



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

**Andrej Angelowski
Jordi Brandts
Carles Solà**

July 2015

Barcelona GSE Working Paper Series

Working Paper n° 839

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

ANDREJ ANGELOVSKI*, JORDI BRANDTS** & CARLES SOLA***

*LUISS Guido Carli, Department of Economics and Finance, Rome.

**Instituto de Analisis Economica (CSIC) and Barcelona GSE.

*** Universitat Autònoma de Barcelona - Department of Business.

July 10th, 2015

Abstract

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. We use laboratory experiments to study whether the way employees are assigned to a manager affects managers' and co-employees' subjective evaluations of employees. Employees can either be hired by the manager, explicitly not hired by him and nevertheless assigned to him or exogenously assigned to him. We present data from four different treatments. For all four treatments we find escalation bias by managers. Managers exhibit a positive bias towards those employees they have hired or a negative one towards those they have explicitly not hired. For three treatments we find that managers' and employees' biases are connected. Exogenously assigned employees are biased in favor of employees hired by the manager and against those explicitly not hired.

Keywords: Escalation Bias, Hiring, Performance Evaluations, Experiments

JEL codes: C92, D83, J63

Postal Address of the Corresponding Author: Campus UAB, 08193 Bellaterra (Barcelona), Spain

Corresponding Author: Jordi Brandts, jordi.brandts@iae.csic.es

The authors would like to thank Tomislav Rimac, Victor Martin, Jonathan Calleja, Marti Sagarra, Laura López Torres, María Teresa Bolívar and Darko Dimitrov for their help and feedback with this paper. We also thank participants in the ESA European Conference 2012 in Cologne, the AP ESA 2013 Tokyo and the SABE/IAREP/ICABEEP 2013 Conference in Atlanta for their comments and suggestions. The authors gratefully acknowledge financial support from the Antoni Serra Ramoneda (UAB – Catalunya Caixa) Research Chair, the Spanish Ministry of Economics and Competitiveness (Grant: ECO2014-59302-P) and the *Generalitat de Catalunya* (Grant: 2014 SGR 510).

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 evaluation of employees as a consequence of hiring decisions. More specifically, we study whether the way employees are assigned to a manager affects managers' and co-employees' subjective evaluations of employees. 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 decision that may be rational, follow it up with an irrational one in order to justify the initial decision and thus make themselves feel better about it. 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.¹ In the context we study, escalation bias consists in the tendency of managers who have made particular hiring decisions to subsequently evaluate employees who they hired in a biased manner

¹ Escalation bias bears some relation to another cognitive bias, confirmation bias, which is the tendency to search for, interpret, or recall information in a way that confirms one's beliefs or hypotheses. It is, indeed, related to escalation bias, since the escalation of commitment is (presumably) based on a biased reading of the case study responses of employees of different types. In our experiment we directly observe whether or not there are biases in evaluations, but not how these reactions come about.

compared to a professionally done evaluation. We will analyze the presence of both positive and negative escalation bias, where the first term refers to an excessively positive evaluation of hired employees and the second to an excessively negative evaluation of employees who were not hired.

We use a lab experiment as our research tool, because it allows us to control for certain factors in a way that otherwise would not have been possible.² 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.

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. The employees' task does not have a quantifiable correct answer and thus warrants a subjective evaluation. There isn't just one strict way to do things. 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. Deviations of participants' evaluation from the evaluations of the experts are the basis for our measurement of biases.

There are two early 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.

² Falk and Heckman (2009) contains an interesting discussion of the merits of lab experiments.

After this participants were given new fictitious data that suggested that the promoted manager was not performing well.

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 that go beyond the two seminal studies just discussed. First, in our experiment decisions will have payoff consequences for participants. Here we follow standard practice in experimental economics. Second, we present data from four 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.

Using four 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 consequences of behavior and of 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. In three of our treatments 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,

we conclude that manager and employee bias may be connected and employee evaluations cannot serve as a counter-balance to those of managers.

2. Background

In this section we discuss some selected contributions to a large literature. Our review does not pretend to be exhaustive. Rather, we point out some connections that we think are interesting.

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.

Carpenter et al. (2010) present an analysis of possible incentive effects of biased evaluations. They show that in a real effort experiment biased peer-to-peer evaluations can cause overall performance in a tournament to be lower than under piece rate. In their set-up participants evaluated others' output by counting it and, subjectively by evaluating the quality of that output. They find that the biased evaluation of others takes place mostly through its subjective part. The biased evaluation leads to employees reducing their effort.

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.³ 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.

³ The instructions can be found in Appendix A.

The experiment consists of four treatments, the *baseline* treatment, the *pay-for-performance* treatment, the *experience* treatment and the *pay-for-accuracy* treatment. In the baseline treatment the manager's evaluations will have payoff consequences for the employees but not for the manager. In the three 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.⁴

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 (PE)* 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. 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. 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.

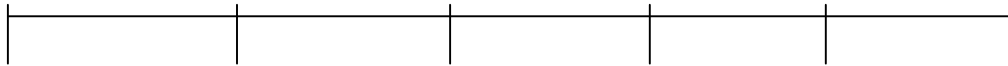
As a result of the process, each manager ends up with three employees: one who was been assigned to him from the start (*given employee, GE*), one who he recommended to hire and was hired (*recommended employee, RE*), and one whom he had the opportunity to hire but didn't hire (*non-recommended employee, 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

⁴ 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.

Initial assignment	Personality Questionnaire	Hiring Process	Task	Evaluation	Compensation
Manager		RE added		Manager	Manager
GE		NRE randomly added		All Employees	All Employees
PE				External	(ev. contingent)

will be discussed in more detail in the following sections. Participants in the role of manager receive a 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. Figure 1 shows the timeline valid for all treatments. 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.⁵

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

⁵ Managers and potential employees knew the content of the questionnaire; given employees did not.

