subpopulations for the other treatments: students who participated to all tests in the BASELINE and in the COMPETITIVE treatment tend to be those who achieved a higher score in the first test (0.7 points higher than the one for those who dropped out in the BASELINE group and 0.85 points higher than the one for those who dropped out in the COMPETITIVE treatment).

tween any two treatments (BASELINE, COOPERATIVE, COMPETITIVE) at any conventional level of confidence (see Table 3.5.1). Figures 3.7.2 and 3.7.3 in Appendix report the empirical probability distribution of the pre-treatment variables (the score in the first test, and the average mark in previous exams). Table 3.5.1 also reports the mean value of several other individual characteristics, obtained from subjects' answers to the questionnaire and p-values of tests aimed at detecting differences in these characteristics across treatments.[19] In general, the overall sample is well balanced across treatments. There are some exceptions: the frequency of use of e-mail is significantly higher in the BASELINE treatment than in the COMPETITIVE and in the COOPERATIVE treatments. Significant differences emerge also in terms of the education level achieved by the students' fathers (but not mothers).

To detect the role of interactions effect between the treatments and the students' ability, we consider several different proxies for student's ability and include interaction terms in a simple regression. Our favorite proxy to control for student's ability is the GPA: students participating in the experiment are third year students taking exams in the last quarter of the third year; therefore, their academic history can be a reliable proxy of their academic skills. In line with the most recent empirical evidence from Italy (AlmaLaurea, 2009), also in our sample females tend to perform significantly better than males in terms of GPA (Females $= 25.2$, Males $= 24.3$, Wilcoxon test $=$ p-value 0.028). We say an individual is a high ability individual if his/her score on the classification variable is above the median for that variable in the sample.[20]

3.5.3 Communication and treatments.

Students under both treatments' schemes had two ways to communicate: they could send text messages or hints.[21] Messages and hints were limited in two ways. On

[19] We contrasted averages across treatments by means of linear and non linear regressions.

[20] By taking the median as reference for the classification, we guarantee that the two groups have similar size. We checked the robustness of our results to different choices of the threshold for the ability level: we consider the 75th and 66th percentile instead of the median. Results are robust to these changes. Regression results are not reported for brevity but are available from the authors upon request.

[21] The hint consisted in a simple message informing the receiver that the sender believes a certain answer to be the right one. The sender can suggest a different answer with respect to the one actually

		mean			p-values		
characteristic	**pooled**	**baseline (BL)**	**cooperative (COOP)**	**competitive (COMP)**	**BL vs. COOP**	**BL vs. COMP**	**COOP vs. COMP**
GPA	24.8	24.8	24.9	24.8	0.808	0.883	0.928
score 1	1.8	1.8	1.9	1.7	0.520	0.564	0.220
age	21.7	21.6	21.9	21.5	0.280	0.736	0.161
gender (Male)	47.4%	40.5%	51.2%	50.0%	0.346	0.418	0.915
freq. mail^{+}	46.7%	62.9%	43.2%	33.3%	0.098	0.017	0.396
freq. chat^{+}	54.3%	54.3%	51.4%	57.8%	0.803	0.785	0.602
freq. pc^{+}	43.8%	45.7%	40.5%	45.5%	0.658	0.983	0.678
father edu.$^{++}$	31.4%	42.9%	13.5%	39.4%	0.008	0.772	0.017
mother edu.$^{+++}$	29.5%	37.1%	24.3%	27.3%	0.241	0.386	0.778
risk aversion	6.0	6.0	6.1	5.8	0.964	0.497	0.515
risk averse	50.9%	48.6%	53.7%	50%	0.659	0.908	0.749
trust 1	4.9	4.7	4.9	5.0	0.567	0.513	0.744
truster (1)	39.5%	27.0%	43.9%	47.2%	0.124	0.077	0.770
trust 2	3.8	3.7	3.8	4.0	0.863	0.533	0.664
truster (2)	27.2%	21.6%	29.3%	30.6%	0.441	0.386	0.902
Observations	114	37	41	36			

Table 3.5.1: Mean value of individual characteristics, by treatment. An individual is risk averse or truster if his answer on the scale is higher or equal to 6. $^{+}$ Binary indicator for whether chat, pc or e-mail are used frequently, i.e. more than once a day. $^{++}$ Binary indicator for whether the father as high education (i.e. if his qualification is equal to college or higher). $^{+++}$ Binary indicator for whether the father as high education (i.e. if his qualification is equal to college or higher). The significance of differences across treatments is estimated by means of simple linear and non linear regressions (logit) for binary indicators. P-values are reported. Sample statistics on GPA, score 1 and gender refer to 114 individuals; the remaining statistics refer to those students who answer the questionnaire, i.e. 105 students.

Table 3.5.2: Use of the chat

Treatment	Acceptance of the chat	Av. num. of messages	Av. message length
Cooperative	98% of subjects	3 (out of 5)	28 words
Competitive	70% of subjects	0.5 (out of 5)	11 words

the one hand students could not send any information useful to identify themselves (under the threat of exclusion from the test); on the other hand, for each of the 5 questions asked in a test, a student can send and receive only one message of both types. Table 3.5.2 together with Table 3.7.3 in the Appendix report descriptive statistics on the use of chat by subjects. The figures suggest that almost everybody under the COOPERATIVE treatment accepted it,[22] and that the average number of exchanged messages is six times higher than in the COMPETITIVE treatment.

The chat tended to be used more frequently than the hint under both schemes. The content of conversations suggests the chat has been actually used to exchange information. Conversely, the chat was not actively used by students under the COMPETITIVE scheme: they declared to be willing to use the chat but only 0.5 messages were exchanged on average. More importantly, students did not believe in the messages of the partner[23]. Indeed, in some cases the chat has been used to deceive the partner (see Table 3.5.4, and Figure 3.7.6 in Appendix for an illustrative example).

Table 3.5.3 reports descriptive statistics on the number of actions taken by students under each treatment. Sending a text message or giving a hint are actions. Under the COOPERATIVE scheme the average number of actions tends to increase from the first test in couples (test 2) to the last (test 4), changing from nearly 5 to above 6, and the correlation between the number of actions taken in different tests is positive, between 0.34 and 0.53, and decreasing with the lag between tests. Some students under the COOPERATIVE scheme used all the available actions (5 text mes-

selected in the test.

[22] At the beginning of the exam the student must input the registration number and then choose if she wants to use the chat or not.

[23] We do not provide descriptive statistics on the extent of sabotage because these statistics would not be comparable across treatments. Indeed, given the low number of individuals that used the chat under the competitive treatment, we will not get reliable statistics for that group.

Table 3.5.3: Number of actions (i.e. use of chat and use of hints) by round and treatment.

	Cooperative				
	mean	sd	median	min	max
Test 2	5.12	3.36	6	0	10
Test 3	5.80	2.92	7	0	10
Test 4	6.37	2.91	6	0	10
	Competitive				
	mean	sd	median	min	max
Test 2	1.47	2.48	0	0	8
Test 3	1	2.51	0	0	10
Test 4	1.67	2.24	0	0	8

Table 3.5.4: Proportion of cases in which the members of the couple give the same answer.

	Test 2	Test 3	Test 4
Cooperative	56.38%	77.26%	84.78%
Competitive	30.5%	52%	56.84%
Difference	25.88	25.26	27.94

sages and 5 hints) and the median number of action is between 6/7: students tended to use at least one of the two available actions in each question of each test and they often used both. Generally, the text message was sent before the hint, and the time lag between the text message and the hint ranges between 1 and 5 minutes in most questions and tests (see Table 3.7.3 in the Appendix). Conversely, under the COMPETITIVE scheme the median number of actions taken is always 0 and the average number of actions remains relatively stable slightly above 1: students tend to use both the chat and the hint for the same question and only once per test. They also tend to send the text message and the hint almost simultaneously or to send the hint before the text message (see Table 3.7.3 in the Appendix). The correlation between the number of actions taken in subsequent tests is weaker (between 0.17 and 0.36) and tends to increase with the lag between tests. The correlation between the exerted effort and the number of actions is negligible under both schemes.

We consider data on the couples in each test and contrast answers of the members: Table 3.5.4 shows that members of the couples under the COOPERATIVE

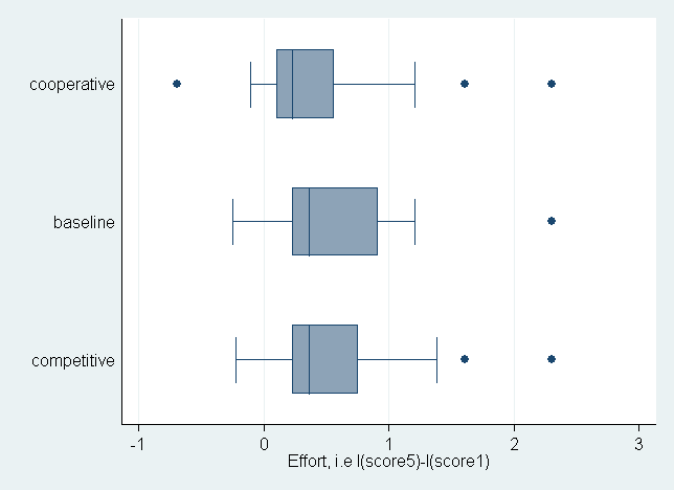
scheme tend to give the same answer much more frequently than their class mates under the COMPETITIVE scheme. The difference is stable across tests and slightly higher than 25%.

We interpret the observed pattern of information exchange across treatments as a positive response to the incentives: students understood the different mechanisms underlying the two different schemes and behaved accordingly as far as exchange of information is concerned.

3.5.4 Treatment effects

Figure 3.5.1 depicts the empirical distribution of effort (i.e. log(net score 5) - log(net score 1)) across treatments. The vertical blue line represents the median of the distribution, the left hinge of the box indicates the 25^{th} percentile, and the right hinge of the box indicates the 75^{th} percentile. Visual inspection suggests that under the COOPERATIVE treatment, subjects perform more poorly respect to the BASELINE treatment, while no sizable differences emerge between the COMPETITIVE and the BASELINE treatments.

Figure 3.5.1: Box-plot showing the distribution of effort across treatments.



Wilcoxon tests do not reject the hypothesis that the distribution of effort is the same across treatments. These tests are not appropriate if we want to establish an ordering across all three treatments. Thus, we also perform a Jonckheere-Terpstra test, a non-parametric test designed to detect alternatives of ordered class differ-

ences.[24]

This test does reject the hypothesis that effort is constant across treatments versus the alternative hypothesis that effort is ordered across treatments according to our main theoretical prediction ($e_i^{coop} \leq e_i^{BL} < e_i^{comp}$) at 10%.
P-values of these tests are reported in Table 3.5.5, together with the mean level of effort in each treatment condition.

Table 3.5.5: Mean level of effort, by gender and treatment.

	pooled	**males**	**females**
Mean effort			
cooperative	0.500	0.377	0.628
baseline	0.583	0.452	0.677
competitive	0.570	0.680	0.459
Wilcoxon tests (p-values)			
base. vs. coop.	0.313	0.135	0.948
base. vs. comp.	0.745	0.442	0.721
coop. vs. comp.	0.190	0.059*	0.713
Jonckheere-Terpstra tests (p-values)			
	0.088*	0.016**	0.624

Legend: One star, two stars, three stars for significant differences at 10%, 5% and 1% level respectively.

It has been pointed out in Section 3.2 that according to the experimental literature, a competitive environment may induce different effects on effort for females and for males. Consistently with these works, we find that the picture indeed changes when we split the sample by gender. Figure 3.5.2 reveals that the treatment effect is substantially different for male and female subjects. The detrimental effect of the COOPERATIVE treatment on effort with respect to the COMPETITIVE treatment only emerges for males, whereas for females no clear treatment effect arises.

One-sided Wilcoxon tests confirms that males' level of effort is significantly lower in the COOPERATIVE treatment than in the COMPETITIVE treatment at 10%

[24]The Jonckheere-Terpstra test is a non-parametric test for more than two independent samples, like the Kruskal-Wallis test. Unlike Kruskal-Wallis, Jonckheere-Terpstra tests for ordered differences between treatments and thus requires an ordinal ranking of the test variable. For a more detailed description of the test, see Hollander and Wolfe (1999). The test is commonly used in experimental economics (Harbring and Irlenbusch 2003; Ferraro and Cummings 2007; Huck, Lunser, and Tyran 2010).

level but no significant difference emerges with respect to the baseline. In contrast, the same test does not reject the hypothesis of equal distribution of effort between any two treatments for the female sample. These tests are not appropriate if we want to establish an ordering across all three treatments. Thus we run the the Jonckheere-Terpstra test for the subsamples of males and females: for the male sample, the test rejects at 5% the null hypothesis that effort is not ordered across treatments against the alternative hypothesis that effort is ordered according to what predicted by the theory; no effect is detected for females. P-values of these tests are reported in Table 3.5.5.

