DOCTORAL DISSERTATION

Experimental Studies on Organizational Behavior

Universitat Autonoma de Barcelona

Department of Business Economics

Doctorate in Economics, Management, and Organizations

Author:

ANDREJ ANGELOVSKI

Dissertation Advisors:

JORDI BRANDTS BERNAD, PH.D.

CARLES SOLÀ BELDA, PH.D.

Barcelona, June 4th, 2014

DOCTORAL DISSERTATION

Experimental Studies on Organizational Behavior

Universitat Autonoma de Barcelona

Department of Business Economics

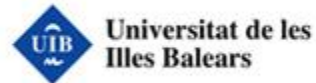Doctorate in Economics, Management, and Organizations

Author:

ANDREJ ANGELOVSKI

Dissertation Advisor:

JORDI BRANDTS BERNAD, PH.D.

CARLES SOLÀ BELDA, PH.D.

Barcelona, June 4th, 2014

ABSTRACT

Organizational behavior looks at how individuals behave in organizational settings, how they interact with other individuals within the setting, as well as how they behave with the organization itself. A growing literature explores how people are affected by the organization, as well as how they affect it with their decision making. My dissertation contributes to this body of literature by investigating topics in this field using experimental methods. More specifically, my work looks at how people are affected by certain biases and how these biases affect the others as well as the organization; ways to eliminate the bias; how individuals behave when faced with the dilemma of contributing to two different public goods of different efficiency; how different levels of communication within an organization can help solve or hinder this dilemma; the effect of interdependence in organizational design on performance; and the effect of payment inequality in this type of organizational design.

In the second chapter of the thesis we conduct an experiment to study whether the way employees are assigned to a manager affects managers' and co-employees' subjective evaluations of employees. Employees are either be hired by the manager, explicitly not hired by him and nevertheless assigned to him, or exogenously assigned to him. For all three we find escalation bias both by managers and by co-employees. Managers exhibit a positive bias towards those employees they have hired or a negative one towards those they have explicitly not hired.

Chapter three is a follow-up study to chapter two. The aim with this chapter is to add to those findings by looking at the bias from the point of view of the employees who have to continue working in an environment where they are positively or negatively biased against, and how it affects their future performance and decisions. Additionally, we try to find whether escalation bias occurs even in a multiple round treatment, in which there is no loss of information, and managers pay could be affected by their evaluation fairness, and find that escalation bias does not occur in such a scenario. We also find that having a manager being positively biased towards an employee has a positive effect on that employee's future performance, even though the employee is aware that the evaluation is not justified. Being negatively biased against, though, does not have any significant impact on future performance, however it increases the likelihood of making a decision to leave or sabotage which is costly to all participants.

Chapter four reports on a set of public goods experiments we conducted in which participants belong to both a smaller "local" group, and a larger "global" group and have various degrees of communication within and between these groups. Participants, in one set of treatments, have the option to contribute just to the more efficient global public good, and in another set of treatments, to both the global public good and the less efficient local public good. The results show that, when participants can only contribute to the global public good, the level of contributions goes up as the level of communication goes up. We also found that when the option to contribute to both public goods is provided with no communication, participants heavily prefer to contribute to the less efficient local public good, however as levels of communication are added both the total contribution and the contribution into the "global" public good go up.

In chapter five we aim to create a paradigm in which a unique yet common combination of sequential and pooled task interdependence can be studied naturally. We do this by designing a production line-like organizational design in which has three stages and every individuals input is essential to the end product of the organization. We also aim to determine the importance of payment equality under two different levels of fair allocation into positions in our highly interdependent task. Our results show that even extreme payment inequality does not significantly affect a work-groups performance in real effort tasks with high interdependence. Additionally, different methods of allocating into positions can effect performances at certain stages of the production line, though not the final performance.

ACKNOWLEDGEMENTS

Over the past four years I have received support and encouragement from a number of individuals. I have come to consider Jordi Brands as a mentor, colleague, and friend. Having his guidance has made this a truly reflective and rewarding journey. Though there were numerous difficulties which I had to face over these past years, knowing that I have his incessant support has made this experience not only tolerable but also fulfilling and enjoyable. It is a truly rare sight in business-economics departments to find a mentor who has such an open mind and an appetite for the different and creative approaches, and for this I feel truly lucky. I need to thank Carles Sola, first and foremost, for helping me realize how much I would like to work on behavioral economics in a class he taught in 2009. He has been a great coauthor and mentor over these years, always full of encouragement and support, and for this I am truly grateful.

I would like to thank Ernesto Reuben for supervising my stay at Columbia Business School, and barring with me even when at times I felt like "fish out of water" in the new environment, with no time to pause and adapt. I have learned a lot from him and played a major role in my development as researcher. My unending gratitude goes to Darko Dimitrov who spent a good portion of a summer helping me pick up basic programming, going as far as learning Z-Tree in parallel with me. I would also like to thank my dissertation committee of Martin Larazza, Arthur Schram, and Pablo Brañas for their great comments and advise on how to improve this work. A special thanks also goes out to all my friends and colleagues who inhabit or once inhabited, what I now know to be "Office B1-129", without you this experience would not have been the same.

Lastly, I would like to thank my entire family for all their support and encouragement. Most importantly, I have always known that I can count on them times get rough, and this type of an endeavor would not have been feasible without it.

TABLE OF CONTENTS

CHAPTER I - Introduction

In the 1930s and 1940s, the attention of organizational behavior moved to finding what effect psychological factors had on organizations. The start of this shift was the detection of the Hawthorne Effect, which states that workers change behavior due to the fact that there is a change in the environment, rather than actually adapting their behavior due to the nature of the change. This brought forward interest on the psychological factors of an individual's behavior in an organizational setting, and research into the field took off. My dissertation contributes to this literature by helping better understanding an individual's decision making in various organizational situations.

In chapter two we present results from a detailed experimental study about *escalation bias* in evaluations as a consequence of previous hiring decisions. Escalation Bias, often referred to as "irrational escalation of commitment", is a common term in psychology, sociology, and finance to refer to a situation in which people who have initially made a rational decision, follow it up with an irrational one in order to justify the initial decision and thus make themselves feel better about it.

In many organizations the measurement of job performance cannot rely on easily quantifiable information. In such cases, supervising managers often use subjective performance evaluations. In our design an employee can be assigned to a manager in three different ways. Employees can either be hired by the manager, explicitly not hired by him and nevertheless assigned to him or exogenously assigned to him. We investigate whether the way employees are assigned to a manager affects managers' and co-employees' subjective evaluations of employees.

Staw (1976) finds that when one makes a decision is one's mind, they unconsciously make a psychological commitment to the decision, so all future decision are biased on the same topic are biased towards the initial decision. Ultimately, there are two precursors to the sets of experiments we ran in chapter two. Bazerman, Beekun, and Schoorman (1982) and Schoorman (1988) who

study and find escalation bias present in promotion decision as well as evaluation decisions. Chapter two adds to this literature with a much more detailed study, and we highlight four major aspects of it. First we follow standard experimental economics practices by making sure that experimental decisions have payoff consequences. Second, we collect and present data from three separate treatments in which we vary the incentives and experience of the manager in order to test for the robustness of the behavior. Thirdly we include peer-to-peer evaluations, in order to find out whether they themselves can be biased by the initial hiring decision even though they are not part of the decision themselves. Lastly, we compare the evaluations of the participants with external experts, who give us an external and unbiased standard of comparison.

Chapter three follows up on the study of chapter two by using and adapting the experimental design of the previous study in order to add to those findings by trying to find ways to eliminate the bias which was initially found and also to look at the bias from the perspective of the employees who have to continue working in an environment in which they receive a positive or negative biased evaluation from their evaluating manager, and find out how this affects their future performance and decisions.

Bol (2010) has found that receiving a higher evaluation than one deserves can lead to an improvement of future performance. Experiments on gift exchange also indicate that having received a gift, one is likely to reciprocate by increasing their effort (ex. Fehr et al. 1993). On the other hand, agency theory predicts that a bias in any direction, whether it is positive or negative, should decrease employees' future performance. This is predicted because bias brakes the important link between effort and performance based compensations.

We find that escalation bias can be eliminated without having to link the manager accuracy of the evaluations to his pay. We also find that having a manager being positively biased towards an employee has a positive effect on that employee's future performance, even though the employee is aware that the evaluation is not justified. We do not find that being negatively biased against has an immediate effect on performance, however it does increase the likelihood of an employee to make the costly decision of sabotaging of leaving the firm.

In chapter four we report on a set of public goods experiments in which participants belong to both a smaller "local" group, and a larger "global" group and have various degrees of communication

within and between these groups. Participants, in half of the treatments, have the just one option available to them; to contribute to the more efficient global public good. In the other hald of the treatments, to both the global public good and the less efficient local public good. We want to find out the effect of having an option to contribute to a local public will have on contributions, as well as to find out if, and in what form, communication can help groups and individuals to contribute more efficiently. This is a typical problem in most medium & large organizations as emplyees have the option to contribute more effort/time to help the smaller branch/work-group they are part of or to the larger firm/organization. Allocating more towards the organization is potentially more profitable and efficient, yet it is harder to achieve.

The results show that, when participants can only contribute to the global public good, the level of contributions goes up as the level of communication goes up. When the option to contribute to both public goods is provided with no communication, participants heavily prefer to contribute to the less efficient local public good, however as levels of communication are added both the total contribution and the contribution into the "global" public good go up.

In chapter five we designed a naturally flowing three-staged production line in which participants work individually, one stage at a time, in order to deliver a final product out to market. This design made sure that each individual effort and performance at all stages has an effect on the performance of the group as a whole. Partcipants in all stages of the experiment multiply numbers, and each stage increases the difficulty, and only the correct answers are passed on to the next level. The final product is complete only if all three stages correctly solve the multiplications passed on to them by the previous stage, thus the group only gets paid by the last stage's (stage 3) correct output.

Our goal is to determine the importance of equality of pay in the scenario of this common yet unique type of task interdependence. Incentives and equality of pay have been very hot topics for quite some time now, yet for the most part they have been studied respectively with the focus on the individual and with a focus on separate hierarchies, which do not necessarily have direct collaboration for the completion of a final good/service. Libby & Thorne (2009) when studying the effects of individual, group, and mixed incentive structures on group performance in assembly lines, among other types of settings, find no difference in group performance only for assembly lines, regardless of the large payment structure differences. Our results go along the lines of this

result, as we find that even extreme payment inequality does not significantly affect a work-groups performance in real effort tasks with high interdependence. Additionally, different methods of allocating into positions can effect performances at certain stages of the production line, though not the final performance.

CHAPTER II

Hiring and Escalation Bias in Subjective Performance Evaluations: A Laboratory Experiment

1. Introduction

Job performance is one of the key aspects of human resources management and has attracted much attention from researchers and practitioners alike. The importance of job performance, though, is parallel to the difficulty of its measurement and consequently performance assessment is considered one of the most important human resources practices (Judge & Ferris, 1993).

It is obviously vital for an organization that performance assessments are done accurately. In many organizations the measurement of job performance can not rely on easily quantifiable information. In such cases, supervising managers often use subjective performance evaluations. After many years of performance assessments in many organizations, and hundreds of papers on the topic, researchers and practitioners, alike, are still trying to find out how to achieve accuracy of subjective performance evaluations. The key issue is that subjective assessments give opportunity for factors like interpersonal relationships, likeability, and affect to have an effect on rating accuracy.

In this paper we present the results from a detailed experimental study about *escalation bias* in the evaluations as a consequence of hiring decisions. Escalation Bias, sometimes referred to as "irrational escalation of commitment", is a term frequently used in psychology, sociology, and finance to refer to a situation in which people who have initially made a rational decision, follow it up with an irrational one in order to justify the initial decision and thus make themselves feel better about it. In this context, we study whether managers' hiring decisions biases subsequent performance appraisal decisions by both managers and employees. Staw (1976) was one of the first to point out that once a decision is made in our minds, we also mechanically make a

psychological commitment to that decision, so that further decisions on the same matter are biased towards the initial decision.

We use a lab experiment as a research method because it allows us to control for certain factors in a way that otherwise would not have been possible. In our stylized setting an employee can be assigned to a manager in three different ways. Employees can either be hired by the manager, explicitly not hired by him and nevertheless assigned to him or exogenously assigned to him. We study whether the way employees are assigned to a manager affects managers' and co-employees' subjective evaluations of employees.[1]

During the experiment all participants in the role of the employees have to perform a task, consisting in reading a short two-page business case study and answering two textual questions on it. This is done in order to mimic real life company problems where there isn't just one strict way to get things done. The employees' task does not have a quantifiable correct answer and thus warrants a subjective evaluation. The employees know that each of them is given the exact same case study which is going to be evaluated by the manager and their peers, the other two employees in their company. A key feature of our experiment is that, in addition to the evaluation by managers and fellow employees, we include the evaluation of the task by three expert outside evaluators. These outside evaluations yield a baseline to which managers' and fellow employees' evaluations can be compared.

There are two precursors of our study. Bazerman, Beekun, and Schoorman (1982) present data from a class-room experiment in which participants were given the role of vice president of a large retail company with numerous stores. Then participants in the experimental group were asked to make a promotion decision to the position of manager of one of the stores, by choosing between three fictitious internal candidates for whom fictitious performance data was provided. After this participants were given new fictitious data that suggested that the promoted manager was not performing well.

---

[1] Falk and Heckman (2009) contains an interesting discussion of the merits of lab experiments.

On the basis of this information participants in the experimental group had to, for the manager they had promoted earlier, 1) recommend an increase in pay and bonuses, 2) make an evaluation of the manager's potential for being promoted, 3) forecast potential sales and profitability. In the control treatment, participants were informed that decision who to promote was completed by a predecessor. They were given the same performance information on the manager and were assigned the same tasks 1), 2), and 3). The results are consistent with escalation bias. Participants that had earlier chosen to promote certain employees were much more likely to later give them higher pay increases, give higher evaluations on managers' potential, and forecast higher sales and profits than managers in the control treatment. Importantly, the experiment involved no incentives. Hence, the bias involves no costs for the evaluators and the evaluated employees.

Bazerman et al. cannot really distinguish between a positive and a negative escalation bias, since the direct comparison between managers involved in hiring and not involved in hiring can not disentangle whether the first kind of manager is too positive or the second too negative. Schoorman (1988) followed up on this study, with a field experiment conducted within a real large public sector organization in which the presence of a positive bias could be separated from that of a negative bias in a particular way. Supervisors were asked to do performance evaluations of their real employees. The experiment involved supervisors evaluating (a) employees who they had not participated in hiring, (b) employees they had participated in hiring and where they did agree with the decision and (c) employees that they had participated in hiring and where they did not agree with the decision. The evaluations were done based on a performance appraisal instrument of actual performance. Comparing (a) and (b) Schoorman finds a positive escalation bias and comparing (a) and (c) he finds an, albeit weaker, negative escalation bias. As in Bazerman et al participants' decisions had no payoff consequences for anybody involved and, in addition, there was no outside standard of comparison for the evaluations.

We think that the possible presence of escalation bias in performance evaluations is an important problem for companies and other organizations and that, therefore, the issue deserves a new more detailed study. We highlight four important aspects of our study. First, in our experiment decisions will have payoff consequences for participants. Here we follow standard practice in experimental economics. Second, we present data from three different treatments in which we vary relevant factors of the environment to test for the robustness of behavior. In particular, we will vary the

incentives and the experience of managers. Third, we not only analyze managers' evaluations of employees, but also peer-to-peer evaluations, that is employees' evaluations of other employees. This will allow us to study whether potential manager bias can be somehow compensated by taking into account employees' evaluations. Fourth, we compare managers' and employees' evaluations to evaluations of experts not directly involved in the experiment, who provide us with an external standard of comparison. These and other elements of our design will allow us to present a rich study of escalation bias.

2. Background

In this section we discuss some selected contributions to a large literature. Our review does not pretend to be exhaustive.

It is widely considered unfortunate by companies that not all parts of job performance can be measured objectively. This is why an ever-growing number of organizations constantly use subjective appraisals in an attempt to provide the best possible measure of the employees' performance. Gibbs et al. (2003) say that some sort of subjective performance evaluation is used in all jobs, and find that there is an apparent weakness in quantitative performance measures which puts employees at risk of downside pay, and thus subjective bonuses can be used to balance the previous effect. This, they state, filters out the effects of bad luck but not of good luck. But as mentioned earlier, these subjective parts of the evaluation can lead to a significantly biased evaluation (Prendergast and Topel (1996)). The reason for this is simple; there is a bias in subjective evaluations due to one's own perceptual biases, but even further there seems to be a systematic bias due to preferences or liking towards the worker. They further state that this can lead to inefficient division of jobs. This is further supported by Milkovich and Wigdor (1991) who find that it's this subjective system that can possibly lead to favoritism. They state that supervisors many times follow their social preferences, which in turn biases the outcome of the evaluation.

After some studies such as Zajonc (1980) and later Dipboye (1985) mentioned that not enough research focus was put on the "affect" variable, many new studies attempted to started looking at interpersonal affect and how it fits in the subject of performance evaluations (e. g. Cardy and Dobbins 1986, Shoorman 1988, Tsui and Barry 1986, Robbins & DeNisi 1994, Robbins & DeNisi 1998, Lefkowitz 2000) and generally concluded that affect has some effect on performance. One paper that is of particular interest for our work is Robbins & DeNisi (1998), who state that affect has a bigger effect on subjective ratings which include things like character traits, than to performance ratings which can be more objectively seen. They further say that the interpersonal affect seems to increase over time between a supervisor and his subordinates, which in the long term can mean an even further escalation of the bias in ratings.

In the last few decades, attention has been placed on variables connected to the individual who does the appraisals. Lefkowitz (2000), in a literature review, summarized his findings by pointing out that a positive affect towards someone leads to higher evaluations, greater halo, better vertical relationships, less inclination to punish subordinates, and less accuracy.

Yet another topic that has received quite a bit of interest is motivation of the raters. Initial general consensus was that people who evaluate are motivated enough to do it accurately. This seems to have changed recently, as researches now for the first time asked the question if evaluators are really motivated to evaluate accurately (Levy & Williams, 2004). One such example, that touches the topic of the motivation of the rater and is relevant to our topic, is Villanova et al (1993). The goal of the paper was to study the level at which raters felt uncomfortable while appraising their subordinates. For this purpose the authors developed the Performance Appraisal Discomfort Scale, and found that evaluators who were higher on the discomfort scale were more likely to give lenient ratings as they didn't want to deal with the confrontation that would arise. Similarly Klimovski and Inks (1990) have found that evaluators tended to give higher evaluations more when they were held accountable to the evaluatees for the given evaluations. This means that, for example, when evaluators expected to need to elaborate their decisions to their evaluatee in a face-to-face meeting, they were much more likely to distort their appraisal.

Other researchers have followed a different approach, namely trying to study the different types of general biases common in organizations, and through it the incentive design of the organizations. Key literature has mostly talked about the two most recognized types of evaluation bias: centrality bias and leniency bias (Ex. Prendergast, 1999). Centrality bias is the inclination to group performance ratings closer together than actual performance dictates which leads to less variance, while leniency bias provides employees with unrealistically high subjective performance ratings. Bol (2011) state that the causes of supervisor bias include: "employee performance, differences in organizational hierarchy, the financial position of the firm, the length of the employee-supervisor relationship, and supervisor characteristics". Further, the author found that supervisor bias affects future employee incentives, in a way that leniency bias improves performance, while centrality affects performance negatively. Bol (2011) also found that each supervisor has their own utility, and thus not all supervisors bias their evaluations equally. Finally, the author finds that supervisors do take into account the possible consequences of communicating performance ratings in order to determine the extent of bias.

3. Design

Our design is meant to capture the essential parts of the internal company processes which we want to study.[2] We frame the experiment and its roles "naturalistically", naming roles, job positions, and tasks as they would be inside a real company.

The situation we are interested in representing is one where a manager has to evaluate three employees who work for him. While one of those employees is assigned to the manager from the outset, the other two are assigned to him after the manager has made hiring decisions pertaining to the two of them. The manager will have hired one of the two employees and will have preferred not to hire the other employee, who will nevertheless be assigned to him. Our focus is on how the manager subsequently evaluates the three employees that have been assigned to him in different ways.

---

[2] The instructions can be found in Appendix A.

The experiment consists of three treatments, the *baseline* treatment, the *pay-for-performance* treatment and the *experience* treatment. In the baseline treatment the manager's evaluations will have payoff consequences for the employees but not for the manager. In the two subsequent treatments we will motivate the manager in two different ways: financially and by letting him experience the employees' task.

3.1. The Baseline Treatment

Each session of the experiment has sixteen participants and consists of four parts. There is only one experimental round.[3]

At the beginning of a session each participant is assigned to one of three *types*. The first type consists of *managers* (who make the hiring recommendations), the second type consists of *given employees*, employees who are told that they are already working for a manager, and the third type consists of *potential employees* who are looking for a job. In each session there are five managers, five already assigned employees, and six potential employees. During the first part of the experiment all potential employees are given a personality questionnaire to fill out. In the second part each manager selects one of the potential employees and recommends that he is hired.[4] This employee is then effectively assigned to him. Subsequently, another of the potential employees whom the manager decided not to recommend is also hired and assigned to him randomly.

As a result of the process, each manager ends up with three employees: one who was been assigned to him from the start (*given employee, GE*), one who he recommended to hire and was hired (*recommended employee, RE*), and one whom he had the opportunity to hire but didn't hire (*non-recommended employee, NRE)*. This can be achieved because the six potential employees can be hired by or assigned to more than one manager, the reasons and implications of which will be discussed in more detail in the following sections. Participants in the role of manager receive a

---

[3] We think that it is appropriate to first focus on escalation bias as a static phenomenon. Dynamic aspects of escalation bias could be studied in future work.
[4] Throughout we will refer to the manager *recommending* to hire a particular employee and *not recommending* to hire the others. In the experiment it later turns out that the recommended employee is effectively hired, but that one of the non-recommended employees will also be hired and assigned to the manager in question by top management.

fixed sum of twelve euros. The way employees are paid is explained below. Participants knew from the start about the four parts of the experiment. We now describe the four parts in more detail.
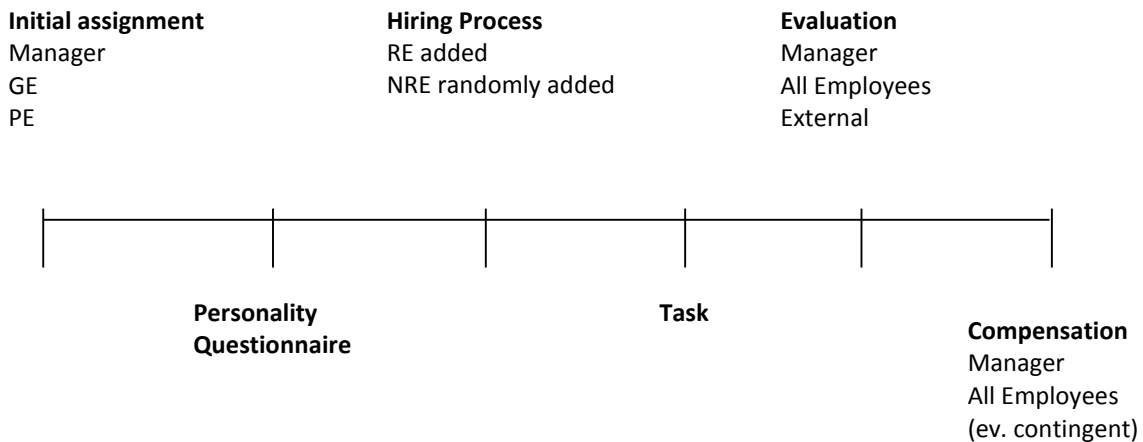
Figure 1. Timeline

M: Managers, GE: Given Employee, PE: Potential Employees, RE: Recommended Employee, NRE: non-recommended employee.

3.1.1. The Personality Questionnaire

After the sixteen participants of an experimental session have found out what type they have been randomly allocated to, the six "potential employees" start filling out a personality questionnaire, while the other ten participants wait. At this point none of the participants has any information about the hiring decisions that come later.

The purpose of the personality questionnaire is to provide the manager with information about the potential employees, so that he has the impression that later he will be able to make an informed hiring decision. In other words, we wanted to give the manager a basis for making their decisions a purposeful one, yet not in a way that he would give them full information on exactly how hired employees would perform in the task ahead. Failer at al. (2013) find that individuals in

environments which have censored information tend to rely too heavily on the censored information they have, causing them to form biased beliefs.

The personality test provided is a BFI-10 test, a 10-item short version of a widely used and recognized Big Five Inventory Test, with the Big Five being: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The official Spanish translation of the test can be found in Appendix B. We opted for this test because of its short length, as this meant that the other participants who didn't have to make a decision at this stage didn't have to sit idly for long. Despite its brevity, the test has been found to possess adequate psychometric properties. (Rammstedt and John (2006)).

3.1.2. The Hiring Process

In the second part of the experiment each of the five managers of a session is given the personality test answers of the six potential employees that are looking for a job in his company. Here it is important to note that the personality test results were purposefully not aggregated (as is often done), so the managers could see all 10 questions and each corresponding answer. The reason we did this is because it further differentiates the six candidates, and mimics a "question and answer" structure akin to that of a real-world interview situation.

The manager is instructed that there are two open positions in his department and that he can recommend one of the six candidates, who will then be hired. Participants are told that the other hiring decision will be made by top management. Since the manager sees the questionnaire responses of all potential employees in his company, it is possible that more than one manager decides to hire the same employee. The reason we allow for this is so that every manager gets to choose from an equal amount of potential employees, while keeping the simplicity of not having to send anyone home in the middle of each session, or have them wait idly for a long time.

Once managers have made their hiring decisions they are told which one of the remaining five potential employees top management has decided to hire and assign to them. The top management hiring decision is in fact a randomly selected potential employee who has not been hired by any of the five managers. In the end all six potential employees are hired either as a first choice by the hiring manager, or as a second choice by top management. Note that the experiment is designed

in such a way that the same potential employee cannot be both a recommended employee and a non-recommended employee in different groups. The given employee is distinct for each of the managers.

At the very end of this stage the six potential employees are informed about whether they have been hired and whether they are first choice picks hired directly by the hiring manager (recommended employee, RE) or not (non-recommended employee, NRE). Now every manager has three employees working for him, one given employee, one recommended employee, and one non-recommended employee.

3.1.3. The Task

During the third part of the experiment all employees, regardless of how they were assigned to their managers, are given the same task to perform individually. The task consists of reading a short two-page business case and answering two questions about it that do not have quantifiable correct answers. This is done in order to mimic real life company problems where there isn't just one strict way to get things done. The responses to this case study warrant a subjective evaluation.

The English translation of the case study can be found in Appendix C. The case study refers to a company and its plan to perform an expansion strategy. At the end of the description of the case, there are two questions each participant in the potential employee role has to answer.

1. Comment on the relationship between quantity and quality. Do you think that there is always an inverse relationship between both of them?

2. Which objectives is this expansion plan pursuing?

The employees know that each of them is given the exact same case study which is going to be evaluated by the manager and their peers, the other two employees in their company. *They are also told that their pay will be based solely on the manager's evaluation of their answers.* In addition, they also know that the manager has no financial incentives in this baseline treatment.

3.1.4. The Evaluation of the Task

In this part of the experiment managers are shown the answers to the two open-ended textual questions on the case study provided in the previous part.

They see the answers of all three of their employees simultaneously and are asked to evaluate each employee's performance on a scale of 1-100. At the same time than the manager, each employee is shown the answers of the other two employees working under the same manager and asked to evaluate them on a scale from 1-100. Importantly, managers and employees know to what type of employee each case study belongs to.

In the case that more than one manager decides to hire the same participant or the same person is hired by more than one top management, that person's task is evaluated by all the managers and co-employees in all groups in which he has been hired. Neither the managers nor the person in question is told that he is now a participant in multiple groups. To determine his final compensation a group is chosen at random (from the ones he is in).