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.

With this design it is possible for both RE's & NRE's to be hired into two different firms. The participants which are hired into two or more firms are randomly allocated to a "primary" firm. They only grade other employees in this firm, and get paid according to their grade in the same firm (as you will see below). These participants are not informed that they have been selected by two or more managers, as this is solely a procedural design decision of the experiment, for the reasons explained above, which has no impact on the research questions at hand. No participant asked us about this issue.

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 privately 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 relative 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 by Managers and Employees

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 0-100. 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 0-100. 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. The evaluations made by employees had no payoff consequences for managers or employees.

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 three of the treatments. They were PhD students in business economics, who do research in and teach courses in business economics, management and related subjects at the Universitat Autònoma 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 0-100. 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.⁶ For the fourth treatment (conducted after the three first ones) one of the evaluators was the same as before and the other two were assistant professors in business.

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

⁶ The external evaluators first evaluated the given employees. Then they evaluated the remaining employees in an effectively random order, which corresponded to the random seating order in the lab.

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 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. We feel that the treatment captures, in a stylized one-shot setting, a phenomenon which in natural environments would take place over time. Each manager will receive 1/3 of the average grade of his employees evaluated by the three external employees. Hence, managers now directly benefit from employees who perform well, as judged by the external evaluators.

In the pay-for-performance treatment the three external evaluators needed to evaluate the employees "on the spot" at the same time that the employees were being evaluated by the managers and by their peers, since their evaluations determine the earnings of the participants in the role of managers. To make this possible we seated the external evaluators in a room adjacent to the lab, but the participants were not aware of this.⁷

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

⁷ In this treatment and in the pay-for-accuracy treatment (see section 3.4) the external evaluators were in a room adjacent to the lab. They were connected to the same network as the participants and received the participants' answers to the case study right away and could grade them at the same time as the managers. In the other two treatments the externals evaluated the case studies after the sessions. We are sure that participants were not aware of this and that it had not effect on the results.

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.

3.4. The Pay-for-Accuracy Treatment

The main difference of this treatment with respect to the baseline is that managers don't obtain a fixed payment. A manager's payment now depends on how the evaluation of an employee by the manager deviates from that of the average of the external evaluators. The rationale behind this treatment is that at some point higher management will find out whether a manager is able to accurately evaluate employees and will be rewarded for it. The payoff function that we use is the following:

Manager Pay = 30 - |Manager's grade of his employee – External's grade of that same employee|(Averaged for all 3 employees).

We chose the value of the first term to be 30, using the data from the first three treatments, so as to obtain average manager earnings of 12, as in the other three treatments.

4. Hypotheses

In this section we propose null and alternative hypotheses both for manager and employee escalation biases. In both cases, the hypotheses we formulate apply to all treatments. However, the rationales behind the hypotheses differ between treatments. The alternative hypotheses we present are not based on any formal model of behaviour. Rather, they encapsulate behavioural biases that have been observed in previous work.

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: Managers evaluate their employees in an unbiased way, independently of how employees were hired into the company.

This hypothesis posits what can be considered perfectly rational behavior in our context. However, previous work has shown that people's decision making process is potentially influenced by their previous decisions. Whenever a manager hires an applicant, he must have certain reasons and criteria by which he has made that decision. This reasoning often remains as an anchor in people's minds, as at least a moderate amount of effort and time was spent into this decision process.

The main reason why we believe that a manager might have a positive bias towards the employee 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 an unfavorable course of action is helped by the need to preserve one-self. Managers are thus likely to distort unfavorable information through 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 proposes that the performance appraising manager is going to show a bias towards his employees.⁸ 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.

⁸ 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.

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. This is a common real-life scenario in which a manager's pay is dependent on his branches performance or on them meeting of a certain goal (Ex. Sales). We conjectured that this would be a factor working against a bias. A manager, who can affect his subordinates pay, knowing that their bad performance is going to cost him money is less likely to reward any of his employees with more pay than they actually deserve. Prendergast & Topel (1993) state that favoritism is stronger when the supervisor is not responsible for the performance of the 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.

In the experience treatment we can study an additional issue. The literature on effort tells us that people value their effort higher than others' efforts, so having put in 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. In particular, 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 were not much effort is put in.

In our framework we can analyse whether the grade that managers obtain in the task has a significant effect on the way they grade others. Ideally, own performance should have no effect on 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. Understanding the task better should help in making better evaluation decisions, and having put more of his own effort into a task likely means that managers could find it harder to grade anyone higher than what they actually deserve. This is captured in the following null hypothesis:

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

A few studies suggest that that the opposite might happen, namely that better performers may not have an advantage when dealing with biases. 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.). Thus, we formulate the following alternative:

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

Finally, in the pay-for-accuracy treatment the manager's payment depends on how the evaluation of an employee by the manager deviates from that of the average of the external evaluators. Our expectation was that this treatment, similarly to the two previous ones, would give a better chance to the null with respect to the biases presented in the alternative hypotheses above.

4.2. Employee Bias

Gómez-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, which pertains to all four treatments:

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.⁹

⁹ 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.

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, which pertain to all four treatments:

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.

5. Procedures

We conducted six sessions with each of the four 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 Autònoma 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 twenty-four sessions there are sixteen participants. We therefore have 384 participants, of which 120 are managers, 120 are given employees, and 144 are potential employees. Of the participants 28% are business or economics students, 16% study natural sciences, 11% study psychology, sociology and anthropology, 8% major in engineering, maths, or computer sciences, 6% major in medicinal studies, and the rest are split between other various disciplines.