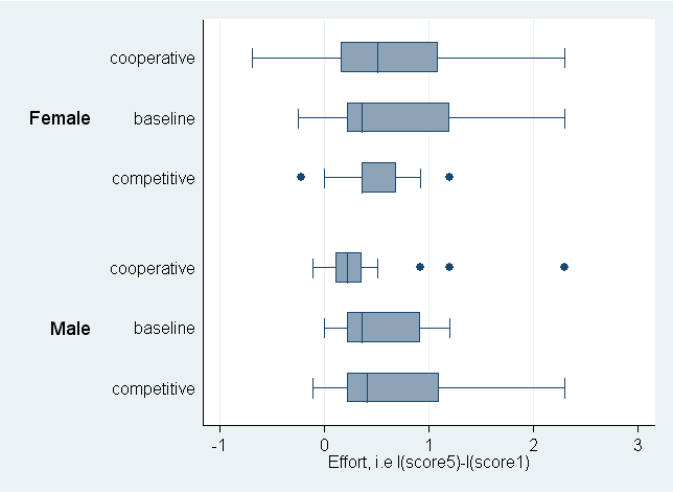


Figure 3.5.2: Box-plot showing the distribution of effort across treatments, by gender.

Our theoretical model predicts heterogeneity in the effect of the incentives' schemes on effort with respect to students' ability, at least for the competitive treatment. To control in a parsimonious way for individual ability, and for other individual characteristics, while assessing the effects of the treatments' scheme on average effort, we use linear regression models.

We run the analysis separately for males and females as previous results suggest that they react differently to incentives.

Table 3.5.6 presents the benchmark results of two baseline specifications for males and females: column (1) and (2) do not allow for heterogeneity in the treatment effects with respect to students' ability while in column (3) and (4) we include

interactions between treatments and the ability indicator based on the average mark at previous exams.[25] All regressions include controls for father education, risk aversion and trust. The top panel of Table 3.5.6 reports coefficients estimates while the bottom reports p-values of both bilateral and unilateral tests: by specifying the direction in which the null hypothesis of no effect is violated (as predicted by theory), we increase the power of the t-test to detect significant deviations.

Results in Table 3.5.6 confirm previous results on the differential effects across treatments: there is evidence of a significant increase in effort under the competitive treatment with respect to the baseline for males but not for females.[26] The effect is statistically distinct from zero at 10% and not-negative at 5%. When we control for ability, we find that : (i) the positive incentive for males is higher for the low ability individuals (still significantly non-negative at 10%) and decreases substantially for high ability individuals; (ii) there is a negative and statistically significant (at 10%) detrimental effect of the cooperative treatment for high ability individuals only. However, the difference in effects of incentives between ability groups is not significant in our sample for the competitive case nor for the cooperative case. The magnitude of the effect ranges from 33% to 49% which is a strong increment of the exerted effort. Notice that this is in line with the findings of Blimpo (2010) who use monetary incentives based on the achievement of a specified score target.

For females, no statistically significant effect can be detected. The pattern of the effect of competitive incentives on effort for females is similar to the one detected for males but in the opposite direction: the point estimate of the effect is negative and, when we control for ability, point estimates of the effect of competitive incentives for females are negative for both low and high ability individuals but less so for high ability individuals.

We detect a significant increase in effort also in the baseline: we attribute this to the fact that students become more familiar with the instruments used for the test

[25] We also run a regression with no covariates: point estimates of the main effects are qualitatively similar to those reported in column (1) and (2) but less precise. The results are not reported for brevity, but are available from the authors upon request. Note that, given the experimental nature of our data, covariates help to improve estimates precision without changing the results substantially, as expected.

[26] Since we include control variates and 9 students do not answer the questionnaire, the sample size relevant for the regressions is 105 instead of 114.

Table 3.5.6: Ordinary Least Squares Estimates of the Treatment Effects: Benchmark Specification. Males and Females.

	(1)	(2)	(3)	(4)
	No heterogeneity with ability		Heterogeneity with ability	
Variables	Males	Females	Males	Females
Constant	0.386**	0.652***	0.239	0.740**
	[0.174]	[0.178]	[0.262]	[0.294]
Cooperative	-0.125	-0.156	0.100	-0.135
	[0.178]	[0.219]	[0.263]	[0.382]
Competitive	0.331*	-0.235	0.492*	-0.363
	[0.189]	[0.211]	[0.272]	[0.359]
Coop · High Ability			-0.486	-0.117
			[0.401]	[0.475]
Comp · High ability			-0.266	0.080
			[0.398]	[0.449]
High ability			0.127	0.073
			[0.280]	[0.319]
High parental education	-0.249	-0.155	-0.317*	-0.219
	[0.151]	[0.185]	[0.173]	[0.210]
Frequent use of e-mail			0.139	-0.224
			[0.161]	[0.206]
Risk averse	0.290*	0.221	0.290*	0.280
	[0.147]	[0.173]	[0.150]	[0.198]
Truster (1)	0.105	0.084	0.136	0.114
	[0.155]	[0.194]	[0.164]	[0.213]
Observations	50	55	50	55
R^2	0.237	0.066	0.311	0.157
P-values for the null of no effect against bilateral or unilateral H_1				
(R)$\equiv H_1$: $\beta > 0$; (L) $\equiv H_1$: $\beta < 0$				
Competitive				
1 sided (R)	0.039**	0.867	0.035**	0.844
2 sided	0.089*	0.266	0.071*	0.312
Cooperative				
1 sided (L)	0.240	0.238	0.649	0.362
2 sided	0.480	0.476	0.703	0.724
Competitive for high ability				
1 sided (R)			0.221	0.837
2 sided			0.441	0.326
Cooperative for high ability				
1 sided (L)			0.092*	0.186
2 sided			0.184	0.373

High parental education is a binary indicator that takes the value 1 if the highest qualification of at least one of the parents of the individual is above high school and 0 otherwise. **Risk averse** is a binary indicator that takes the value 1 if the answer of the individual on the risk aversion scale is above 6 and 0 otherwise. **Truster (1)** is a binary indicator that takes the value 1 if the answer of the individual on the trust 1 scale is above 6 and 0 otherwise. Standard errors in brackets. Three stars, two stars and one star for significant effect at 1%, 5% and 10% level respectively.

(*learning*). Students' ability does not play any role in determining the increase in effort in the baseline. Few regressors are relevant in determining changes in students effort: risk aversion and parental background attract significant coefficients in some specifications, suggesting that individuals who are risk averse tend on average to increase effort, while males with higher socio-economic background (here proxied by highly educated parents) tend to decrease effort, other things equal.

Previous experiments have shown that relevant gender differences emerge in terms of risk aversion, trust and trustworthiness (see Buchan, Croson, and Solnick 2008 and **?**; see also Croson and Gneezy 2009 for an extensive review). These factors could interact with the incentives in different ways for males and females: unfortunately, we do not have enough statistical power to detect these gender specific differential effects in our sample.

3.6 Conclusions

Our study investigates how two alternative incentive schemes affect students' effort, both from a theoretical and from an empirical point of view. To test the theoretical predictions, we run a field experiment in an undergraduate course at the University of Bologna (Italy). We randomly assign students to either a tournament, where coupled students compete to get the reward, a cooperative scheme where information sharing is allowed, or a baseline treatment in which students can neither compete, nor cooperate. Differently from previous studies, none of our treatments involves pecuniary incentives but consists in extra points for their final grade. By doing so, we provide incentives to students in "the same currency" in which they are usually rewarded.

The field-experiment data we collected confirm the theoretical predictions: we observe a weak ordering between the effort exerted by students under the different treatments with students in the competitive treatment exerting on average more effort with respect to students in the baseline and in the cooperative treatment.

We break down our results by gender and show that a significant difference emerges: only males react to incentives to compete while we cannot detect sig-

nificant effect for females. Cooperation seems not to foster effort exertion and no gender effect emerges. In contrast with theoretical predictions we find that students' ability plays little role in determining the effectiveness of the incentives.

Our experimental results suggest that non-pecuniary incentives based on competition have the potential to increase students' effort as pecuniary incentives do (see Blimpo 2010) but potentially at a much lower financial cost (the one of grading more tests per student). In our case competition proves to work on males which is line with findings in several other contexts (see for example Gneezy and Rustichini 2004, Niederle and Vesterlund 2010) where it has been shown that males are more prone to compete with respect to females. The estimated increase in effort induced for males in the competitive treatment ranges from 33% to 49%, meaning that, for example, if a student in the baseline spends 3 afternoons in preparing the test (roughly 10 hours), a student under the competitive scheme will spend one more afternoon. Moreover, highlighting the different effect of incentives to compete depending on gender, we complement the results in Angrist and Lavy (2009) who show that monetary incentives based on absolute performance are more effective for females.

Our study represents a first exploration of the effects of alternative non-monetary incentives based on grading rules on students' performance and effort. These results are relevant for teachers and policy makers who aim at improving the efficiency of the schooling system, since they suggest that pedagogical practices – that can be implemented by faculty members – can increase students' effort, at least in our setting. It would be interesting to extend the inquiry to different samples, to verify whether our result holds for students with different majors (such as literature of philosophy), who are probably less trained to optimization, and for younger students at high school and middle-high school.

3.7 Appendix

3.7.1 Laboratory

Figure 3.7.1: The laboratory arrangement

3.7.2 Additional tables

Table 3.7.1: Descriptive statistics

			Descriptive Statistics- Stayers Predetermined controls		
	assigned	stayers	score 1	exams' avg	score 5
Baseline (control)	47	37	1.80 *(0.81)*	24.76 *(1.8)*	2.91 *(0.24)*
Cooperative	42	41	1.92 *(0.84)*	24.88 *(2.3)*	2.80 *(0.50)*
Competitive	41	36	1.69 *(0.74)*	24.83 *(1.6)*	2.69 *(0.53)*
Full sample	130	114	1.81 *(0.80)*	24.83 *(1.9)*	2.80 *(0.45)*

Score 1: score at the first mock exam. Score 5: score at the last mock exam. Exams' avg: average score at previous exams. Stayers: students who participated to 5 experimental sessions.

In Table 3.7.2, we report the precise definition of questionnaire data used in the analysis.

Table 3.7.2: Description of questionnaire data.

Variable	**Corresponding question**	**Range**	**Coding**
gender	*gender*	0, 1	1 = male
age	*age*	0-100	age in years
freq. mail	*how frequently do you check your e-mail?*	1-5	1="more than once per day" 2="at least once per day"
freq. pc	*how frequently do you use the pc to study/work?*	1-5	3= "at least once per week" 4="less than once per week"
freq. chat	*how frequently do you exchange text messages via chat (msn, facebook, google talk, skype, etc.)?*	1-5	5="Never"
father edu.	*please, indicate the education level achieved by your father*	1-5	1="junior high school" 2="high school"
mother edu.	*please, indicate the education level achieved by your mother*	1-5	3="bachelor" 4="master" 5="Ph.D."
risk aversion	*I would describe myself as a risk-averse person.*	1-10	1="fully agree" 10="fully disagree"
trust 1	*Do you think that most people try to take advantage of you if they got a chance or would they try to be fair?*	1-10	1="people would try to take advantage" 10="people would try to be fair"
trust 2	*Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?*	1-10	1="you can never be too careful" 10="most people can be trusted"

Table 3.7.3: Descriptive statistics -mean, [median] and (standard deviation)- on lag between the use of chat and use of hints, by treatment and round. Questions 1-5

Lag & proportion of user of both chat and hint (seconds). Test 2.

Treatment	Question 1	Question 2	Question 3	Question 4	Question 5
Cooperative	107.6 [21.7] (449.0)	322.8 [252.2] (378.3)	58.6 [6.3] (565.9)	151.0 [114.1] (513.5)	54.2 [6.4] (495.6)
Users (count)	14	12	11	13	16
Users (%)	35.0%	30.0%	27.5%	32.5%	40.0%
Competitive	80.6 [54.1] (67.9)	-123.3 [-123.3] (131.1)	130.4 [23.9] (1113.5)	13.3 [13.3] (n.a.)	25.3 [16.9] (21.3)
Users (count)	3	2	3	1	3
Users (%)	8.8%	5.9%	8.8%	2.3%	8.8%

Lag & proportion of user of both chat and hint (seconds). Test 3.