At the end of the session 36 euros are proportionally split between each of the three employees working under a manager, with the proportions based solely on the score of each of the employees evaluated by each manager. For example, if the given employee received a grade of 80, and the recommended and non-recommended employees both received a grade of 40, the given employee will receive 18 euros in compensation while the other two will receive 9 euros each. Recall that in this first baseline treatment the manager is paid a fixed sum of 12 euros.

3.1.5. The External Evaluation

With this evaluation the experimental sessions of the baseline treatment formally finish. After the session all of the employees' case study answers were taken, secretly coded, and given to three external evaluators to grade. The external evaluators were the same for all sessions of all three treatments. They are PhD students in business economics, who do research in and teach courses in business economics, management and related subjects at the Universitat Autonoma de Barcelona. In particular, they have experience in evaluating students' answers to case studies like the one we use in the experiment. The external evaluators do not participate in the experiment and have no

way to know which answer belongs to which person, type, or group. Just like the managers within the experiment, they grade the answers with a score from 1-100. Because these external evaluators are experts and have no means to be biased in any way we regard their score to be a proxy variable for employees' real performance.

The comparison between the average grade given by the three external evaluators and by the managers and employees will be the basis for our analysis of escalation in the results section. This is an important feature of our experiment.

3.2. The Pay-for-Performance Treatment

The three treatments are very similar in design, so the second and third treatment will be explained by highlighting the differences to the baseline treatment. Everything that is not clearly specified as different from the baseline treatment should be assumed to be identical.

The first difference of the pay-for-performance treatment with respect to the baseline treatment is that managers don't obtain a fixed payment. Their payment now depends on how the external evaluators evaluate the three employees that work for them. This procedure is meant to represent the fact that in organizations and companies managers' income may in part depend on the actual performance of those employees that they give support to within the organization. Each manager will receive 1/3 of the average grade of his employees evaluated by the three external employees.

The second difference with respect to the benchmark treatment is that in the pay-for-performance treatment the sixteen participants and the three external evaluators are all present in the session at the same time, separated in two rooms. The sixteen participants, with the same roles as in the baseline treatment are in one room and the three external evaluators are in the second room. The reason is that the external evaluators now need to be present during the experimental sessions, since their evaluations determine the earnings of the participants in the role of managers. They evaluate the employees "on the spot" at the same time that the employees are being evaluated by the managers and by their peers. All participants were informed of the presence of the external evaluators, but could not see them. In contrast, in the baseline treatment the evaluations of the

outsiders had no payoff consequences for the participants. They were only used to check for escalation bias.

3.3. The Experience Treatment

Our third treatment, the experience treatment, is identical to the baseline treatment introduced in section 3.1, with the exception that managers perform the same task as the employees and at the same time. Once the managers have performed the task they evaluate their employees, and the employees do the peer-to-peer evaluations. Managers and employees are paid just as in the baseline treatment. After the experiment all tasks (case study answers), including the managers', are evaluated by the external evaluators, who are not present during the sessions. These evaluations don't have any payoff consequences, but are the basis for our analysis of escalation bias.

4. Hypotheses

In this section we propose hypotheses both for manager and employee escalation bias. In both cases, the hypotheses we formulate apply to all three treatments. However, the rationales behind the hypotheses will differ between treatments.

4. 1. Manager Bias

Ideally a performance evaluation should be a true measure of performance of an individual employee. That is, any external factor, in particular, the fact that a manager has hired a particular employee should not affect the performance evaluation. Thus the following null hypothesis:

H10: M*anagers evaluate their employees in an unbiased way, independently of how employees were hired into the company.*

However, we know from previous work that people often suffer from escalation bias. Whenever a manager hires an applicant, he must have certain reasons and criteria by which he has made his

decision. This reasoning may remain as an anchor in people's minds, as at least a moderate amount of effort and time was spent into this decision process.

As discussed above, the main reason why we think that a manager might have a bias towards the ratee he has personally hired is because that employee represents his choice, which he may feel obliged to defend. Staw (1981) finds that one's decision to stick to a unfavorable course of action is helped by the need to preserve one-self. Managers are thus likely to distort unfavorable information though self-preservation defense mechanisms. Following the notion that humans find it hard to admit their mistakes yet are relatively quick to praise their good decisions, the alternative hypotheses posit that the performance appraising manager is going to be biased towards the employee.[5] We formulate two alternative hypotheses, one for positive bias and one for negative bias:

H1a: *Managers have a positive bias towards the employees they decided to hire.*

H1b: *Managers have a negative bias towards the employees they decided not to hire but were hired by a third party.*

As mentioned in the introduction, both the pay-for-performance and the experience treatments introduce what a priori could be seen as factors moderating the tendency towards biased evaluations.

In the pay-for-performance treatment managers are not paid a fixed sum. They are paid according to the average of the grades which their three employees have received by the external evaluators. This means that if his employees performed badly the manager will make less money. We conjectured that this would be a factor working against a bias. Prendergast & Topel (1993) state that favoritism is bigger when the supervisor is not responsible for the performance of the

---

[5] Another reason for a manager bias is likability. Managers make decisions on the basis of likeability and we know from earlier studies by psychologists that there is a clear connection between likability, attribution, and ratings. Regan, et al. (1974) examined the link between liking and attribution. They found that rater's attributions for ratee performance varied, depending on their liking for the ratee. In our setting, it is not easy that managers develop a liking for the employee he recommends. However, it can not be excluded.

subordinate. They continue on to say that the supervisors' incentives need to be aligned to those of the organization, something that could be accomplished by tying rewards to promotion and making supervisors responsible for the performance of the promoted subordinates.

In the experience treatment managers perform the task at the same time as the employees. We conjectured that by performing the task they would get a better feel for what it entails to perform it well, and would therefore be less likely to evaluate employees in a biased way. Tyler et al. (1999) find that the psychology of preexisting preference and post-experience evaluations will differ. This suggests to us that managers may be less lenient towards the recommended employees in the experience treatment because by going through the same experience they themselves have had to think more deeply about what the correct answer might be and at the end are more knowledgeable about the question, leading to more unbiased evaluations.

Additionally, literature on effort tells us that people value their effort higher than others' efforts, so having put a higher effort to solve the task themselves, they are likely to be less lenient towards employees who don't put much effort into solving the task. Franco-Watkins et al. (2011) found that when more effort is put in, there is a tendency to put a higher monetary value on that effort exerted as well as to compensate oneself and others differently in comparison to say windfall gains where not much effort is put in.

The experience treatment gives us the opportunity to study an additional issue. We can analyse whether the grade that managers obtain in the task has a significant effect on the way they grade, but also whether managers' performance score in the task affects their grading. Ideally, own performance should have no effect on evaluations. This is captured in the following null hypothesis:

H20: Managers' performance score in the task does not affect their biases in the performance evaluations.

One may also conjecture that managers who perform better in the task are less likely to be biased (both positively and negatively) towards their employees. If a manager performs better in the task it means that a) he understands the topic better and/or b) he has put more effort into the task.

However, some studies suggest that that the opposite might happen. It has been found that it is much harder to realize a bias in one's own decisions and action, than it is in the decision and actions of others. This is what is called a "bias blind spot," explored in Pronin, Lin and Ross (2002). West et al. (2012) find that bias blind spots are not lessened by measures of cognitive ability (cognitive scores, SAT scores etc.). Our data allow us to check for the presence of bias blind spot in a managerial context. Thus, we formulate the following alternative:

H2a: Managers' performance score in the task does affect their biases in the performance evaluations.

4.2. Employee Bias

Gomez-Mejía et al. (2005) discuss the extensive use in companies of peer-to-peer as well as 360 degree evaluations. Moreover, Baron and Kreps (1999) and Lazear (1998,) consider that including peers, clients and subordinates increases validity, reliability and legitimacy of the evaluation system. Its wide use and its supposed benefits justify our decision to include peer-to peer evaluation in this experiment. We were interested in seeing whether employees could also be affected by the way their co-workers have been hired into the company, even though they are not directly involved. If this were the case, it would suggest that such evaluations should be used with caution.

Similarly to the case of manager evaluations, ideally peer-to-peer evaluations should not be influenced by anything other than the performance of their co-employees. As for managers we posit a null hypothesis of no bias:

H30: Employees evaluate their co-employees in an unbiased way, independently of how co-employees were hired.

However, there are several potential influences of the hiring process on employees' peer-to-peer evaluations. One of these influences is conformism, the tendency to follow others' opinions and decisions when there is real or perceived pressure by these others (Cialdini and Trost, 1998). Since

the classic experiments reported by Asch (1955), conformism has been a topic analyzed for its wide implications in economics and management. Akerlof (1997) introduces a model where individuals want to conform, more in the sense of normative influence, when the concern to obtain approval of others is important. Even in the context of recruiting, Granovetter (2005) discusses several studies where workers entering a firm through recommendations appear to be more productive.[6]

In our experiment, even though employees didn't make hiring decisions, they were informed about the hiring decisions of their manager Conformity would lead to all employees favoring the recommended employee and disfavoring the non-recommended employee.

Another influence is that non recommended employees are likely to be dissatisfied because they weren't hired by the evaluating manager and therefore may exhibit this dissatisfaction by punishing the other employees. This is a form of displaced aggression. If NRE perform better in the task a positive bias towards them may be found in the GE ratings, if that causes them to believe that a correct hiring decision was not made by the manager. Due to the contradicting factors that may affect the peer-to-peer evaluations we formulate the following alternative hypothesis:

H3a: *Recommended Employees are biased by the way they and other employees have been hired into the company.*

H3b: *Non Recommended Employees are biased by the way they and other employees have been hired into the company.*

H3c: *Given Employees are biased by the way other employees have been hired into the company.*

---

[6] Bikhchandani et al. (1992) and Banerjee (1992) study a kind of conformity that arises rationally. In their models, agents make decisions sequentially observing both a private signal and the decisions of those who go before them. They found that agents choose to put aside the signals which they receive and follow their predecessors' decisions, even when their own signals offer a much stronger clue of what the correct decision is, an indication of informational influence.

5. Procedures

We conducted six sessions with each of the three treatments. The average running time of the first and third treatments was one hour, while it was an hour and 30 minutes for the second treatment, due to the presence of the outside evaluators in the session. The experiment was conducted at two of the computer rooms of the Universitat Autonoma de Barcelona and participants were undergraduate students from the university. The z-tree software was used to run the experiment (Fischbacher, 1999).

In each of the eighteen sessions there are sixteen participants. We therefore have 288 participants, of which 90 are managers, 90 are given employees, and 108 are potential employees.

6. Results
We discuss the results of manager and employee evaluations treatment by treatment.

6.1. Results of the Baseline Treatment

Table 1 reports the means of the external evaluations, manager evaluations as well as the three distinct peer-to peer-evaluations, where each type of employee does not evaluate the own type.[7]

As can be seen from the last row in the table there are differences in the mean grades for the different types of evaluators. One can see that grades given by the externals are lower than those for the managers, perhaps due to higher standards of the experts. This difference between the outside evaluators and the managers holds for all three types of employees, with the difference for the recommended employee being the largest.

Comparing the evaluations of the three types of employees one can see that the average grade given by the non-recommended employee is lower than for the other two types of employees, reflecting perhaps a general dissatisfaction for not having been recommended by the corresponding manager.

---

[7] Appendix D contains additional tables with information about the grading in the three treatments.

We now move to the statistical tests for escalation bias. Given the features of our design we need to standardize the evaluation grades. We use the evaluations of the given employees as the baseline. The given employees are not part of the previous hiring process, and managers have been told that they were assigned to them from the start. We think that their evaluations are a natural standard of comparison to which to compare the (possibly biased) evaluations of the recommended and non-recommended employees..

For manager evaluations, the standardization gives rise to two new variables, namely "RE–GE" and "NRE–GE". The first variable corresponds to the difference of the manager evaluation of the recommended employee and that of the given employee. The second variable similarly corresponds to the difference of the evaluation of the non-recommended employee to that of the given employee.

Table 1. Mean Grades (Baseline Treatment)					
	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NR Evaluates
GE	59,88	64,97		59,21	52,18
RE	55,96	72,00	67,10		55,06
NRE	56,19	62,29	53,11	62,00	
Mean Grade	57,34	66,42	60,11	60,61	53,62

Figures 2 and 3 show average differences in the standardized evaluations for managers and employees respectively, together with the corresponding comparisons with the evaluations of the external evaluators. Starting with Figure 2, there appears to be a large difference between the means of the RE-GE variable of the managers and external evaluators. In fact the differences have opposite signs. Managers grade recommended employees higher than the given employees, whereas the external evaluators consider that the given employees performed better. Recall that external evaluators do not know which case study response belongs to which type of employee, so that they have no basis for discrimination.

The statistical backing for the manager bias in this treatment comes from the results of a one-tailed Wilcoxon signed-rank test. The test finds a significant difference in the grading of the RE-GE variable with a $p=0.02$. Remember that the given employee did not take part in the hiring process and, therefore, there was no obvious basis for any bias towards them. Additionally we know from

Table 1 that on average managers did not evaluate the given employees lower than the external evaluators, so that the higher difference between the evaluations of the recommended employee and the given employee comes from the higher grade given to the recommended employee by the managers. For the baseline treatment our evidence is consistent with H1a; managers have a positive bias towards employees who they have personally decided to hire.

We move on to the last two bars of Figure 2 to check for the presence of a negative bias. It seems clear that there is not much difference between managers' evaluations and the actual performance (external evaluation) of the NRE-GE difference. The one-tailed Wilcoxon signed-rank test for our second hypothesis finds a p=0.406. We can not reject the null hypothesis in favor of H1b, there is no negative bias of evaluating managers towards employees explicitly not hired by managers but nevertheless assigned to them.

Figures 2&3

Manager Evaluation Means (Baseline Treatment)	RE-GE	NRE-GE
Manager	7.03	-4.32
External	-3.92	-4

Peer-To-Peer Evaluation Means (Baseline Treatment)	NRE-GE (RE)	RE-GE (NRE)	RE-NRE (GE)
Employee	2.5	3.06	17.55
External	-0.96	-5.27	1.56
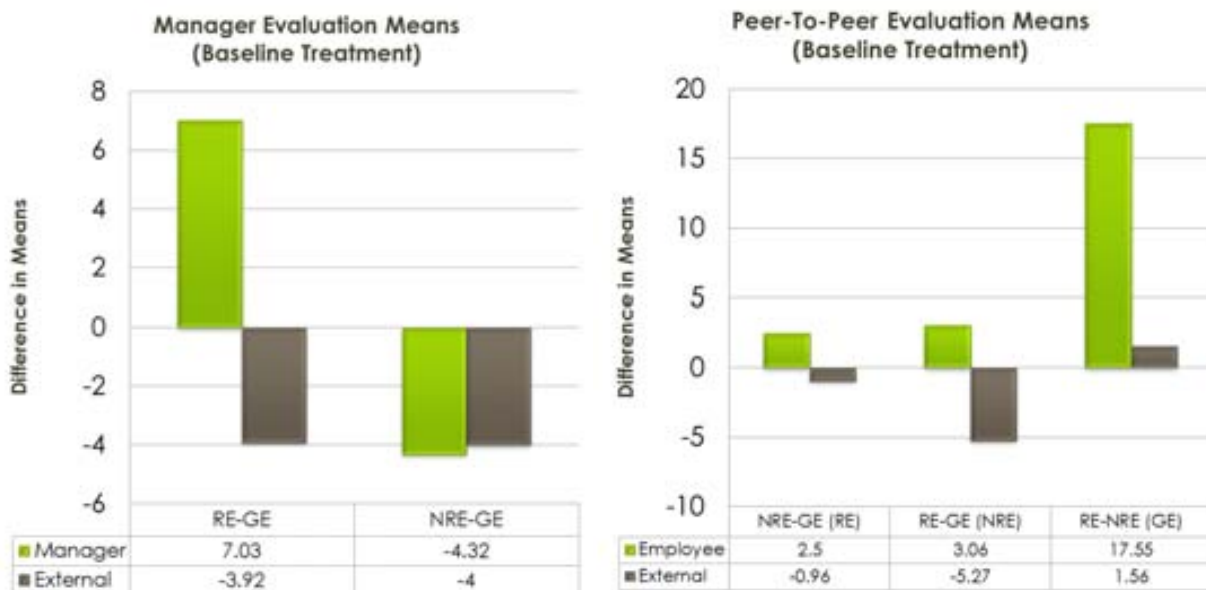
In Figure 3 one can see the information pertaining to the peer-to-peer evaluations. Each employee evaluates the other two types of employees who work under the same manager; therefore by taking

their difference three peer-to-peer variables can be formed. The variable for recommended employees' evaluations can be formed by taking the difference between their evaluations of non-recommended employees and that of given employees ("NRE-GE(RE)"). In this same way we form the other two peer-to-peer variables, namely "RE-GE(NRE)" and "RE-NRE(GE)". These are again compared to the external evaluators' scores of the identical pairs that each employee evaluated, where in parentheses we denote which type of employee has performed the evaluations in each case.

The biggest difference in the peer-to-peer evaluations can be seen in the last two bars of Figure 3 which represent the ratings of given employees. Given employees graded the recommended employees, 17.55 points higher than the grade of the non-recommended employees. Outside evaluators find an average difference of only 1.56 points.

We ran a two-tailed Wilcoxon signed-rank test of the RE-NRE difference evaluated by the given employees compared to those exact matched pairs evaluated by the external evaluators. We found that given employees have evaluated the difference of RE-NRE higher than the true difference, a result which is statistically significant (p=0,002). The result shows that existing employees working under a manager tend to be affected by decisions made by their manager, to a point that their evaluations become very biased, consistent with the notion of conformity discussed above.

For the evaluations done by the recommended and non-recommended we don't find a significant difference with respect to outside evaluators. We ran a two-tailed Wilcoxon signed-rank test of NRE-GE evaluated by the RE compared to those exact matched pairs evaluated by the external evaluators.[8] It shows that there doesn't seem to be a clear direction in the way the Recommended Employees have graded their two co-workers. Therefore it is safe to conclude that employees hired by the manager are not affected by this hiring decision when evaluating their peers.

We next look at the evaluations done by the NRE, specifically the difference in scores of RE-GE. Although the pattern of differences is the same as for the manager, the Wilcoxon test shows that the result is not significant though (p=0,224). This could be due to the tendency towards

---

[8] The number of observations is different between managers and given employees on one side and the other types of employees on the other side.

conformity being compensated by a pull in the other direction, because of the non-recommended employees' dissatisfaction with not having being selected.

In summary, for the baseline treatment we find that for the manager the evidence is consistent with positive escalation bias but not with negative escalation bias. In addition we find a bias in the peer-to-peer evaluations: the given employee biase s his evaluation in favor of the recommended employee and against the non-recommended employee.

The next treatment is meant to be an environment where the biases are more difficult to arise.

6.2 Results of the Pay-for-Performance Treatment

Table 2 shows the mean grades of the second treatment. Compared to Table 1, we see a decrease in the mean grade obtained by both the recommended and the non-recommended employees when looking at the manager evaluations. Here it is important to note that their actual performance (as proxied by the external evaluations) has in fact increased. In Table 2 we also see for the first and only ti me manager evaluations be lowe r than the re al performance, wh en looki ng a t the non - recommended employees.

What seems to remain constant between Ta ble 1 a nd Ta ble 2 (the ba seline a nd the pay-for-performance tre atments) is that the biggest difference in the evaluations done by the employees pertains to the given employees and it goes in the same direction. On average, given employees evaluated the recommended employees much higher than the non-recommended ones even though their real performance doesn't seem to be that different.

Table 2. Mean Grades (Pay-for-Performance Treatment)					
	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NR Evaluates
GE	60,33	69,20		61,95	62,00
RE	62,06	67,30	65,43		68,31
NRE	60,00	57,83	54,43	54,65	
Mean Grade	60,80	64,78	59,93	58,30	65,16

Figure 4 shows the means of the standardized evaluations of managers and external evaluators in the pay-for-performance treatment. Unlike in the baseline treatment (Figure 2), we now don't find

a big difference in the RE-GE variable between the managers' and the external evaluations. Indeed, there has been an elimination of the positive bias managers previously exhibited towards the employees they personally recommended for hire. The one-tailed Wilcoxon signed-rank test finds p=0.375.[9] We can therefore not reject the null hypothesis of no bias towards the recommended employee in this case.

However we now find that managers graded the NRE much lower than the GE, whereas the external evaluators find that in fact there isn't much difference in the real performance between the two types.

The statistical backing for the manager bias again comes from the results of a one-tailed Wilcoxon signed-rank test which finds p=0.026. Hence, in this case we can reject the null in favor of H_{1b}. An interpretation of this result is that, since the manager's pay now depends on the performance of his employees, he is less lenient in his evaluation of the recommended and the non-recommended employee. This overall decrease of the grade level leads to the disappearance of a positive bias and the appearance of a negative one. Instead, the bias is displaced. Hence, incentives do have an effect but not that of eliminating the bias.

We now move to the peer-to-peer evaluations. The pay-for-performance treatment does not introduce any change of employees' incentives, who still performed the same tasks and were evaluated and paid the same way as before. The only change was that employees now learned, from the beginning, that their real performance as well as that of the other employees had an impact on the pay of their manager. The perhaps natural conjecture in this case is that peer-to-peer evaluations would not change, since nothing has directly changed for the employees. However, an indirect effect can not be excluded by which, in a kind of mimetic reaction, given and recommended employees also become more demanding with respect to the non-recommended employees.

Figure 5, shows the results. As before, we don't find any significant results in the peer-to-peer evaluations of the recommended employees and of the non-recommended employees. The only notable change in employees' grading pattern has been the direction in which the recommended

---

[9] In fact, from the 30 observations, fifteen managers graded the RE higher than the GE and the other fifteen graded the GE higher than the RE.

employee have graded, with the non-recommended employees now receiving more than seven points less than given employees, while their true performance difference is , again, close to 0. However, this result only has a significance of p=0.156 using the two-tailed Wilcoxon signed-rank test.

Figures 4 & 5
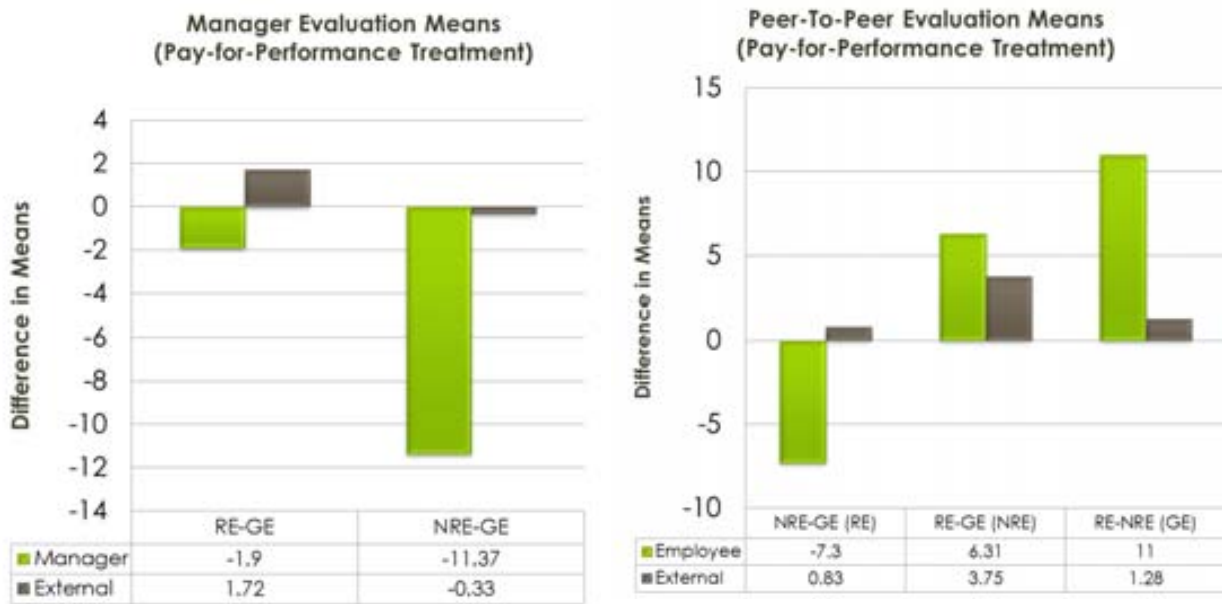
The evaluations of the NRE didn't differ a lot from their real performance and thus the two-tailed Wilcoxon signed-rank test result of p=0.756.

The biggest difference in means, again, comes from the given employees who evaluate the recommended employees considerably higher than the non-recommended like in the baseline treatment, again consistent with conformist behavior with respect to their manager. Table 2 and Figure 5 show us that GE have on average evaluated RE employees 11 points higher than NRE employees, whereas the difference in true performance is almost equal to 0. Using a two-tailed Wilcoxon signed-rank test we have found the significance to be just over the 5% significance mark, with p=0.054.

6.3. Results of the Experience Treatment

Table 3 shows the mean grades of the experience treatment. As with the other two treatments managers grade on average higher than the external evaluators. Looking at the managers grading we can see that the pattern of the means is very similar to that of the second treatment. The non-recommended employees receive by far the lowest average grade.[10] Looking at the peer-to-peer evaluation means one can see that they are also similar to those of the second treatment. The biggest two differences come from the evaluations of RE and GE, with both evaluating the NRE far lower than then each other. The main discrepancy with respect to the pattern for the pay-for-performance treatment is the non-recommended employees' evaluation of the recommended employee.

Table 7. Mean Grades (Experience Treatment)					
	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NR Evaluates
GE	60,00	71,83		65,31	60,20
RE	59,28	65,57	63,27		57,45
NRE	56,39	59,57	52,53	53,75	
Mean Grade	58,56	65,65	57,90	59,53	58,83

For managers' evaluations, the significant results are the same as for the previous treatment. Looking at the first two bars of Figure 6 the manager's mean evaluations don't even favor the RE in comparison to the GE, and the external evaluator's scores are again almost equal (close to zero). Again, we found no significant results regarding a bias towards the RE, the 1 tailed Wilcoxon signed-rank test significance is 0.216. From the 30 observations, showed that fourteen managers had scored a higher difference in favor of the RE and fourteen have scored a higher difference in favor of the GE, and there were 2 ties. In summary, when managers perform the task beforehand they lose the positive bias towards the RE.

Now moving on to the last two bars of Figure 6 we see a big difference between the variable which represents the managers grading difference and the variable which represents the difference in true performance. Again as in the second treatment we find that with our measures when the positive bias towards the RE disappears, a negative bias towards the NRE develops. Again the result is

---

[10] Managers' average grade in their task was 56.17, which was slightly lower than the mean grade.

significant; the one-tailed Wilcoxon signed-rank test finds a statistical significance of p=0.04. We therefore reject the null of no bias in this case.

Overall, we find that, as for the pay-for-performance treatment managers' evaluation bias is displaced from a positive bias towards the RE to a negative bias towards the NRE.

As discussed in section 4.1 we can also study whether the manager's own performance has an effect on the level of bias he shows in his evaluations. Perhaps surprisingly, we find that when managers perform better in the task they are more likely to be biased. An OLS regression of RE-NRE on the managers' grade in the case study finds a positive coefficient of 0.387 (p=0.021). Hence, we reject the null in favor of H_{2a}.