6. Results

We start with a brief analysis of the characteristics of recommended employees, in terms of their responses to the questionnaire. This is done for all treatments together, since the first stage of the experiment is the same for all treatments. After that we discuss the results of manager and employee evaluations treatment by treatment.

6.1. Who Gets Recommended

Table 1 shows the results of two ordered probit regressions. In both regressions the dependent variable “Recommendedby2+” corresponds to employees who have been recommended by two or more managers. In specification 1 the exogenous variables correspond to the ten questions of the personality questionnaire. We show the results of this specification, but the more meaningful results correspond to specification 2. This specification uses as exogenous variables the big five categories which can be computed from the responses to the ten original questions. Here one can see that participants in the role of the manager recommend employees in a meaningful way. Agreeableness and conscientiousness have a strongly significant positive effect, while neuroticism and, perhaps more surprisingly, extraversion have significant negative impacts.

[Table 1 about here]

Table 2 shows the results of OLS regressions. In both regressions the dependent variable “Performance” corresponds to employees’ score (between 0 and 100). The independent variables

are the same as in Table A. In the regression using the original categories of the questionnaire there are some significant effects, but not in the regression using the big five.

[Table 2 about here]

Our summary of this evidence is that managers seem to make their recommendation choices in a thoughtful manner, but that the responses to the questionnaire do not have an impact of employees' performance.¹⁰

6.2. Results of the Baseline Treatment

Table 3 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.¹¹

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. All mentioned results so far are statistically significant, but also irrelevant as these results are not standardized. They do, however, help in understanding what is going behind the standardized results we are going to present.

We now move to the statistical tests for escalation bias. Given the features of our design we need to standardize the evaluation grades. The way we standardize is by using proportions of grades with respect to an appropriate aggregate.¹² For manager evaluations we express their grade of

¹⁰ To avoid an additional effects of liking we did not ask managers to fill out the personality questionnaire.

¹¹ Appendix D gives more information on all the evaluations of the externals.

¹² We thank an anonymous reviewer for suggesting this way of standardizing.

specific employees as a proportion of the sum of the grades for all three employees. This standardization gives rise to two new variables, namely “ $RE/(RE+NRE+GE)$ ” and “ $NRE/(RE+NRE+GE)$ ” The first variable corresponds to the proportion of the manager evaluation of the recommended employee with respect to that of the sum of all employees. The second variable similarly corresponds to the proportion of the evaluation of the non - recommended employee with respect to the sum of the evaluations of all three employees. Given the directional character of the alternative hypotheses H1a and H1b, we will use one-sided tests to evaluate them.

For the peer-to-peer evaluations, we base our statistical analysis on three proportions. The first pertains to the grading by the recommended employee of the non-recommended employee and is given by $NRE/ (NRE+GE)$, where the grade is now standardized by the sum of the grades of the two other employees. Similarly, we will study the grades given by the non-recommended employee to the recommended employee by using the proportion $RE/ (RE+GE)$ and the grade given by the given employee to the recommended employee by using $RE/ (RE+NRE)$. Given the more exploratory character of the hypotheses pertaining to employee behavior we will use two-sided tests in this case.

Table 3. Mean Grades (Baseline Treatment)						
	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NRE Evaluates	Mean
GE	59.88	64.97		59.21	52.18	59.06
RE	55.96	72.00	67.10		55.06	62.53
NRE	56.19	62.29	53.11	62.00		58.40
Mean Grade	57.34	66.42	60.11	60.61	53.62	59.62

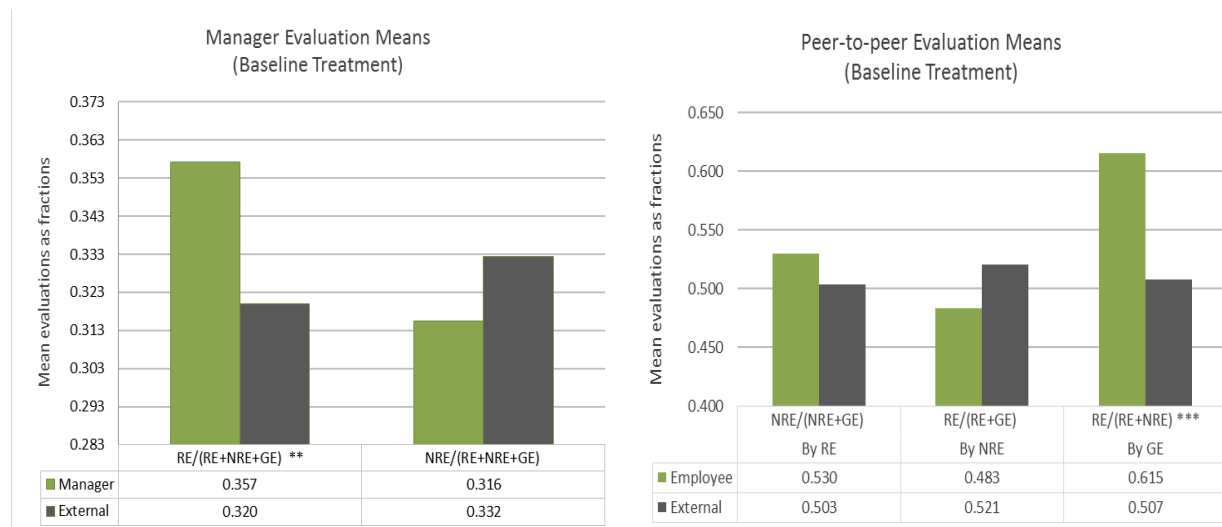
Figures 2 and 3 show average proportions in the standardized evaluations for managers and employees respectively, together with the corresponding comparisons with the evaluations of the external evaluators, where the stars next to the different proportions express the level of significance.¹³ Starting with Figure 2, on average managers grade recommended employees relatively higher than 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.