Treatment	Question 1	Question 2	Question 3	Question 4	Question 5
Cooperative	492.5 [374.8] (693.1)	496.2 [368.1] (577.4)	-56.0 [-3.2] (423.0)	167.6 [71.7] (427.3)	76.4 [5.2] (303.5)
Users (count)	14	16	14	17	15
Users (%)	35.0%	40.0%	35.0%	42.5%	37.5%
Competitive	73.8 [7.6] (131.3)	-164.2 [-164.2] (951.9)	720.1 [720.1] (223.6)	-194.0 [165.4] (380.5)	97.0 [97.0] (123.7)
Users (count)	3	2	2	4	2
Users (%)	8.8%	5.9%	5.9%	11.7%	5.9%

Lag & proportion of user of both chat and hint (seconds). Test 4.

Treatment	Question 1	Question 2	Question 3	Question 4	Question 5
Cooperative	146.9 [55.9] (482.1)	119.8 [20.8] (342.8)	-15.8 [-3.3] (522.4)	169.7 [40.8] (355.0)	95.0 [4.1] (240.2)
Users (count)	14	17	17	17	22
Users (%)	35.0%	42.5%	42.5%	42.5%	55.0%
Competitive	194.3 [12.3] (365.8)	8.5 [8.5] (n.a.)	180.1 [2.1] (314.1)	458.8 [72.6] (721.4)	322.3 [322.2] (449.4)
Users (count)	4	1	3	3	2
Users (%)	11.7%	2.9%	8.8%	8.8%	5.9%

3.7.3 Additional figures

Figure 3.7.2: Empirical probability distribution of score 1 by treatment

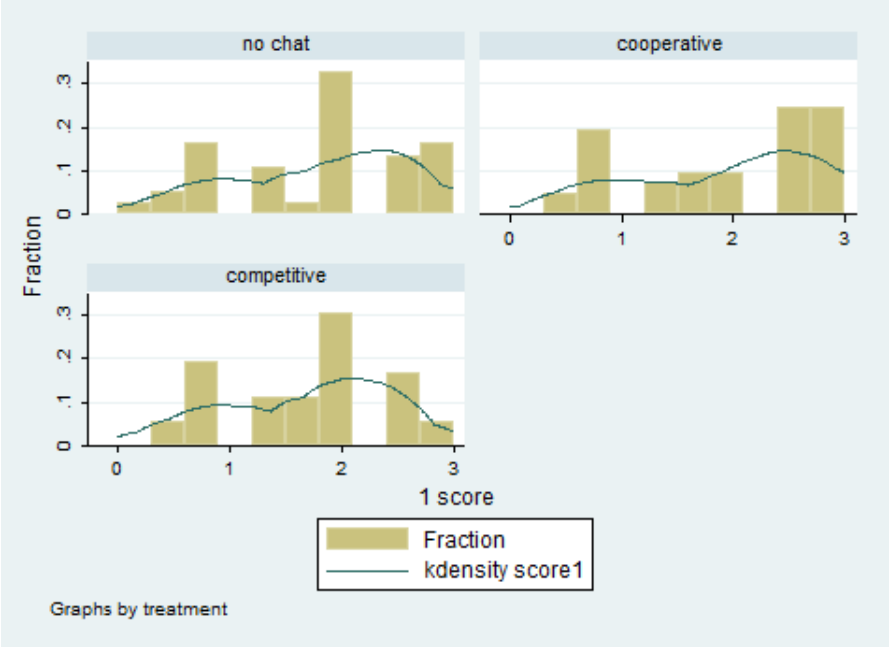


Figure 3.7.3: Empirical probability distribution of average score at previous exams by treatment

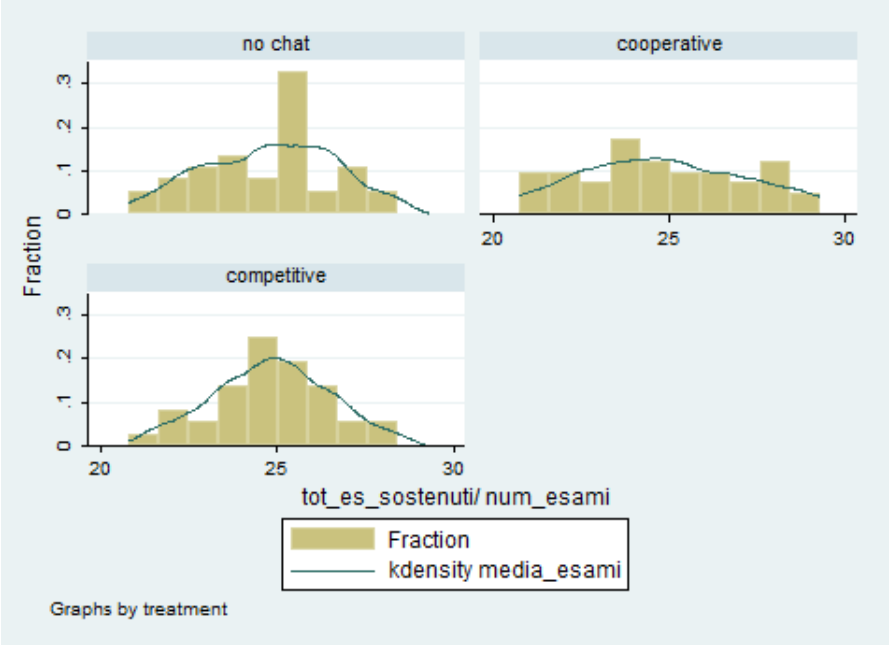
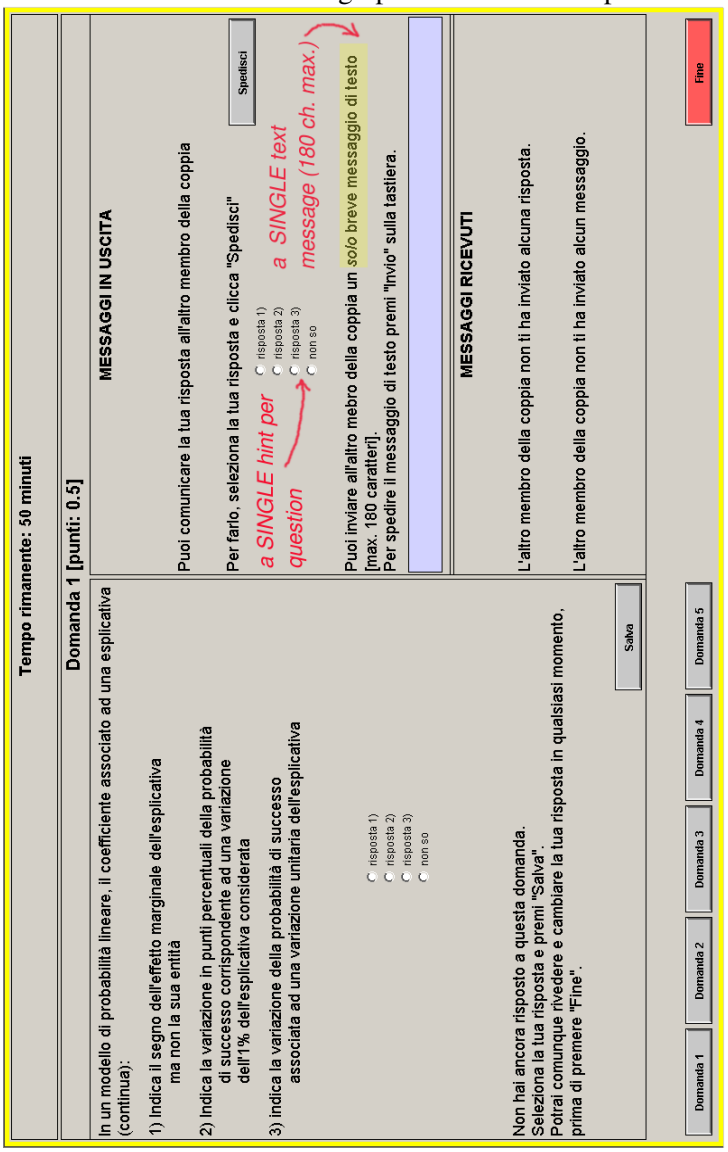


Figure 3.7.4: Screen-shot of the graphical interface for partial exams.



3.7.4 Examples of chat messages

Figure 3.7.5: Example of use of the chat under the cooperative scheme

A: Come on! Tell me which answers do you need. If you don't get 1.5 points, we will lose the bonus.

B: In my opinion the right one is the 2nd.
C (replies): OK! I trust you.

Figure 3.7.6: Example of use of the chat under the competitive scheme

D: In this case the 4th is the best answer!
E (replies): Why do you pass me this solution? Are you trying to screw me?

F: I know that you are going to pass me the wrong answers.

G: I'm not sure...probably the right answer is the 1st *[she choses the 3rd]*

Chapter 4

The Credit Crunch and Fertility in the United Kingdom

ABSTRACT

This paper aims to test the apparently counterintuitive demography hypothesis that predicts an increase of the birth rate during an economic crisis. For a working woman, the switch from a period of economic growth to a recession causes a direct decrease in the maternity leave opportunity cost. From an economic point of view, the change in the relative costs provides a rational incentive to plan childbearing during a recession period optimizing the career-fertility program. In order to investigate empirically this optimizing behaviour, we focus on the recent UK credit crunch. Adopting a Difference-in-Differences identification strategy, we analyze a salient industry division of the UK labour market. We investigate whether employed woman directly exposed to the adverse economic condition are more prone to undertake childbearing that comparable colleagues only indirectly affected by the crisis. The data analysis reveals positive evidence in support of the conjectured career-fertility optimizing behaviour.

Keywords: Fertility, Opportunity Cost, Economic Crisis.

JEL code: J13.

4.1 Introduction

Past research has revealed that economic recessions can directly affect the demography of a country, by influencing the dynamics of family formation, fertility, migration and mortality (for an interesting review of the literature, see Sobotka, Skirbekk, and Philipov 2010).

Focusing on the effects generated by economic crises in terms of changes in fertility outcomes, the existing literature has mainly revealed pro-cyclical negative birth patterns with respect to the negative variation of the main macroeconomic aggregates (Bengtsson, Campbell, and Lee 2009; Lee 1990; Tzannatos and Symons 1989; Van Bavel 2001 Macunovich 1996).

Only Butz and Ward (1979a) and Butz and Ward (1979b) have provided some tentative empirical evidence in support of an alternative counter-cyclical theory. In particular, they described a mechanism that could lead to a positive demographic spillover in periods of economic downturn. The basic intuition is the following: for working women, economically good times would be the relatively most expensive periods to have a maternity leave since be good job market conditions. Following this reasoning, seasons of growth would be characterized by low birth rates, viceversa in case of less favorable economic conditions since the adverse job prospects. In late seventies they conjectured that this effect would have become relevant in the future since the increasing involvement of the female population in the labour market.

The recent financial crisis represents a nice environment to test this conjecture as it occurs in a time where women are much more involved in the labour market than in the past. The rising educational levels joint to the availability of reliable contraceptive methods, allow them to rationally plan childbearing according to both their career and family plans. This may suggest that the current crisis is likely to differ from the previous ones (Sobotka, Skirbekk, and Philipov 2010) and its effects could differ from those observed in the past.

On one side, the higher female labour participation may suggest that the demographic effects of the ongoing economic crisis could be even more negatively

marked than in the past as it directly impacts on both the main household members' earnings and job prospects.

On the other side, in line with Butz and Ward (1979a) and Butz and Ward (1979b), since the opportunity cost of the maternity leave is lower during an economic downturn than during a season of growth, it could be rational to plan childbearing during the crisis optimizing in this way the career-fertility process. This rational behaviour would lead to positive effects in terms of fertility rate during the economic downturn.

This counterintuitive side effect of the economic crisis has recently received considerable attention and anecdotal support from media of different countries. It is especially true for countries hardly hit by the crisis such as the United Kingdom, Ireland and Iceland[1].

Despite the anecdotal evidence, a significant positive effect does not emerge from yearly macro census statistics. This does not mean that a local positive effect does not exist at all. At aggregate level, it could be that complementary negative pro-cyclical effects compensate for it.

Exploiting a micro-level perspective, the aim of this paper consists in providing a more rigorous assessment of the hypnotized counter-cyclical optimizing behaviour.

The United Kingdom represents a unique context where to reveal the fertility effects of the current economic crisis. The first reason lies in the country's exposure to the 2008 credit crunch: the economy was highly affected by the credit crunch since the prominent role of the financial industry. In addition to that, the UK Quarterly Labour Force Survey offers detailed households and individual labour market

[1]Some columnists have hypothesized how the positive fertility effect could be driven by a more simply house-production story. Due to the economic crisis, people are less willing to spend money for leisure services purchased on the market such as dining, concerts, cinemas etc, therefore couples switch to more cheap home-produced "entreating activities". This paper focuses its attention only on the labour market transmission channel.
UK: http://dl.dropbox.com/u/8796618/aUK.london.hamhigh.pdf
UK: http://dl.dropbox.com/u/8796618/aUK.telegraph.uk.pdf
Iceland: http://dl.dropbox.com/u/8796618/aIC.guardian.uk.pdf
Ireland: http://dl.dropbox.com/u/8796618/aIE.indipendent.ie.pdf
Russia: http://dl.dropbox.com/u/8796618/aRU.moscow.news.pdf

data at a micro-level.