Figures 6&7

Manager Evaluation Means (Experience Treatment)		
	RE-GE	NRE-GE
Manager	-6.27	-12.27
External	-0.72	-3.61

Peer-To-Peer Evaluation Means (Experience Treatment)			
	NRE-GE (RE)	RE-GE (NRE)	RE-NRE (GE)
Employee	-11.56	-2.75	10.73
External	-6.46	-1.25	2.28
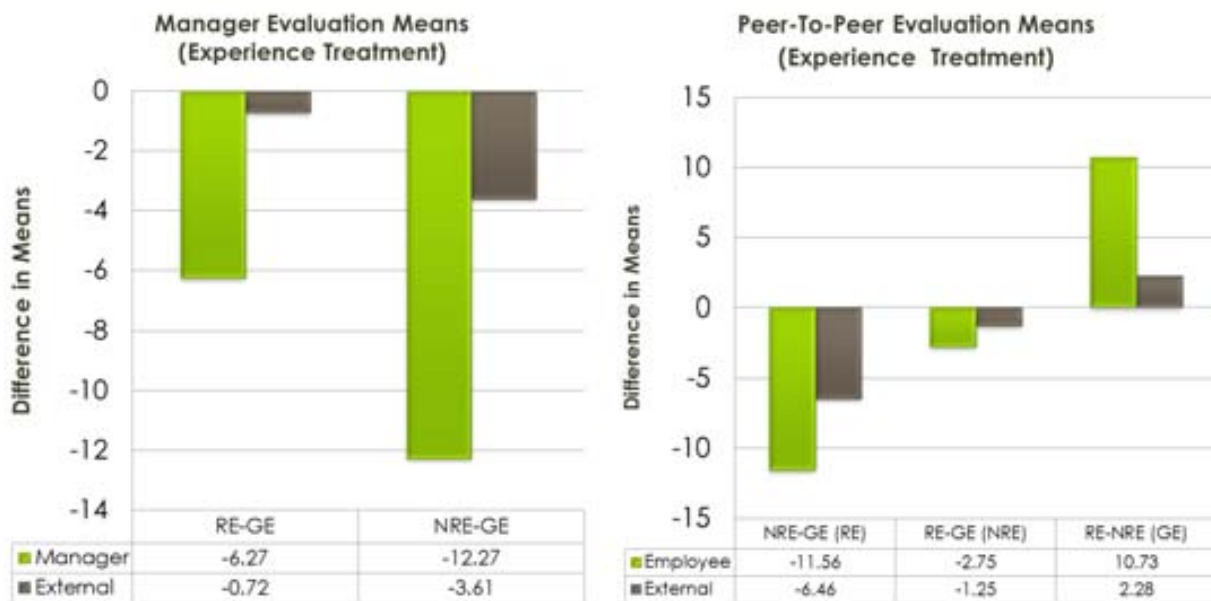
Figure 7 shows the peer to peer evaluation descriptives for the third treatment. From it yet again we see very similar results to the previous treatments. The only difference which is significant is the grading of the given employees, which again favors the RE over the NRE. The statistical backing for the GE bias comes from the results of a two-tailed Wilcoxon signed-rank test, which finds p= 0.044. The grading of the GE has been the most constant result in this experiment, not

changing throughout the 3 treatments. The given employees have constantly graded the NRE lower than the GE, in true conformist fashion, compared to the grading of the outside evaluators.

7. Conclusions

We set out to produce a detailed experimental study of hiring and escalation bias in subjective performance evaluations. More broadly, we want to contribute to the growing experimental and behavioral literature that studies managerial problems as in recent work by Brandts and Solà (2010), Harbring and Irlenbusch (2011), Berger, Harbring and Sliwka (2013) and Corgnet and Hernan (forthcoming).

Using three treatments we find that managers exhibit either positive escalation bias towards the employees they decided to hire or negative escalation bias towards those employees they decided not to hire but were nevertheless assigned to them. Both the introduction of material incentives for managers and experience of managers with the same task that employees have to perform leads managers to become less lenient with both recommended and not-recommended employees. The bias is displaced from a positive to a negative one.

Our results also show that the managers' own performance in the task has an influence on the escalation level. Somewhat counter intuitively managers are more biased (escalate more) when their performance is higher. We believe this could be due to the blind spot bias and a kind of over-confidence, as managers who feel confident with their performance in the task may automatically feel more confident with their initial hiring decision.

Another contribution of our research is that it shows that escalation bias doesn't only affect the people who made the initial decision, but that it can also affect others in the organization possibly due to conformity. Employees who were not part of the original decision consistently give more weight to the information coming from the manager that one person had been hired over another one, than to the in principle more important information coming from their own analysis of their co-employees' performance. In this sense, manager and employee bias are connected and employee evaluations can not serve as a counter-balance to those of managers.

8. References

Akerlof, G.A., (1997) "Social distance and social decisions" *Econometrica* **65**(5) 1005-1027.

Asch,S.E., (1955) "Opinions and social pressure" *Scientific American* **193**(5), 31-35.

Banerjee, A., (1992) "A simple model of herd behavior" *Quarterly Journal of Economics* **107**(3), 797–817.

Baron, J. N., and D. Kreps (1999) "Strategic Human Resources" John Wiley & Sons.

Bazerman, M. H., R. I. Beekun, and F. D. Schoorman (1982), "Performance evaluation in a dynamic context: a laboratory study of the Impact of a prior commitment to the ratee" *Journal of Applied Psychology* **67**(6) 873-76.

Berger, J., Harbring C. and D. Sliwka (2013), "Performance Appraisals and the Impact of Forced Distribution—An Experimental Investigation" *Management Science* **59**(1), 54-68.

Bikhchandani, S., D. Hirshleifer, I. Welch (1992), "A theory of fads, fashion, custom, and cultural change as informational cascades" *Journal of Political Economy* **100**(5), 992–1026.

Bol, J. (2011) "The determinants and performance effects of managers' performance evaluation biases" *The Accounting Review*: September 2011, Vol. 86, No. 5, pp. 1549-1575.

Brandts, J. and C. Solà (2010) "Personal relations and their effect on behavior in an organizational setting: an experimental study", *Journal of Economic Behavior and Organization*, **73**, 2010, 246-253.

Cardy, R.L. and D. Dobbins (1986) "Affect and appraisal accuracy: liking as an integral dimension in evaluating performance", *Journal of Applied Psychology*, **71**(4): 672–8.

Cialdini, R. B., and M. Trost (1998) "Social influence: social norms, conformity and compliance" In Gilbert, D.T., Fiske, S:T., and Lindzey, G. (Eds), *The Handbook of social psychology, Vol. 2*. Fouth Edition, McGraw-Hill.

Corgnet, B. and R. Hernan (forthcoming), "Don't ask me if you will not listen: the dilemma of participative decision making," *Management Science.*

Dipboye, R.L. (1985) "Some neglected variables in research on discrimination in appraisals", *Academy of Management Executive*, 10(1): 116–27.

Feiler DC, Tong JD, Larrick RP (2013) "Biased judgment in censored environments" *Management Science*. **59**(3), 573-591.

Falk, A., and J. J. Heckman. (2009) "Lab experiments are a major source of knowledge in the social sciences" *Science,* **326**(5952), 535-538.

Franco-Watkins, A. M., B. D. Edwards, and R. E. Acuff (2011) "Effort and Fairness in Bargaining Games". *J. Behav. Decis. Making* doi: 10.1002/bdm.762.**26**(1), 79-90.

Fischbacher, U. (2007) "Z-tree: zurich toolbox for ready-made economic experiments" *Experimental Economics* 10.2 171-78. Print.

Gibbs, M., K. A. Merchant, W. A. Van der Stede, and M. E. Vargus (2004) "Determinants and effects of subjectivity in incentives" *Accounting Review* **79**(2), 409.436.

Gomez Mejía, L. , Balkin, D. and Cardy, R. (2005)"Managing Human Resources" Pearson Education.

Granovetter, M. (2005) "The impact of social structure on economic outcomes" *Journal of Economic Perspectives* **19**(1), 33-50.

Harbring, C., and B. Irlenbusch (2011) "Sabotage in Tournaments: Evidence from a Laboratory experiment" *Management Science*, **57**, 611-627.

Judgde, T. A., and G. R. Ferris (1993) "Social context of performance evaluation decisions." *Academy of Management Journal* **36**(1), 80-105.

Klimoski, R.,& L. Inks (1990) "Accountability forces in performance appraisal" *Organizational Behavior&Human Decision Processes*, **45**(2): 194.

Lazear, E. P. (1998) "Personnel economics for managers" John Wiley & Sons.

Lefkowitz, J. (2000) "The role of interpersonal affective regard in supervisory performance ratings: a literature review and proposed causal model." *Journal of Occupational and Organizational Psychology* **73**: 67-86.

Levy, P. E., and J. R. Williams (2004) "The social context of performance appraisal: a review and framework for the future." *Journal of Management* **30**(6), 881-905.

Milkovich, G. T. and A. K. Wigdor (1991) "Pay for performance" *National Academy Press.*

Prendergast, C. J. (1999) "The provision of incentives in firms" *Journal of Economic Literature* **37**, 7-63.

Prendergast, C., and R. Topel. (1993) "Discretion and bias in performance evaluation" *European Economic Review* **37**: 355-365.

Prendergast, C., and R. H. Topel (1996) "Favoritism in organizations." *Journal of Political Economy* **104**.5: 958.

Pronin, E., Lin, D. Y., & Ross, L. (2002) "The bias blind spot: perception of bias in self versus others" *Personality and Social Psychology Bulletin 28,* 369–381.

Rammstedt, B., and O. John (2007) "Measuring personality in one minute or less: a 10-item short version of the big five inventory in english and german" *Journal of Research in Personality* **41**(1): 203-12.

Regan , D. Straus , E. Fazio , R. (1974) "Liking and the attribution process" *Journal of Experimental Social Psychology*, **10**, 385-397.

Robbins, T.L. and DeNisi, A.S. (1994) "A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations", *Journal of Applied Psychology*, **79**(3): 341–53.

Robbins, T.L. and DeNisi, A.S. (1998) "Interpersonal affect: identifying process and rating distortions in performance appraisal*", Journal of Business and Psychology*, **12**(3): 313–25.

Schoorman, F. D. (1988) "Escalation bias in performance appraisals: an unintended consequence of supervisor participation in hiring decisions" *Journal of Applied Psychology* **73**(1), 58-62.

Staw, B. (1976) "Knee-deep in the big muddy: a study of escalating commitment to a chosen course of action*1" *Organizational Behavior and Human Performance* **16**(1), 27-44. Print.

Staw, B. M. (1981) "The escalation of commitment to a course of action" *Academy of Management Review*, **6**(4), 577-587.

Tsui, A.S. and Barry, B. (1986) "Interpersonal affect and rating errors" *Academy of Management Journal*, **29**(3): 586–99.

Tyler, T. Y. J. Huo and E. A. Lind. (1999) "Pre-experience choices and post-experience evaluations" *Group Processes Intergroup Relations* **2**(2), 99-118

Villanova, P., Bernardin, H. J., Dahmus, S. A., & Sims, R. L. (1993) "Rater leniency and performance appraisal discomfort" *Educational &Psychological Measurement* **53**(3): 789–799.

West, R.F., Meserve, R. J., and Stanovich, K. E. (2012) "Cognitive sophisticaion does not attenuate the bias blind sport" *Journal of Personality and Social Psychology*, 1-14

Zajonc, R.J. (1980) "Feeling and thinking: preferences need no inferences", *American Psychologist*, **35**(2), 151–75.

Appendix A. Instructions

ALL TEXT IN CAPITAL LETTERS (LIKE THIS ONE) IS ADDED FOR READERS AND DOES NOT BELONG TO THE ORIGINAL INSTRUCTIONS

1. GENERAL. ALL PARTICIPANTS

Instructions

Welcome and thank you very much for your disposition to participate in this experiment. You will receive a minimum of 5 euros for participating in this experiment. Any contact to other participants in this room is from now on not allowed any more. If you have questions, raise your hand and we will come to your seat.

General information

For the purpose of this experiment you will be randomly split into 3 groups: Managers, Employees, and Potential Employees. The experiment starts off with 5 different companies all of which have a separate manager and 1 employee already working for each of those managers, the Assigned Worker. Each firm will end up with a Manager, a Given Employee, and two additional workers selected between the Potential Employees.

First Stage

The first step of the experiment is a hiring process where managers will choose which of the Potential Employees to hire, having the results of a personality test that will be conducted. Each firm will ultimately select two of the Potential Employees, but the Manager will only be able to select one Potential Employee.

Each potential employee will answer a standard test asking how the person identifies herself with respect to some statements concerning personality traits. Nobody in the experiment will know who gave what answers, we will identify answers with an anonymous code. Once Managers see the answers, each Manager will be able to select one Potential Employee for her firm, the Recommended Employee.

Finally the firm will be composed by a Manager, a Given Employee, A Recommended Employee and a non-recommended employee, randomly assigned among those no selected by the manager of the firm.

Following this all employees (newly hired and old) will have to do a task which will be evaluated by the manager.

Second stage

[IN BASELINE TREATMENT AND PAY-FOR-PERFORMANCE TREATMENT]

In this stage, each employee will develop an activity that will be later evaluated by the Manager in her firm as well as by the coworkers.]

[IN EXPERIENCE TREATMENT]

[In this stage everybody, Managers and Employees will develop an activity]

AL TREATMENTS

This activity will consist of answering two questions on a case study. These answers would allow the firm and the Manager specifically take the right decisions in his activity.

Third stage

In this stage Managers will evaluate the answers of their employees. Compensation for employees will depend only upon the relative evaluation of the answers to the case study by the Manager. Specifically, 36 euros will be distributed among the three workers under the evaluation of a manager considering the three evaluations. The split is not going to be in equal shares but relative to their performance, meaning how the manager evaluates the task they have just done. Every employee will receive at least 5 euros.

In addition, each employee will also evaluate her coworkers, but this evaluation will not affect compensation.

Answers by employees will also be evaluated by an external organization.

[BASELINE TREATMENT

Compensation to the manager will be a fixed amount of 12 euros.]

[PAY-FOR-PERFORMANCE TREATMENT

Compensation to the manager will be one thid of the mean evaluation that the external organization will do of the three employees. That is, if the mean evaluation by the external organization was 60 points, the manager will receive 20 euros.]

[EXPERIENCE TREATMENT

Compensation to the manager will be a fixed amount of 12 euros.]

ALL TREATMENTS

Concluding remarks

Keep in mind that your answers to the questionnaires as well as during the subsequent experiments will of-course be treated anonymously. As codes are used for identity nobody except the experimenters will know exactly which task results, and personality questionnaires belong to you.

The actual experiment starts now. Please continue to be quiet and avoid any communication with the other participants. If you have questions, please raise your hand.

INSTRUCTIONS INCLUDED IN THE PROGRAM AS THE EXPERIMENT UNFOLDS

2. BEFORE PERSONALITY QUESTIONNAIRE
POTENCIAL EMPLOYEES

You have been randomly selected as a potential employee. Your first task is to fill out the personality test provided. Please answer the questions from your personal perspective, by writing 1-5 in the space provided, to what degree that statement applies to you. Depending on these results, the managers will choose to recommend you for hire or not. Please answer all questions seriously and honestly. Your answers to the questionnaires as well as during the subsequent experiments will of-course be treated anonymously.

GIVEN EMPLOYEE

You have been randomly selected as an employee. As mentioned in the first part of the instructions you are an employee in a company that has two vacancies which are going to be filled in the first part of the experiment. As part of your job you will be asked to do a task which will be evaluated by the manager. The task will involve reading a case study and answering 2 questions related to it. The manager is then going to evaluate your answers. At the end you will be asked to evaluate the performance of the remaining two employees in your company. Your final payment is going to be based on the manager's evaluation of your answers as well as his evaluation of your colleagues, the new employees that are going to be hired in the first part of the experiment. More precisely 24 euros are going to be split between all 3 employees working for each manager. The split is going to be relative to how well the manager has evaluated you in comparison to the other 2 employees. Remember that the minimum you can receive is 5 euros.

MANAGERS

As all managers in this experiment you are a new middle-level manager in your company. There is already one employee working in your department, who has been hired by your predecessor. However there are still are two positions open in your department. Your task as manager will be to recommend one person to hire; who you think would do the best job in your department. The tool that is given to you to possibly help you make this decision is a personality questionnaire that the potential employees have filled. Further you will be asked to evaluate the task performance of all your employees, according to which they will be paid.

3. **AFTER PERSONALITY TEST**

MANAGERS

Step 1 -Your first task as manager is to hire a potential employee. Six people have applied for the job. As you know, they have been asked to do a personality test. You are now required to recommend which applicant you want to hire, based on their personality test questionnaire and the actual personality test results.

So please wait for all candidates to answer the personality tests. Then please make a recommendation as to who you would like to hire. That person will be hired into your company.

Even though there are two vacancies in your department you only have the right to recommend one person.

After recommending a person for hire, and having handed in your recommendation sheet you may go on to step two.

4. AFTER SELECTION AND ASSIGNMENT
RECOMMENDED EMPLOYEES

Congratulations, you have been recommended for hire by one of the managers, and have been hired in a company. One more employee has been hired into the company along with you, who hasn't been recommended by the manager in this experiment.

NOT RECCOMENDED EMPLOYEES

Even though you were no directly recommended for hire by any of the managers you have been hired into a company. Congratulations. One more employee has been hired into the company along with you, who has been recommended by the manager in this experiment.

ALL EMPLOYEES

As part of your job you will be asked to do a task which will be evaluated by the hiring manager. The task will involve reading a case study and answering 2 questions related to it. Later you will be asked to evaluate the performance of the remaining two employees in your company. Your final payment is going to be based on the manager's evaluation of your answers as well as his evaluation of your colleagues, the newly hired that was recommended by the manager and the existing employee. More precisely 24 euros are going to be split between all 3 employees working for each manager. The split is not going to be in equal shares but relative to how well the manager has evaluated you in comparison to the other 2 employees. Remember that the minimum you can receive is 5 euros

Employees please read carefully the case study presented and on the answer sheets provided answer the 2 questions at the end of the case study to the best of your ability.

MANAGERS

Please find the case study provided to you and read it carefully.

[IN EXPERIENCE TREATMENT

Now you are required to answer the two questions concerning the case study to the best of your ability.]

You will soon receive 2 questions and answers regarding this case study from all three of your employees. Based on this you will have to evaluate your employees. With the Q&A of your employees you will receive an evaluation sheet where you need to write the evaluation score for each of your employees.

5. **AFTER CASE ANSWERS**

ALL EMPLOYEES:

You are now required to evaluate the performance of your two co-employees. Wait to receive their case study answers and evaluate each person with a total score from 0 to 100 based on the answers given. Remember that your final payment doesn't depend on the peer to peer evaluation but solely on how the manager evaluates you.

MANAGERS

You are now required to evaluate the performance of your employees in their answers to the case study. You are to give each employee a score from 0-100. Your employee's payment depends on your evaluation. Precisely 24 euros are going to be split between all 3 employees working for you. The split is not going to be in equal shares but relative to their performance, meaning how you evaluate the task they have just done.

Appendix B. Personality Questionnaire

English version.
Instruction: How well do the following statements describe your personality?

I see myself as someone who ...	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
... is reserved	(1)	(2)	(3)	(4)	(5)
... is generally trusting	(1)	(2)	(3)	(4)	(5)
... tends to be lazy	(1)	(2)	(3)	(4)	(5)
... is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
... has few artistic interests	(1)	(2)	(3)	(4)	(5)
... is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
... tends to find fault with others	(1)	(2)	(3)	(4)	(5)
... does a thorough job	(1)	(2)	(3)	(4)	(5)
... gets nervous easily	(1)	(2)	(3)	(4)	(5)
... has an active imagination	(1)	(2)	(3)	(4)	(5)

Scoring the BFI-10 scales:
Extraversion: 1R, 6; Agreeableness: 2, 7R; Conscientiousness: 3R, 8; Neuroticism: 4R, 9;
Openness: 5R; 10 (R = item is reversed-scored).

Appendix C. Case Study

The Spanish company Lladró was born in 1953 when Juan, José and Vicente Lladró , three brothers sons of farmers with great artistic talent , founded a small family business in the Valencian town of Almácera . Lladró since then has undergone a huge metamorphosis from a craft workshop to a large international company in continuous expansion.

Company management, as in its origins, is under the control of the members of the Lladró family , who owns the company. Initially, the team just tried to adapt a vintage style, but in a short time certain traits appeared that would be recognized later on as the Lladró style. Since the beginning, the public was infatuated with their creations. In little time the workshop was expanded several times and an increasing number of collaborators multiplied the work of the brothers. As of today sculptures which are born in the City of Porcelain do so in a completely handmade process in which 2500 employees participate, exporting to more than 120 countries of five continents: the Netherlands, USA, United Kingdom and Japan being the most important.

Since 1955, year in which the first shop was opened in Valencia, Lladró has been increasing its network of stores in all major shopping malls in the world: Valencia , Madrid , London, New York , Beverly Hills , Singapore , Hong Kong, Las Vegas , Sydney . 6,900 authorized dealers exhibit art of Lladró porcelain.

However, the company's large expansion has a point of inflection at the end of 2001, when the Lladró announced the closing of 2,000 points of sale. In this way Lladró got rid of those dealers who were not taking care of the luxury image of its figures, and were exhibiting them together them with figures of the competition and even with imitations. The objective of this measure, with an effect of decreasing its billing by 17%, is to prove the company's commitment to quality, instead of quantity.

To recover from this measure, Lladró has put together an expansion plan consisting in opening 50 own points of sale to sell its new and innovative designs and to create a new image for the firm. This strategy will allow the firm to have a direct contact with its customers and it will complement the already established relationship with collectors through the Prestige Club Lladró, allowing

them to access exclusive collections and having privileged shopping options of limited edition products.

Among the objectives of this plan, one should point out the reinforcement of the quality image of the organization, the increase the value of sales by 9% of billings and the maintenance of the return on sales.

QUESTIONS:

1. - Discuss the relationship between quantity and quality. Is it always an inverse relationship?

2 . - What objectives Lladró pursues with the expansion plan based on Lladro's own shopping points? Would it be possible to attain these objectives with external distributors?

Appendix D. Descriptive Tables

D1. Manager Evaluation Descriptives (Baseline Treatment)			
	Recommended - Given		Not Recommended - Given	
	Manager Evaluates	External Evaluates	Manager Evaluates	External Evaluates
Mean	7,03	-3,92	-4,32	-4,00
Median	10,00	-4,17	10,00	-6,00
Minimum	-55,00	28,33	-90,00	-45,00
Maximum	70,00	20,00	51,00	21,67

D2 Peer to Peer Evaluation Descriptives (Baseline Treatment)					
	Not Recommended - Given		Recommended - Given		Recommended – Not Recomended	
	RE Evaluates	External Evaluates	NRE Evaluates	External Evaluates	GE Evaluates	External Evaluates
Mean	2,50	-0,96	3,06	-5,27	17,55	1,56
Median	10,00	-0,50	3,00	6,66	15,00	0,00
Minimum	-30,00	-21,67	-55,00	-28,33	-60,00	-25,00
Maximum	40,00	21,67	50,00	20,00	90,00	58,33

D3 Manager Evaluation Descriptives (Pay-for-PerformanceTreatment)			
	Recommended - Given		Not Recommended - Given	
	Manager Evaluates	External Evaluates	Manager Evaluates	External Evaluates
Mean	-1,90	1,72	-11,37	-0,33
Median	-4,50	4,17	-17,50	-1,67
Minimum	-50,00	-28,33	-75,00	-21,67
Maximum	60,00	20,00	45,00	35,00

D4. Peer to Peer Evaluation Descriptives (Pay-for-Performance Treatment)					
	Not Recommended - Given		Recommended - Given		Recommended – Not Recomended	
	RE Evaluates	External Evaluates	NRE Evaluates	External Evaluates	GE Evaluates	External Evaluates
Mean	-7,30	0,83	6,31	3,75	11,00	1,28
Median	-10,00	-1,67	7,50	9,17	15,00	4,17
Minimum	-70,00	-18,33	-40,00	-23,33	-45,00	-38,33
Maximum	75,00	35,00	70,00	20,00	55,00	25,00

D5. Manager Evaluation Descriptives (ExperienceTreatment)			
	Recommended - Given		Not Recommended - Given	
	Manager Evaluates	External Evaluates	Manager Evaluates	External Evaluates
Mean	-6,27	-0,72	-12,27	-3,61
Median	0,00	-0,83	-11,00	-5,00
Minimum	-60,00	-18,33	-75,00	-38,33
Maximum	47,00	25,00	50,00	25,00

D6. Peer to Peer Evaluation Descriptives (ExperienceTreatment)					
	Not Recommended - Given		Recommended - Given		Recommended – Not Recomended	
	RE Evaluates	External Evaluates	NRE Evaluates	External Evaluates	GE Evaluates	External Evaluates
Mean	-11,56	-6,46	-2,75	-1,25	10,73	2,28
Median	-10,50	-6,67	-5,50	-1,67	12,50	0,83
Minimum	-60,00	-36,67	-40,00	-18,33	-55,00	-28,33
Maximum	40	11,67	40,00	25,00	50,00	38,33

CHAPTER III

Escalation Bias and its Effects on Employees Performance

1. Introduction

Angelovski et al. (2014) in a lab experiment find escalation bias coming from both evaluating managers and evaluating co-workers in all three treatments they perform. They find that managers have a positive bias towards those employees they have hired or a negative one towards those they have explicitly not hired. The authors also find a connection between the managers' and employees' biases. Employees who are exogenously hired are biased in favor of employees who are hired by the manager and against those explicitly not hired by their manager.

The authors aim with the paper was to prove that escalation bias does occur in subjective evaluations, with the escalation being based on previous decisions made by the manager, which in their case was the hiring decision.

One of our aims with this paper is to add to their findings by looking at bias problems from the point of view of the employees who have to continue working in an environment in which they receive a positive or negative bias from their evaluating superiors. It has been shown that receiving a higher evaluation than one deserves can lead to a positive performance improvement (e.g. Bol, 2010). This is predicted to be true because it has been shown that workers tend to overestimate their performance, therefore, when being over-evaluated, employees tend to believe that they are evaluated fairly. We, however, do not know the impact of receiving positive bias when one is aware that the high grade, and thus payment, they have received, is not justified by the performance. Experiments on gift exchange indicate that the employees might reciprocate (Fehr et al. 1993; Fehr et al. 1998; Fehr and Falk 1999; and Charness 2004) by increasing the effort, yet agency theory predicts that both positive and negative biases that may arise from the bias of managers should have a negative effect on their employees' performance. The simple reasoning

being the agency theory version is that any bias brakes the sacred link between effort and the performance based compensation which is in place.

Additionally we aim to report whether escalation bias can still persist in a multi-stage scenario where the managers' biases are no longer hidden from the employees.

In order to find the answer to these questions we have taken parts of the design from Angelovski et al. (2014) and added to it in order to explore these new questions. In that paper the authors designed an experiment (the baseline treatment) in which every manager has an exogenous employee in the company, and is told that there are two vacant positions in his department, for which there are 6 applicants. The manager is asked to hire one of the 6 applicants, while from the remaining 5 another will be hired by upper management. After this decision is made participants perform a task that is to be evaluated subjectively by the manager and the other 2 employees in the company. The employees are paid based on the evaluation of the manager, and the experiment ends. All tasks are later evaluated externally (and anonymously) by "experts" to get a proxy for their real performance and thus find whether a bias does exist.

We take this same design, and adapt it slightly to our needs. First, we remove the employee evaluations, as we want to eliminate the noise that feedback from different sources may create. Secondly, we add 2 more rounds, therefore, two more tasks and two more evaluations. We do this in order to find the employees reaction in subsequent rounds. Thirdly, we use the external evaluators (the same three used in Angelovski et al. 2014) as part of the experiment, as the initial study did in the pay-for-performance treatment. They are placed in a separate room and evaluate the employees at the same time as the managers do. The fourth change is that now employees receive feedback between each round in a monetary form. Lastly employees have a choice to make at the end of round 2 whether to stay, sabotage, or quit the company. The decision could lead to an immediate effect on their pay as well as the pay of the others of the firm, additionally, it could have an effect on their own evaluations and pay in the subsequent round.

2. Background

For many years researchers and managers have tried to figure out how to exactly motivate and empower employees in order to perform better in their job environment. Spreitzer (1995) indicates that feelings of empowerment can be positively affected by two effects: feedback on the employees' feedback, as well as performance-based reward systems. Drake at al. (2007) follows up this study, and finds that it might be more complicated than initially thought. More specifically, they find that the techniques that work to increase manager perceptions of empowerment do not seem to work at lower organizational levels; and, even in situations when they do work, the actual increase in employee motivation is likely to not be significant.

Silveman and Waxley (2006) find that not being able to participate in the development of the appraisal or feedback process will lead to negative perception regarding the process and its negative outcomes. Jone and Lyman (1986) hypothesized and found that attitudes toward the performance appraisal systems and organizational commitment will decrease and remain low for those receiving satisfactory ratings, whereas the attitudes of those receiving higher appraisal ratings will remain unchanged. Leung et al. (2001) found that "criticism delivered with greater interpersonal fairness resulted in more favorable dispositional attributions about the supervisor, more acceptance of the feedback, and more favorable reactions towards the superior and the organization."