¹³ Appendix E shows figures corresponding to an alternative method of standardization based on differences.

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 proportions of the grades of managers and externals with a $p=0.017$. 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 3 that on average managers did not evaluate the given employees lower than the external evaluators, so that the higher difference in the proportions comes from the higher grade given to the recommended employee and the lower 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. Managers evaluate non-recommended employees lower than the externals. However, the one-tailed Wilcoxon signed-rank test for our second hypothesis finds a $p=0.1706$. We can not reject the null hypothesis in favor of H1b, there is no significant negative bias of evaluating managers towards employees explicitly not hired by managers but nevertheless assigned to them.

Figures 2&3



In Figure 3 one can see the information pertaining to the relevant proportions for the peer-to-peer evaluations where in parentheses we denote which type of employee has performed the

evaluations in each case. These are again compared to the external evaluators' scores of the identical pairs that each employee evaluated. 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.

For the differences in proportions we ran two-tailed Wilcoxon signed-rank tests. All three types of employees grade the other two employees in their branch, all of which may be affected differently by the manager's decision in the hiring stage. For this reason we believe that a two-tailed test is the most appropriate for all peer-to-peer evaluations since a priori a possible bias can go in either direction. We found that given employees evaluated recommended employees relatively higher, a result which is statistically significant ($p=0,0015$). 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 we don't find significant differences with respect to outside evaluators ($p=0.7020$).¹⁴ Recommended Employees have graded their two co-workers in an unbiased way. We next look at the evaluations done by the NRE. Although the pattern of differences is the same as for the manager, the Wilcoxon test shows that the result is not significant though ($p=0,2342$). This could be due to the tendency towards conformity being compensated by a pull in the other direction, because of the non-recommended employees' dissatisfaction with not having being selected.

Therefore it is safe to conclude that employees hired by the manager are not affected by this hiring decision when evaluating their peers.

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 biases 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.

¹⁴ The number of observations is different between managers and given employees on one side and the other types of employees on the other side.

6.3. Results of the Pay-for-Performance Treatment

Table 4 shows the mean grades of the second treatment. Compared to Table 3, 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 also important to note that their actual performance (as proxied by the external evaluations) has in fact increased. In Table 4 we also see for the first and only time manager evaluations be lower than the real performance, when looking at the non-recommended employees.

What seems to remain constant between Table 3 and Table 4 (the baseline and the pay-for-performance treatments) 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 4. Mean Grades (Pay-for-Performance Treatment)						
	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NRE Evaluates	Mean
GE	60.33	69.20		61.95	62.00	63.37
RE	62.06	67.30	65.43		68.31	65.78
NRE	60.00	57.83	54.43	54.65		56.73
Mean Grade	60.80	64.78	59.93	58.30	65.16	61.79

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 significant difference in the $RE/(RE+NRE+GE)$ 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 $p=0.3071$.¹⁵ We can therefore not reject the null hypothesis of no bias towards the recommended employee in this case.

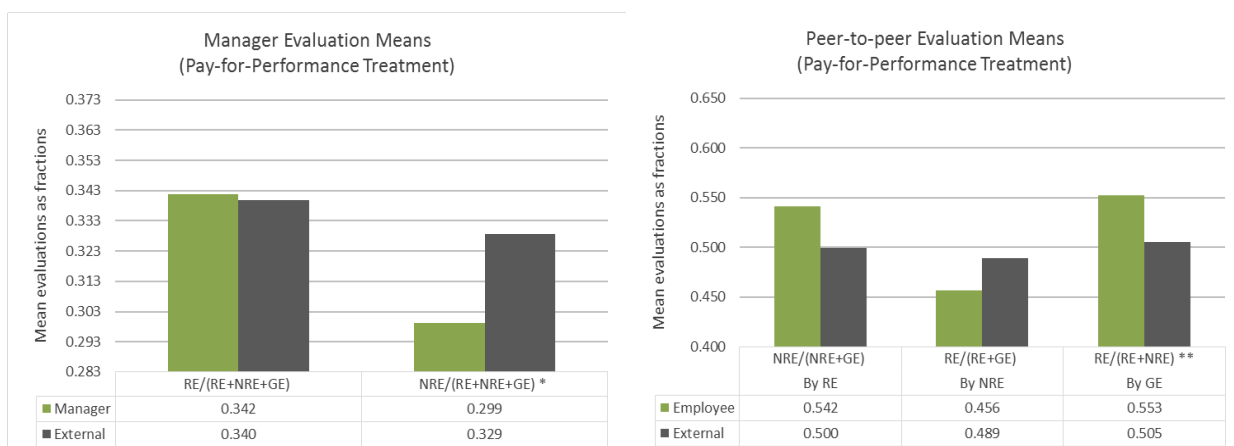
Figure 4 shows that on average managers graded relatively lower the NRE lower than the GE, compared with the externals. A one-tailed Wilcoxon signed-rank tests finds $p=0.0639$. Hence, in this case we can reject the null in favor of H_{1b} . The incentive change inherent in the pay-for-performance treatment leads to a displacement of the managers' biases.

¹⁵ 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.

We now move to the peer-to-peer evaluations. The pay-for-performance treatment does not introduce any change of employees' incentives, which 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 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 $p=0.1214$ 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 $p=0.5520$.

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 $p=0.0385$.

6.4. Results of the Experience Treatment

Table 5 shows the mean grades of the experience treatment. As in the two previous 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.¹⁶ 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 peer-to-peer 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.