In order to identify our effect of interest, we exploit a Difference-in-Differences identification strategy. We focus our analysis on two well defined and comparable populations of female workers in fertile age working the same industry (Health and Social Care) where only one of the two groups is actually exposed to the potential negative effects of the crisis (treated group) therefore subjects to a change in the relative cost of the maternity leave. Following this framework, the exposure to the exogenous shock provided by the credit crunch, in terms of higher job market uncertainty that translates in a lower maternity leave opportunity cost, represents the *treatment*. Private employed workers constitute the *treated group* directly exposed to the economic downturn. Public employed workers represent the unaffected *control group* since their employment contracts are highly protected and stable.

From our analysis emerges evidence supporting the actual existence of a such countercyclical positive fertility effect. In 3 out of 4 post-crisis quarters, we identify sizable and significative – both from a demographic and statistical point of view – spikes in the birth rate in the treated population that faces a direct decrease in the maternity leave opportunity cost.

The paper is structured as follows. Section 2 describes the empirical framework. In section 3 the data are presented. Section 4 reports the data analysis. In section 5 we discuss the results.

4.2 Empirical Framework

In this study we are interested to identify the hypnotized positive effect in terms of higher fertility output triggered by an economic downturn. A negative performance of the economic system brings negative forecasts and – as a direct consequence – an increase of uncertainty in the labour market. The uncertainty over the future (e.g. unemployment or lower expectations concerning future income) contributes to decrease the relative cost of a maternity leave during a recession than during a more favorable season of economic expansion. It is well know that workers employed in

the public sector can enjoy positions and labour contracts pretty stable while comparable positions in the private sector are much more flexible and cycle-dependent. When the economy crashes, workers engaged in the private sector suffer severe negative consequences. This consideration lead us to claim that private sector employees are directly exposed to all the negative effects generated by the crisis while public employees are only indirectly affected by the economic crisis.

According to this analysis, only workers engaged in the private sector can experience a real decrease in the childbearing opportunity cost when the economy goes down. We exploit the distinction between public and private sector employees, engaged in the very same industry, to define a Difference-in-Differences measurement strategy. The exogenous change in the maternity leave opportunity cost, via the increase in job uncertainty caused by the economic crisis, represents the treatment that affects only the private sector employees (treated group) while the crisis leaves unaffected the group composed by the public sector employees (control group). Comparing the net relative differential changes of the fertility outcomes in the two different groups pre and post the economic shock, we will be able to isolate the potential positive effect in terms of women's childbearing propensity in the private sector.

$Y_{d,t}$ binary outcome: 1 =new childbearing, 0 =no childbearing

$D = \{0, 1\}$ treatment-variable: 1 =treated (private sector), 0 =control (public sector)

$T = \{0, 1\}$ time-variable: 1 =post-treatment (post-crisis), 0 =pre-treatment (pre-crisis)

$Y_{0,0}$ public sector worker's outcome realized before the credit crunch

$Y_{1,0}$ private sector worker's outcome realized before the credit crunch

$Y_{0,1}$ public sector worker's outcome realized after the credit crunch

$Y_{1,1}$ private sector worker's outcome realized after the credit crunch

$\delta = (\overline{Y}_{1,1} - \overline{Y}_{1,0}) - (\overline{Y}_{1,0} - \overline{Y}_{0,0})$

$Y_{idt} = \beta_0 + \alpha T + \gamma D + \delta(T \times D)$

4.3 Data

The empirical exercise is based on UK Quarterly Labour Force Survey data.

To observe the fertility behavior of the two reference groups pre and post-crisis, we focus on the time window between 2007 and 2010. It is composed by 16 consecutive quarterly waves.

The UK LFS is a quarterly rotating panel survey based 60,000 households per wave corresponding to 65,000,000 weighted observations . It is conducted on a stratified random sample and carried out by the UK Office of National Statistics (ONS). Each household is followed for 5 consecutive waves (15 months), collecting a wide range of individual data such as demographics, employment status, wages, job characteristics, education, health conditions. Thanks to a key matching variable at household level provided on purpose by the ONS[2], we are able to retrieve the complete structure of each single household in every wave. This allow us to identify the birth events and to match the newborns with the parents. In addition to that to that, knowing the complete household's structure and its relevant features (e.g. partner's details, number of kids etc.) it is possible check whether the two reference populations (treated and controls) are balanced in terms of pre-treatment characteristics.

Following the Difference-in-Differences framework, in order to have a clear identification – in addition to the common trend assumption – it is necessary to prove that the assignment to the treatment (exposure to the crisis/private sector employment) is a random process. Please note that the treatment is represented by the interaction between the time dimension (post vs. pre exogenous shock/credit crunch period) and the fact of belonging to the treated group (private vs. public sector). As far it concerns the time dimension, since the crisis was unpredicted by the economic/politic establishment and by media, it is pretty unlikely that workers engaged in the private sector migrated strategically to the public one in order to reduce the job uncertainty caused by the forthcoming economic crunch.

A bit more demanding is to argue that workers engaged in the public sector are

[2]We acknowledge the kind assistance and cooperation provided by Colin Hewat.

good counterfactuals for the ones engaged in the private sector. Firstly, in the real word it is difficult to observe an industry sector having comparable size and relevance both in the private and public sector. Secondly, often the work arrangement for analogous job positions are very different in the two different sectors. This could constitute a source of self-section into the two different sectors according to individual-specific characteristics that might be non-orthogonal to the fertility preferences. To address this issue we restrict on purpose our attention to the "Health and Social Care" (HSC) industry [3] as in the UK labour market, this industry fits particularly well the needs of our analysis.

We further restrict our focus on coupled women(married or cohabiting, we drop singles and women living at parents' place) since partnership stability assures for a higher degree in terms of rational/conscious reproductive choices. Give these restrictions, the pool of observations in each wave results to be around 1,000,000 frequency weighted units.

Summarizing, our reference population is composed by age-fertile coupled women employed in the "Health and Social Care" sector by public or private providers.

This specific sample is quite well balanced between public and private sector. The 56% of the HSC manpower is employed in the public sector while the 44% works for private provides. In both the cases, the share of female employees is the 78%. Most of the private sector employees are hired by HSC providers working in close partnership with the National Health System or under its national monitoring activity. For these reasons, the educational and professional requirements are the same in both the sectors.

Table 4.6.3 reports a battery of control characteristics for the two sectors both for the pre and post crisis periods. As reference period for the pre-crisis scenario we focus on the 2nd quarter 2007 while the 2nd quarter 2008 is considered as post-crisis reference period. On average, both in the pre and post crisis periods,

[3] It is coded using the UK Standard Industrial Classification Of Economic Activities (SIC(92)N:85(54)). It is a hierarchical 5-digit Industry Classifications code that conforms to the European Community Classification of Economic Activities (NACE) Version 1 codes). This industry classification includes: (i) Human Health Activities like hospitals, nursing homes, dental practices, opticians, etc.; (ii) Social Work Activities, such as social work services with and without accommodation, as detailed above.

fertile women working in the HSC public sector result to be 1 year older than their colleagues working in the private sector (36 vs. 35 years old). From a statistical point of view, the difference is significative but its low size in real terms does not lead us to claim that this difference strongly qualify the to samples. Independently by the the sector of employment and the reference period (pre-post crisis), 90% of the workers belong to the British ethnic group and on average each woman has 1.5 children.

On the contrary it is necessary to notice how workers in the private sector gain significantly less generous wages than their colleagues engaged in the public one. On average, public employees enjoy a +25% wage premium compared to the private sector ones. At a first look, this difference appers as a very strong difference. Nevertheless, following the analysis put forth by Rutherford (2010) that focuses on the UK HSC industry, it is possible to learn how the wage gap observed between public and private HSC sectors results to be less severe in real terms. He shows that higher wages in the public sector compensate for frequent unpaid overtime. The required flexibility is not explicitly contracted, but forms a part of the labour contract that is enforced through organizational norms.

Both in the private and public sector, workers provide their regular service for about 30.5 hours per week (part-time contracts are frequent). After the credit crunch, both public and private HSC workers increase by 30 minutes their average weekly working time.

Since the childbearing decision is the outcome of a common decision made by both the partners, it is important to verify that the two groups of women are balanced also in terms of partners' characteristics.

Partners of women working in the HSC public sector are on average 6 months older that the average partner of a private sector employee. It is true for both pre and post crisis reference periods. The difference is weakly significant (at 5% level in the pre-crisis period, at 10% level in the post-crisis period) and its low size in real terms does not lead us to claim that this difference qualify the to samples. This small difference in the average partners' age is in line with the 1 year age difference observed between public and private HSC sector female workers. Partners mainly

belong to the British ethnic group, and the share of 88% is constant across reference periods and sectors of emplyment.

The hourly wage gained by partners of women working in the HSC public sector is on average more generous (+1.5 £) – in particular after the crisis (+2.3 £)– compared to the average hourly salary received by partners of women working in the HSC private sector. The wage gap results to be significant from a statistical point of view but its size in real terms is rather small.

Partners' labour supply is pretty stable across groups. They provide around 42 hours per week. On average, partners of women working in the public sector tend to work 20 minutes less per week than partners of women employed by private providers. After the crisis, the labour supply of both the groups decreases by 24 minutes per week. Despite the statistical significance of these differences, their small size (about 5 minutes per day) suggests a weak economic relevance.

In the pre-crisis period, 93% of women working in the HSC public sector are coupled with employed partners of which the 27% of them works in the public sector as well. In the same reference period, 90% of women working in the HSC private sector are coupled with employed partners of which 18% of them works in the public sector. In the post-crisis period, the share of employed partners decresed by 3 p.p. in both the gruoups.

4.3.1 Timing

According to the financial reporters, the UK credit crunch took "officially" place on 9th August 2007 (3rd quarter 2007) when the European Central Bank pumped 95bn €(63bn £) into the banking market to try to improve liquidity[4]. Despite this precise date, the financial crisis actually affected the UK real economy with some lag, between the 1st and the 2nd quarter 2008[5]. According to that, from a demographical point of view, the first effects generated by the crisis (in terms of newborns) should be observed starting from the 1st quarter 2009. As pointed out by Sobotka, Skirbekk, and Philipov (2010) some time lag should be expected even if couples

[4](i) http://dl.dropbox.com/u/8796618/aUK.creditcrunch3.guardian.pdf
[5](ii) http://dl.dropbox.com/u/8796618/aUK.creditcrunch1.bbc.pdf

reacted very rapidly in adjusting their plans according to the changed economic conditions, considering the time between the initiation of pregnancy attempts and achieving a conception and between conception and childbirth.

For this reason, our analysis considers all the births registered in the 2007 and the 2008 as pre-treatment outcomes while births registered during the 2009 (and the 2010) are considered as outcomes affected by the crisis.

Concerning the different degrees in terms of job market uncertainty in the public and private sector, Table (4.6.1) provides chronological evidence on this point. Consistently to our reference population, we focus on redundancies affecting active workers aged between 18 and 46. The pre-crisis redundancy ratio registered in the year 2007 for the private sector is 11.4 [6] while the ratio in the public sector is 2.5.

After the credit crunch, the ratio for the private sector increases to 16.2 in the year 2008 and up to 23.3 in the year 2009. For the public sector, the redundancy ratio tends to be pretty stable after the credit crunch. It decreases to 1 in the year 2008 and to 2 in the year 2009.

These figures suggest that, after the credit crunch, private sector employees are exposed to a significantly higher job market uncertainty with respect to their own pre-crisis situation and the post-crisis scenario faced by public sector employees.

4.4 Analysis

As a first step, we proceed computing the "group-specific birth rate" for both public and private reference populations for all the 16 waves (2007 - 2010).

The group-specific birth rate is defined as the number of newborns every 1,000 members of the reference population (figures reported in Table 4.6.2).

$$\frac{\text{number of newborns in the group}}{\text{total group members}} \times 1000$$

[6] The yearly redundancy ratio is based on the summatory of all the redundancies registered in each quarter divided by the number of employed workers in previous quarter, multiplied by 1,000.