Even though many studies exist which confirm the existence of various types of biases in job appraisals (E.g. Prendergast and Topel 1996; Milkovich and Wigdor 1991; Klimovski and Inks 1990), research has yet to search for substantial evidence which shows the effects of biased evaluations on future performance by those evaluated. This gap in the literature is where this paper finds its motivation from. Using the method of inducing bias in evaluations used by Angelovski et al. (2014), we aim to go a step further in this study and find the effects on performance which may arise from potential positive and negative biases.

Agency theory predicts that both positive and negative biases that may arise from the escalation bias of managers should have a negative effect on their employees' performance. The reasoning behind this is that from the point of view of the employee, having received a biased evaluation

(either positively or negatively) destroys the incentives to perform well under performance based compensation. This is somewhat countered by behavioral economics, which believe that either positive or negative bias could have a positive effect on motivation to improve performance by increasing the perceived fairness of an incentive system. (Bol, 2010)

As mentioned earlier Bol (2010) finds that leniency bias has a positive effect on performance improvement. The author further argues that this result shows consistency with arguments from behavioral theory that bias can positively affect performance improvement by increasing the perceived fairness of an incentive system.

Bernardin and Villanova (1986), and Klimovski and Inks (1990), have found that more lenient ratings are given in order to avoid conflict and confrontation. This is consistent with the findings of Friedrich (1993) who states that managers have a tendency to bias evaluations upwards because employees could possibly perceive their ratings as being too low. This is particularly true when managers do not have complete information on employee's performance available to them. Villanova et al (1993) studies the level at which raters felt uncomfortable while appraising their subordinates. For this purpose the authors developed the Performance Appraisal Discomfort Scale, and found that evaluators who were higher on the discomfort scale were more likely to give lenient ratings as they didn't want to deal with the confrontation that would arise. Woods (2012) also finds that evaluations are more likely to be raised rather than lowered, even when these raises are not easily justifiable.

One of the reasons why we don't expect to find leniency bias across the board of all evaluations is that, with our design, confrontation between managers and employees is not possible. The design eliminates the possibility of this type of bias, and thus provides the means for escalation bias to be observed without much noise from other biases, as was shown by Angelovski et al. (2014). Also, as our system of pay is budget based, even if a small degree of leniency existed across all ratings, it would not have any effect on our results. What is important for our research though is the connection between leniency bias and positive escalation bias. In both cases employees get a more favorable grade than they deserve and the way this is perceived from the point of the employees is expected to be similar.

Garland (1973) has found that good performers tend to become less motivated when they receive the same compensation as someone who doesn't perform as well. This leads to them decreasing their efforts as they feel they have no incentives to continue working at that level of effort. Even though, in our design, employees cannot see the tasks performed by the other employees in their firm, it is likely that this same principle may apply if they compare their external feedback to the manager's feedback and find that they have been biased against by their manager. This would likely lead to a decrease in performance due to lack of incentives to continue working at the same level of effort.

Holmstrom (1979) shows that when performance is linked to payment, employees are motivated to increase effort, as per agency theory. This finding has been confirmed many times E.g. Rees, (1985) who states that one of the core reasons of using performance based compensations with ones employees is to motivate them to put in more effort. Employees should not be expected to increase effort if they see that improved performance does not necessarily translate to improved pay (Holmstrom and Milgrom, 1991). This is exactly where the largest problem of biases comes from; it destroys the perceived connection between performance and pay (Pendergast, 1999; Bol, 2000).

An employee receiving a positive bias from his manager and thus receiving higher pay, in theory, is a blow to the incentives put in place, as the employee knows that they get paid more for exerting less effort. Research has also shown that that perceived fairness of compensation based on performance has an impact on incentives to exert effort. Behavioral research has likewise long argued that employees not only care about their own compensation received, but also how it compares to their expectations as well as to the compensation received by their co-workers as well (Greenberg 1990; Fehr et al. 1993; Fehr et al. 1998; Fehr and Falk 1999; Colquitt et al. 2001; and Charness 2004); though, the payment equality theory has been shown to not always hold truth.

In our design, employees could be affected by both of these additional effects, comparing their feedback with the other employees and comparing it to their expectations. Even though employees do not get feedback on the evaluations of the other employees in their firm, there is a budget type of a payment in place and employees receive feedback on how much they were positively or negatively biased against monetarily. So if they receive feedback that they have earned a euro less

than they should have, they know that their euro they should have received has gone to someone else's pocket.

However, comparing their received income to their expected income in this design is more complex. Employees have a tendency to evaluate their own work higher than their managers would evaluate that same work (McFarlane et al. 1986; Harris and Schaubroeck 1988), nonetheless, in this experiment employees receive an evaluation from the external evaluators who have no means to be biased and therefore it is used as a proxy for real performance. The employees never receive this evaluation individually in the form of a score from 1 to 100 but as a monetary value; how much of the budget for that round they should have received based on theirs and their coworkers real performance (evaluations from the external evaluators). As it is explained to all participants that the external evaluators are experts on the questions and have no means to be biased in any way or form, we expect that the employees will accept their evaluations as being correct. Additionally, employees never see the answers given by their coworkers, thus they can never really feel that the externals did them an injustice as they do not know what they are being compared to. Still, we cannot rule out the possibility that a few individuals may believe that they performed much better than the results from both their own managers and external evaluators indicate.

Another topic of interest in this study is the motivation of the raters. When subjective evaluations began to be used massively, the general consensus was that people who evaluate are motivated enough to do it accurately. This seems to have changed recently, as researches now, for the first time, asked the question whether evaluators are really motivated to evaluate accurately (Levy & Williams, 2004). Raters' motivation to evaluate accurately is a topic of great interest as it is the most likely solution to finally solving biases that come from subjective evaluations, since we know that objective evaluations are not possible in all jobs. This experiment provides an extreme scenario in which employees are made aware whether they have been positively or negatively biased, as well as how much that bias will cost or gain them. Furthermore, at one point, it gives them an option to quit or sabotage the firm. One of the reasons for this type of design is to find out how employees behave in such scenarios, since in companies employees do eventually learn whether someone favors them or whether someone works against them. Equally so, to find out whether such an extreme design will eliminate the escalation bias which was persistent in Angelovski et al. (2014).

3. Design

The design of this experiment is meant to capture the essential parts of the internal company processes, which we want to study, as well as the effects they have on the employees. For this purpose, we both design and frame the experiment and its roles "naturalistically", naming roles, job positions, and tasks as it would be likely for one to find inside a company.

In this experiment we create a scenario in which a manager has to make a hiring decision and then evaluate three employees who work for him. This has been shown to create an escalation bias coming from the managers when they evaluate their employees. We then have employees continue working for the manager and at one point make a decision about their future in the company. We are interested in finding whether this bias persists in this scenario, as well as, the effects a biased evaluation may have on the employees being evaluated.

One of those employees is assigned to the manager from the outset, the other two are assigned to him after the manager has made hiring decisions pertaining to the two of them. The manager will have hired one of the two employees and will have preferred not to hire the other employee, who will nevertheless be assigned to him. This is a similar design to that of Angelovski at al. (2014) in which the authors find a positive bias towards employees they recommended for hire and a negative one for the ones which they did not. Our goal for using a part of their design is not to replicate those results, but to use a method which has been proven to lead to a bias naturally, with the purpose of looking at the effect that managers' biases, in evaluation, have on the future work and actions of their employees.

For this purpose we have adapted the design to have three stages where after each stage, the work of the employees will be evaluated both by their manager, as well as by external evaluators, which acts as a proxy for their real performance. After the second stage and evaluation, the employees are given an option to leave the company or sabotage the company. Each session of the experiment has sixteen participants and consists of multiple parts. The part in which the participants do a real-life task has three rounds.

At the beginning of a session each participant is assigned to one of three *types*. The first type consists of *managers*, who make the hiring recommendations; the second type consists of *given*

employees, employees who are told that they are already working for a manager; and the third type consists of *potential employees,* those who are looking for a job. In each session there are five managers, five employees which have already been assigned, and six potential employees. During the first part of the experiment, all potential employees are given a personality questionnaire to fill out. In the second part, each manager selects one of the potential employees and recommends that they are hired.[11] This employee is then effectively assigned to him. Subsequently, another of the potential employees whom the manager decided not to recommend is also hired and assigned to him randomly.

As a result of the process, each manager ends up with three employees: one who had been assigned to him from the start (*given employee, GE*), one who he recommended to hire and thus was hired (*recommended employee, RE*), and one whom he had had the opportunity to hire but didn't hire (*non-recommended employee, NRE*). Participants in the role of manager receive a fixed sum of 21 euros. The way employees are paid is explained below. Participants knew from the start about the four parts of the experiment. We now describe the parts in more detail.
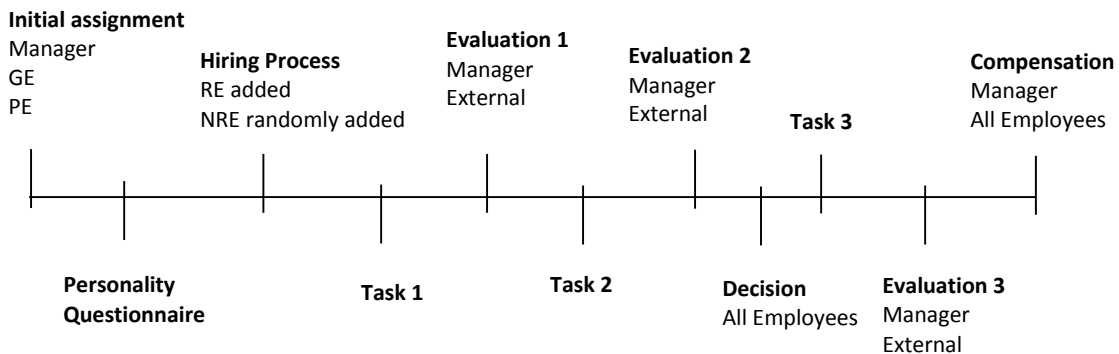
Figure 1. Timeline

M: Managers, GE: Given Employee, PE: Potential Employees, RE: Recommended Employee, NRE: non-recommended employee.

---

[11] Throughout we will refer to the manager *recommending* to hire a particular employee and *not recommending* to hire the others. In the experiment it later turns out that the recommended employee is effectively hired, but that one of the non-recommended employees will also be hired and assigned to the manager in question by top management.

3.1 The Personality Questionnaire

After the sixteen participants of an experimental session have found out what type they have been randomly allocated to, the six "potential employees" start filling out a personality questionnaire, while the other ten participants wait. At this point, none of the participants are given any information about the hiring decisions that come later.

The purpose of the personality questionnaire is to provide the manager with information about the potential employees, so that he has the impression that later he will be able to make an informed hiring decision. In other words, we wanted to give the manager a basis for making their decisions a purposeful one, yet not in a way that they would be given full information on exactly how hired employees would perform in the task ahead. Failer at al. (2013) find that individuals in environments which have censored information tend to rely too heavily on the censored information they have, causing them to form biased beliefs.

The personality test provided is a BFI-10 test, a 10-item short version of a widely used and recognized Big Five Inventory Test, with the Big Five being: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The official Spanish translation of the test can be found in Appendix A. We opted for this test, again, as it was used by Angelovski et al. (2014). The authors of the original study use it because of its short length, as this meant that the other participants who didn't have to make a decision at this stage didn't have to sit idly for long. Despite its brevity, the test has been found to possess adequate psychometric properties. (Rammstedt and John (2006)).

3.2. The Hiring Process

In the second part of the experiment each of the five managers of a session is given the personality test answers of the six potential employees that are looking for a job in his company. Here it is important to note that the personality test results were purposefully not aggregated (as is often done), so the managers could only see the 10 questions and each corresponding answer. The reason

we did this is because it further differentiates the six candidates, and mimics a "question and answer" structure akin to that of a real-world interview situation.

The manager is instructed that there are two open positions in his department and that he can recommend one of the six candidates, who will then be hired. Participants are told that the other hiring decision will be made by top management. Since the manager sees the questionnaire responses of all potential employees in his company, it is possible that more than one manager decides to hire the same employee. The reason we allow for this is so that every manager gets to choose from an equal amount of potential employees, and so we don't have to send anyone home in the middle of each session.

Once managers have made their hiring decisions, they are told which one of the remaining five potential employees top management has decided to hire and assign to them. The top management hiring decision is in fact a randomly selected potential employee who has not been hired by any of the five managers. In the end, all six potential employees are hired either as a first choice by the hiring manager, or as a second choice by top management. Note that the experiment is designed in such a way that the same potential employee cannot be both a recommended employee and a non-recommended employee in different groups. The given employee is distinct for each of the managers.

At the very end of this stage, the six potential employees are informed about whether they have been hired and whether they were the first choice picks, hired directly by the hiring manager (recommended employee, RE) or not (non-recommended employee, NRE). Now, every manager has three employees working for him, one given employee, one recommended employee, and one non-recommended employee.

3.3. The Task – Round 1

During the third part of the experiment, all employees, regardless of how they were assigned to their managers, are given the same task to perform individually. The task consists of reading a short, two-page business case and answering one question about it that does not have quantifiable

correct answers. This is done in order to mimic real life company problems where there isn't just one strict way to get things done. The responses to this case study warrants a subjective evaluation.

The case study refers to a company and its plan to perform an expansion strategy. At the end of the description of the case, there are three questions each participant in the potential employee role has to answer, one in each round.

3. Comment on the relationship between quantity and quality. Do you think that there is always an inverse relationship between both of them?
4. Which objectives is this expansion plan pursuing?
5. How can the company use the current process of regularization and the new megaproject that is addressed to improve its position?

The employees know that each of them is given the exact same case study which is going to be evaluated by the manager, as well as externally. *They are also told that their pay will be based solely on their manager's evaluation of their answers.* In addition, they are also told that the manager has no financial incentives to grade a certain way in this experiment.

3.4. The Evaluation of the Task

In this part of the experiment managers are shown the answers to an open-ended textual question that employees answer in the first task round

They see the answers of all three of their employees simultaneously and are asked to evaluate each employee's performance on a scale of 1-100. Importantly, managers do know to what type of employee each case study belongs to.

In the case that more than one manager decides to hire the same participant, or that the same person is hired by more than one top management, that person's task is evaluated by all the managers and in all groups in which he has been hired. Neither the managers nor the person in question is told that he is now a participant in multiple groups. To determine his final compensation in cases like this, as soon as the hiring process is finished, a group is chosen at random (from the ones he is in).

At the end of the round, 21 euros are proportionally split between each of the three employees working under a manager, with the proportions based solely on the score of each of the employees evaluated by each manager. For example, if the given employee received a grade of 80, and the recommended and non-recommended employees both received a grade of 40, the given employee will receive 14 euros in compensation while the other two will receive 7 euros each for a round. The manager is paid a fixed sum of 21 euros for the experiment, or 7 per round as the average for the employees.

3.5. The External Evaluation

At the same time that the managers evaluate their three employees, in an adjacent room the external employees receive the case study answers, which automatically are secretly coded so that the external evaluators cannot know which answer corresponds to which group or type of employee. The external evaluators remain the same for all sessions of the experiment. They are PhD students in business economics, who do research in and teach courses in business economics, management and related subjects at the Universitat Autonoma de Barcelona. In particular, they have experience in evaluating students' answers to case studies like the one we use in the experiment. Again, the external evaluators do not participate in the experiment and have no way of knowing which answer belongs to which person, type, or group. Just like the managers within the experiment, they grade the answers with a score from 1-100. Because these external evaluators are experts and have no means to be biased in any way, we regard their score to be a proxy variable for an employees' real performance.

The difference between the average grade given by the three external evaluators and by the managers will be the basis for our analysis of the way bias (positive or negative) affects employees in a company.

3.6. The Feedback – Round 1

After both the managers review their three employees' answers, as well as the three external evaluators grade all employees' answers, the first feedback begins. At this time, the employees see on their screen the payment they have received for the initial round by the manager as well as the

payment they should have received based on the average grade they and the other two employees under the same manager were given by the three externals. Recall that the pay for each round is a split of 36 euros between the three employees working under the same manager. The percentage of each employees' share of the total is equal to their grade received by the manager divided by the sum of the three grades given by the manager. Therefore each employees' pay is not only affected by any bias the manager has towards him but also by bias the manager may have towards the other two employees. After the employees have seen what their pay for the round is as well as what it should have been, we go to round two of the task.

3.7. Task, Evaluation by Manager & Externals, and Feedback (Round 2)

As soon as all employees have seen their feedback for round one, they immediately start with the task of round two. The second task is comprised of answering a different question regarding the same case study they were given before. The question is regarded to be of equal difficulty as the previous one.

Then just as before, once the employees have submitted their answers, they will be evaluated by the managers as well as the three externals and will be given feedback in terms of how much of the split they earned this round, as well as how much they should have learned.

3.8. Employees Choice

After the feedback of round two, all employees are given three options that could affect their future pay, as well as the pay of other employees under their same manager. They are the following:

1. Do Nothing: The first option is that they do nothing. They continue working under the same manager in the last of the three rounds, as they have done so far, and earn in the same manner as before by splitting the amount in the same way that they did in the first two rounds. This option has no cost.

2. Quit: The second option available is that they quit the job. If they decide this option, they immediately have a cost equal to 110% of the amount lost due to any negative bias incurred by them over the two rounds, and everyone else in the group (The evaluating manager and the other two employees) looses 120% of this amount. Ex. If in total they have incurred a cost of 1euro due to negative biases against them, by choosing this option they lose 1.1 euros more while everyone else loses 1.2 euros. However, if they do this, they know that in the third round they will be paid according to their true performance and not their current managers' evaluation. This is a representation of the real life situation of changing jobs, and thus has a cost which is inevitable in every job change.

3. Sabotage: The third option is to remain in the company and sabotage it. If the employees choose this option they stay in the same position in the next round and are paid the same way they have been paid so far but hurt the both themselves and the company. The cost of choosing this option is 55% of the total loses suffered so far to the employee who decides to sabotage and 120% of these loses to everyone else in the company, including the manager.

3.9. Task, Evaluation by Manager & Externals, and Feedback (Round 3)

Once the decision is made by all employees they start the third round of the task, which is comprised of another question of equal difficulty from the same case study. They are then evaluated by the externals and their manager and given feedback for the second task. The employees who chose options 1 and 3 are paid the same way as before, whereas the employees who chose option 2 are paid according to their true performance; the evaluation received by the externals rather than the Manager.

At the end, the employees are shown the total profit for the three stages, and are paid their total amount. With this the experiment comes to an end.

4. Hypothesis

Employees are expected to put in their best effort at their jobs, regardless of externalities such as job evaluations. Whether the evaluations end up going fairly, in their favor, or against them, the employees' performance should not be affected significantly. Thus the following null hypothesis:

H1₀: *A positive or negative bias by the manager, in the evaluation of task one, will not have an effect on the performance of his employees in task two.*

On the other hand, employees like being valued, even if overvalued, and are likely to try to do their best in order for that trend to continue. This holds even truer when evaluation is being linked with the pay; participants who earned more than they should have are likely to want to repeat such good earnings and are therefore likely to try even harder in the next round. Additionally, experiments on gift exchange indicate that since employees know that they are being unjustly favored, they are likely to reciprocate by increasing the effort (Fehr et al. 1993; Fehr et al. 1998; Fehr and Falk 1999; and Charness 2004). Even though agency theory predicts that both positive and negative biases are likely to lead to decreases in performance, behavioral theory counters this, particularly on the point of positive biases.

In the opposite case, when employees realize that they are being negatively biased against, they have a tendency to get demotivated and therefore decrease their efforts. This effect is likely escalated with the employees knowing that what income they lost from being negatively biased against someone else in the company has earned, even though their performance didn't justify it. This goes hand in hand with the finding of Garland (1973) on above average performers. Thus:

H1a: *A positive bias by the manager in the evaluation of task one, will have a positive effect on the performance of his employees in task two.*

H1b: *A negative bias by the manager in the evaluation of task one, will have a negative effect on the performance of his employees in task two.*

The sabotage option is an economically irrational decision, therefore economic theory would argue that regardless of the circumstances, it is not an option that should be ever chosen. This argument was purposefully helped by the design of the experiment which called the option to sabotage:

"sabotage", which Harbring & Irlenbusch (2009) have found to be the main factor which reduces destructive behaviour. The option to quit can be a rational choice if one is being heavily biased against and has a reason to believe that they will be heavily biased against again in the next round; however, it is heavily linked to risk-aversion. Thus, the null hypothesis:

H20: *The manager's evaluation in tasks one and two, does not have an effect on the decision managers make between tasks 2 and 3.*

On the other-hand, being biased in the same direction twice in a row is likely to make employees believe that the next evaluation will remain in the same direction. Thus, if employees receive two positive biases and see no reason why this would change in the third round, it would be irrational to do anything else other than to stay in the company and incur no extra costs. If employees have received two negative biases and thus believe that they might receive a negative bias again, they might opt for the costly option of sabotage or quitting. Quitting may be risky, but may pay-off if the employee does get negatively biased against again in the third round. Sabotage would be a fully irrational decision. In our experiment, employees make a one-time decision and therefore, the option to sabotage loses the signalling effect that it could have in real life. In a real company it may not be irrational because by sabotaging an employee is sending a message to the manager who is working against him, that he will continue to sabotage if the situation does not change. However, many behavioural experiments have shown that costly sabotaging is chosen by participants in order to punish others for not being fair, and Abbink & Sadrieh (2008) even go as far as to show that when participants are anonymous, they do choose to act destructively for no apparent reason. We construct the following alternative hypotheses:

H2a: *Consequent positive biases in tasks one and two leads to more managers deciding to stay with the firm given the choice to stay, leave, or sabotage.*

H2b: *Consequent negative biases in tasks one and two leads to more managers deciding to either leave or sabotage the firm.*

The changes to the design of the experiment used by Angelovski et al. (2014) have been numerous. Participants now play more rounds with more evaluations, which opens the possibility that the bias does not persist through all rounds. Additionally, employees now have full information on the bias

towards them and the manager, knows this as well. Multiple experimental evidence suggests that when negative or unfair behavior is found out, it imminently decreases, even when the other does not have the option to retaliate; people don't generally continue acting unfairly once they have been found out. Lastly, employees now do have the option to retaliate by sabotaging and quitting which is likewise costly for the manager. In Angelovski et al. (2014) the managers pay was linked to the performance of his employees, however, due to the one shot nature of the experiment whether the manager was actually biased or not was not going to affect his/her monetarily. Therefore, the combination of these changes could be more than enough to fully eliminate escalation bias. Thus the following null hypothesis:

H30: M*anagers evaluate their employees in an unbiased way, independently of how employees were hired into the company.*

Angelovski et al. (2014) has shown that escalation bias persists through both the experience of the task, and through linking the manager's pay to the actual performance of his employees. Regardless of the extremes in our design, the manager is still getting paid an almost inflexible fee, thus making it possible that the bias persists. Therefore:

H3: *Managers retain a positive bias towards the employees they decided to hire and/or a negative bias towards the employees they decided not to hire but were hired by a third party.*

5. Procedures

We conducted six sessions of this experiment. The average running time of the experiment was 90 minutes, of which 30 minutes was spent on reading the instructions and the hiring process, while the remaining 60 were spent on the 3 tasks, their evaluation, the decision and the results. The experiment was conducted at two of the computer rooms of the Universitat Autonoma de Barcelona and participants were undergraduate students from the university. The z-tree software was used to run the experiment (Fischbacher, 1999).

In each of the six sessions there are sixteen participants. We therefore had 96 participants, of which 30 are managers, 30 are given employees, and 36 are potential employees.

6. Results

Table 1 looks at the descriptive statistics of the mean evaluations performed by the managers as well as the three external evaluators. In it we can see that managers evaluate slightly higher than the actual performance of the employees, which is particularly true for tasks 2 and 3. This is consistent with the theories of leniency and centrality bias found by many researchers (Ex. Landy and Farr 1980, Murphy and Cleveland 1991, Bol 2011). Leniency bias is the tendency for evaluators to be more lenient towards the people they evaluate, especially when their wage depends on them, while centrality bias is the tendency to compress the ratings. Additionally, in table 1 we see that the average manager evaluations start at the same level as actual mean performance, yet begin to increase and deviate starting from the second task.

Recall that we use external evaluation grades as a proxy for real performance, since the external evaluators 1. Are experts at the particular case study, having read and evaluated answers coming from it in previous studies, and 2. Do not know which answer belongs to which employee and cannot see the participants, thus have no means to in anyway favor any of the employees.

Table 1: Descriptive statistics of the mean evaluations by the managers and external evaluators

	Mean Manager Evaluation	*Std. Deviation (Manager)*	*Mean External Evaluation*	*Std. Deviation (External)*
Task 1	61.32	20.45	61.06	7.58
Task 2	69.68	24.13	62.62	6.83
Task 3	65.98	20.82	61.21	8.51

Table 2 looks at the performances of the employees in task one and two, separated by the direction of the bias (positive or negative) they received in the evaluation of the first task. In round one, from the 66 total observations, 30 received a positive bias and 33 received a negative bias, meaning that only 3 employees got the same evaluation by the manager and the externals. It is important to note that both the positive and negative biases vary from very small to very large, and that the employees know exactly how much the bias towards them has earned or cost them monetarily.

 As can be seen from table 2, when employees are evaluated higher than they should have been in the evaluation of task one, thus getting paid more than they should have been, they increase their

average performance from 60.50 to 64.83. In the case where they are evaluated lower than they should have been, they slightly decrease their average performance from 61.77 to 60.15.

Table 2: Performance of employees in task 1&2 by type of bias

	Positive Bias in round 1			Negative Bias in Round 1		
	Observations	Mean	Std. Deviation	Observations	Mean	Std. Deviation
Performance in Round 1	30	60.50	7.87	33	61.77	7.26
Performance in Round 2	30	64.83	5.42	33	60.15	7.23

The increase of the performance in round 2, stemming from a positive bias in round 1, is significant using 1-tailed Wilcoxon Signed Rank test (p=0.003). Therefore we can confirm H1a: A positive bias by the manager in the evaluation of task one will have a positive effect on the performance of his employees in task two. Even though performance decreases when employees are negatively biased against in round one, we cannot confirm H1b, as the results are not significant enough using the same test as above.

Table 3 looks at the effect of the number of positive and negative biases on the decision employees need to make between rounds two and three. When employees receive a positive bias from the manager in both of the rounds, and are faced with the 3 options, 17 decided to do nothing and 1 decided to sabotage. When employees received one positive and one negative bias in the two rounds, 27 of them did nothing, 5 decided to quit and 2 decided to sabotage. Finally when they were negatively biased against in both of the rounds only 4 decided to do nothing, 9 quit, and 1 decided to sabotage.

Table 3: Effect of number of biases on the "Employee Decision"

	2 Positive Biases	*1 Positive and 1 Negative Bias*	*2 Negative Biases*
Do Nothing	17	27	4
Quit	0	5	9
Sabotage	1	2	1

The picture becomes clearer when we convert these numbers into percentages. When receiving two consequent positive biases, only 5.6% of the employees are willing to incur a cost to themselves and the firm by either sabotaging or quitting. On the other hand, when they received one of each bias 20.6% of employees decide to sabotage or quit, and when they receive two negative biases 71.4% of the employees decided to either sabotage or quit. Thus it should be no surprise that we found that the number of positive and negative biases to be significant in determining the decision of the employees using a logistics regression. (p=0.015)

Table 4: Descriptive Statistics of "RE-GE" and "NRE-GE" variables by Manger and Externals

	Recommended – Given				*Not Recommended – Given*			
	Manager		External		Manager		External	
	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Round 1	-8.47	37.53	1.72	12.13	0.90	30.03	-1.00	8.74
Round 2	-6.77	34.37	-1.04	11.24	-2.53	29.46	-2.33	8.73
Round 3	-4.16	28.79	1.00	12.04	-3.16	28.16	1.44	11.23

Table 4 looks at the descriptive statistics of the differences in the grading of the manager and the externals. In our analysis of this segment we have followed the features of the design of Angelovski et al. (2014) in order to get more comparable results. Therefore, in this paper as well, we standardize the evaluation grades. We use the evaluations of the given employees as the baseline. The given employees are not part of the previous hiring process, and managers have been told that they were assigned to them from the start. We think that their evaluations are a natural standard of comparison to which to compare the (possibly biased) evaluations of the recommended and non-recommended employees.