	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NRE Evaluates	Mean
GE	60.00	71.83		65.31	60.20	64.34
RE	59.28	65.57	63.27		57.45	61.39
NRE	56.39	59.57	52.53	53.75		55.56
Mean Grade	58.56	65.66	57.90	59.53	58.83	60.09

For managers' evaluations, the statistical results are qualitatively the same as for the previous treatment. The first two bars of Figure 6 show the managers' and the external evaluator's mean evaluations of the RE in relation to the sum of evaluations. We find no significant difference between managers' and external evaluators' evaluations; a one-tailed Wilcoxon signed-rank test finds $p=0.4346$. For the 30 managers of this treatment, fourteen managers scored a higher

¹⁶ Managers' average grade in their task was 56.17, which was slightly lower than the mean grade.

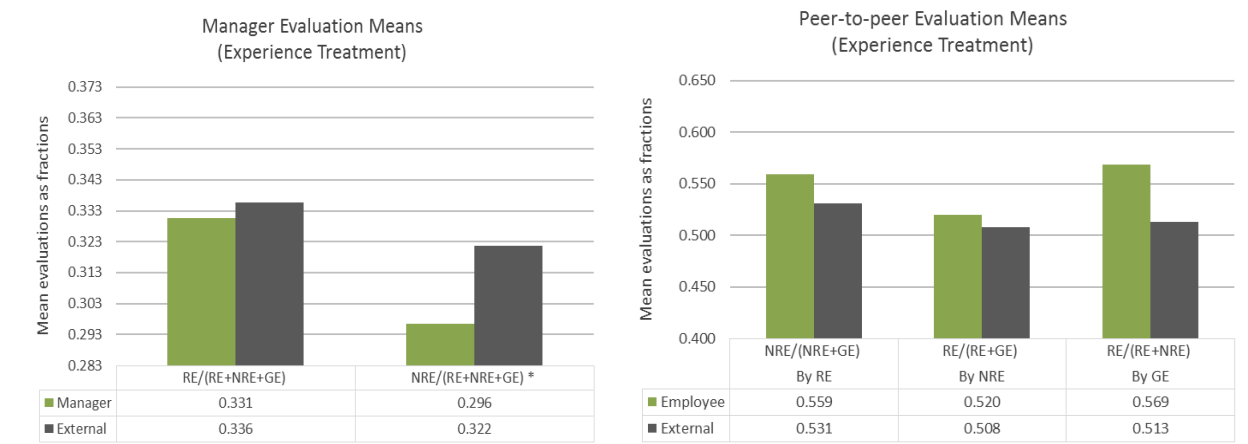
proportion in favor of the RE than the externals and fourteen scored a lower proportion than the externals, and there were 2 ties. In summary, when managers perform the task beforehand they lose the positive bias towards the RE.

The last two bars of Figure 6 show the average proportions for the evaluations of the NRE. As for the second treatment we find that the positive bias towards the RE observed in the baseline treatment disappears and a negative bias towards the NRE develops. Again the result is significant; the one-tailed Wilcoxon signed-rank test finds a weak statistical significance of $p=0.0749$. 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 the sum of the RE and the NRE biases on the managers' grade in the case study finds a positive coefficient of 0.003 ($p=0.005$).¹⁷ Hence, we reject the null in favor of H_{2a} .

Figures 6&7



¹⁷ More specifically we regress $|\text{[RE/(RE+NRE+GE) by Man.]} - \text{[RE/(RE+NRE+GE) by Ext.]}| + |\text{[NRE/(RE+NRE+GE) by Man.]} - \text{[NRE/(RE+NRE+GE) by Ext.]}|$ on the manager's grade for the case study. We have 30 observations and the R-squared = 0.244.

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 $p=0.011$. 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. This will change in the fourth treatment to which we move now.

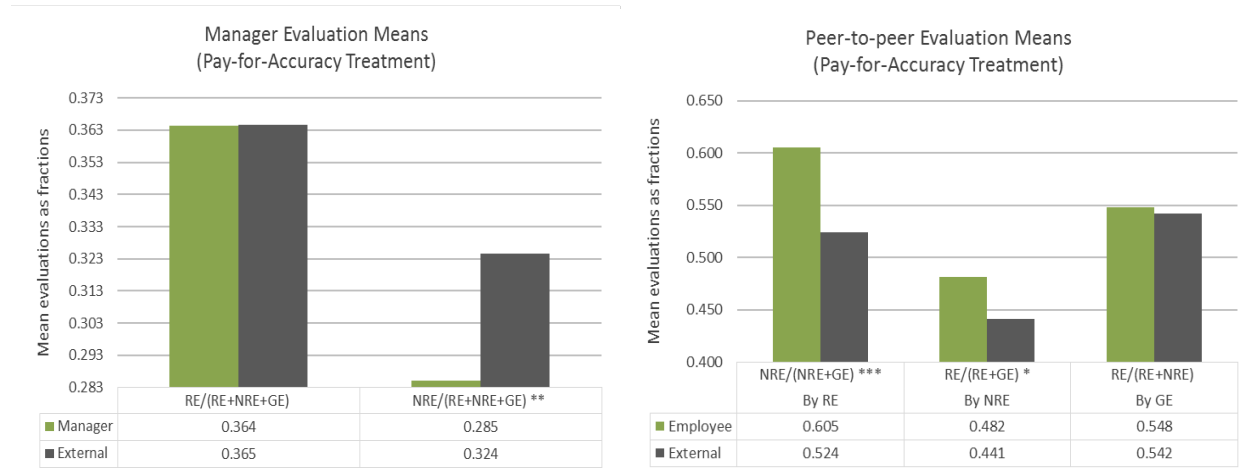
6.5. Results of the Pay-for-Accuracy Treatment

Table 6 shows average grades for the pay-for-accuracy treatment and Figures 8 and 9 show the average proportions that we use in our statistical tests.