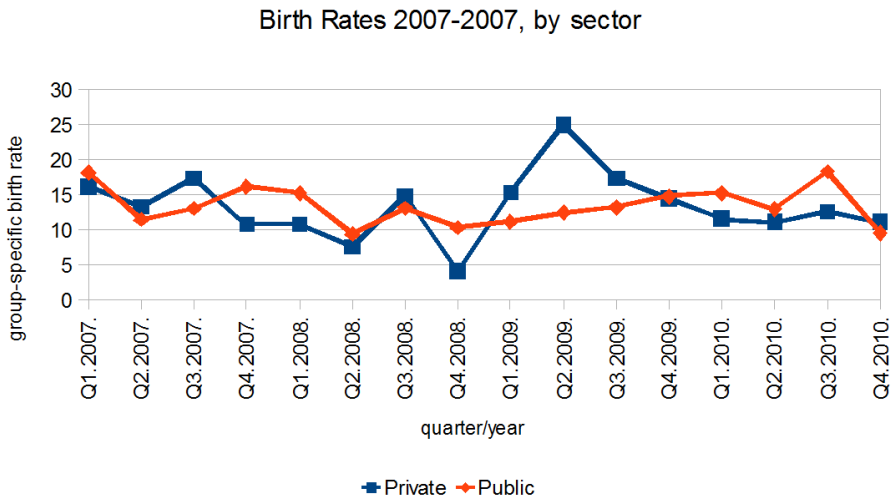


Figure 4.4.1: General birth rate trend, by sector

4.4.1 Common trend assumption

To verify the pre-treatment (2007-2008) common trend assumption in terms of birth ratios in the two different groups, Figure 4.4.2 provides 4 different scatter plots grouped by quarters to take into account seasonality.

From a qualitative visual inspection of the plots, it is possible to verify that for all the 4 quarters and in both the pre-treatment period 2007-2008, the common trend assumption is satisfied. Keeping constant the season, the group-specific birth ratios in both private and in public employed populations co-move proportionally and in the same directions for the 2007 and the 2008. This means that the within seasons birth ratio differences are constant between sectors for the 2007 and the 2008. To provide a further statistical test for common trend assumption, we run a Difference-in-Differences estimation to check whether the deviations from exact common trend have any meaning from a demographic point view (estimates reported in Table 4.6.4) in terms of childbearing probability variation.

The outcome variable is Y_{idt}. It is a dummy equal to 1 in case of finalized pregnancy (newborn) for woman i working in sector D (1: private; 0: public) in period T (1: Qt.2008; 0: Qt.2007), 0 otherwise. In particular, we are interested to assess the net relative change in the childbearing probability, for each quarter in the 2008 compared its correspondent quarter in the 2007, for private sector workers

with respect to public sector workers.

This net relative effect is captured by the interaction term $\delta(T \times D)$ that characterizes the following Difference-in-Differences estimation.

$$Y_{idt} = \beta_0 + \alpha T + \gamma D + \delta(T \times D)$$

Over the pre-crisis period, the relative differential changes in childbearing probability, between sectors and within each corresponding quarter of the 2007 and the 2008, result to be negligible. The size of the deviations from the perfect trend ranges between 0.02 and 0.03 p.p.

The negligible magnitudes of such differences provide evidence in support of the common trend assumption.

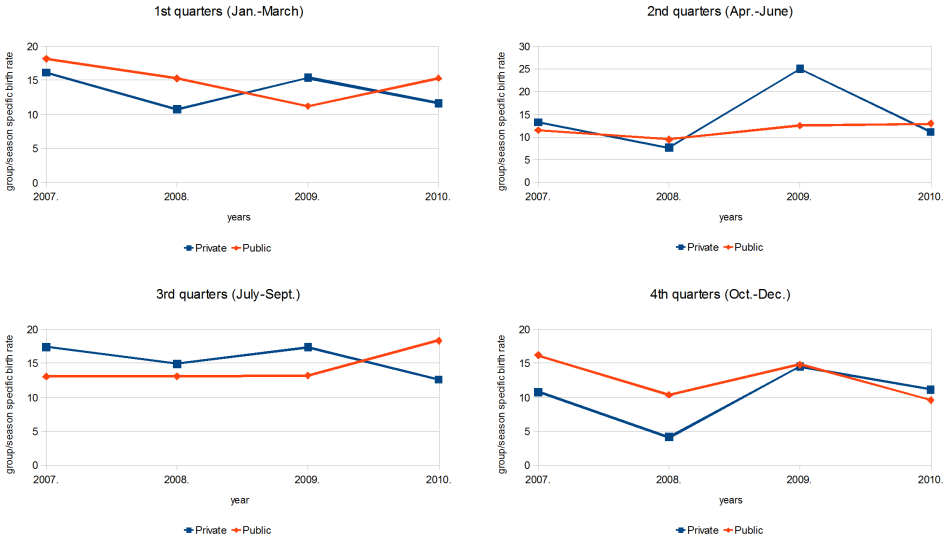


Figure 4.4.2: Seasons-specific trends, by sector

4.4.2 1st Quarter: post-treatment analysis

Focusing our attention on the post-treatment period, Graph 4.4.3 shows that starting from the 1st quarter 2009 the common trend pattern changed. In the 1st quarters of 2007 and 2008 the birth ratio in the public HSC is higher than in the private one, 18.17 vs. 16.15 in 2007 and 15.25 vs. 10.76 in 2008, (see also Table 4.6.2). During the 1st quarter 2009 the scenario is reversed. After the crisis we observe

15.37 births every 1,000 fertile woman in the private sector and 11.20 in the public one. This relative change is consistent with the hypnotized opportunity-cost theory. Since the realization of the economic crisis and the consequent higher job market uncertainty in the private sector, the cost of a maternity leave decreases and workers rationally react to the change in the relative prices.

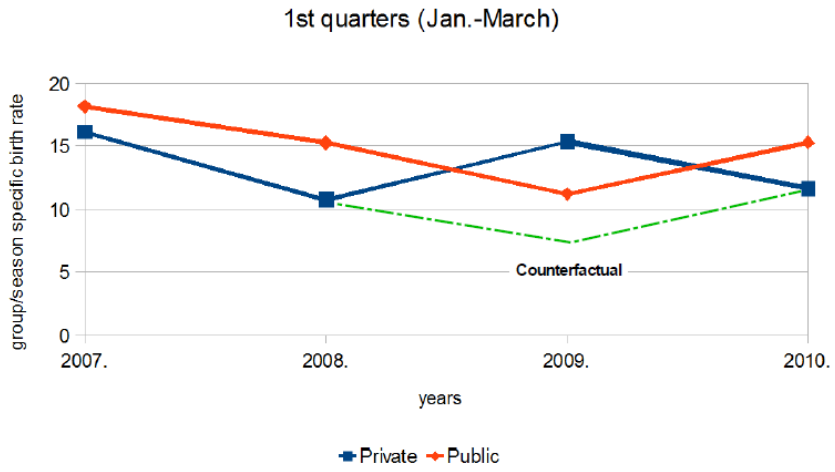


Figure 4.4.3: 1st quarters analysis (winter season - Jan.-March)

To assess the net relative magnitude of the effect, we run a Difference-in-Differences estimation. The outcome variable is Y_{idt}. It is a dummy equal to 1 in case of finalized pregnancy (newborn) for woman i working in sector D (1: private; 0: public) in period T (1: post-crisis exposure; 0: pre-crisis), 0 otherwise. In particular, we are interested in observe the net relative change in childbearing probability for a representative woman employed in the private sector being exposed to the crisis. This net effect is captured by the interaction term $\delta(T \times D)$ that characterizes the following Difference-in-Differences estimation.

$$Y_{idt} = \beta_0 + \alpha T + \gamma D + \delta(T \times D)$$

In Table 4.6.5 we report the Difference-in-Differences interaction term estimates for Q1.2009 w.r.t. Q1.2008. We present both a Probit[7] and LPM[8].

The interaction term results to be highly statistically significant at 1% level and sizable from a demographic point of view. In the Probit specification, the childbearing probability for a representative women engaged in the private sectors being exposed to the crisis, increases by 1.02 p.p. Since the predicted baseline probability is 1.29 % an increase of 1.02 p.p. represents a relevant change. The LPM delivers a qualitatively similar result predicting an increase of 0.86 p.p.

4.4.3 2nd Quarter: post-treatment analysis

Moving to the analysis of the 2nd quarter 2009, from Graph 4.4.4 we can easily detect a sharp discontinuity that characterized the first post-treatment spring quarter. In 2007 and 2008 the birth rates in the public and private HSC populations are essentially overlapped, 13.27 vs. 11.49 in 2007 and 7.65 vs. 9.46 in 2008, respectively (see Table 4.6.2). During the 2nd quarter 2009 data reveal a huge spike equal to a births ratio of 25.00 in the private sector while the ratio for the public one remains essentially stable at a level of 12.51 as registered in the year before. Also in this case, the huge boost in the HSC private sector birth rate is consistent with the hypnotized opportunity-cost theory.

In Table 4.6.6 we report the Difference-in-Differences interaction term estimates for Q2.2009 w.r.t Q2.2008. The interaction results to be highly statistically significant at 1% level and extremly significant from a demographic point of view. In the Probit specification, the childbearing probability for a representative women working in the private sector being exposed to the crisis, increases by 1.45 p.p. Since the predicted baseline probability is 1.21 % an increase of 1.45 p.p. represents a strong

[7] According to Puhan (2012) and differently from Ai and Norton (2003), we interpret the delivered Difference-in-Difference interaction term as proper treatment effect.

[8] Following the approach put forth by Abrevaya and Hamermesh (2012).

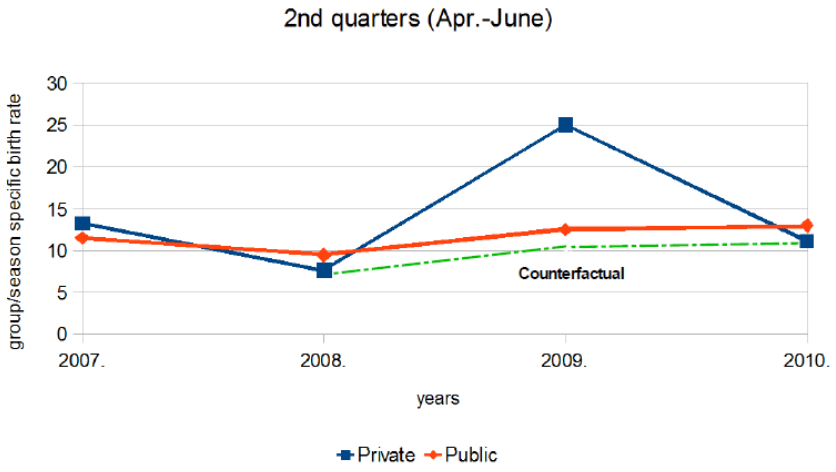


Figure 4.4.4: 2st quarters analysis (spring season - Apr.-June)

change. The LPM delivers an analogous result predicting an increase of 1.45 p.p.

4.4.4 3dr Quarter: post-treatment analysis

Focusing on the analysis of the 3rd quarters (summer seasons), looking at Graph 4.4.5 we can observe how the birth rates both in the public and in private employed reference populations are essentially stable across 2007, 2008 and 2009 – 13.04 vs. 17.38 in 2007; 13.09 vs. 14.91 in 2008; 13.20 vs. 17.34 in 2009 – (see Table 4.6.2) .

In Table 4.6.7 we report the Difference-in-Differences interaction term estimates for Q3.2009 w.r.t Q3.2008. The interaction results to be highly statistically significant at 1% level, nevertheless its magnitude has negligible meaning in demographic terms. In the Probit specification, the childbearing probability for women working in the private sectors being exposed to the crisis, increases by 0.02 p.p. Since the predicted baseline probability is 1.43 % an increase of 0.02 p.p. represents a non-effect. The LPM delivers a result predicting an increase by 0.02 p.p. confirming the the smoothness of the trends visualized in Graph 4.4.5. Therefore,

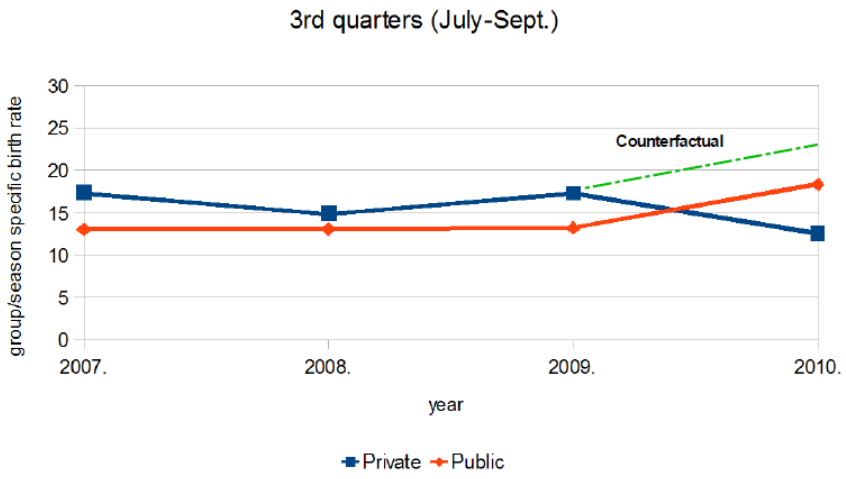


Figure 4.4.5: 3rd quarters analysis (summer season - July.-Sept.)

in the 3rd quarted 2009 it can not be observed any fertility spillover in the private sector workforce.