For manager evaluations, the standardization gives rise to two new variables, namely "RE–GE" and "NRE–GE". The first variable corresponds to the differences in the evaluation of the recommended employee and that of the given employee. The second variable similarly

corresponds to the differences in the evaluation of the non-recommended employee to that of the given employee.

As can be seen from Table 4, the RE-GE variable is not larger for the manager than for the external, which would indicate a possible positive escalation toward the Recommended Employees. We checked for statistical backing for the difference in evaluation by the manager and external of the RE-GE variable using a two-tailed Wilcoxon signed-rank test for all three rounds. The test finds no significant results, meaning that there is no positive or negative bias towards the Recommended Employees. Moving on to the results of the NRE-GE variable in Table 4 we find an even less of a difference in the grading between the Managers and the Externals. Again, we check the statistical backing for the difference in evaluation by the Manager and External of the NRE-GE variable using a two-tailed Wilcoxon signed-rank test for all three rounds, and again found no significant results. Therefore, we can also conclude that there is no positive or negative bias towards the Non-Recommended Employee, thus confirming the null hypothesis H30 which states that Managers evaluate their employees in an unbiased way, independently of how employees were hired into the company.

7. Conclusion

We set out to produce a detailed experimental study while looking to add to a previous study on the topic of escalation bias using subjective evaluations. The aim was to find a way of both eliminating bias as well as taking an opportunity to look the issue from an opposite perspective; from the point of view of those being evaluated.

We tried to eliminate the bias without having to go via the most obvious route of linking the manager's pay directly to the accuracy of his evaluation. Our results have shown that a multiple-round treatment, where there is no loss of information, and where managers' pay could potentially be affected by their evaluation fairness, is enough to eliminate the bias. We found this design to work while managing to keep the potential monetary loss for the manager very small. The differences in bias between the rounds do indicate that the overall bias may be decreasing with each round, but it also shows signs that a negative bias towards the NRE's may have started to

appear had there been more rounds or participants. This seems to indicate that perhaps the elimination of the escalation bias may come more from the combination of effects, rather than a single change in the design of Angelovski et al. (2014). One of the main limitations of this finding is that we cannot yet conclude which addition to the original design has more of an effect on the end result. However, the we believe that we have made the initial step by showing that the bias can be eliminated, without having to link the correctness of the evaluations by the manager to their actual pay.

The findings also indicate that an employee's future performance is not immune to bias differences. Having a manager be positively biased towards an employee has a positive effect on that employee's future performance. What is interesting here is that this increase is not perceived fairness due to the employee's high self-assessment, as other studies have indicated when studying increased performance in connection to leniency bias. Employees are aware of their actual performance due to the information from the externals, they are also aware that the extra pay they have received is not justified, and have gone on in the next round to justify it. Remember that the manager's pay is not affected by the increased or decreased performance of his employees after this bias, so it cannot be said that they are repaying him in terms of a classical gift exchange scenario.

Being negatively biased against has no significant impact on performance in the next round, but we have found that subsequent negative biases do lead to more costly decisions for both the managers and their employees.

8. References

Abbink, K, and A Sadrieh (2008): The Pleasure of Being Nasty. Working paper, University of Amsterdam and University of Magdeburg.

Angelovski, Andrej and Brandts, Jordi and Solà, Carles, Hiring and Escalation Bias in Subjective Performance Evaluations: A Laboratory Experiment, Working Paper (December 11, 2013).

Bernardin, H. J., and P. Villanova. 1986. Performance appraisal. in Generalizing From Laboratory to Field Settings edited by E. A. Locke. Lexington Books D.C. Heath and Company: Lexington.

Bol, J. (2011) "The Determinants and Performance Effects of Managers' Performance Evaluation Biases" *The Accounting Review*: September 2011, Vol. 86, No. 5, pp. 1549-1575.

Charness, G. (2004), "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics*, **22**, 665-688.

Colquitt, J. A., D. E. Conlon, M. J. Wesson, C. O. L. H. Porter, and K. Yee Ng. 2001. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology* 86(3):425-445.

Drake, A. R., Wong, J. and S.B. Salter (2007) "Empowerment, Motivation and Performance: Examining the Impact of Feedback and Incentives on Non-management Employees", *Behavioral Research in Accounting*, Vol. 19, pp. 71–89.

Fehr, E. and A. Falk (1999), "Wage Rigidity in a Competitive Incomplete Market," *Journal of Political Economy*, **107(1)**: 106-134.

Fehr, E., E. Kirchler, A. Weichbold and S. Gächter (1998), "When Social Norms Overpower Competition: Gift Exchange in Experimental Labor Markets," *Journal of Labor Economics*, **16(2)**: 324-351.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993), "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, **108(2)**: 437-459.

Feiler DC, Tong JD, Larrick RP (2013) "Biased judgment in censored environments" *Management Science*. Vol. 59, No. 3

Friedrich, J. 1993. Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review* 100 (2):298-319.

Garland, H. 1973. The effects of piece-rate underpayment and overpayment on job performance: A test of equity theory with a new induction procedure. *Journal of Applied Social Psychology* 3:325-334.

Greenberg, J. 1990. Organizational Justice: Yesterday, today, and tomorrow. *Journal of Management* 16(2): 399-432.

Harbring, C. and B. Irlenbusch (2010). "Sabotage in Tournaments: Evidence from the Lab." *Management Science*.

Harris, M. M., and J. Schaubroeck. 1988. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology* 41(1):43-62.

Holmstrom, and P. Milgrom. 1991. Multi-task principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7:24- 52.

Holmstrom, B. 1979. Moral hazard and observability. *Bell Journal of Economics* 10(1):74-91.

Klimoski, R.,&Inks, L. (1990) "Accountability forces in performance appraisal" *Organizational Behavior&Human Decision Processes*, 45(2): 194.

Leung, K., S. Su, and M. W. Morris. "When is Criticism Not Constructive? The Roles of Fairness Perceptions and Dispositional Attributions in Employee Acceptance of Critical Supervisory Feedback." *Human Relations* 54 (2001): 1155-187.

Levy, P. E., and J. R. Williams. 2004. The social context of performance appraisal: A review and framework for the future. *Journal of Management* 30(6): 881-905.

McFarlane Shore, L., and G. C. Thornton. 1986. Effects of gender on self-and supervisory ratings. *Academy of Management Journal* 29(1):115-129.

Milkovich, G. T. and A. K. Wigdor (1991) "Pay for Performance" *National Academy Press.*

Pearce, Jone, and Lyman Porter. 1986. Employee Responses to Formal Performance Appraisal Feedback. *Journal of Applied Psychology* 71 (2):211-218.

Prendergast, C., 1999. The provision of incentives in firms. *Journal of Economic Literature*37(1):7-63.

Prendergast, C., and Robert H. Topel (1996) "Favoritism in Organizations." *Journal of Political Economy* 104.5: 958.

Rammstedt, B., and O. John (2007) "Measuring Personality in One Minute or Less: A 10-item Short Version of the Big Five Inventory in English and German" *Journal of Research in Personality* 41.1: 203-12.

Spreitzer, G. M. (1995). Psychological empowerment in the workplace: Dimensions, measurement, and validation. *Academy of Management Journal*, *38*(5), 1442-1465.

Villanova, P, H. J. Bernardin, S. A. Dahmus, and R. L. Sims. 1993. Rater leniency and performance appraisal discomfort. *Educational and Psychological Measurement* 53:789- 799.

Woods, A. 2012. Subjective adjustments to objective performance measures: The influence of prior performance. Accounting, Organizations and Society 37 (6): 403-425.

Appendix A. Personality Questionnaire

English version.
Instruction: How well do the following statements describe your personality?

I see myself as someone who …	Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
… is reserved	(1)	(2)	(3)	(4)	(5)
… is generally trusting	(1)	(2)	(3)	(4)	(5)
… tends to be lazy	(1)	(2)	(3)	(4)	(5)
… is relaxed, handles stress well	(1)	(2)	(3)	(4)	(5)
… has few artistic interests	(1)	(2)	(3)	(4)	(5)
… is outgoing, sociable	(1)	(2)	(3)	(4)	(5)
… tends to find fault with others	(1)	(2)	(3)	(4)	(5)
… does a thorough job	(1)	(2)	(3)	(4)	(5)
… gets nervous easily	(1)	(2)	(3)	(4)	(5)
… has an active imagination	(1)	(2)	(3)	(4)	(5)

Scoring the BFI-10 scales:
Extraversion: 1R, 6; Agreeableness: 2, 7R; Conscientiousness: 3R, 8; Neuroticism: 4R, 9;
Openness: 5R; 10 (R = item is reversed-scored).

Appendix B. Case Study

The Spanish company Lladró was born in 1953 when Juan, José and Vicente Lladró , three brothers sons of farmers with great artistic talent , founded a small family business in the Valencian town of Almácera . Lladró since then has undergone a huge metamorphosis from a craft workshop to a large international company in continuous expansion.

Company management, as in its origins, is under the control of the members of the Lladró family , who owns the company. Initially, the team just tried to adapt a vintage style, but in a short time certain traits appeared that would be recognized later on as the Lladró style. Since the beginning, the public was infatuated with their creations. In little time the workshop was expanded several times and an increasing number of collaborators multiplied the work of the brothers. As of today sculptures which are born in the City of Porcelain do so in a completely handmade process in which 2500 employees participate, exporting to more than 120 countries of five continents: the Netherlands, USA, United Kingdom and Japan being the most important.

Since 1955, year in which the first shop was opened in Valencia, Lladró has been increasing its network of stores in all major shopping malls in the world: Valencia , Madrid , London, New York , Beverly Hills , Singapore , Hong Kong, Las Vegas , Sydney . 6,900 authorized dealers exhibit art of Lladró porcelain.

However, the company's large expansion has a point of inflection at the end of 2001, when the Lladró announced the closing of 2,000 points of sale. In this way Lladró got rid of those dealers who were not taking care of the luxury image of its figures, and were exhibiting them together them with figures of the competition and even with imitations. The objective of this measure, with an effect of decreasing its billing by 17%, is to prove the company's commitment to quality, instead of quantity.

To recover from this measure, Lladró has put together an expansion plan consisting in opening 50 own points of sale to sell its new and innovative designs and to create a new image for the firm. This strategy will allow the firm to have a direct contact with its customers and it will complement the already established relationship with collectors through the Prestige Club Lladró, allowing

them to access exclusive collections and having privileged shopping options of limited edition products.

Among the objectives of this plan, one should point out the reinforcement of the quality image of the organization, the increase the value of sales by 9% of billings and the maintenance of the return on sales.

1. Comment on the relationship between quantity and quality. Do you think that there is always an inverse relationship between both of them?
2. Which objectives is this expansion plan pursuing?
3. How can the company use the current process of regularization and the new megaproject that is addressed to improve its position?

CHAPTER IV

Social structure and the provision of intergroup public goods

1. Introduction

People, every day, are part of and interact with different groups and on different levels. More often than not, they are also part of a smaller group (local) within a larger group (global); departments within a university, work groups within a department, branches within a company, co-ethnicities in a village, etc. This means that many if not most, face the dilemma of how to allocate both time and effort between these groups in which they belong. As allocating time and effort into one group, decreases the possibility of allocating it in the other, it is of essential value that we understand how these decisions are made as well as how to promote the most efficient decision on allocating time and effort.

Researchers tend to agree that allocation into the global group is generally more efficient than allocation into the local one. The larger or "global" group, in which all the smaller or "local" groups belong to, generally have bigger impacts on the end outcome on the goals set. The obvious reason for this is that global groups affect more people, which certainly makes sense knowing that the local groups all belong to it. Another reason why global groups have the potential to be comparatively more efficient is that high end decisions tend to be made on a global level rather than a local one. It is not an easy option as high levels of coordination are needed, but once this is established the benefits could potentially be very big.

On the other hand contributing locally seems to be a much safer, risk-averse option of contribution. Local groups are smaller and thus require much less social effort and coordination between the participants then global ones. Some form of tribalism, also seems to have an effect on the decision of where people contribute more. Hirshleifer's (1983) makes the point that people respond more to rescue efforts that are of local importance rather than global ones. Additionally, due to the size and structure, in local groups, members are much likely to get acknowledgement for their

contribution from their teammates than in global ones. All these reasons lead to why we see everyday examples of people that prefer to contribute to local groups rather than global ones. Importantly for us, communication is a tool that has been found to be very effective in solving common-pool dilemmas. Ostrom and Walker (1991) have found that when costless communication was added, players used it to: 1) calculate coordinated strategies, 2) make verbal non-binding agreements to implement the strategies, and 3) deal with non-performing players. Additionally there has been extensive research showing that that even non-binding communication between decision makers significantly reduces free riding behavior (Ex. Isaac and Walker, 1988; Cason et al., 2012). Even further, Sally (1995) and Pavit (2011) find that the opportunity to communicate has been found to be the most important factor in group cooperation decisions.

Even though the is numerous work that has showed that communication is a very good tool for solving common pool dilemmas, previous research does not take into account that communication networks and public good externalities do not always match. Additionally, communication has not been used in combination with two public good options, one of which is more efficient than the other. In our experiment we chose to go with chat-based verbal communication, as our social enforcement tool, as it closely resembles face-to-face communication without having the problem of the removal of anonymity.

Our aim is with this paper is to add to the body of literature by combining communication, as a social enforcement tool, with the possibility to contribute to public goods of different levels of efficiency. We want to find out the effect of having an option to contribute to a local public will have on contributions, as well as to find out if, and in what form, communication can help groups and individuals to contribute more efficiently. We attempt to do this by testing is the interaction of different levels of communication on the various levels of PG's available.

2. Literature Review

Various experiments have shown us that higher marginal returns lead to higher contributions (see, for instance, Isaac et al. 1984; Isaac and Walker, 1988; Fisher et al., 1995). Thus, it has been argued

that if a global PG has significantly higher marginal return than the local PG, it may counter the effect people preferring to contribute locally. The answer to this dilemma has been sought out twice so far in Blackwell and Mckee (2001) and Fellner and Lünser (2008). Blackwell and Mckee (2001) found that as the MPCR of the global good increases, the contributions into it also increase, but this does not lead to a decrease of the contribution into the local public good. Fellner and Lünser (2008) also found that when the global PG had a higher MPCR players started of contributing into the more efficient global PG. However, in their research, players, with repetition, quickly settled back to their comfort zone of contributing more locally.

Another key issue that has risen from the literature of public goods is enforcement. Rational choice theory states that each individual balances costs and benefits in order to maximize personal gain (Friedman 1953). Economic rationality, interpreted in the form of "wanting more rather than less of a good" (Becker 1976) is one of the most common suppositions of subjects behaviors in economic theory. So, when it comes to collective goods, rational choice theory without any type of enforcement predicts under-provision.

When it comes to the free rider problem in Public Goods games, some sort of enforcement is needed in order to counter it. The enforcement we use in our experiment is communication, as it has been found to solve issues that arise when collective actions are needed (Isaac and Walker, 1988). Ledyard (1995), among many others, confirmed this.

Kerr and Kaufman-Gilliland (1994) state that the reason why communication is so effective in solving common-pool dilemmas is manifold, including; "(1) it can help in understanding of the game, (2) it encourages coordination of cooperative action, (3) it alters expectations of others' likely behavior, (4) it enhances general norms of benevolence, (5) it creates social norms of cooperation, (6) it humanizes fellow group members, (7) it promotes group solidarity."

Out of all these it is the norm-psychology, defined by Chudek and Henrich (2011) as: "a suite of psychological adaptations for inferring, encoding in memory, adhering to, enforcing and redressing violations of the shared behavioral standards of one's community", which makes communication as excellent enforcement device.

Reviews of relevant research, hypothesized that communication is essential when it comes to the emergence of cooperative social norms, enhancement of mutual trust, establishment of joint action plans, and establishment of group identity (Bornstein, 1992; Pavit, 2011). Lapinski and Rimal (2005), and Pavit (2011) also state that communication is instrumental to social influence and the development of norms. The first studies of group norm formation are Sherif's auto-kinetic studies (Sherif, 1935, 1936). Members of groups use the behavior of other group members as information or signal in order to develop a group norm. It is the need of members to stick to the social norms of the group in which they are placed that makes communication an excellent enforcement device.

Even though when one thinks of enforcement in a PG game, the first thing that comes to mind is direct material punishment or binding decisions between groups, non-binding communication has been found to be at least an equally good enforcement device itself (Bochet et al. 2006; Andrighetto at al., 2013).

Bochet et al. (2006) have found that adding communication to a PG game drastically increases contributions. Further, they find that there is no statistically significant difference to contributions once a punishment option is added to the game and that the treatments which only had a form of open ended communication had considerably higher earnings then the treatments that had punishment but no communication. They find that communication itself is a better enforcement tool than punishment. Oprea et al. (2013) find that cooperation is not helped by continuous time itself and that communication is required for effective enforcement mechanisms. They find that when subjects are allowed to communicate, median cooperation rises to 100% and remains there. They further state that communication is self-enforcing and that it doesn't require any outside intervention or enforcement, as well as, that communication as an enforcement and coordination device doesn't lead to any inefficiencies that usually arise with punishment. Andrighetto at al. (2013) find that the combination of norm communication and material punishment lead to a higher and more stable cooperation than when used separately. They state that norm communication is the factor that actually boosts cooperation and that material punishment helps maintain it. This further suggests that norms prescribe how people behave, and not material punishment. In fact they also sum up ethnographic literature on the topic and state that evidence has shown us that the most common punishment in human ecologies have mostly been part of social norm communication (gossip, mockery, criticism, and blame) and until recent history have not necessarily been combined with material punishment.

Past experiments have tried to test different means of communication as enforcement in order to find the most effective one. Isaac and Walker (1988) focused on the effect of face to face communication in public goods games, and reported that it increased both efficiency and contributions. For many years it was argued that face-to-face communication is by far the best way to enforce cooperation, which if proven true over time made it the most effective type of communication in public good games. The argument certainly made sense as face-to-face incorporated other types of unique cues unavailable to other types of communication, such as: body language, tone, and expressions.

Later on researchers found that other means of communication are also effective if not as effective as face-to-face communication when it comes to enforcement in public good games. Riechmann and Weimann (2008) make the argument that one of the main reasons why communication is so effective is that it gives an insight into the others behavior, thus eliminating one's uncertainty. Looking at it from that point of view it certainly makes sense that face-to-face would be the most effective type of communication as it potentially gives the most insight. And research has proven this to be true. (Frohlich and Oppenheimer, 1998)

On the other hand, numerous studies have found that other means of communication are also effective in improving in various experiments. Bochet et al. (2006) found that verbal chat room communication was almost identically efficient as face to face communication. Charness (2000) and Charness and Grosskopf (2004) have found that costless pre-play communication (cheap talk) is effective in coordination games, and Van Huyck et al. (1993) have reported that even pre-play tacit communication enhances efficiency.

3. Experimental Design

Here we describe the public goods game used in the experiment in more detail. The public goods game is played by nine participants for fifteen periods. All participants are part of the same global group but are split into three local groups, each local group consisting of 3 participants. Each participant receives an endowment of 20 points at the start of each period, and each individually

chooses how much of it, if any, they contribute into the public good. As in all public goods games, the total payoff is maximized when everyone contributes all of their points into the pool. However, individually, there is likely to be a temptation to contribute little or zero and free ride on the contributions of others.

All participants share the same information throughout the experiment; they each know the contributions of the other players in their own local group, as well as the average contributions of the other local groups. For the first three treatments, participants can only contribute to the global public good, at an MPCR rate of 0.3. At this rate at least four participants are needed to contribute a similar amount, so that the contribution ends up profitable.

$$\pi_i = 20 - c_i + 0.3 \sum_{j=1}^{9} c_j^F$$

The above payoff formula is valid for treatments 1, 2, and 3. The difference between the first three treatments is the level of communication participants have available to them. Such that treatment one has no communication, treatment two has only within-group (local) communication, and treatment three has both within-group communication and between-group (global) communication. Participants in treatment two can chat but only to the other 2 members of their local group, but unlike in treatment three they cannot chat with anyone else in the other two groups in the global group. In the third treatment, one "representative" of each local group is randomly chosen at the beginning of the experiment to chat with the representatives of the other two groups. Chatting is simultaneous and occurs for five minutes between periods 1, 6, and 11.

In treatments 1-3 we were interested in finding out what effect the different levels of social structures and communication between them have on the cooperation and contribution into the public goods game. Most experimental studies focus on games where any player knows the contribution of any other player and can enforce cooperation on them, through punishment or communication. In real life, however, most social networks do not have this commonly used structure.

In treatments four to six the notion of local public good is added. This means that in these periods participants can contribute to the global PG and/or the local PG. The payoff for the global PG stays the same and the MPCR rate for the local PG is 0.6, but it only benefits members of the local group. This makes the local public good potentially less efficient as if all 9 participants were to contribute 1 point to the Local PG they would each get a return of 1.8 whereas if they all contributed the same amount to the Global PG they would be getting a return of 2.7 points. However contributing into the local public good requires less coordination and trust. This goes in line with real life scenarios, which have shown that global contribution is harder, as more coordination is needed, however it could also be potentially much more beneficial than local contribution. The new payout for treatments four to six is:

$$\pi_i = 20 - c_i^F - c_i^B + 0.3 \sum_{j=1}^{9} c_j^F + 0.6 \sum_{k=1}^{3} c_k^B$$

Treatments 4 - 6 are equal to treatments 1 - 3 in everything except the option to contribute into both public goods, so they have the same difference in communication structure. Treatment four has no communication, treatment five has only within-group communication, and treatment six has both within-group communication and between-group communication. This design provides for a multi layered analysis, as with it we can not only see how an additional communication possibility changes affects the contribution of a public good, but also we can differentiate between the two public goods scenarios available, when the level of communication is the same.

4. Theoretical predictions

The first dimension of our research is that we have three treatments, with different levels of communication, in which participants have the option to contribute only to the global PG and three treatments in which they can contribute to both the global PG as well as the less efficient local PG. Even though contributing globally can have a much more efficient outcome, research by Blackwell and Mckee (2001), and Fellner and Lünser (2008) has shown that when given the option participants seem to choose the easier and safer option of local contribution. Thus, it may be expected that the addition of a local PG good would make the option to contribute to the more efficient global PG a harder one to make. In the treatments with no communication we expect to see the contributions to the global PG to decrease, however we expect an increase in total contributions once the option to increase locally is available in treatment four. The reason for this is twofold; 1) people generally prefer to have a choice, and 2) it provides a less risk-averse option for the participants who might think that contributing into the global PG is a too risky strategy.

The second dimension of our research is that we have 3 levels of communication available in separate treatments. Though the literature on the effects of communication is extensive, the effects of communication on global and local PG games have not been well established until now. We believe that communication could be a crucial element in solving the efficiency problem in having the choice to contribute both locally and globally because communication can have more than one effect on this game, most importantly for our research; as an enforcement device, as well as a coordination device in the traditional sense of helping solve a game of multiple equilibria (Ostrom and Walker, 1991; Kerr and Kaufman-Gilliland, 1994).

The key reason why communication is considered a great enforcement device is because communication has been found to be instrumental in the grounding of social norms (Bornstein 1992; Lapinski and Rimal, 2005; Andrighetto at al., 2013). Communication provides the mechanism for participants to feel like they are part of a group, even though the participants are part of a group even when no communication is available. It is the human need to be part of a group, and abide by the group's norms that make it a great enforcement device and make it key in increasing contributions in most PG games as well as solving the efficiency problems which arise with the inclusion of a less efficient PG.

If communication is thought as simply a form of punishment or enforcement device (i.e., promises) then we expect it to work pretty well in the first three treatments where only global PG are available, since people can enforce provision within their local group and adding representatives would have an additional effect. On the other hand, if we consider that communication also has a coordination element then local enforcement might not be enough and representatives would help solve the problem as they can now coordinate better with other representatives. The inclusion of representatives in treatment 3 also, to some extent, might shift the local group mentality into a global one which contribute in slightly more contributions.

With the addition of the local public good in treatment 4 we expect coordination to become harder, as well as have the mentality shift even more to a local one. Therefore one could say that local enforcement is expected to work relatively less and representatives would have a larger effect when introduced in treatment 6.

Participants may coordinate on the local public good to the detriment of the more efficient global public good. This would mean that treatment five might not have great contributions into the global PG, but that the contributions into the global PG would rest on the shoulders of the representatives (treatment 6), and it would be up to them to increase contributions. Thus we expect treatment six to have a significant increase into the global PG, but with larger variance.

5. Experimental procedures

The experiment was computerized and typical procedures of anonymity were used. Once the instructions had been read, the computerized experiment (programmed in z-Tree, Fischbacher, 2007) started. After the game ended, subjects were confidentially paid their earnings in cash. Additionally it used neutrally worded instructions, and monetary incentives. In total, 432 students participated in the experiment which lasted for an average of 45 minutes. Each session of the experiment consisted of 36 subjects who were randomly assigned to one of four global groups. Two sessions were run for each treatment to bring the total participants for each treatment to 72. This gave us 8 global groups to analyze for each treatment. Points were converted to Euros at a rate of 30 points = 1€ and as the design itself guarantees some winnings there were no show-up fees. Average earnings equaled 17€.

Upon arrival, each subject drew a card to be randomly assigned to a seat in the laboratory. Once everyone was seated, subjects read the instructions for the experiment (a copy of the instructions of one of the treatments is available in the appendix). The participants were told that they will be randomly assigned to a group of 3 persons. To distinguish themselves, each person was assigned a random number, so each group will have a member #1, a member #2, and a member #3. Each group of 3 is part of a larger global group. The global group consists of the group of three each member is in plus another two groups of three, and these groups are going to be the same two groups throughout the experiment. In treatments three and six, members #1 of each of the 3 groups can communicate with each other at the same time they are communicating with the other 2 members of their local group.

6. Results

This section consists of three section. In the first section (Section 3.1), we analyze the results of the public goods game where participants can only contribute to the global public good (Treatments 1, 2, and 3). In the first part of section 3.2 we do the same for treatments four and five, in which participants can choose to contribute to the global PG or local PG and have either no communication or within-group (local) communication only. Then towards the end of section 3.2 we look at the results of the public good game in treatment 6 where participants can contribute to either global PG or local PG, and have both means of communication (global and local) available. Lastly in section 3.3 we look at the results we got from coding the chat communication the participants had in the four treatments in which chat was available.

We aim to find whether and how different levels of communication affect the contributions to the global public good in a multi-local group scenario. As social enforcement (communication) occurs mostly within local groups, we compare the cooperation when there is no communication to the cooperation when within-group communication is available. We then look at the additional effect on cooperation when between-group (global) communication is added.

6.1. Global Contribution with Communication (No Local PG)

For each of the first three treatments, Table 1 shows the descriptive statistics of the contribution in the public goods game. In it we can see the mean contribution of all 15 periods played, the mean standard deviation of contributions within a group, the mean standard deviation of contributions within groups, and the mean standard deviation between local groups.