Table 6. Mean Grades (Pay-for-Accuracy Treatment)						
	External Evaluates	Manager Evaluates	GE Evaluates	RE Evaluates	NRE Evaluates	Mean
GE	54.93	65.93		75.65	63.74	65.06
RE	63.00	70.33	67.20		65.63	66.54
NRE	54.09	56.27	57.47	51.94		54.94
Mean Grade	57.34	64.18	62.33	63.79	64.68	62.47

As for treatments two and three, two-tailed Wilcoxon rank-sum tests find that differences between managers' and externals' evaluations are not significant for $RE/(RE+NRE+GE)$, with a $p=0.3365$, while they are significant for $NRE/(RE+NRE+GE)$, $p=0.010$. For peer-to-peer evaluations we do not find results consistent with those of the previous three treatments. We find that the RE now grades the NRE significantly higher than the GE, with $p=0.0012$. The NRE also grades the RE significantly higher than the GE, although with $p=0.0765$. Finally, the GE does not grade the RE higher than the RE. We do not have a good explanation for why the pattern of the peer-to-peer evaluations is different than for the other three cases. Employees may realize that managers now have clear incentives to not be biased and that leads to the elimination of the influence of managers' evaluations on those of the employees.

Figures 8&9



7. Conclusions

We set out to produce a detailed experimental study of hiring and escalation bias in subjective performance evaluations. More broadly, our aim was 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 four 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 consequences and of experience of managers with the same task that employees have to perform does not eliminate managers' bias. What we observe is that the bias moves from being a positive bias towards recommended employees to a negative one for non-recommended employees.

Our results in one of the treatments also show that 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 at the same time feel more confident with their initial hiring decision. As mentioned above, this evidence is in accordance with some recent evidence pertaining to biases of development professionals reported in the recent World Development Report (2015) of the World Bank.

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. In three of the treatments, 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. This suggests that manager and employee bias are connected and employee evaluations cannot serve as a counter-balance to those of managers. However, in the fourth of our treatments we find a different pattern for the peer-to-peer evaluations, which is not easy to explain.

The implications of the findings go beyond a potential bias in performance evaluations. They show just how easily previous decisions can contribute to making us make irrational decisions subsequently. In addition, we find that the actions of others, at least of those perceived to have higher positions, can influence people around them to also be biased, without there even being direct contact between them.

Even though this paper covers several scenarios in which escalation bias might have an effect on performance evaluations, we believe that it is only the first step in order to understand the full effect that escalation bias has on the performance and dynamics of a work-group. With this aim in mind we propose that further research is done with a focus on three main issues: 1) testing escalation-bias's strength over time in a repeated game scenario; 2) testing employees' behaviour

once they have been the subject of a positive or negative bias by an evaluating manager; and 3) studying the effects of different 360 degrees evaluation schemes involving incentives.

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* **86**(5), 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**, 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.
- Carpenter, J., P.H. Matthews and J. Schirm (2010) "Tournaments and office politics: evidence from a real effort experiment", *American Economic Review* **100**(1), 504-517.
- 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. Fourth 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" *Journal of Behavioral Decision Making* **26**(1), 79-90.
- Fischbacher, U. (2007) "Z-tree: zurich toolbox for ready-made economic experiments" *Experimental Economics* **10**(2), 171-78.
- 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.
- Judge, T. A., and G. R. Ferris (1993) "Social context of performance evaluation decisions" *Academy of Management Journal* **36**(1), 80-105.
- Klimoski, R. and L. Inks (1990) "Accountability forces in performance appraisal" *Organizational Behavior & Human Decision Processes* **45**(2), 194-206.
- 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), 958-998.
- 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.
- 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 sophistication does not attenuate the bias blind spot" *Journal of Personality and Social Psychology*, 1-14.
- World Bank (2015), "World Development Report 2015: Mind Society and Behavior."

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

Table 1: Ordered Probit regressions relating the number of recommendations to the responses in the personality questionnaire.

(1) <i>Independent variables</i>	(1) <i>Recommby2+</i>	(2) <i>Independent variables</i>	(2) <i>Recommby2+</i>
<i>Reserved</i>	.176 (.117)	<i>Extraversion</i>	-.174** (.071)
<i>Trusting</i>	.210* (.118)	<i>Agreeableness</i>	.245*** (.078)
<i>Lazy</i>	-.210 (.159)	<i>Conscientiousness</i>	.277*** (.088)
<i>Relaxed</i>	.640*** (.160)	<i>Neuroticism</i>	-.230*** (.065)
<i>Artistic</i>	-.105 (.087)	<i>Openness</i>	.101 (.068)
<i>Outgoing</i>	-.126 (.134)	-	-
<i>Finds faults</i>	-.262** (.106)	-	-
<i>Thorough</i>	.347** (.158)	-	-
<i>Nervous</i>	.073 (.136)	-	-
<i>Imaginative</i>	.218 (.151)	-	-
<i>N</i>	144	-	144
<i>R-square</i>	0.200	-	0.239

Table 2: OLS regressions relating the answers in the questionnaire to performance.