4.4.5 4th Quarter: post-treatment analysis

Moving to the analysis of the 4th quarters, from a visual inspection of Graph 4.4.6 we can easily detect a strong discontinuity that characterized the first post-treatment fall quarter. In the 2007 and 2008 the birth rate trends of both public and private HCS employees are essentially parallel and dominate by the public sector – 16.17 vs. 10.79 in 2007; 10.45 vs. 4.14 in 2008 – (see Table 4.6.2). During the 4th quarter 2009 data describe a boost in the private sector birth rate up to a ratio of 14.80 that results to be slightly higher than the 14.56 ratio registered in the public sector populations. Also in this case, the relative boost in the HSC private sector birth rate is consistent with the conjectured opportunity-cost story.

In Table 4.6.8 we report the Difference-in-Differences interaction term for Q4.2009 w.r.t. Q4.2008. The interaction results to be highly statistically significant at 1% level and sizable from a demographic point of view. In the Probit specification, the

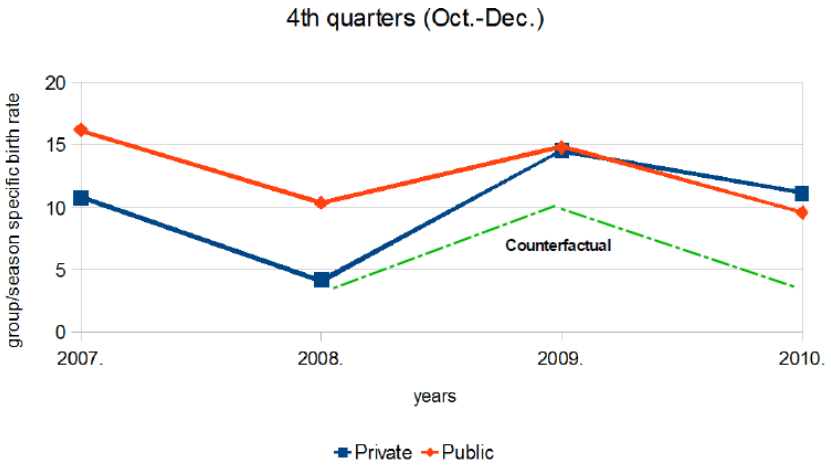


Figure 4.4.6: 4st quarters analysis (autumn season - Oct.-Dec.)

childbearing probability for a representative women working in the private sector and exposed to the crisis, increases by 1.08 p.p. Since the predicted baseline probability is 1.02% an increase of 1.08 p.p. represents a significant change. The LPM delivers a qualitativelly less sharp but still remarcable effect of 0.61 p.p.

4.5 Discussion

We analyzed two comparable sets of workers engaged in the public and in the private sector, respectively.

Before the economic crisis, the childbearing propensity in the two groups results to be characterize by a common pattern. After the economic shock represented by the crisis, this regularity suddenly changed.

From an economic point of view, the crisis affects directly only the employment outlook in the private sector while public employees do not have any real reason to fear the risk of falling into unemployment. Since the two groups are composed by workers having similar characteristics, we claim that the change in the birth rate pattern is caused by the only variable that has changed in the two groups since the credit crunch occurred. This variable is represented by the increasing job market uncertainty. The employment uncertainty affects substantially private

sector workers and leaves unaffected public sector employees. According to that, we have good reasons to claim that the change in the common trend is due to the change in the fertility propensity of private sector employees. The change in the fertility preferences for private employed women is triggered by the lower relative cost of the maternity leave caused by the economic crisis. Viceversa, since we have no reasons to conjecture any real change into the more stable public sector labour market, we sustain that the public sector birth rates observed in the second period of analysis (post-crisis, 2009) are not affected by the crisis and they are determined by the standard demographic process. This framework of analysis lead us to claim that after the credit crunch, in 3 out of 4 quarters of the 2009, we properly identified positive and significant birth rate spillovers in the exposed-to-the-crisis private sector workforce.

As far as it concerns the analysis of the year 2010, looking at Graphs 4.4.3 and 4.4.4 that refer to 1st and 2nd quarters, we can observe that the seasonal fertility rates of the private sector workers re-aligns to the pre-crisis common trend with respect to the public sector fertility ratios. In relative terms, in the 3rd quarter 2010 we observe a net decrease in the private sector birth rate. Even though it is weaker than the one observed in the 2009, looking at the 4th quarter 2010 it is still possible to detect a relative positive fertility trend in the private sector population.

The analysis focused on the year 2009 provides evidence in support of the consistency of the "childbearingopportunity cost" conjecture put forth by Butz and Ward (1979a).

For future research, it would be interesting to deeper analyze the seasonality dimension, but more historical data are needed in order to purse this goal. A further interesting development of this study could drive its attention on the study of "quantum vs. tempo" effects (Bongaarts and Feeney 1998).

Do the extra births, that we have observed in the 2009 after the crisis, represent proper *extra babies* (quantum effect) or more simply do their births have been anticipated (tempo effect) by the crisis?

4.6 Appendix

Table 4.6.1: Redundancy Ratio by sector/year

year	Public	Private
2007	2.5	11.4
2008	1	16.2
2009	2	23.3
2010	4	13.6

Number of workers made redundant
every 1,000 workers in the reference sector

Table 4.6.2: Group-specific Birth Rates by quarters

quarter/year	Public	Private
Q1 2007	18.17	16.15
Q1 2008	15.25	10.76
Q1 2009	11.20	15.37
Q1 2010	15.28	11.61
Q2 2007	11.49	13.27
Q2 2008	9.46	7.65
Q2 2009	12.51	25.00
Q2 2010	12.96	11.15
Q3 2007	13.04	17.38
Q3 2008	13.09	14.91
Q3 2009	13.20	17.34
Q3 2010	18.32	12.60
Q4 2007	16.17	10.79
Q4 2008	10.35	4.14
Q4 2009	14.80	14.53
Q4 2010	9.59	11.14

$$\frac{\text{number of newborns in the group}}{\text{total group members}} \times 1,000$$

Figure 4.6.1: 1st quarters - COUNTERFACTUAL

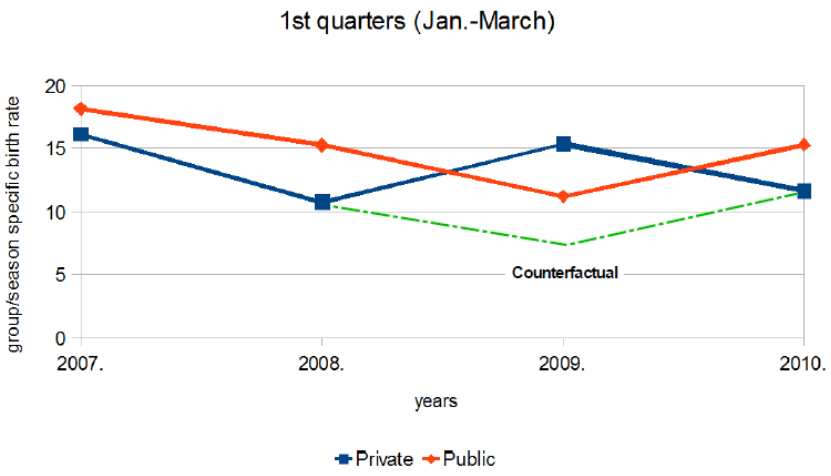


Figure 4.6.2: 2nd quarters - COUNTERFACTUAL

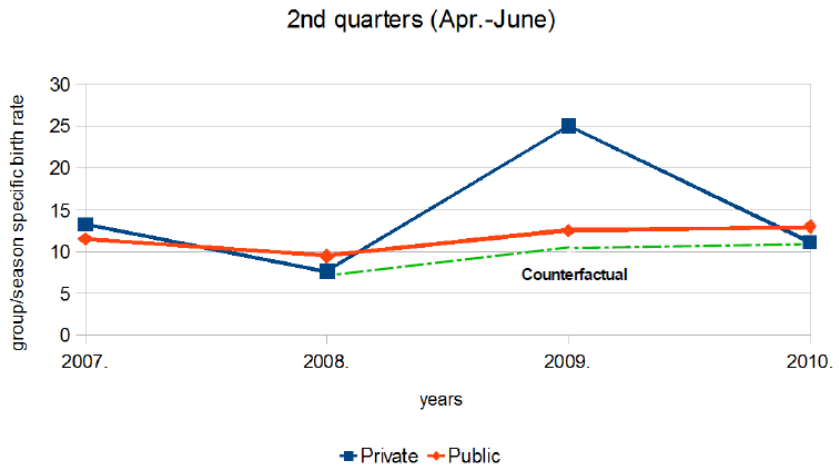


Figure 4.6.3: 3rd quarters - COUNTERFACTUAL

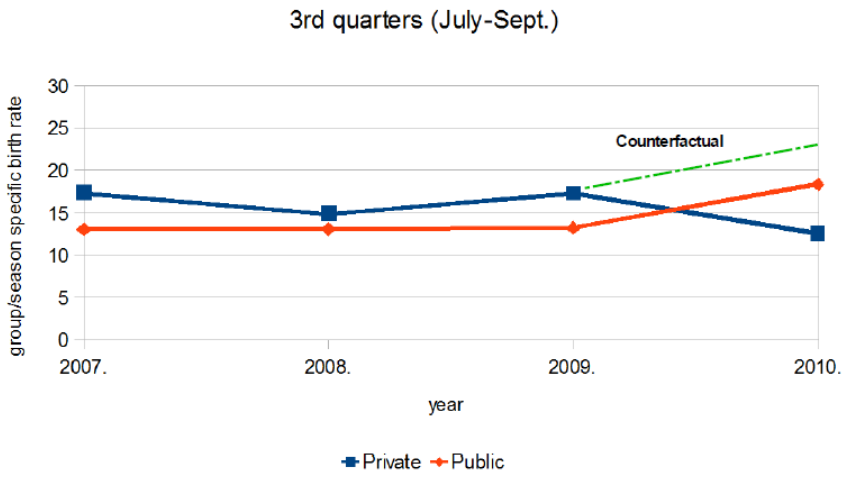


Figure 4.6.4: 4th quarters - COUNTERFACTUAL

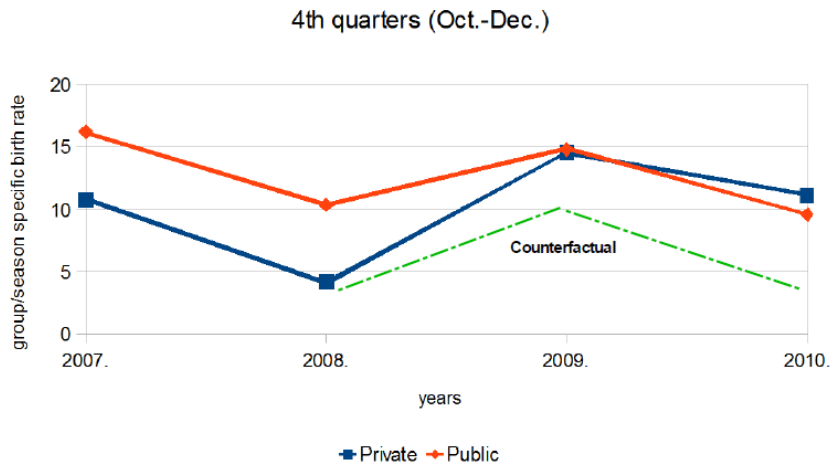


Table 4.6.3: Control Variables, by sector/pre-post crisis

sex: F, age: 18-46, industry: HSC

	Pre-crisis (Q2.2007)			Post-crisis (Q2.2008)		
variable	Public	Private	Δ	Public	Private	Δ
age	36.3 (6.6)	35.4 (7.1)	+0.9***	36.4 (6.6)	35.1 (7.1)	+1.3***
average number of children	1.5 (1.0)	1.5 (1.1)	0	1.5 (1)	1.4 (1.1)	+0.1
British ethnicity	90%	90%	0	90%	90%	0
hourly wage(£)	12.1 (5.7)	9.2 (5.0)	+2.9***	13.6 (7.7)	8.7 (4.1)	+4.9***
weekly working hours	30.4 (9.9)	30.5 (12.3)	+0.1	31.2 (9.9)	31.1 (12.3)	+0.1
partner's age	38.9 (7.6)	38.4 (8.1)	+0.5**	39.2 (7.6)	38.3 (8.1)	+0.9*
partner British ethnicity	88%	88%	0	87%	88%	-1%
partner's hourly wage (£)	14.5 (11.1)	13.0 (7.1)	+1.5***	15.4 (10.1)	13.1 (6.9)	+2.3***
partner's weekly working hours	41.9 (9.4)	42.2 (9.3)	-0.3***	41.5 (9.1)	41.8 (10.2)	-0.3***
partner employed	93%	90%	+3%	91%	88%	+3%
partner public employment	27%	18%	+9%	28%	17%	+11%

Average values of the control variables by sectors, pre and post crisis.