Table 1. Descriptive Statistics – Treatments with Global PG

Treatment	Mean Contribution	Trend	Std. Dev. Within Local Groups	Std. Dev. Between Local Groups	Std. Dev. Between Global Groups
1	8.00	-2.52	2.31	1.88	2.89
2	11.43	-2.52	1.10	2.62	4.38
3	16.09	-1.19	0.90	1.17	3.94

As can be seen from second column of Table 1, the contribution to the global public good (cooperation) significantly increases between the 3 treatments as more levels of communication are added. In treatment 1 when there is no communication between players, there is no social enforcement and we see that players on average contribute less than eight points. This result goes to show how standard contributions into a public good are when there is no enforcement. The results of treatment one are along the lines of the findings of (Davis and Holt, 1993; Ledyard, 1995; Bochet et al., 2006; Fellner and Lünser, 2008); who all find that when there is no enforcement contributions tend to start of slightly above the 50% mark and decrease over time. In the second treatment as within-group communication is added we see the mean contribution increase to 11.433. This increase has a statistical significance of 0.045 a random-effects tobit model. This is a clear example of social enforcement in action; now that participants can talk to each other they are encouraged to contribute more to the public good. Additionally participants of a local group know that if no one outside their group contributes they are better off not contributing anything, however if all 3 local group members contribute only 1 more global group member (from the 6 remaining) needs to contribute a similar amount in order for the contributions to become

profitable. Then, as between-group communication is added we see the mean contribution increase even further to 16.088. The increase of contribution between the second and the third treatments has a statistical significance of 0.036 using the previously mentioned model. In the third treatment, for the first time, participants can enforce social pressure to all the global members. Though this enforcement is non-binding, we see that it heavily contributes to more cooperation. Looking at the trend in table 1 as well as the graphs in figure 1 below we see that there is a clear negative trend of contribution throughout the 15 periods. The negative trends for treatments 1 and 2, when between-group communication isn't available, is much larger than period 3, both of these results are significant with p= 0.012. When between-group communication is added there are two separate social enforcements on commitment, and thus we see a weaker decline of contributions.

Figure 1. Global Contributions in Treatments 1-3.

Vertical Axis: Contributions, Horizontal Axis: Period
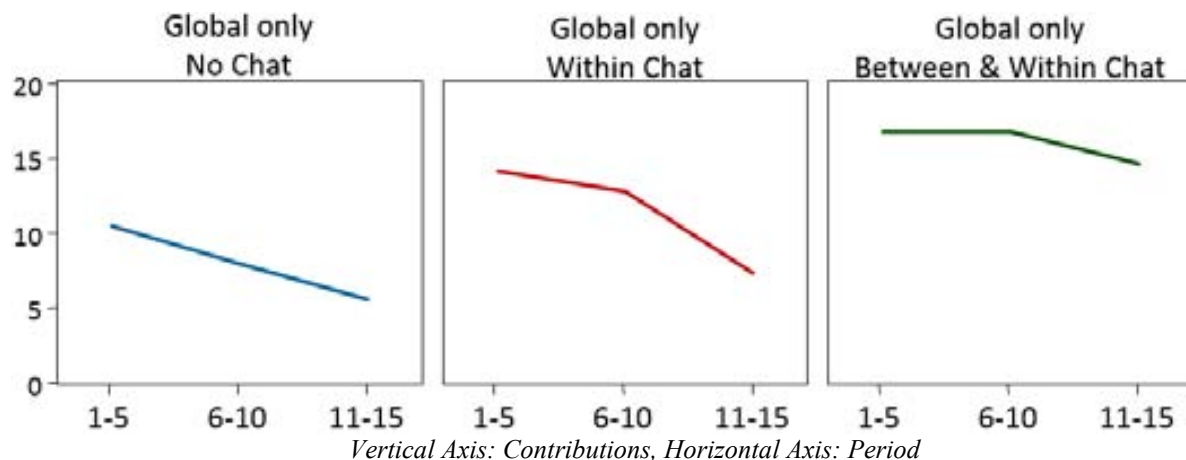
It might be instructive to look at the standard deviation to understand the effects of communication on social structure. As communication between the local members is added, the standard deviation of the mean contribution of the local members more than halves. The decrease of the mean standard deviation within-groups between treatments 1 and 2 is backed by a statistical significance of 0.000 using a two-sample Fligner-Policello Robust Rank Order Test. As the addition of communication in the 3rd treatment is between the local groups, it does not contribute to a better coordination within a local group and thus we don't see a statistical difference between treatments 2 and 3 for the standard deviation within the local groups variable.

Looking at the 4th column of Table 1, we find that the only statistically significant difference of the mean standard deviation between the local groups is between treatments 2 and 3. This is backed by a two-sample Fligner-Policello Robust Rank Order Test which showed a significance of 0.04. This again, goes hand in hand with our predictions as it is in the 3rd treatment that local groups can communicate between each other and coordinate. We can see the effect of this in the table as the standard deviation of the contributions decreases by more than half between these two treatments.

The last column of table 1 looks at the mean standard deviation of the different sets of eight "global groups" which participated in each of the three treatments. Though we do not don't find any significant differences in this result, it is curious that the variance in contribution is smallest when there is no communication.

Experimental Results 1: *As different levels of communication are added between the 3 treatments, the mean contribution increases dramatically. Within-group communication significantly increases contribution to the global public good. Between-group communication further increases contributions to the global public good. Furthermore when communication between same local group members is added, the standard deviation of the mean contribution of the members in that group decreases. When communication between the different local groups is added, the standard deviation of the mean contribution between the groups decreases.*

6.2. Global and Local Contribution with Communication

As explained in the experimental design section treatments four to six, are the same as treatments 1-3, with the key difference being than now participants can choose whether to contribute to the global public good or the less efficient local public good. For each of the three treatments, Table 2 shows the descriptive statistics of both global and local public good contributions. In it we can see the mean contribution of all 15 periods played the mean standard deviation of contributions within a local group, the mean standard deviation of contributions between local groups, and the mean standard deviation between global groups.

The three columns under Mean Contribution in Table 2, show the contributions to the global public good, the contributions to the local public good and the total contributions. By comparing the two

types of contribution for each treatment we can see the effect the different types of communication have on both the total contribution as well as the two individual contributions. With the possibility to contribute to the local PG as well as to the global one, we see in treatment four that almost two thirds of the total contributions go to the local PG. This leads to our next finding; the option to be able to contribute to the local PG leads to a significant decrease in global PG contribution when there is no communication available (significance of 0.039). The results show that when there is an option to contribute to both types of public goods, they choose to contribute to the local PG rather than the global. Considering that the local PG is a less efficient PG, this result perhaps seems to indicate that the participants trust smaller groups such as their local one more then they trust larger groups. Treatment four also sees the greatest drop in contributions (15 periods) in comparison to the other two treatments that have the possibility to contribute to the local PG, and is the only treatment whose drop in total contribution is significant with p = 0.05. The other two treatments also have a negative trend in table 2, however looking at the graphs above it is clear that most of the negative trend comes from the end game effect.

Figure 2. Global, Local, and Total Contributions in Treatments 4-6.

Vertical Axis: Contributions, Horizontal Axis: Period
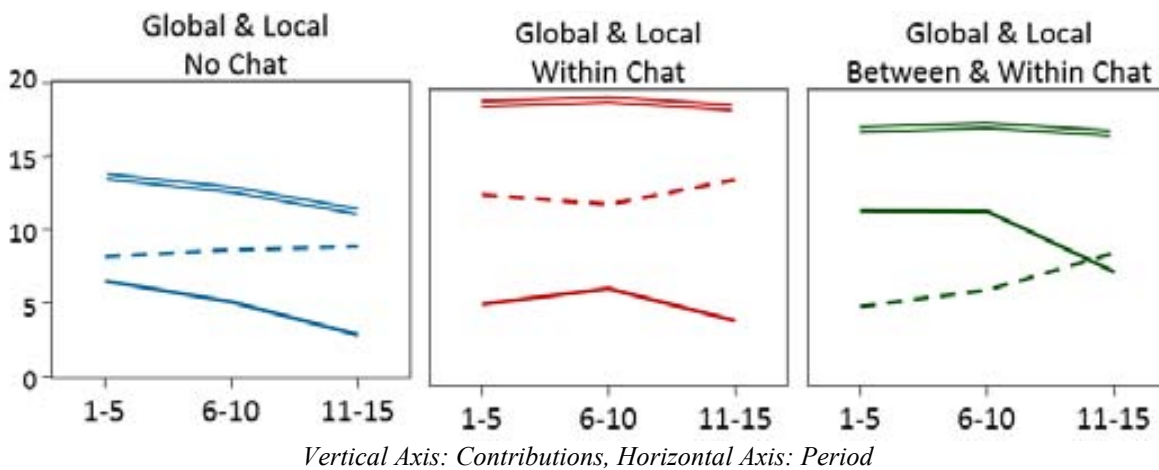
When within local group communication is added in the 5th treatment we see the total contribution go up significantly from 13.37 to 18.52 (p=0.000). However most of the increase in contribution is in the less efficient local public good. We see contributions to the global public good increase by an insignificant amount (p=0.775) whereas the contribution in local public goods sees a large increase from 8.54 to 13.07 points (P=0.001). These results go again hand in hand with our

hypothesis, as local-group members can now communicate with each other they are able to socially enforce cooperation between themselves. This leads to higher local PG contributions, which in turn, leads to higher total contributions. As between-group communication is still unavailable, just as in treatment four, we do not find a significant change in the contributions to the global public good.

Table 2. Descriptive Statistics - Treatments with Global & Local PG

T	Mean Contribution			Trend	Std. Dev. Within Local-Groups		Std. Dev. Between Local-Groups		Std. Dev. Between Global-Groups	
	Global PG	Local PG	Total PG		Global PG	Local PG	Global PG	Local PG	Global PG	Local PG
4	4.83	8.54	13.37	-1.96	2.05	1.70	0.96	3.60	2.52	1.50
5	5.45	13.07	18.52	-1.26	0.65	0.88	1.28	1.60	3.74	3.32
6	10.43	6.88	17.31	-1.54	0.70	1.20	0.97	1.93	7.61	5.58

T = Treatment; Trend = The trend of increase/decrease of contributions over time (periods)

We next look at the standard deviation for treatments four and five in order to recognize the effects of communication on what is now a more complex social structure. Looking at the two sub-columns of the standard deviation within local groups, we can see how coordinated individual members within a local group were throughout the treatments. As period four has no communication between the members, it is expected that this is where we see the largest variance in contribution. When within-group (local) communication is added in treatment five, we see a large decrease in the standard deviation of contributions within a local group. As local group members can now communicate between each other, they coordinate better. Both global PG and local PG see a larger drop in standard deviation in treatment five compared to treatment four, and looking at the mean contributions of treatment five, it is safe to assume that the reason for this is that local members decide not to contribute much to the global public good and contribute to their own local public good instead. The drop of standard deviation in both the global PG and the local PG contributions once within-group communication is added is backed by a strong statistical significance of 0.000 using a two-sample Fligner-Policello Robust Rank Order Test.

Moving on to the standard deviation between local groups, right away we can see that the only big difference in variance is in the contribution to the local PG between the 4th and 5th treatment. As communication between local group members is added in the 5th treatment, the standard deviation of the local PG decreases from 3.6 to 1.6. As local members are able to communicate with each other, but not to the other groups they decide to contribute more to the local PG. From this result we can see that all local groups made the same decision, even though they didn't have means to communicate with each-other. This result is in fact the only statistically significant result (p = 0.027) when looking at the standard deviation between local groups.

Experimental Result 2: *When within-group communication is introduced in the 5th treatment the total contribution goes up. This increase comes from the significant increase in contribution of the less efficient local public good. This also leads to a significant drop in standard deviation of contributions to both the global PG and the local PG. Additionally within-group communication leads to a significant decrease of standard deviation between local groups in the local PG contributions.*

In treatment six between-groups communication is added. As explained in the design section, one representative of each local group can talk to the other representatives of a local group within the same global group once every five periods. One result of the between-group communication is that we see the contribution to the global public good almost double from 5.45 to 10.43 (p=0.15). As local groups have a way to communicate with each other they have the opportunity to exert social enforcement, and trust each other more than when communication isn't available. Even though this result is only weakly significant with a 1-tailed random effect tobit model test (p=0.075), it is clear that contributions have substantially increased into the more efficient form of PG. The reasons why this result isn't strongly significant is due to the large variance between different global groups in treatment six, as well as the end game effect which played a role in decreasing the contributions. Treatment six also sees the contribution to the local public good almost halve, this result has a significance of 0.012 using the random effect tobit model. When looking at the change in contributions from treatment 5 to treatment 6 as a whole (figure) it is clear that there is a major change in the contribution pattern when between-group communication is available. The contribution shifts from being primarily into the local PG to being primarily in the global PG,

suggesting that between-group communication is critical for more efficient results in PG contributions.

The addition of between-group communication in the 6th treatment doesn't lead to better communication within a local group so we do not see a significant difference in standard deviation within local groups between treatments 5 and 6, which is consistent with the findings of table 1. However, the introduction of between-group communication doesn't result in a significant difference in the between-groups standard deviation of global PG's either. We believe that the reason for this is a ceiling effect of the standard deviation of global PG in treatment five, due to the very low contributions into the global PG in that treatment.

The last two sub-columns of table 2 look at the mean standard deviation to contribution of the different sets of eight "global groups" which participated in each of the last three treatments. The only significant result we find here is an increase in variance of the global public good between periods five and six (p=0.02). This result shows that once local groups in a given global group are able to communicate with each other, even though the actual mean contribution into the global PG doubled, not all global groups are able to coordinate and utilize this optimal strategy.

Additionally, we find that the introduction of the local public good in treatments four to six increases the total contribution in all three treatments compared to the treatment in which participants could only contribute to the global PG and had the same level of communication. This goes in line with the findings of Blackwell and Mckee (2001) who reported that total contributions went up once they had a global PG with a higher MPCR.

Experimental Result 3: *The introduction of between-group communication in treatment six, almost doubles the contribution to the global public good, however this result is not strongly significant. At the same time the contribution to the local public good significantly drops to almost half.*

6.3 Results from the Chat coding

We coded the entire 19,000 lines of chat which we accumulated during the running of this experiment in order to find out exactly how the participants used the chat to agree contributions

into the public goods available to them. Looking at Table 3 what we find is that actual contributions do differ from what they agreed to do in the chat-room but not by huge amounts, and the differences between agreed and actual contributions do not vary greatly between treatments. Participants generally contribute a slightly less than they agreed to in the chat room, around 10% less for the Global PG and around 5% less for Local PG.

Table 3. Mean Actual Contribution vs. Agreed Contribution – By Conditional /Non-Conditional Agreements

	T2		T3		T5				T6			
	Global		Global		Global		Local		Global		Local	
	C	NC	C	NC	C	NC	C	NC	C	NC	C	NC
Actual Contribution	8.34	14.67	13.31	17.30	7.93	4.24	13.18	14.02	12.16	10.05	5.45	7.67
Agreed Contribution	11.19	14.81	16.64	18.57	9.56	3.79	10.03	14.81	12.96	11.15	6.16	8.61

C = Conditional Agreement; NC = Non-Conditional Agreement

Some of this difference comes from occasional individuals who are unwilling to cooperate and from the end game effect. This differences do increase in some treatments however when we separate the agreements made in chat into two categories: conditional agreements and unconditional agreements. We classified conditional agreements, as agreements in which local group members agreed on a contribution amount, but based on some condition being met (that nobody defaults, that the other local groups also contribute a certain amount etc.). Unconditional agreements do not require any condition being met, participants simply agree to contribute a certain amount. As can be seen from table 3 conditional agreements have a much larger difference between agreed and actual contributions. To a small extent this was simply because the condition in the agreement was not being met, but at least equally as much this type of agreement was used as an excuse to not cooperate and free-ride buy certain individuals. The smallest hint that a vague condition may not be fully met, was used in order to stray from the agreement. This in turn lead to others straying from the agreement and thus the difference we see in the table. The exception to this is treatment 6, which has local and global PG and between group chat, where we see that conditional agreements are more respected than non-conditional. The main reason for this is that the complexity of that treatment required that well defined conditional agreements were agreed

upon. So in T6 we have more conditional agreements than any other treatment, around 45%, and due to the complexity of the treatment those who did not agree conditionally are more likely to be targeted by an opportunist.

7. Conclusion

In our first (baseline) treatment we find the typical PG contribution. Contributions start just above the halfway point and then drop significantly over time. This showed, yet again, that enforcement is needed for PG to be highly efficient even when participants don't have the option to contribute to more than one public good. As we added different levels of communication between the first 3 treatments, the mean contribution increases dramatically. Within-group communication significantly increases contribution and between-group communication further increases contributions to the global public good. We also find that when we added communication between the local group members, the standard deviation of the mean contribution of the members in that group decreases. Additionally, when communication between the different local groups is added, the standard deviation of the mean contribution between the groups decreases.

But as we know that in real life most people belong to groups within groups, and more often than not they have the option to contribute to more than one level of public goods, regardless whether the contribution is monetary, effort, or time. Thus, we wanted to test what happens when this lower level of PG is introduced, as well as, to see how the different levels of communication affect this new scenario. When within-group communication is introduced in the 5th treatment the total contribution goes up. This increase comes from the significant increase in contribution of the less efficient local public good. This also leads to a significant drop in standard deviation of contributions to both the global PG and the local PG. Additionally within-group communication leads to a significant decrease of standard deviation between local groups in the local PG contributions. Finally we find that the introduction of between-group communication in treatment six, almost doubles the contribution to the global public good, however this result is not strongly significant. At the same time the contribution to the local public good significantly drops to almost half. These findings, to some extent, go well with our predictions. We showed that within group communication is not going to be enough to promote global PG contribution, and that between-

group communication, through the representatives is the key to solving the efficient public good contribution in a multi-level PG scenario.

8. References

Andrighetto G, Brandts J, Conte R, Sabater-Mir J, Solaz H, Et Al. (2013) Punish And Voice: Punishment Enhances Cooperation When Combined With Norm-Signalling.

Becker S. Gary (1976). *The Economic Approach To Human Behavior*. Chicago.

Blackwell, Calvin And Mckee, Michael (2001): Only For My Neighborhood? Preferences And Voluntary Provision Of Local And Global Public Goods. Journal Of Economic Behavior And Organization, 52, 115-131.

Bochet, O., Page, T., Putterman, L., 2006. Communication And Punishment In Voluntary Contribution Experiments. Journal Of Economic Behavior & Organization 60, 11{26.

Bornstein , R. F. (*1992*). *Subliminal* mere *exposure effects*. In R. F. *Bornstein* & T. S. Pittman (Eds.), *Perception without awareness: Cognitive, clinical and social Perspectives.* (pg. 191-210)

Cason T., Sheremeta R., Zhang J., Communication And Efficiency In Competitive Coordination Games, Games And Economic Behavior, Volume 76, Issue 1, September 2012, Pages 26-43

Charness, G. (2000). Self-Serving Cheap Talk: A Test Of Aumann's Conjecture. Games And Economic Behavior, 33, 177-194.

Charness, G., & Grosskopf, B. (2004). What Makes Cheap Talk Effective? Experimental Evidence. Economics Letters, 83, 383-389.

Chudek M, Henrich J (2011) Culture-Gene Goevolution, Norm-Psychology And The Emergence Of Human Prosociality. Trends In Cognitive Sciences 15: 218–226

Davis And Holt (1993) Experimental Economics, Princeton University Press, , To ... Jan 23-25 Methodology: Davis And Holt Chapter 1, Pp.1-45

Fellner, Gerlinde And Lünser, Gabriele K., (2008), Cooperation In Local And Global Groups, Department Of Economics Working Papers, Vienna University Of Economics, Department Of Economics.

Fischbacher, Urs (2007): Z-Tree: Zurich Toolbox For Ready-Made Economic Experiments. Experimental Economics, 10, 171-178.

Fisher, Joseph; Isaac, R. Mark; Schatzberg, Jeffrey W. And Walker, James M. (1995): Heterogeneous Demand For Public Goods: Behavior In The Voluntary Contributions Mechanism. Public Choice, 85, 249-266.

Friedman, Milton. *Essays In Positive Economics*. Chicago, IL: University Of Chicago, 1953. Print.

Frohlich, Norman And Oppenheimer, Joe, (1998), Some Consequences Of E-Mail Vs. Face-To-Face Communication In Experiment, *Journal Of Economic Behavior & Organization*, **35**, Issue 3, P. 389-403.

Hirshleifer, J., 1983. From Weakest-Link To Best-Shot: The Voluntary Provision Of Public Goods. Public Choice 46, 221–231.

Isaac, R.M., Walker, J., Thomas, S., 1984. Divergent Evidence On Free Riding: An Experimental

Isaac, R. Mark, And James M. Walker. "Communication And Free-Riding Behavior: The Voluntary Contribution Mechanism." *Economic Inquiry* 26.4 (1988): 585-608. Print.

Isaac, R.M., Walker, J.M., 1988. Group Size Effects In Public Goods Provision: The Voluntary Contributions Mechanism. Quarterly Journal Of Economics 103, 179–199.

Kerr, Norbert L., And Cynthia M. Kaufman-Gilliland. "Communication, Commitment, And Cooperation In Social Dilemma." *Journal Of Personality And Social Psychology* 66.3 (1994): 513.

Lapinski, M. K., & Rimal, R. N. (2005). An Explication Of Social Norms. Communication Theory, 15, 127–147.

Ledyard, J.O., 1995. Public Goods: A Survey Of Experimental Research. In: Kagel, J., Roth, A. (Eds.), The Handbook Of Experimental Economics. Princeton University Press, Princeton, NJ, Pp. 111–181.

Oprea, Ryan, Charness, Gary & Friedman, Dan, 2012. "Continuous Time And Communication In A Public-Goods Experiment"

Ostrom, E., Walker, J., Gardner, R., 1992. Covenants With And Without A Sword: Self Governance Is Possible. American Political Science Review 86, 404–416.

Pavitt S., 2011. Communication, Performance, And Perceptions In Experimental Simulations Of Resource Dilemmas. *Small Group Research*

Riechmann, Thomas and Weimann, Joachim, (2008), Competition as a coordination device: Experimental evidence from a minimum effort coordination game, *European Journal of Political Economy*, 24, issue 2, p. 437-454.

Sally, D., 1995. Conversation And Cooperation In Social Dilemmas: A Meta-Analysis Of Experiments From 1958 To 1992. Rationality And Society 7, 58–92.

Sherif, M. (1935). A Study Of Some Social Factors In Perception. Archives Of Psychology, 27, 1–60.

Sherif, M. (1936). The Psychology Of Social Norms. New York: Harper.

Van Huyck, J. B., Battalio, R. C., & Beil, R. O. (1993). Asset Markets As An Equilibrium Selection Mechanism: Coordination Failure, Game Form Auctions, And Tacit Communication. Games And Economic Behavior, 5, 485–504.

Appendix A. Instructions

You are participating in an experiment about economic decision-making. During the experiment, you can earn money. Your earnings will depend on your decisions and the decisions of others. These instructions describe the decisions you and other participants should be taken and how your earnings are calculated. Therefore, it is important to read them carefully.

During the experiment, all the interaction between the participants will take place through computers. It is forbidden to communicate with other participants by other means. If you have any questions, please raise your hand and one of us will come to answer it. Keep in mind that the experiment is anonymous, i.e., your identity will not be disclosed.

During the experiment, your winnings will be calculated in points. At the end of the experiment the points will be converted to euro at the following Exchange rate:

30 points = 1 euro

General instructions

In the experiment, the participants will be randomly assigned to a **Group of 3 persons**. As the experiment is anonymous, no one will come to know the identity of the other members of the group. To distinguish itself, each person in the group will be assigned a random number, so each group will have a member of Group #1, a member of #2 group, and a member of #3 group.

The groups will remain constant throughout the experiment. In other words, you're going to be a member of the same group, and to keep the same number within your group throughout the experiment.

During the experiment, your group will interact with **two other groups**, and these groups are going to be the same two groups throughout the experiment.

The experiment is divided into 15 periods. In each period, you will have to make a decision.

Your decision

At the beginning of each period, each participant receives an initial income of 20 points. Your decision is to choose: points are going to invest in the project to your group, how many points are going to invest in the project B, and points are going to stay to yourself. For example, you can invest 5 points in project A, 10 points in project B, and stay 5 points for yourself. Everyone is going to make the decision at the same time.

Only members of your group can invest in project a to your group. Note that each of the two other groups have their own project, to which only the members of that group can invest. On the other hand, all the members of your group over all members of the other two groups can invest in project B. Therefore, only members of your group receive profits from the project to your group, but all the members of your group and all members of the other two groups receive profits from the project B.

How to calculate your winnings

Your earnings consist of three parts:

1. The points you get for yourself, by which you receive 1 point in earnings for every point you get.
2. Earnings from the project to your group, which are calculated in the following way:

   Your project A profit = 0.6 times the total investment of your group in A project

3. Earnings of project B, which are calculated in the following way:

Your winnings from the project B = 0.3 times the total investment of all stakeholders in the project B

Therefore, your total winnings are:

(20 - your investment in project A - your investment in project B) + 0.6 × (total investment of your group in A project) + 0.3 × (total investment of all stakeholders in the project B)

Remember: the total investment in A project is the sum of points invested in the project by all the members of **your group**, and the total investment in the project B is the sum of points invested in project B by all members of **the three groups** (i.e. points invested by your group(, more the sum of points invested by members of the other two groups).

One way to see the decision is thinking about how to invest 1 point. You have three possibilities:

1. Can stay point for yourself, which increases your winnings by 1 point and does not affect the earnings of others.
2. You can invest in project A point, which increases the total investment of the project to your group by 1 point. Therefore your winnings increase $0.6 \times 1 = 0.6$ points, and the other two members of your group profits also increased 0.6 points. In other words, the total profit of your group increased 1.8 points.
3. You can invest in project B point, which increases the total investment of the project B by 1 point. Therefore your winnings increase $0.3 \times 1 = 0.3$ points, the profits of the other two members of your group also increase 0.3 points, and earnings of each Member of the other two groups increase 0.3 points. In other words, the total profit of your group increase 0.9 points and earnings total of the three groups (yours and the other two) increased 2.7 points.

Keep in mind that as well as you can increase earnings of others by investing in the project A and B, your earnings will also increase if other members of your group to invest in A project or if anyone of the three groups invested in project B.

All winnings are calculated in the same way.

Example:
Suppose that:

☐ You get 5 points, you invest ten points in A project and 5 points in project B.
☐ The other two members of your group invested a total of 20 points in A project and a total of 15 points in project B.
☐ Participants in the other two groups invested a total of 30 points in project B.

In this case, the total invested in A project is 10 + 20 = 30 points, which means that your project A profit are 0.6 × 30 = 18 points. The total invested in the project B is 5 + 15 + 30 = 50 points, which means that your project B gains are 0.3 × 50 = 15 points. Therefore, your total earnings are the sum of the points that you stay (5 points), the profits from the project (18 points), and the profits of the project B (15 points), which is equivalent to 5 + 18 + 15 = 38 points.

Communication

At the beginning of **periods 1, 6, and 11**, they will be able to communicate with the members of his group, and also, some people are going to be able to communicate with members of other groups.

In particular, the members of the group will be able to communicate among themselves via a chat window. No one outside your group will be able to see the conversation of your group.

In addition, a person (and only one) in your group will be able to communicate with a person in each of the other two groups. We call these people the group **representatives** . In addition to the chat window used to communicate with members of their group, the representative of the Group has another chat window in which it can communicate with the other representatives. Only members can view this window. In each group, and throughout the experiment, representative will be assigned person to be the Member of Group #1.

You can chat freely. The only exception is that, to remain anonymous, **is forbidden to transmit information that serves to identify you**. If you transmit a message of this type we would not pay for your experiment.

Screens and the order in which decisions are made

The periods 1, 6, and 11, the first screen is the chat window. Members of Group #2 and #3 see the following screen.

In the chat window, you can communicate with other members of your group. Note that to save space the phrase "member of group" is has abbreviated "Mg". To send a message, write it on the bottom of the window and press the "Enter" key. At the bottom of the screen you can see the time remaining to chat. Anyway, if you want to end the chat session before that is exhausted the time, you can press the button "Ended up talking". The chat session will end if all members of the group press the button. As the Member of Group #1 is the representative, this member has a chat window additional which can communicate with other representatives. The Member of Group #1 see the following screen.

Periodo #1	Tú eres el miembro de grupo #1	Tú eres el representante Y

Tu ingreso inicial es de 20 puntos

Tu inversión al proyecto A: _____

Tu inversión al proyecto B: _____

OK

To send a message to representatives of the other groups, write it on the bottom of the window of the representatives and press the "Enter" key. After the chat screen is the screen where you decide to invest points. This screen looks the following way:

As you can see, on this screen you decide how many points like to invest in your group project A and project B. The difference between what you invest in two projects and your initial income is the amount of points you get for yourself. To make your decision, type a number in the respective field and press the "OK" button.