(1) <i>Independent variables</i>	(1) <i>Performance</i>	(2) <i>Independent variables</i>	(2) <i>Performance</i>
<i>Reserved</i>	-.065 (1.014)	<i>Extraversion</i>	.546 (.638)
<i>Trusting</i>	.769 (1.026)	<i>Agreeableness</i>	.287 (.699)
<i>Lazy</i>	1.840 (1.290)	<i>Conscientiousness</i>	.911 (.779)
<i>Relaxed</i>	1.105 (1.173)	<i>Neuroticism</i>	-.373 (.569)
<i>Artistic</i>	-.903 (.774)	<i>Openness</i>	.039 (.636)
<i>Outgoing</i>	2.329** (1.126)	-	-
<i>Finds faults</i>	-.496 (.891)	-	-
<i>Thorough</i>	4.209*** (1.307)	-	-
<i>Nervous</i>	.038 (1.119)	-	-
<i>Imaginative</i>	-2.266* (1.290)	-	-
<i>Constant</i>	33.395*** (11.416)		46.641*** (10.080)
<i>N</i>	144	-	144
<i>R-square</i>	0.122	-	0.031

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 third 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 36 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 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 RECOMMENDED EMPLOYEES

Even though you were not 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 36 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 Almacera . 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

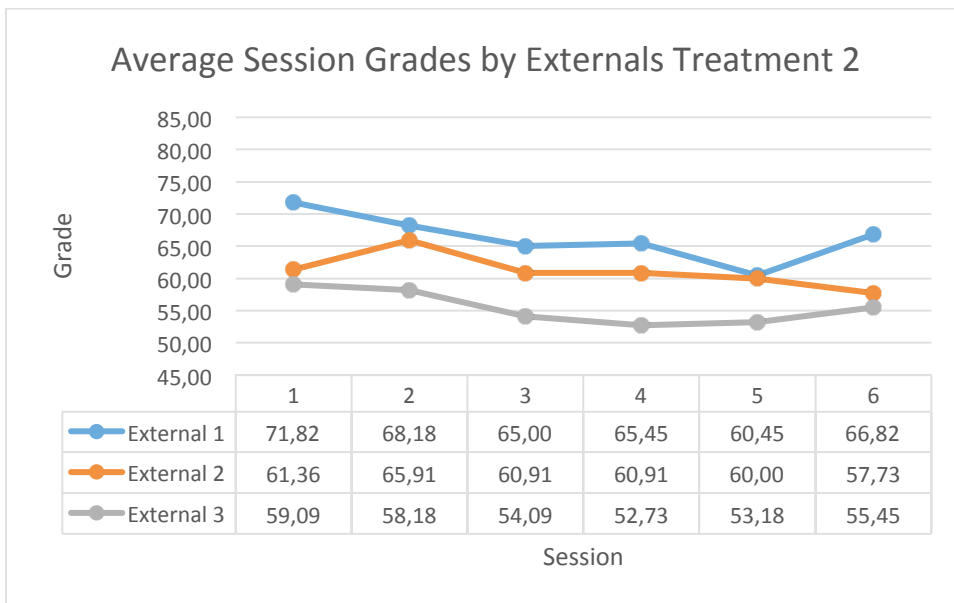
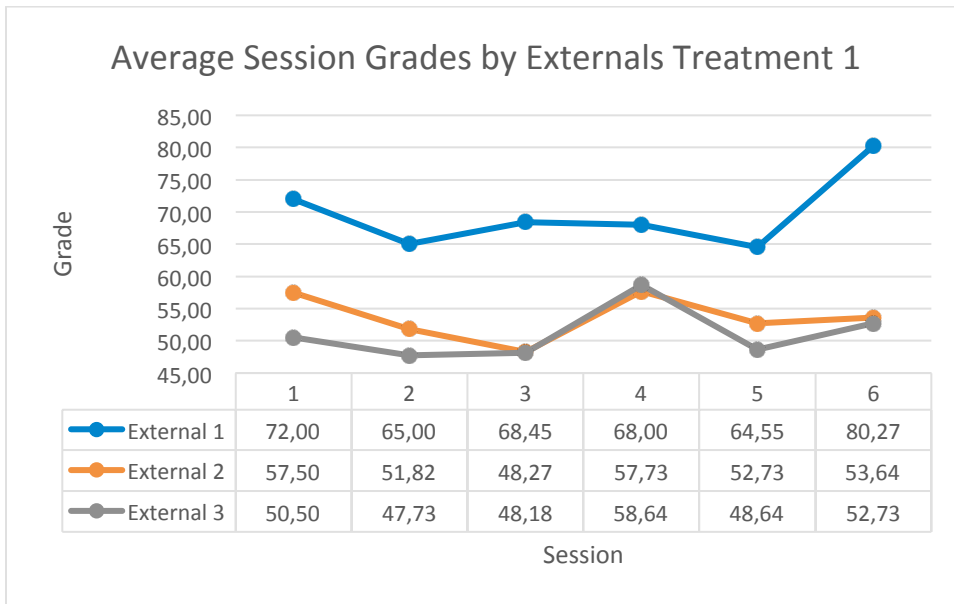
Lladró, allowing them to access exclusive collections and having privileged shopping options of limited edition products.

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.

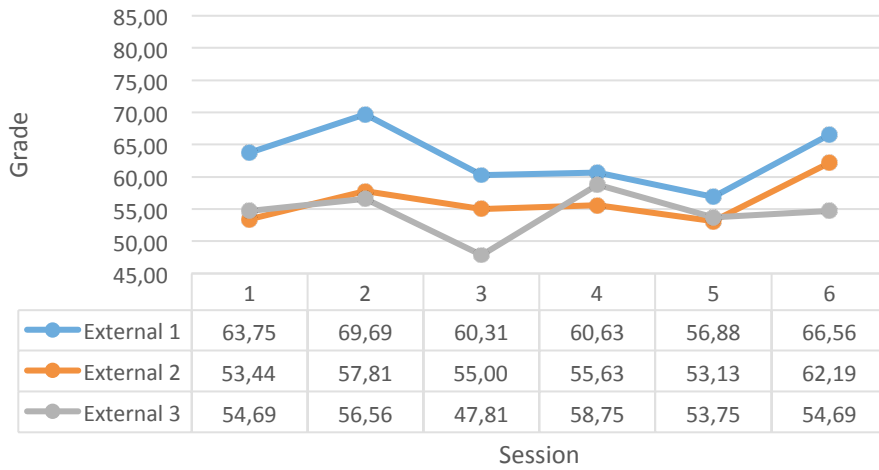
QUESTIONS:

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?