Standard deviations reported in parenthesis.

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.10$.

Table 4.6.4: Common Trend Test - Difference-in-Differences estimation (Probit)

variable	**Probit**$^{(a)}$
outcome variable: $Birth(1)$	
Q1 2008 vs. Q1 2007	
$Pr(birth)$	0.015
$(T \times D)$ ***DiD interaction***	-0.003 (0.0003)
obs.	2,008,320
Q2 2008 vs. Q2 2007	
$Pr(birth)$	0.010
$(T \times D)$ ***DiD interaction***	-0.003 (0.0002)
obs.	2,032,896
Q3 2008 vs. Q3 2007	
$Pr(birth)$	0.014
$(T \times D)$ ***DiD interaction***	-0.002 (0.0003)
obs.	2,046,325
Q4 2008 vs. Q4 2007	
$Pr(birth)$	0.009
$(T \times D)$ ***DiD interaction***	-0.003 (0.0002)
obs.	2,064,501

$^{(a)}$ Marginal Effects reported (AME).
Standard Errors reported in parentheses.
T: 2008; D: Private

Table 4.6.5: Difference-in-Differences estimation (LPM - Probit)

Q1 2008 *vs.* Q1 2009		
variable	**LPM**	**Probit**$^{(a)}$
outcome variable: $Birth(1)$		
β_0 / $Pr(birth)$	0.0152 (0.0001)	0.0129
(T) *post-crisis*	- 0.005 (0.0002)	- 0.004 (0.0002)
(D) *private sector*	-0.004 (0.0002)	-0.004 (0.0002)
$(T \times D)$ ***DiD interaction***	0.0086 (0.0032)	0.01029 (0.0045)
obs.	2,059,444	2,059,444

$^{(a)}$ Marginal Effects reported (AME).
Standard Errors reported in parentheses.

Table 4.6.6: Difference-in-Differences estimation (LPM - Probit)

Q2 2008 *vs.* Q2 2009		
variable	**LPM**	**Probit**$^{(a)}$
outcome variable: $Birth(1)$		
β_0 / $Pr(birth)$	0.009 (0.0001)	0.0121
(T) *post-crisis*	0.003 (0.0002)	0.003 (0.0002)
(D) *private sector*	0.001 (0.0002)	0.002 (0.0002)
$(T \times D)$ ***DiD interaction***	0.0142 (0.0003)	0.0142 (0.0003)
obs.	2,071,855	2,071,855

$^{(a)}$ Marginal Effects reported (AME).
Standard Errors reported in parentheses.

Table 4.6.7: Difference-in-Differences estimation (LPM - Probit)

Q3 2008 *vs.* Q3 2009		
variable	**LPM**	**Probit**$^{(a)}$
outcome variable: $Birth(1)$		
β_0 / $Pr(birth)$	0.0131 (0.0001)	0.01431
(T) *post-crisis*	- 0.0001 (0.0002)	- 0.0001 (0.0002)
(D) *private sector*	-0.001 (0.0002)	-0.002 (0.0002)
$(T \times D)$ ***DiD interaction***	0.0023 (0.0003)	0.0021 (0.0003)
obs.	2,088,325	2,088,325

$^{(a)}$ Marginal Effects reported (AME).
Standard Errors reported in parentheses.

Table 4.6.8: Difference-in-Differences estimation (LPM - Probit)

Q4 2008 *vs.* Q4 2009		
variable	**LPM**	**Probit**$^{(a)}$
outcome variable: $Birth(1)$		
β_0 / $Pr(birth)$	0.0103 (0.0001)	0.0102
(T) *post-crisis*	- 0.0044 (0.0002)	- 0.0037 (0.0002)
(D) *private sector*	-0.0061 (0.0002)	-0.0085 (0.0002)
$(T \times D)$ ***DiD interaction***	0.0061 (0.0002)	0.0108 (0.0004)
obs.	2,111,952	2,111,952

$^{(a)}$ Marginal Effects reported (AME).
Standard Errors reported in parentheses.

Bibliography

ABELER, J., A. FALK, L. GÖTTE, AND D. HUFFMAN (2011): "Reference Points and Effort Provision," *The American Economic Review*, 101(2), 470–492.

ABREVAYA, J., AND D. HAMERMESH (2012): "Charity and Favoritism in the Field: Are Female Economists Nicer (To Each Other)?," *The Review of Economics and Statistics*, 94(1), 202–207.

ADAMS, J. (1965): "Inequity in Social Exchange.," *Advances in Experimental Social Psychology*, 2, 267–299.

AI, C., AND E. NORTON (2003): "Interaction terms in logit and probit models," *Economics Letters*, 80(1), 123 – 129.

AKERLOF, G. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economic*, 84(3), 488–500.

AKERLOF, G. (1982): "Labor Contracts as Partial Gift Exchange," *The Quarterly Journal of Economics*, 97(4), 543–569.

AKERLOF, G. (1984): "Gift exchange and efficiency-wage theory: Four views," *The American Economic Review*, 74(2), 79–83.

AKERLOF, G., AND R. KRANTON (2010): *Identity Economics: how our identities shape our work, wages, and well-being*. PUP.

ALMALAUREA, . (2009): "XI Rapporto Almalaurea 2009," Report, Consorzio Interuniversitario AlmaLaurea - MIUR.

ANDREONI, J. (1989): "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *The Journal of Political Economy*, pp. 1447–1458.

——— (1990): "Impure Altruism and Donations to Public Goods: a Theory of Warm-Glow Giving," *The Economic Journal*, 100(401), 464–477.

ANGRIST, J., D. LANG, AND P. OREOPOULOS (2009): "Incentives and Services for College Achievement: Evidence from a Randomized Trial," *American Economic Journal: Applied Economics*, 1(1), 136–63.

ANGRIST, J., AND V. LAVY (2009): "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial," *American Economic Review*, 99(4), 1384–1414.

ARIELY, D., A. BRACHA, AND S. MEIER (2009): "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially," *American Economic Review*, 99, 1–1.

ARON, D., AND P. OLIVELLA (1994): "Bonus and Penalty Schemes as Equilibrium Incentive Devices, with Application to Manufacturing Systems," *Journal of Law, Economics, and Organization*, 10(1), pp. 1–34.

AZMAT, G., AND N. IRIBERRI (2010): "The Provision of Relative Performance Feedback Information: An Experimental Analysis of Performance and Happiness," Working Papers (Universitat Pompeu Fabra. Departamento de Economía y Empresa), No. 1216.

BAKER, G. (1992): "Incentive contracts and performance measurement," *Journal of Political Economy*, 100(3), 598–614.

BAKER, G., R. GIBBONS, AND K. J. MURPHY (1994): "Subjective Performance Measures in Optimal Incentive Contracts," *The Quarterly Journal of Economics*, 109(4), pp. 1125–1156.

BARIGOZZI, F., AND G. TURATI (forthcoming): "Human Health Care and Selection Effects. Understanding Labor Supply in the Market for Nursing," *Health Economics*.

BECKER, W. E. (1997): "Teaching Economics to Undergraduates," *Journal of Economic Literature*, 35(3), pp. 1347–1373.

BENABOU, R., AND J. TIROLE (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70(3), 489–520.

——— (2006): "Incentives and prosocial behavior," *The American Economic Review*, 96(5), 1652–1678.

BENGTSSON, T., C. CAMPBELL, AND J. Z. LEE (2009): *Life Under Pressure: Mortality and Living Standards in Europe and Asia, 1700-1900*, vol. 1. The MIT Press.

BENZ, M. (2005): "Not for the Profit, but for the Satisfaction? - Evidence on Worker Well-Being in Non-Profit Firms," *Kyklos*, 58(2), 155–176.

BESLEY, T., AND M. GHATAK (2005): "Competition and Incentives with Motivated Agents," *American Economic Review*, 95(3), 616–636.

BLIMPO, M. (2010): "Team Incentives for Education in Developing Countries. A Randomized Field Experiment in Benin," mineo, March 2010 - Stanford University.

BLOUNT, S. (1995): "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior and Human Decision Processes*, 63(2), 131–144.

BOLTON, G. E., AND A. OCKENFELS (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *The American Economic Review*, 90(1), 166–193.

BOLTON, P., AND M. DEWATRIPONT (2005): *Contract Theory*. The MIT Press.

BONGAARTS, J., AND G. FEENEY (1998): "On the Quantum and Tempo of Fertility," *Population and Development Review*, 24(2), pp. 271–291.

BRATTI, M., D. CHECCHI, AND A. FILIPPIN (2011): "Should You Compete or Cooperate with Your Schoolmates?," *Education Economics*, 19(3), 275–289.

BREKKE, K. A., AND K. NYBORG (2010): "Selfish Bakers, Caring Nurses? A Model of Work Motivation," *Journal of Economic Behavior & Organization*, 75(3), 377 – 394.

BRUNI, L., M. GILLI, AND V. PELLIGRA (2008): "Reciprocity: theory and facts," *International Review of Economics*, 55(1), 1–11.

BUCHAN, N. R., R. T. CROSON, AND S. SOLNICK (2008): "Trust and Gender: An Examination of Behavior and Beliefs in the Investment Game," *Journal of Economic Behavior and Organization*, 68, 466–476.

BULL, C. (1987): "The existence of self-enforcing implicit contracts," *The Quarterly Journal of Economics*, 102(1), 147–159.

BUTZ, W. P., AND M. P. WARD (1979a): "The Emergence of Countercyclical U.S. Fertility," *The American Economic Review*, 69(3), pp. 318–328.

——— (1979b): "Will US Fertility Remain Low? A New Economic Interpretation," *Population and Development Review*, 5(4), pp. 663–688.

CAPLAN, A., D. AADLAND, AND A. MACHARIA (2010): "Estimating Hypothetical Bias in Economically Emergent Africa: A Generic Public Good Experiment," *Agricultural and Resource Economics Review*, 39(2), 344–358.

CARDY, R., AND G. DOBBINS (1986): "Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance.," *Journal of Applied Psychology; Journal of Applied Psychology*, 71(4), 672.

CHARNESS, G., AND P. KUHN (2010): "Lab labor: What can labor economists learn from the lab?," *NBER Working Paper*.

CHARNESS, G., AND P. KUHN (2011): *Lab Labor: What Can Labor Economists Learn from the Lab?*

CROSON, R., AND U. GNEEZY (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47(2), 448–474.

DE PAOLA, M., V. SCOPPA, AND R. NISTICO (2010): "Monetary Incentives and Student Achievement in a Depressed Labour Market: Results from a Randomized Experiment," Working Paper 06 - 2010, UNICAL, Dipartimento di Economia e Statistica.

DELFGAAUW, J., AND R. DUR (2007): "Signaling and Screening of Workers' Motivation," *Journal of Economic Behavior and Organization*, 62(4), 605–624.

ENDERER, F., AND G. MANSO (2009): "Is Pay-For-Performance Detrimental to Innovation?," *UC Berkeley: Department of Economics, UCB*.

FALK, A., AND E. FEHR (2003a): "Why Labour Market Experiments?," *Labour Economics*, 10(4), 399–406.

FALK, A., AND E. FEHR (2003b): "Why labour market experiments?," *Labour Economics*, 10(4), 399 – 406, Special Issue on Labour Market Experiments.

FALK, A., AND U. FISCHBACHER (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54(2), 293–315.

FALK, A., AND S. GÄCHTER (2008a): "Experimental Labour Economics," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf, and L. E. Blume. Palgrave Macmillan, Basingstoke.

FALK, A., AND S. GÄCHTER (2008b): "Experimental Labour Economics," *The New Palgrave Dictionary of Economics, Basingstoke: Palgrave Macmillan*.

FALK, A., AND J. HECKMAN (2009a): "Lab Experiments Are a Major Source of Knowledge in the Social Sciences," *Science*, 326(5952), 535.

FALK, A., AND J. J. HECKMAN (2009b): "Lab Experiments Are a Major Source of Knowledge in the Social Sciences," *Science*, 326(5952), 535–538.