The results screen will appear once all have decided how much to invest. The results screen looks in the following way:

Periodo #1   **Tú eres el miembro de grupo #1**   **Tú eres el representante Y**

Tu Grupo

Inversión en:	Proyecto A	Proyecto B
Tú (miembro de grupo #1)	10	5
Miembro de grupo #2	10	10
Miembro de grupo #3	10	5
Total	30	20
Promedio de los otros miembros del grupo	10.0	7.5

Todos los Grupos

Inversión en:	Proyecto B
Tu Grupo	20
Grupo X	10
Grupo Z	20
Gran Total	50
Promedio de los otros grupos	15.0

Tus Ganancias

Cantidad no invertida	5.0
Ganancias de A	18.0
Ganancias de B producidas por tu grupo	6.0
Ganancias de B producidas por otros grupos	9.0
Ganancias totales de B	15.0
Tus ganancias totales	38.0

OK

In the left part of the screen you have a table called "Your group". This table contains the investments of each Member of your group, the average investment of the two other members of your group, and your group's total investment. Only members of your group can see the information in this table. In the center of the screen you have a table called "All groups". This table contains the total investment of your group and the other two groups in project B, the average investment of the other two groups in project B, and grand total invested in project B. Everyone can see the information in this table. Finally, on the right side of the screen you have a table called 'Your earnings'. This table contains your winnings in this period and explains how they were calculated. You can only view the information in this table (but keep in mind that the information in the table on the left is sufficient to calculate the profits of all the members of your group).

Raise your hand if you have any questions. Otherwise press the button "Finished reading".

CHAPTER V

Equality of Pay in Highly Interdependent Work Groups:

1. Introduction

Every work group, small or large, organizes employees in a way that they are, at least to some extent, interdependent. Since the dawn of industrialization, firms have put in place tasks and processes that expect employees to work interdependently of each other in order to more efficiently complete final goods and services. Though interdependence in the active running of work groups has been a topic of interest for considerable amounts of research in the past, the focus of most of this research has been the relation of the level of interdependence in work groups to performance and issues related to performance. (e.g., Campion, Medsker, & Higgs, 1993; Campion, Papper, & Medsker, 1996; Guzzo & Shea, 1992; Saavedra et al., 1993; Van der Vegt et al. 2005).

Task interdependence, along with goal interdependence and feedback interdependence are considered the key components to interdependent work environments. Task interdependence has been defined as the level to which an individual team member's job design necessitates coordination activities and information exchanges with others on the team, in order to fully complete a task (Brass, 1981; Kiggundu, 1983; Van der Vegt et al. 2001). Perhaps the best definition comes from Daft (1983), who defines it as "the dependence of one unit on another for materials, resources, or information".

The levels of task interdependence stem back to Thompson (1967) and his typology of interdependence, which separated task interdependence to three different types: pooled, sequential, and reciprocal interdependence. Pooled interdependence is the lowest level of interdependence, as each unit (individual, organizational department, group etc.) performs completely separate functions. The interdependence here comes solely because they work towards the same goal, such as the profit of the company of which both units are part of. Under pooled interdependence, each member contributes to the final output without a need for any interaction between different members, and each member completes their entire task individually. According to Saavedra at al.

(1993) under pooled interdependence group performance is defined as simply the sum of all individual performances. An example of this type of loose interdependence is salesmen in a firm, who work independently of each other, and independently contribute towards the same goal of their firm. Sequential interdependence, is a more complex and demanding type of interdependence in which one unit in the overall process produces an output which is necessary for the performance of the next unit. (Thompson, 1967) This type of interdependence requires that units perform different parts of a task in a clearly arranged, one-way, order. The most famous example of this type of interdependence is the assembly line, in which the output of a worker on the line is heavily dependent on the workers in the earlier parts of that very same assembly line. Lastly reciprocal interdependence is the highest level of interdependence and thus the most complex of the three. It is very similar to sequential interdependence with the addition of cyclicity. That is, tasks do not move in only one direction on a line, but go back and forth between different members multiple times.

Goal interdependence has been defined as when individuals or units are in a situation in which they can achieve their goals only if the other individuals or units they are connected with also attain their goals (Johnson & Johnson, 2009). The benefits of having goal interdependence, as opposed to unique individual goals for each unit, is that the first results in units which encourage and facilitate each-others' efforts, all for the purpose of obtaining the interdependent goals. Feedback independence is the interconnection between members of a work-group based on the performance feedback they receive (Saavedra at al. 1993).

Our initial goal with this paper was to create a paradigm in which this very common combination of sequential and pooled task interdependence can be studied naturally in the lab, in relation to a variety of related topics. For this purpose we constructed a naturally flowing three-staged production line in which participants work individually, one stage at a time, in order to deliver a final product out to market. The design we created made sure that each individual effort and performance at any stage has an effect on the maximum individual effort and performance in all subsequent stages, as well as a general effect on the final performance of the group as a whole. We have individuals who are part of a production line multiply numbers at all three stages of the experiment, with each stage increasing in difficulty (more digits to multiply), and with only the

correct answers passing on to the next level. The final product is complete only if all three stages correctly solve the multiplications passed on to them by the previous stage, thus the group only gets paid by the last stage's (stage 3) correct output.

Further, our goal is to determine the importance of equality of pay in the scenario of this common yet unique type of task interdependence. Incentives and equality of pay have been very hot topics for quite some time now, yet for the most part they have been studied respectively with the focus on the individual and with a focus on separate hierarchies, which do not necessarily have direct collaboration for the completion of a final good/service. Yet, this most common type of task interaction, sequential with some pooling, is where one would expect the topic to be of most importance, as it is here that different units depend on each other the most, and thus any factor that may negatively impact the will to collaborate with others would be more detrimental to the final outcome of the work group. To support this Fisher (1994) has shown that in the presence of high task interdependence, the performance of the whole group could be identical to the output of the worst performing group-member. Though this is not a typical scenario where one would expect to find free-riding problems, any lack of effort or sabotage coming from one individual due to dissatisfaction, could potentially be disastrous for the performance of the entire work-group.

One of the main reasons why we are interested in equality of pay in a sequentially interdependent task, regardless whether the situation is one of an industrial assembly line or a service, such as delivering a marketing campaign, is because not all stages of a production line are of equal difficulty or of equal supply and demand of workers who can complete it. For example, in the US assembly line workers who work on the manufacturing and assembly of the motor of the car are paid around 3 times more than the average assembly line worker.

In this experiment we use our own version of what Saavedra at al. (1993) called complex interdependence, namely a combination of two tasks interdependencies (pooled and sequential), and to some extent goal interdependence. These are the most common types of interdependencies seen in many work-groups, and yet the most understudied. Even though some work-groups have a fixed level of task interdependence throughout all stages (eg. certain types of assembly lines), in the majority of work environments this is not the case. Task interdependence can vary at different

stages of the process, and it can also change over time. Olson et al. (2001) states that the amount of interaction and exchange of information between various stages, which are part of product development, increases as the product comes closer to completion. This means that in the process of product development the task interdependence moves from pooled interdependence at the beginning to the higher levels of interdependence at the end.

We designed four treatments with two different payoff sharing schemes, two treatments with equal payoff sharing and two with an exponentially different share of the payoff between the three stages of production. The way in which the treatments with the same payment scheme differ from each-other is in the way that participants are allocated to their positions on the production line; in two treatments they get allocated randomly, and in the other two they get allocated by a pre-experimental test which ranks them according to performance. The reason we implemented different allocations into position is threefold. First, in order to see whether a participant's response to the payment schemes differs if participants feel that the difference in the equality or inequality of pay is justified rather than not. Second, we wanted to see whether allocating participants by ranking is enough to offset or even trump the difference that may arrive due to pay inequality, even with such a general task as multiplications. Lastly, we wanted to find out whether participants would behave differently in the ranked pre-experimental round under the two payment conditions; namely, whether they will self-select themselves into positions differently by adjusting their effort.

2. Background

Due to researchers' realization of the importance of task interdependence, they have over time managed to develop a wide variety of categorizations on this topic. Thompson (1967) categorized task interdependence into three types: pooled, reciprocal, and sequential; Kiggundu (1983) then separated initiated task interdependence and received task interdependence; Saavedra at al. (1993) works on complex interdependence, and Wageman (1995) clarified the difference between task and outcome interdependence. Researches in management have put to use Thompson's interdependence typology to investigate a variety of related issues. Thus, it has been used to investigate the difference between individual versus group rewards (Ex. Chow, Shields, &

Chan, 1991), to look at its effect on team performance (Saavedra et al. 1993; Langfred & Shanley, 2001), to study task uncertainty (Ex. Hirst & Yetton, 1999), to look at mechanisms of coordination (Macintosh and Daft, 1987), and to study budget targets (Fisher, 1994).

Of the many positive effects task interdependence has been found to have on work-groups, one that must be singled out is the correlation that task interdependence has with cooperation (e.g. Pearce & Gregersen, 1991; Wageman & Baker, 1997). Kiggundu (1983) has implied that this comes from the interactive nature of the jobs related to task interdependence, namely that they increase the responsibility felt for others in the work-group. Anderson & Williams's (1996) explanation of this positive correlation is that perhaps larger amounts of task interdependence increase the understanding that there is a need for better coordination and problem solving. Thus it should not be surprising that task interdependence is also regarded as part of the key variables that effects the performance of a team; Guzzo & Shea (1992) have defined it as the extent to which team members are required to interact and coordinate in order to finish a given task.

Yet, research has acknowledged that there is a limit and that sometimes too much task interdependence can actually have an opposite effect and hinder intragroup cooperation, timing, and coordination (Steiner, 1972; Saavedra et al., 1993). Allen et al. (2003) in an experimental study also found that the high task complexity which is required for high task interdependence can counter the positive performance effects of high levels of helping behavior and effort.

Van der Vegt & Van de Vliert (2005), conducted a panel study of senior business students who signed up for a management game. They found that when task interdependence is low, perceived skill dissimilarity, leads to decreases in both self-reported and peer-rated helping behavior, but high task interdependence increases a person's helping behavior. The correlation between helping behavior and task interdependence has also been found in a lab experiment by Allen et al. (2003). Group-level task interdependence has also been found to positively relate to job and team satisfaction (Van der Vegt et al. 2001). Further, Langred (2005) demonstrated that team performance depends on both individual and team autonomy, stating that the optimal combination of individual and team autonomy are contingent on the level of task interdependence in a team. Hertel et al. (2004), in their field study with virtual teams, showed that high task

interdependence, high outcome interdependence, and quality of goal setting are correlated with more effective teams.

Libby et al. (2001) using an experiment approach found that that there is no incentive structure that leads to a dominant performance across all task interdependencies. More importantly for our work though, the authors found that the effectiveness of a budget based incentive structure varied according to the degree of task interdependency. Interestingly Wageman and Baker (1999) find that task and reward interdependence is important to performance but only task interdependence drives observed cooperative behavior, and that rewards don't make a significant difference. Libby & Thorne (2009) study the effects of individual, group, and mixed incentive structures on group performance in both assembly lines and on teams. They find that performance is higher under group incentives for teams, however they find no difference in group performance for assembly lines, regardless of the large payment structure differences. This result seems to be a recurring theme when it comes to payment schemes in work-groups with high interdependence. The small amount of research that looks at different payment schemes in sequential and reciprocal task interdependence settings, seems to find that the difference between the payment schemes do not lead to a significant improvement in performance. Similarly, Guyman (2006) compares group performance under a group piece-rate incentive contract to a group performance under a group budget-based incentive contract in two production settings, a dependent production setting and an interdependent production setting. The author states that "group budget-based incentive contract leads to higher levels of performance only for groups that perform independent tasks".

Even though not extensive, evidence seems to show that highly interdependent tasks, seem to improve cooperation and helping behavior so much that different payment schemes do not seem to make as big a difference in performance. This effect increases when goal interdependency is added. Our experimental design does not set goals or targets to members of a work group, but as their payment is based on the revenue sharing principle, their goals are automatically aligned and due to the design dependent on each other.

Knowing this, we are interested to see the effects of payment equality under two different levels of fair allocation to position on our highly interdependent task. It has been argued that a key element in the various models of efficiency wage, is the idea that workers will be unwilling to exert full effort when they realize that they are not being paid fairly (Akerlof and Yellen, 1990).

Additionally, there is the idea that what employees see as fair payment, heavily depend on the wages paid to their co-employees (e.g. Frank, 1984; Lazear, 1989).

Charness and Kuhn (2007) argue that the previous might not be the case. Their argument, which goes hand in hand with their results, is that the little evidence available which finds that worker effort depends on whether they perceive to be treated fairly is mostly done through experiments of gift exchange. Studies like Fehr et al. (1993), Fehr et al. (1998), Fehr and Falk (1999), and Charness (2004), all show that when employers offer a gift to employees (wages), employees seem to reciprocate (effort), even when this is far from the dominant strategy. Charness and Kuhn (2007) argue that even though this shows that employees do respond to the perceptions of fairness, it doesn't not necessarily prove that employee's respond to the wages of their co-workers. Their own result goes along with this argument. In a laboratory experiment they find that workers effort decisions are highly sensitive to their own wages, but largely unresponsive to co-workers wages. They also argue that wage compression can be harmful and lead to paying equal wages to workers of unequal productivity, which is far from profit maximizing.

We are using these arguments and findings to see the effect of payment equality on our unique interdependent design. We have two types of payment sharing in this experiment. The first is equal pay where every member of a work-group is paid an equal share of the total earned by the work group. In comparison to the first , the second type of payment sharing decreases the payment of level one participants by 400%, and increases the payment of the level three participant by 400%, which leads to a difference of pay which is 16 times larger for one participant in comparison to the lowest paid workers. Even according to the findings of Charness and Kuhn (2007) this should lead to a significant decrease in performance for the level one participants, if not because of the huge wage disparity, due to the 400% decrease in their pay. However, their experiment does not have a real effort task with high interdependence, which we believed to be the key in eliminating payment effects on performance. This is the exact reason why we chose such extreme payment differences. That is, to find out, in a realistic scenario where effort from every member of the work-group is needed in order for anyone to get paid, whether extreme pay inequality will have a significant impact on the performance, knowing that high interdependence also leads to less payment sensitivity. The fact that we use a real effort task is also crucial because, effort means

exerting real effort, in this case multiplying numbers, and not a direct monetary cost. So, in the treatments with unequal pay, level-one participants do actually have to decide whether they are going to put in maximum effort into solving multiplications when their total payment for their 90 minutes of varied effort would likely be measured in cents.

3. Procedures

We conducted multiple sessions for each of the four treatments. The average running time of all four treatments was one hour and 40 minutes, which includes the time for reading the instructions out loud before the session as well as the time for paying all of the participants at the end of the session. Participants are paid a 4 euro show up fee plus whatever they end up making during the experiment. The experiment was conducted at two of the computer rooms of the Universitat Autonoma de Barcelona and participants were undergraduate students from the university. The z-tree software was used to run the experiment (Fischbacher, 1999).

In each of the sixteen sessions there were either fourteen or twenty-one participants, meaning two or three work-groups. We therefore have 322 participants in total, of which 84 took part in treatment 1, 77 took part in treatment 2, 77 took part in treatment 3, and 84 took part in treatment 4.

4. Experimental Design

Our design is meant to capture the essential parts of multiple stages of a production line process which we want to study.

The situation we are interested in representing is one in which a group works on completing a real effort task, which is in form of a production process and has combination of pooled and sequential task interdependence, a design which is common in real life practices. In our production process, there are 3 sequential stages in which individuals solve multiplication equations. Each subsequent round consists of harder multiplications than the previous, and the maximum number of possible multiplications is fully dependent on the performance of the previous round. The final pay for all participants in the production process group depends on the number of correctly answered multiplications at the final stage of production. Our focus is on the effort and

performance of the participants in all stages of the interdependent sequential task and whether and how these change under different payment schemes and methods of allocation.

Multiplications, as a task, should be part of all university student's general knowledge. Further, they are easy to manipulate at different stages, easy to increase or decrease in difficulty, and are a task which provides a level playing field regardless of field of study and academic performance. Dohmen and Falk (2011), who also use multiplications in their experiment on the impact of incentives on self-selection, state that: "As a task, multiplying numbers is also well suited for our purposes because it requires no previous knowledge, is easy to explain, and guarantees a sufficient degree of heterogeneity in productivity". The authors further cite Roth (2001), who states that multiplications are a good proxy for general cognitive ability, and that the learning effects of this type of task are expected to be small.

The experiment consists of four treatments, all of which have a different combination of 2 allocation methods and two payment schemes, differentiated solely by the two extremes of payment equality they represent. Treatment 1 has random allocation and equal pay, treatment 2 has random allocation and exponentially unequal pay, treatments 3 has ranked allocation and equal pay, and treatment 4 has ranked allocation and exponentially unequal pay.

Figure 1: Treatments Matrix

	Equal Pay	**Unequal Pay**
Random Allocation	Treatment 1	Treatment 2
Ranked Allocation	Treatment 3	Treatment 4

We will use the first treatment as a baseline for describing the experimental design in detail and then explain the differences of the other 3 treatments in comparison to the first.

While participants do the tasks both individually and sequentially, the interdependence in the production process dictates that if the performance of only one stage is lacking it will have a negative effect on the entire production and thus the entire groups pay.

4.1 Treatment 1 – Random allocation & Equal Pay

The experiment starts by assigning participants randomly into groups of seven and consists of four rounds, of which three are experimental (paying rounds) and one is a practice round. The participants then all do a practice "pre-experimental" round in which they have to solve multiplications of 2, 3, and 4 digit numbers. Once this period starts, all participants receive a different and random 2x2 digit multiplication. They are provided with paper and pencil in order to solve the multiplications, but ultimately they have to input the answer on the screen and press an "OK" button. It is important to note that all pieces of paper provided to the participants, on which they are to initially solve the multiplications, already have on them one correctly solved 4x4 digit multiplication as an example. At the bottom of the screen there is a counter telling participants how many correct answers they have so far, therefore, they receive feedback after each input. Once a participant inputs the answer, he/she then receives a random 3x3 digit multiplication; after this is answered, they receive a random 4x4 digit multiplication, and then the screen goes back to a random 2x2 digit multiplication. This process goes on until the 6 minutes run out, after which all participants are given feedback on their results of the practice rounds and separated for each of the 3 levels of difficulty (2,3,4 digit). However, they do not receive any feedback about how the others in the group performed during this round.

Figure 2: Design of interdependent work-group

First-stage workers   Second-stage workers   Third-stage worker

P1 x * x = y1

P2 x * x = y2   P5 y1 * y2 = z1

P7 z1 * z2 = w

P3 x * x = y3

P4 x * x = y4   P6 y3 * y4 = z2

P = Participants, x = Double Digit Numbers, y = Triple Digit Numbers, z = Four digit numbers
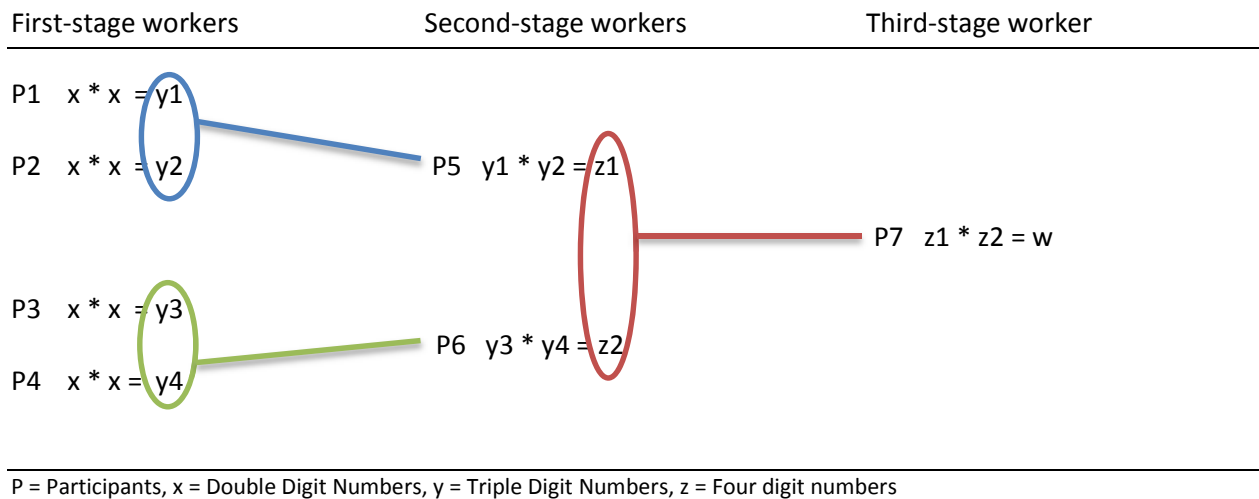
After this, participants are randomly assigned to one of the seven positions in the group. In one of the three different stages of the production line. Four are allocated to a position on the first

stage of the production line, two are allocated to a position on the second stage of the production line, and one is allocated to the final, third stage, of the line. They are told in which of the three stages of the production they have been randomly assigned to, and the exact number of their position in the production line.

The first experimental round starts off by the four group members allocated to the 1st stage. They are given 4 minutes to individually solve as many two-digit, 2 number multiplications (ex. 25*30=?) as they can. The input and generation of new multiplications works the same as in the practice round explained above, but now only two digit numbers are generated. Every answer that is correct goes into one of two "pools". The correct answers of participants 1 and 2 go into the pool of participant no. 5, and the correct answers of participants 3 and 4 go into the pool of participant no. 6. This is the small pooled component of our design. Even though the design in general is sequential, and bad performance at any stage can be detrimental to the entire work-group, there is a small cushion in the form of pooling. Players 1 & 2, players 3 & 4, and players 5 & 6 pool their answers together, thus bad performance (deliberate or not) from one player, in the first two stages can, to some extent, be saved by a great performance from his/her "pooling partner", though it would still lead to a large impact in the final result. The reason we designed the experiment this way is that it most closely represents what happens in real companies, even those that have a sequential conveyor-belt type of task interdependence. The fact of the matter is that even the most extreme sequential task, the conveyor belt, uses multiple workers at one stage of the belt doing the same task, and a poor performance by one of the workers can be, to some extent, made up by the other (not without a cost to productivity, but without a failure to the entire process). It is only the most difficult tasks, which require highly specialized skills, that cannot be supported by other workers' efforts in a particular stage of production.

All numbers in the 2x2 digit multiplications consist of two random numbers between 18 and 31, the reason for this is that any combination of these numbers when multiplied make a 3 digit number, which is not too low for manipulation at a later stage. The correctly answered inputs are stored by the program into one of the two pools for the next stage of the production.

Once the 4 minutes run out, stage one participants stop working and stage two participants begin their task. The two participants in the second stage have 6 minutes to solve 3x3 digit multiplications. However, they do not solve multiplications of random three-digit numbers, but multiplications of the correct answers given in stage one. Participant no. 5 (P5) gets multiplications

of correct answers inputted by participants 1 & 2, and participant no. 6 gets multiplications of correct answers inputted by participants 3&4. The correct answers of P1 and P2 get pooled into one pool so it doesn't matter if P5 gets to multiply two correct answers of P1, two correct answers of P2, or one of each. Each correct answer from the first stage gets used only once in a 3x3 digit multiplication in stage two. For example: If P1 solves 13 multiplications correctly and P2 solves 4 multiplications correctly, P5 should only be given 8 multiplications to solve (13+4 =17, 17/2=8.5). The period ends for both participants of the stage when the 6 minutes run out, or for one of them when he/she runs out of multiplications to solve (if it happens before the 6 minutes run out). All the correct answers given by player 5 and player 6 are divided by 100 and rounded, in order to make a four digit number, and then pooled together in yet another pool and passed off to stage three.

As soon as stage 2 finishes, stage 3 starts. The sole participant in the 3rd stage of the production line now has 8 minutes to solve 4x4 digit multiplications, which were pooled by the correctly answered multiplication questions by P5 and P6. The stage finishes when the 8 minutes run out or when P& runs out of possible multiplications to solve.

With this, the first experimental round ends, and all participants learn in detail about the performance of their production line group and how much they have earned in the round. They are shown the number of correct answers and incorrect answers at each stage of the production, but cannot see the performance of any individual participant (except P7 who is the sole participant of the 3rd stage). The total earnings of the round depend solely on the correct number of 4x4 digit multiplications minus a varied amount for all wrong inputs in all three stages of production. These earnings are then split evenly among all 7 members of the group, which equates to 14.29% each.

Revenue for the work-group = (Number of Correct w's * 8 euros) - [(Number of incorrect y's * 0.15 euros) + (Number of incorrect z's * 0.30 euros + (Number of incorrect w's * 0.60 euros)]

As in real life, a company makes revenue only on a final good on their production line and not on a semi completed good, which is why the income comes only when a group manages to complete a good correctly though all 3 stages of production. Companies have a cost for all mistakes made throughout all stages of production, with the mistakes generally being more costly the closer the product is to completion. Additionally, we preferred there to be an additional cost to getting

answers wrong, other than the time wasted, as this brings the three stages closer together in terms of effort that needs to be exerted. Estimating or guessing low 2x2 digit multiplications is possible whereas doing the same with 3x3 or 4x4 digit multiplications is much more difficult, and by adding a cost to the wrong answers we aim to eliminate additional difference in effort needed between the three rounds.

The production line process is repeated for a total of three rounds, after which participants are shown a final table with the earnings of each round, and a sum total earned for the experiment which consists of the sum of the earnings for the three experimental rounds. If the total earning after the three rounds is negative, the program tells them that their total income is 0. This did not happen even once, though many groups did get a negative income (loss) for at least one of the three experimental rounds, in which case this loss is subtracted from the rounds with positive income. With this, the experiment ends and all participants are paid the total amount earned for the three rounds plus the 4 euros show up fee.

4.2 Treatment 2 – Random allocation & Exponential Pay

Treatment two is identical to treatment one, except for the way in which a work-group splits the final revenue between its members. In treatment one all participants split the revenue equally between them, meaning they get 14.29% each, but in treatment two the revenue is exponentially higher for the participant in stage 3 than for the participants in stage 1. In this treatment participants in stage 1 receive four times less than in treatment one, and participants in stage 3 receive four times more than in treatment 1. This equates to the participants of the first stage receiving 3.57% each of the total income, participants in stage two receiving 14.29% of the total income (same as in treatment 1), and the participant in stage three receives 57.14% of the total income, or 16 times more than the participants in stage one. The participants know the way the profit is split from the very beginning.

4.3 Treatment 3 – Ranked Allocation & Equal Pay

Treatment three differentiates itself from the previous two treatments by the way that now the participants are allocated to one of the 7 positions on the production line. In treatments one & two the allocation is done randomly by the z-tree program, which is not the case in this treatment. In this treatment the allocation depends on the performance of the "pre-experiment" round, which is now not only used as a practice round. This round is in all other ways exactly the same as it was in the first two treatments. After the "pre-experiment" round finishes the program ranks the 7 participants in terms of performance. It does this by awarding 1 point for each correctly solved 2x2 digit multiplication, 2 points for each correctly solved 3x3 digit multiplication, and 4 points for each correctly solved 4x4 digit multiplication. It also subtracts 0.1 point for each incorrectly solved 2x2 digit multiplication, 0.2 points for each incorrectly solved 3x3 digit multiplication, and 0.4 points for each incorrectly solved 4x4 digit multiplication. The participant with the most points is allocated to P7 in stage three of the production line, the 2nd and 3rd highest scoring participants are randomly placed into positions 5 & 6 in the second stage of the production line, and the 4 lowest scoring participants get randomly allocated to one of the four positions in stage one of the production line. All participants are made aware of this from the beginning of the experiment and are informed about their position as soon as the pre-experimental round finishes. Participants in this treatment split the final income equally, just as in treatment 1.

4.4 Treatment 4 – Ranked Allocation & Exponential Pay

Treatment four is a combination of the changes introduced in treatments 2 and 3. With everything else being the same as in all treatments, it has ranked allocation based on the pre-experimental round, just as in treatment 3, as well as the exponential pay used in treatment 2. Again, as in all treatments, participants know all of the information about the allocation as well as the income sharing from the very beginning of the experiment.