FALK, A., AND A. ICHINO (2006): "Clean Evidence on Peer Effects," *Journal of Labor Economics*, 24(1), 39–57.

FEHR, E., A. FALK, AND U. FISCHBACHER (2003): "On the nature of fair behavior," *Economic Inquiry*, 41(7).

FEHR, E., AND K. SCHMIDT (1999): "A theory of Fairness, Competition, and Cooperation," *Quarterly journal of Economics*, 114(3), 817–868.

——— (2003): "Theories of fairness and reciprocity: Evidence and economic applications," in *Advances in economics and econometrics: Theory and applications, Eighth World Congress*, vol. 1, pp. 208–257.

FEHRLER, S., AND M. KOSFELD (2010): "Contracts for Motivated Agents," *Mimeo*.

FERRARO, J., AND R. CUMMINGS (2007): "Cultural Diversity, Discrimination, and Economic Outcomes: an Experimental Analysis," *Economic Inquiry*, 45(2), 217–232.

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

——— (2007b): "z-Tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, 10(2), 171–178.

FOLGER, R. (2001): "Fairness as deonance," *Theoretical and cultural perspectives on organizational justice*, pp. 3–33.

FRANCOIS, P., AND M. VLASSOPOULOS (2008): "Pro-social Motivation and the Delivery of Social Services," *CESifo Economic Studies*, 54(1), 22–54.

FRYER, R. J. (2010): "Financial Incentives and Student Achievement: Evidence from Randomized Trials," *working paper Harvard University and NBER*.

GIBBONS, R. (1998): "Incentives in organizations," *The Journal of Economic Perspectives*, 12(4), 115–132.

GIEBE, T., AND O. GUERTLER (2012): "Optimal contracts for lenient supervisors," *Journal of Economic Behavior & Organization*, 81(2), 403 – 420.

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

GNEEZY, U., M. NIEDERLE, AND A. RUSTICHINI (2003): "Performance In Competitive Environments: Gender Differences," *The Quarterly Journal of Economics*, 118(3), 1049–1074.

GNEEZY, U., AND A. RUSTICHINI (2004): "Gender and Competition at a Young Age," *American Economic Review*, 94(2), 377–381.

GREENWALD, B. (1986): "Adverse selection in the labour market," *The Review of Economic Studies*, 53(3), 325.

GREINER, B. (2004): "The online recruitment system ORSEE 2.0-A guide for the organization of experiments in economics," *University of Cologne, Working paper series in economics*, 10.

GROVE, W. A., AND T. WASSERMAN (2006): "Incentives and Student Learning: A Natural Experiment with Economics Problem Sets," *AER: AEA Papers and Proceedings*, 96(2), 447–452.

GRUND, C., AND J. PRZEMECK (2012): "Subjective performance appraisal and inequality aversion," *Applied Economics*, 44(17), 2149–2155.

HANSMANN, H. (1987): "The Role of Nonprofit Enterprise. The Economics of Nonprofit Institutions: Studies in Structure and Policy. S. Rose-Ackerman," .

HARBRING, C., AND B. IRLENBUSCH (2003): "An Experimental Study on Tournament Design," *Labour Economics*, 45(2), 443–464.

HARRISON, G. W., AND J. A. LIST (2004): "Field Experiments," *Journal of Economic Literature*, 42(4), pp. 1009–1055.

HEYES, A. (2005): "The Economics of Vocation or 'Why is a Badly Paid Nurse a Good Nurse'?," *Journal of Health Economics*, 24(3), 561–569.

HIGGINS, C., T. JUDGE, AND G. FERRIS (2003): "Influence tactics and work outcomes: a meta-analysis," *Journal of Organizational Behavior*, 24(1), 89–106.

HOLLANDER, M., AND D. A. WOLFE (1999): *Nonparametric Statistical Methods*. Wiley and Sons, Inc.

HÖLMSTROM, B. (1979a): "Moral Hazard and Observability," *The Bell Journal of Economics*, 10(1), pp. 74–91.

HÖLMSTROM, B. (1979b): "Moral hazard and observability," *The Bell Journal of Economics*, pp. 74–91.

HOLT, C., AND S. LAURY (2002): "Risk Aversion and Incentive Effects," *The American Economic Review*, 92(5), 1644–1655.

HUCK, S., G. LUNSER, AND J. R. TYRAN (2010): "Consumer Networks and Firm Reputation: A First Experimental Investigation," *Ecomonics Letters*, 108, 242–244.

JENSEN, M., AND W. MECKLING (1976): "Theory of the firm: Managerial behavior, agency costs and ownership structure," *Journal of financial economics*, 3(4), 305–360.

JUDGE, T., AND G. FERRIS (1993): "Social context of performance evaluation decisions," *Academy of Management Journal*, 36(1), 80–105.

KAMBE, S. (2006): "SUBJECTIVE EVALUATION IN AGENCY CONTRACTS*," *Japanese Economic Review*, 57(1), 121–140.

KANE, J. (1994): "A model of volitional rating behavior," *Human Resource Management Review*, 4(3), 283–310.

KERR, S. (1975): "On the folly of rewarding A, while hoping for B," *Academy of Management Journal*, 18(4), 769–783.

KOSFELD, M., AND S. NECKERMANN (2011): "Getting More Work for Nothing? Symbolic Awards and Worker Performance," *American Economic Journal: Microeconomics*, 3, 1–16.

KREHBIEL, P., AND R. CROPANZANO (2000): "Procedural justice, outcome favorability and emotion," *Social justice research*, 13(4), 339–360.

KREMER, M., E. MIGUEL, AND R. THORNTON (2009): "Incentives to Learn," *Review of Economics and Statistics*, 91(3), 437–456.

KREPS, D. M. (1997): "Intrinsic Motivation and Extrinsic Incentives," *American Economic Review*, 87(2), 359–64.

LAFFONT, J., AND D. MARTIMORT (2002): *The theory of incentives: the principal-agent model*. Princeton Univ Pr.

LAZEAR, E. (1999): "Personnel economics: past lessons and future directions," .

LAZEAR, E. P., U. MALMENDIER, AND R. A. WEBER (2012): "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 4(1), 136–163.

LEE, R. (1990): "The Demographic Response to Economic Crises in Historical and Contemporary Populations," *Population Bulletin of the United Nations*, 29, 1–15.

LEETE, L. (2001): "Whither the Nonprofit Wage Differential? Estimates from the 1990 Census," *Journal of Labor Economics*, 19(1), 136–170.

LEUVEN, E., H. OOSTERBEEK, AND VAN DER KLAAUW B. (2010): "The Effect of Financial Rewards on Students' Achievements: Evidence from a Randomized Experiment," *Journal of the European Economic Association*, 8(6), 1243–1265.

LEVENTHAL, G. (1980): "What Should Be Done with Equity Theory? New Approaches to the Study of Fairness in Social Relationships.," in *Social exchanges: Advances in theory and research*, ed. by K. Gergen, J. Greenberg, and R. Willis, pp. 27–55, New York. Plenum Press.

LEVIN, J. (2003): "Relational incentive contracts," *The American Economic Review*, 93(3), 835–857.

LINARDI, S., AND M. MCCONNELL (2011): "No Excuses for Good Behavior: Volunteering and the Social Environment," *Journal of Public Economics*, 95(5-6), 445–454.

MACLEOD, W., AND J. MALCOMSON (1989): "Implicit contracts, incentive compatibility, and involuntary unemployment," *Econometrica: Journal of the Econometric Society*, pp. 447–480.

MACLEOD, W. B. (2003): "Optimal Contracting with Subjective Evaluation," *The American Economic Review*, 93(1), pp. 216–240.

MACUNOVICH, D. (1996): "Relative Income and Price of Time: Exploring their effcts on U.S. Fertility and Female Labor Force Participation, 1963-1993," Department of Economics Working Papers 174, Department of Economics, Williams College.

MAS-COLELL, A., M. WHINSTON, AND J. GREEN (1995): *Microeconomic Theory*, vol. 981. Oxford Univiversity Press.

NIEDERLE, M., AND L. VESTERLUND (2007): "Do Women Shy Away from Competition? Do Men Compete Too Much?," *The Quarterly Journal of Economics*, 122(3), 1067–1101.

NIEDERLE, M., AND L. VESTERLUND (2010): "Explaining the Gender Gap in Math Test Scores: The Role of Competition," *Journal of Economic Perspectives*, 24(2), 129–44.

PEARCE, D., AND E. STACCHETTI (1998): "The interaction of implicit and explicit contracts in repeated agency," *Games and Economic Behavior*, 23(1), 75–96.

PRENDERGAST, C. (1999): "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37(1), pp. 7–63.

——— (2002): "Uncertainty and incentives," *Journal of Labor Economics*, 20(2; PART 2), 115–137.

PRENDERGAST, C., AND R. TOPEL (1993): "Discretion and bias in performance evaluation," *European Economic Review*, 37(2-3), 355–365.

PRENDERGAST, C., AND R. H. TOPEL (1996): "Favoritism in Organizations," *Journal of Political Economy*, 104(5), pp. 958–978.

PUHAN, A. (2012): "The treatment effect, the cross difference, and the interaction term in nonlinear Difference-in-Differences models," *Economics Letters*, 115(1), 85 – 87.

RABIN, M. (1993): "Incorporating fairness into game theory and economics," *The American Economic Review*, pp. 1281–1302.

RABIN, M. (2000): "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica*, 68(5), 1281–1292.

RABIN, M., AND J. L. SCHRAG (1999): "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics*, 114(1), 37–82.

REBITZER, J. B., AND L. J. TAYLOR (2011): *Extrinsic Rewards and Intrinsic Motives: Standard and Behavioral Approaches to Agency and Labor Markets*.

RUTHERFORD, A. (2010): "Where is the Warm Glow? Donated Labour in the Health and Social Work Industries," Mimeo, Department of Economics, University of Stirling.

RYAN, R. M., AND E. L. DECI (2000): "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," *Contemporary Educational Psychology*, 25(1), 54 – 67.

SAUERMANN, H., AND M. ROACH (2010): "Not All Scientists Pay to Be Scientists: Heterogeneous Preferences for Publishing in Industrial Research," *DRUID Working Paper No. 11-03r*.

SCHMIDT, K., AND M. SCHNITZER (1995): "The interaction of explicit and implicit contracts," *Economics Letters*, 48(2), 193–199.

SEITHE, M. (2010): "Introducing the Bonn Experiment System," *Discussionpaper*.

SHAPIRO, C., AND J. STIGLITZ (1984): "Equilibrium unemployment as a worker discipline device," *The American Economic Review*, 74(3), 433–444.

SOBOTKA, T., V. SKIRBEKK, AND D. PHILIPOV (2010): "Economic Recession and Fertility in the Developed World: a Literature Review," Research Note February 2010, Vienna Institute of Demography.

STEERS, R., R. MOWDAY, AND D. SHAPIRO (2004): "Introduction to special topic forum: The future of work motivation theory," *The Academy of Management Review*, 29(3), 379–387.

STERN, S. (2004): "Do Scientists Pay to Be Scientists?," *Management Science*, 50(6), pp. 835–853.

STRAUSZ, R. (1997): "Collusion and Renegotiation in a Principal–Supervisor–Agent Relationship," *The Scandinavian Journal of Economics*, 99(4), 497–518.

THIELE, V. (2011): "Subjective Performance Evaluations, Collusion, and Organizational Design," *Journal of Law, Economics, and Organization*.

THOMAS, J., AND H. MEEKE (2010): "Rater error," *Corsini Encyclopedia of Psychology*.

TIROLE, J. (1986): "Hierarchies and bureaucracies: On the role of collusion in organizations," *Journal of Law Economics and Organization*, 2, 181.

TONIN, M., AND M. VLASSOPOULOS (2010): "Disentangling the Sources of Prosocially Motivated Effort: A Field Experiment," *Journal of Public Economics*, 94(11-12), 1086 – 1092.

TZANNATOS, Z., AND J. SYMONS (1989): "An Economic Approach to Fertility in Britain since 1860," *Journal of Population Economics*, 2, 121–138.

VAFAÏ, K. (2010): "Opportunism in organizations," *Journal of Law, Economics, and Organization*, 26(1), 158–181.

VAN BAVEL, J. (2001): "Malthusian Sinners: Illegitimate Fertility and Economic Crisis. A case study in Leuven, 1846-1856," *Tijdschrift voor Nieuwste Geschiedenis/Revue Belge d'Histoire Contemporaine*, 31, 371–410.

VARMA, A., A. DENISI, AND L. PETERS (1996): "Interpersonal affect and performance appraisal: A field study," *Personnel Psychology*, 49(2), 341–360.

YELLEN, J. (1984): "Efficiency wage models of unemployment," *The American Economic Review*, 74(2), 200–205.