5. Results

We discuss the individual and group performances in this section of the paper. The section consists of three parts. In the first part (Section 6.1), we analyze the results of the pre-experiment for all 4 treatments. In the first part of section 6.2 we do the same for all stages of the experimental rounds, looking at the performances of the groups in each treatment. Lastly, towards the end of section 6.2, we look at the profits of the groups for each treatments and discuss the findings.

We aim to find whether and how an individual's effort and productivity, in a production line type of setting, are affected by different types of pay for their production. Knowing that multiplying is one of the real-effort tasks which all university students should be more than familiar with and it has been shown that it does not have a large learning effects, we look at how the performance changed when we gave them the opportunity to try to place themselves into one of the three positions in the production line according to merit.

5.1. Results – Pre-Experiment

Table 1 reports the mean number of correct multiplication and the mean number of total multiplications attempted by each player in the experiment for all four treatments. In the first two treatments the players are told that the pre-experimental round is a good opportunity to practice multiplications, for which they are given 6 minutes. In the 3rd and 4th treatments, the participants are told the same thing with the key difference that their performance during the pre-experimental round will be used as a basis for allocating each player to a position for the paying portion of the experiment.

As explained above, in this round each participant is first given a 2x2 digit multiplication to solve, followed by a 3x3 digit multiplication, then by a 4x4 digit multiplication, with the circle starting again with a new 2x2 digit multiplication and so on until the given time runs out.

Among the four treatments there are two very different payment schemes used, one pays equally to all 7 players in a group, while the other pays the player at level 3 sixteen times more than the four players at level one get paid. Additionally there is a very different connotation added to the pre-experimental round between the "unranked" and "ranked" treatments, with one being

solely a practice opportunity for the upcoming paying rounds, whereas in the other it is also a means of getting in to a better position and thus adding an additional way to help yourself and the group receive a better pay. Regardless of this, by looking at Table 1, we see that there is little difference between the pre-experimental results of the four treatments. Correct answers, total answers attempted in the six minutes, and percentage of correctly answered multiplications do not vary greatly between the four treatments.

That, however, is not to say that there are not any differences in the pre-experiment results between the treatments. As can be seen from table 1, treatment two has slightly lower number answers attempted for all types of multiplications, in comparison to all the other treatments. The result seems to show that participants seem disheartened by knowing that only one random member of the group is going to get paid substantially more than the rest, and decide to put less effort into the practice round. Even though they know that it might be them who is allocated randomly to the 3rd level on the production line, they also know that the probability of that happening is one in seven. This result is significant between treatment two and treatment three (p = 0.009), and more importantly between treatments two and four (p = 0.009), using a one-tailed using a 1-tailed Mann Whitney rank-sum test. The result is, however, just shy of significant between treatments one and two, which is understandable, as treatments three and four have an additional motivation for exerting higher effort.

Table 1: Pre-experiment: Average number of correct and attempted multiplications per player

	Equal Pay - Unranked			Exponential Pay - Unranked			Equal Pay - Ranked			Exponential Pay - Ranked		
	Correct	Attempted		Correct	Attempted	%	Correct	Attempted	%	Correct	Attempted	%
2x2	1.50	2.02	4.3%	1.39	1.84	6.5%	1.56	2.14	3.9%	1.49	2.11	1.6%
3x3	0.90	1.75	1.4%	0.84	1.59	3.8%	0.97	1.81	4.6%	0.96	1.79	4.6%
4x4	0.51	1.19	3.9%	0.45	1.09	1.3%	0.51	1.30	9.2%	0.66	1.35	9.9%

The only other statistically significant result comes from the correctly answered four digit multiplications, where as expected we find that treatment four has a higher average of four times four digit multiplications correctly answered. Using the 1-tailed Mann Whitney rank-sum test we find a significance of p=0.025 when comparing treatments two and four, but no full significance

when comparing treatments three and four (p=0.1). In treatment 4 participants have a higher motivation (incentive) to solve the pre-experiment multiplications correctly than in the other three treatments. One explanation of this result is that the 4x4 multiplications are the ones where the most effort is needed, and it could be that it is the only time when some participants lacked that extra incentive to but in that extra bit of effort. This explanation would go hand in hand with the general outcome of the pre-experiment results, in which, as mentioned above, we do not find large differences between the four treatments. It seems that the intrinsic motivation to practice in order to improve your performance in the paying rounds, is in most cases enough of a motivation.

What we found to be surprising here is the result of correctly and incorrectly solved 4x4 multiplications in treatment 3. As can be seen from Table 1, treatment 3 has the same number of average correct answers but also more attempted answers than the other equal-pay treatment (treatment 1). Due to the fact that in treatment 3 the participants are put into a position as a result of the ranking of the pre-experiment, we expected there would be an additional motivation more for participants to improve in comparison to treatment 1. The reasoning behind this was that if participants feel that they are good at doing multiplications they should put in more effort, thus be placed in the higher positions in the group and improve the profit and pay for everyone in the group. Even though the results don't disprove this theory, they seem to indicate that the opposite might have happened as well, namely that individuals who did not feel confident enough or did not want to feel responsible for the entire payment of the group might have purposefully not done as well as they could have in the 4x4 digit multiplications. As you can see from the table above participants in treatment 3 have the highest amount of correctly solved 2x2 and 3x3 digit multiplications yet they fall very short of the mark in the 4x4 digit multiplications. Additionally in Treatment 3 only 5% of the participants never reached at least one 4x4 digit multiplication, whereas the percentage for the other 3 treatments ranges from 11% (Treatment 4) to 19% (Treatments 1). These differences between T3 and T1 as well as T3 and T2 are statistically significant using a logistic Regression with categorical predictor variables, with the p value being 0.013 & 0.044 respectively. The difference between T3 and T4 are only weakly significant (p=0.096).

5.2. Results –Experimental Rounds

Table 2 shows the results of the average number of correct and attempted multiplications per group, per period. Looking first at the results of the 2x2 digit (first stage in the production line) multiplications, we see that there is a clear trend showing of a decrease of correct answers, increase of incorrect answers, and a decrease in overall 2x2 digit multiplications attempted. The difference between T1 and T2 is solely in the way that participants are paid, so the decrease in correctly solved 2x2 digit multiplications likely comes due to the fact that the participants in stage one now get paid 16 times less than the person in stage three of the production line. In treatment 3 we introduce the placement of participants into position based on the pre-experiment results and we see an even further decrease in the amount of correctly answered 2x2 digit multiplications and an increase in the incorrect answers. This happens even though the participants in the group will now all be paid equally, just as in T1. The reason for this is likely because now the best performers, those willing to exert the most amount of effort, are in stages 2 & 3, even though the pre-experimental results of T3 showed that some people are unwilling to be at the top in this type of payment scheme. Treatment 4, which is also ranked by the pre-experimental round and has a very unequal pay, as in T2, sees an even further decrease in the performance of participants in the first stage of the production line group. The decrease in correct answers from T3 to T4 is around the same as the decrease from T1 to T2, but T4 has an even higher increase in incorrect answers. From the results, it is safe to conclude that there is a small decrease in correctly solved 2x2 digit multiplications that comes from the participants who feel that they are not paid fairly even though in actual terms they work the most, and a small decrease that comes due to the ranking by the pre-experiment where the "best" participants are being placed at the higher stages of production.

It is important to emphasize here that these constant decreases in the performance are small, regardless whether they arise due to the difference in the payment schemes (Ex. T1 and T2) or because of the difference of allocation into position (Ex. T1 and T3). Though the differences between all treatments are close to significant, the only truly significant result comes from the difference in performance (2x2 digit) between treatment 1 and treatment 4 ($p=0.05$), which is the only combination of treatments that have both payment and allocative differences between them. The result was found using a linear regression with categorical predictor variables and clustered for firms. The fact that the differences that arise from the latter are small shouldn't be a surprise knowing

that the real effort task is doing multiplications, a task which participants at university level should master. Additionally, participants were given an example which showed them how to multiply, with the participants at this stage having the easiest task of all three stages. Ultimately, the difference in allocation has made a bigger difference in performance of participants in the first stage than the difference in equality, which is a surprising result. The fact that the large difference in the payment schemes has such a small effect, means that participants are unwilling to underperform when they are part of a group in which the pay for the entire group depends partly on themselves even if their own part of the payment is extremely small, in this case 3.57%.

Table 2: Average number of correct and attempted multiplications per group per period

	Equal Pay - Unranked			Exponential Pay - Unranked			Equal Pay - Ranked			Exponential Pay - Ranked		
	Correct	Attempted	%	Correct	Attempted	%	Correct	Attempted	%	Correct	Attempted	%
2x2	44.44	49.36	89.9%	42.61	47.67	87.9%	41.79	47.76	85.8%	38.22	45.41	82.8%
3x3	9.44	13.41	69.7%	9.18	12.15	75.3%	12	14.30	82.9%	10.69	13.38	79.4%
4x4	2.11	3.53	57.3%	2.42	3.72	63.4%	3.21	4.90	62.4%	3.28	4.75	70.6%

Moving on to the results of the participants of the second stage on the production line, who need to solve three by three digit multiplications, it is important to remember that participants at this stage are the only participants that get the same percentage of the profits regardless of the payment scheme (14.29%). Again, as in stage one, we see that in the treatments with random allocation the payment scheme does not seem to greatly affect the performance of the participants in this stage. Participants in T2 get an almost significant decrease of wrong multiplications in comparison to T1, but this is to some extent countered by the lower amount of correct answers, as it is the amount of correct answers that ultimately adds more possible production to stage 3 where the group's profit comes from. In fact both treatments with equal pay have more correctly solved 3x3 digit multiplications than their corresponding treatments with unequal pay, with the difference between the two treatments, which allocate positions by the pre-experiment, being even larger. This result also seems to be related to the result of the pre-experiment, as some players in T3 seemed to not be keen on ending up in the highest position, it is quite likely that at least in a few groups the person that should have ended up at the top, had they tried their best, actually ended up in stage two. Additionally, it is important to note that stage one of the treatments with equal pay performed better

than their corresponding treatments with unequal pay, which equates to the second stage of treatments 1&3 having a higher cap on the maximum amount of multiplications that could be solved.

Again, as with the results of stage 1, the considerable differences in results come when one compares the treatments with positions allocated by the pre-experiment to the treatments without allocation. Both treatments 3 & 4 have a higher amount of correctly solved multiplications, as well as fewer incorrectly solved ones, than treatments 1 & 2. Most importantly, the difference in correct 3x3 digit multiplications between T1 and T3 is significant (p=0.012) and between T2 and T4 is just shy of significant (p=0.077) using a Linear Regression with Categorical Predictor Variables and clustered by groups. The difference between incorrect 3x3 digit multiplications is also significant between T1 and T3 with a p value of 0.009.

Lastly we look at the last row of table 2, which shows the results of the third stage participants, and find it to be almost the exact inverse of the first row of the table (stage one). In this stage with each subsequent treatment the performance improves, again with the allocation of the positions making a bigger difference than the payment scheme. Differences in the average correct number of 4x4 digit multiplications are practically significant both between T1 & T3 (p=0.051) as well as between T2 & T4 (p=0.052).

Table 3: Percentage of 3rd level participants who ran out of multiplication before the period ended.

	Equal Pay - Unranked		Exponential Pay - Unranked		Equal Pay - Ranked		Exponential Pay - Ranked	
	Ran Out of Multiplications	Multiplications Remaining	Ran Out of Multiplications	Multiplications Remaining	Ran Out of Multiplications	Multiplications Remaining	Ran Out of Multiplications	Multiplications Remaining
%	55.56%	44.44%	57.78%	42.42%	63.64%	36.36%	83.78%	16.22%

Differences in the number of correct 4x4 digit multiplications between treatments that have the same allocation type, but different payment scheme, are not significant. However, in both cases the treatments with unequal pay had a lower number of multiplications to solve to begin with, yet they still outperformed the corresponding equal pay treatments. Table 3 shows that third level participants in T4 ran out possible multiplications to solve in 83% of the cases where as in T3 only 63% ran out of multiplications before the stage ended due to lack of time remaining. The difference in real performance between treatments 3 & 4 can also be clearly seen by the percentage of correctly answered 4x4 digit multiplications in Table 2.

Table 4: Average group profit for all 3 periods.

	Equal Pay - Unranked	Exponential Pay - Unranked	Equal Pay - Ranked	Exponential Pay - Ranked
Average Firm Profit	42.34	49.81	69.30	71.09

Table 4 shows the average group profit for all four treatments. As the profit is for the most part based on the results of stage 3, the profit differences strongly mimic those of the last row of Table 2. Exponential pay treatments have higher profits than their corresponding treatments with equal pay. Again though these differences are smaller and not fully significant. Allocating positions based on the pre-experiment makes a substantial difference in terms of profit, as can be seen from table 4, there are large differences between the profits of T1 and T3 as well as the profits of T2 and T4. These results are significant using a Linear Regression with Categorical Predictor Variables and clustered by groups with the p-values being 0.054, and 0.049, respectively. As with the rest of the results, these significances should be stronger with more treatments.

6. Conclusion

We set out to create a paradigm which is common in the private sector, both in an industrial setting and outside of it, in order to help study complex task and work-group functions in the lab. Our complex task is a combination of pooled and sequential task interdependence, which makes sure that every stage has to perform at a high level in order maximize output and therefore pay, yet has a cushioning in the form of small pooling so that a bad performance by one person at any given time does not mean immediate disaster and shut down.

We build on the small literature that has so far indicated that work-groups with very high task interdependence are less sensitive to different pay systems. Our results show that even extreme payment inequality does not significantly affect a work-groups performance in real effort tasks with high interdependence. With both types of allocation into position we use in our treatments, we find only small and insignificant changes to performance when comparing equal pay and the exponentially unequal pay treatments. We find this result even though the amount of pay for two of the three stages in the production line between the two treatments differ by 400%, and the

comparative difference in pay between the two stages (one and three) is 1600%. Additionally, in the two treatments with ranked allocation by the pre-experiment, even the small insignificant differences we see between the equal pay and unequal pay treatments to some extent comes from the different self-selection in position that occurs in the pre-experiment. Furthermore, we find different methods of allocating into positions can effect performances of certain stages of the production line, but not the final performance.

7. References:

Akerlof, G. A., and J. L. Yellen. 1988. Fairness and unemployment. *The American Economic Review* 78(2):44-49. E.g. Frank, 1984; Lazear, 1989

Allen, B. C., Sargent, L. D. y Bradley, L. M. (2003). Differential effects of task and reward interdependence on perceived helping behavior, effort, and group performance. *Small Group Research*, *34*, 716-740.

Anderson, S. E., &Williams, L. J. 1996. Interpersonal, job, and individual factors related to helping processes atwork. *Journal of Applied Psychology*, 81: 282-296.

Brass, D. J. 1981. Structural relationships, job characteristics, and worker satisfaction and performance. *Administrative Science Quarterly*, 22: 331-348.

Campion, M. A., Medsker, G. J., & Higgs, A. C. 1993. Relations between work group characteristics and effectiveness: Implications for designing effective work groups. *Personnel Psychology*, 46: 823-850.

Campion, M. A., Papper, E. M., & Medsker, G. J. (1996). Relations between work team characteristics and effectiveness: A replication and extension. *Personnel Psychology, 49,* 429–452.

Charness, G. (2004), "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics*, 22, 665-688.

Charness, Gary and Kuhn, Peter 2007. "Pay Inequality, Pay Secrecy, and Effort: Theory and Evidence." Journal of Labor Economics, Vol. 25, No. 4: pp. 693-723.

Daft, R. L. 1983. *Organization Theory and Design.* St. Paul, MN: West Publishing Company (pg. 430).

Dohmen, Thomas and Armin Falk. 2011. Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. American Economic Review, 101, 556-590. Roth (2001

Fehr, E. and A. Falk (1999), "Wage Rigidity in a Competitive Incomplete Market," *Journal of Political Economy*, 107(1): 106-134.

Fehr, E., E. Kirchler, A. Weichbold and S. Gächter (1998), "When Social Norms Overpower Competition: Gift Exchange in Experimental Labor Markets," *Journal of Labor Economics*, 16(2): 324-351.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993), "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, 108(2): 437-459.

Fischbacher, Urs. (2007) "Z-Tree: Zurich Toolbox for Ready-made Economic Experiments." *Experimental Economics* 10.2 171-78. Print.

Fisher, J. G. 1994. Technological Interdependence, Labor Production Functions, and Control Systems. *Accounting, Organizations and Society* 19: 493-505.

Guymon, R. N., R. Balakrishnan and R.M. Tubbs (2008). The Effect of Task Interdependence and Type of Incentive Contract on Group Performance. *Journal of Management Accounting Research* 20: 1-18.

Guzzo, R. A., &Shea, G. P. 1992. Group performance and intergroup relations in organizations. In M. D. Dunnette& L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 3): 269-313.Palo Alto, CA: Psychological Press.

Hertel, G., Konradt, U., and Orlikowski, B. (2004) Managing distance by interdependence: Goal setting, task interdependence, and team-based rewards in virtual teams, *European Journal of Work & Organizational Psychology*, 13, 1, 1-28.

Hirst, M. K., and Yetton, P. W. 1999. The effects of budget goals and task interdependence on the level of and variance in performance: a research note. *Accounting, Organizations and Society* 24: 205-216.

Kiggundu, M. N. 1983. Task interdependence and job design: Test of a theory. *Organizational Behavior & Human Decision Processes*, 31(2): 145-172.

Langfred, C. W. (2005). Autonomy and performance in teams: The multilevel moderating effect of task interdependence. *Journal of Management, 31*, 513-529.

Langfred, C.W., & Shanley, M. T. 2001. Small group research: Autonomous teams and progress in issues of context and levels of analysis. In R. Golembiewski (Ed.), *Handbook of organizational behavior* (2nd ed.): 81-111. New York: Marcel Dekker

Libby, Therresa, Steven Lee, and Linda Thorne. "Task Interdependence, Budget-Based Incentives, and Performance." Group and Team Interactive Processes.

Macintosh, N. B., and Daft, R. L. 1987. Management Control Systems and Departmental Interdependencies: An Empirical Study. *Accounting, Organizations and Society* 12: 49-61.

Olson, E. M., Walker, O. C., Ruekert, R. W., & Bonner, J. M. 2001. Patterns of cooperation during new product development among marketing, operations and R&D: Implications for project performance. *Journal of Product Innovation Management*, 18: 258-271.

Pearce, J. L. y Gregersen, H. B. (1991). Task interdependence and extra-role behavior: A test of the mediating effects of felt responsibility. *Journal of Applied Psychology*, *76*, 838-844.

Saavedra, R. P., Earley, P. C., & Van Dyne, L. 1993. Complex interdependence in task-performing groups. *Journal of Applied Psychology*, 78: 61-72.

Steiner, I. D. (1972). *Group processes and productivity.* New York: Academic Press.

Theresa L. and Linda T. (*2009*) The Influence of Incentive Structure on Group Performance in Assembly Lines and Teams. *Behavioral Research in Accounting*: Fall, Vol. 21, No. 2, pp. 57-72.

Van der Vegt, G. S. y van de Vliert, E. (2005). Effects of perceived skill dissimilarity and task interdependence on helping in work teams. *Journal of Management,* 31, 73-89.

Van der Vegt, G. S., Emans, B. y van der Vliert, E. (2005). The congruence hypothesis of intrateam interdependence in organizations. En L. L. Neider y Ch. A. Schriesheim (Eds.) *Understanding Teams*. Greenwich, CN: Information Age Publishing.

Van derVegt, G. S., Emans, B. J. M. &Van de Vliert, E. 2001. Patterns of interdependence in work teams: A two-level investigation of the relations with job and team satisfaction. *Personnel Psychology*, 54: 51-69.

Wageman, R. 1995. Interdependence and group effectiveness. *Administrative Science Quarterly*, 40: 145-180.

Wageman, R. B. y Baker, G. (1997). Incentives and cooperation: the joint effects of task and reward interdependence on group performance. *Journal of Organizational Behavior*, *18*, 139-158.
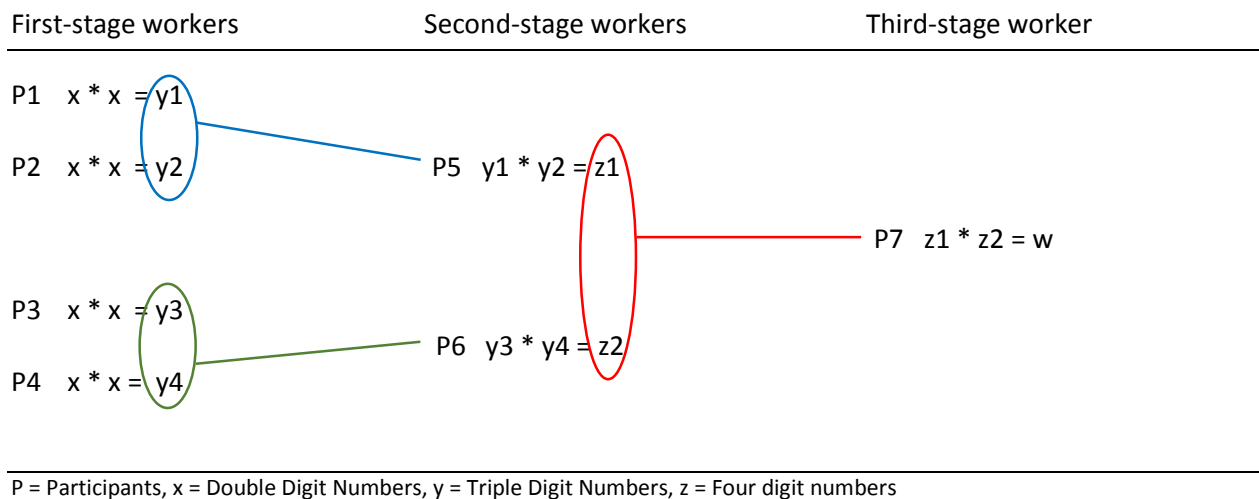
Appendix A. Instructions – Treatment 3

Welcome and thank you for your willingness to participate in this experiment. You will receive 4 euros for introducing you to the experiment. In addition, you can earn money during the experiment. From now on, contact is not allowed, in any way, with other participants in the room. If you have any doubts, please raise your hand and we will approach you. This is a sequential experiment so not everyone will act at the same time. While you're waiting your turn to participate you can do anything you want, without leaving your place. And, please, check your screen constantly in order to make sure when it is your turn to participate.

For the purpose of this experiment you are all workers in a production process of different companies that have no relationship between them. In every company there are 7 workers. You work for a company and you will be assigned to one of the 3 groups in the same company: 1st stage workers, 2nd stage workers, and 3rd stage workers. In each company, four of you will be workers of stage 1, two of you will be workers in stage 2, and one will be a worker at stage 3. You will be informed of your place in the company (on the display you will see your number and level) as soon as the experiment begins. No one will know the identities of other workers in the company.

Before the start of the experiment you will have 5 minutes to make multiplications of 2, 3 and 4 digit numbers. Your position within the company depends on the total number of correct and incorrect answers. The correct answers with multiplication of 4 digits (4 points) have more value than the 3-digit (2 points), and they have more value than the 2 digits multiplications (1 point). And the incorrect answers lead to points being subtracted: 0.1 points for an incorrect 2-digit multiplication, 0.2 points for an incorrect 3-digit multiplication, and 0.4 for an incorrect 4-digit multiplication. At the end of these 5 minutes, the person who gets the most points will be assigned to the 7 position, the next two will be assigned to positions 5 and 6 (not in order), and the rest will be assigned to positions 1 to 4 (not in order). You can see the different positions in the graph that below.

The experiment begins with the 4 workers in the first sage. They will have 4 minutes to individually solve multiplications of 2 digits (ex. 25 * 30 =?) that will appear on the screen. Each correct answer will be sent to another worker in stage 2. The correct answers of workers 1 and 2 will be pooled and sent to worker no. 5, and the correct answers of workers 3 and 4 will be pooled and sent to worker no. 6. Wrong answers will be lost, i.e. not be used later and will have a cost for the company. Remember that only the organizers of the experiment will know which position you occupy as a worker and in which company you are.

First-stage workers	Second-stage workers	Third-stage worker

P1 $x * x = y_1$

P2 $x * x = y_2$ P5 $y_1 * y_2 = z_1$

                                                          P7 $z_1 * z_2 = w$

P3 $x * x = y_3$

P4 $x * x = y_4$ P6 $y_3 * y_4 = z_2$

P = Participants, x = Double Digit Numbers, y = Triple Digit Numbers, z = Four digit numbers

Once workers of stage 1 are finished, stage 2 workers will have 6 minutes to solve as many 3-digit multiplications as possible available to them (Ex. 234 * 197 =?). The multiplications of stage 2 will be based on the numbers obtained in the previous step with the correct answers of the corresponding workers of stage 1. Workers 5 and 6 will solve these multiplications until the end of the 6 minutes or until they run out of available multiplications (obtained from the correct answers of the level 1). All the correct answers of workers 5 and 6 will be pooled together and will be sent to level 3. Again, the incorrect answers of participants 5 and 6 will be lost and not be used later,), and incur a cost to the company.

At stage 3 worker no. 7 has 8 minutes to solve multiplication of 4 digits (3422 * 7324 =?). These 4-digit multiplications shall be obtained from the correct multiplications of the

workers of stage 2. Stage 3 ends after the 8 minutes run out or when the worker finished the available operations (the correct answers of the level 2). Remember that to have multiplication available to solve in stage 3 it is required that there are correct answers solved in stages 1 and 2. Stage 3's correct answers generate the income of the company and the wrong answers of each level generate costs. An incorrect answer's cost will be higher the greater the level. The following equation specifies the relationship between revenues and benefits:

Function of total profits of the company.

Final profit = (number of correct answers * 8 euros)-[(number of wrong answers in the stage 1* 0.15 euros) + (number of wrong answers in stage 2 * 0.30 euros + (number of wrong answers in stage 3 * 0.60 euros))]

I.e. the full profit that the company receives will depend on the correct answers of worker 7 minus an amount per incorrect answer of each worker at every level. You notice that the incorrect answers at level 1 have a cost of 0.15 euros; an incorrect answer in the level 2 has a cost of 0.30 euros and an incorrect answer at 3 level costs 0.60 euros.

At the end the profit will be shared between all 7 workers equally in such a way that each one will receive a 14.29%. At this point the round will end.

There will be a total of 3 rounds. At the end of the experiment you will receive 4 euros plus what you have won in 3 rounds.

CHAPTER VI

Concluding Remarks

An organization's base can stay strong or crumble depending on its goals, values, vision, and philosophy. These factors have a large impact on the organizational culture and thus behavior. It is how employees judge the quality of the working environment which in turn hugely impacts their satisfaction, motivation, performance and development. Additionally, we know that the feelings and thus behavior of one individual within an organization can, direcetly or indirectly, impact the behavior of others in the same organization.

Therefore it is not surprising that literature over the years has become increasingly focused on finding out all and intricate details of how individuals affect and are affected by the organization, as well as how they affect eachother in a variery of organizational scettings and situations. The work in this thesis adds to the vast body of literature trying to disentangle and better understand behavior in organizations.

All the research seen in the thesis was conducted using experimental methods in a lab in the UAB's Business Economics Department using Fischbacher's Z-Tree program. We use a lab experiment as a research method because it allows us to control for multiple factors, which are nessesary for studying organizational and individual behavior, and in a way that otherwise would not have been possible.

To sum, my research has so far: 1) found that evaluators are biased based on previous decisions they have made regarding the same individual; 2) shown how employees are affected by these biases and how they affect the others as well as the organization; 3) looked at various ways to eliminate the bias, and managed to succeed without having to link the manager accuracy of the evaluations to his pay; 4) shown how individuals behave when faced with the dilemma of contributing to two different public goods with different efficiency; 5) shown how different levels of communication within an organization can help solve or hinder the previous dilemma; 6) created

a paradigm in which a unique yet common combination of sequential and pooled task interdependence can be studied naturally in the lab; 7) shown the effect of interdependence in organizational design on performance; 8) found the effect of payment inequality in an organizational design with high task interdependance.