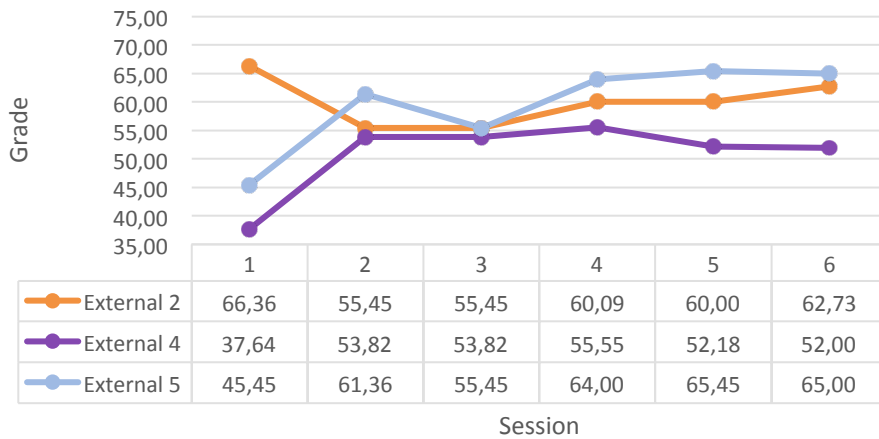
Appendix D. Average Evaluations by External Evaluator and Session.



Average Session Grades by Externals Treatment 3

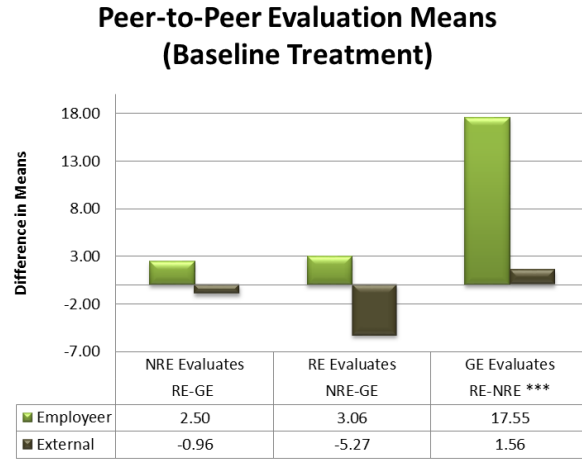
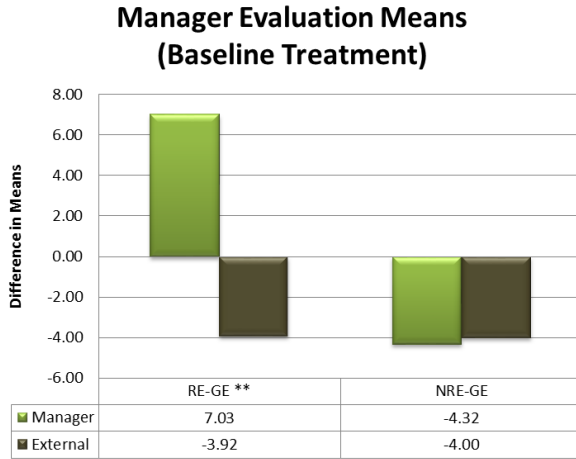


Average Session Grades by Externals Treatment 4

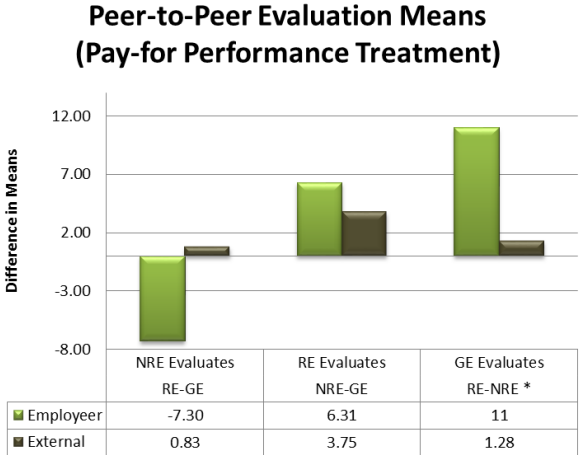
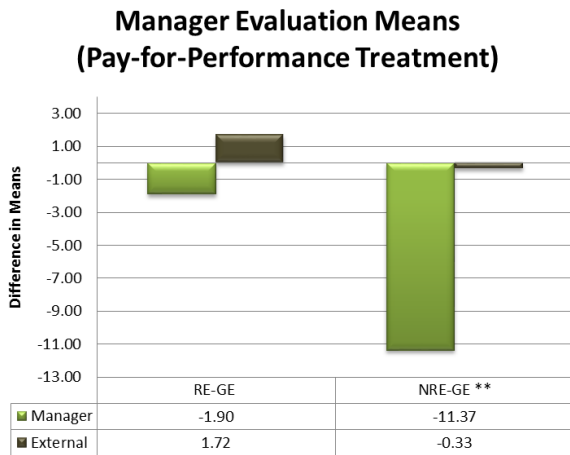


Appendix E

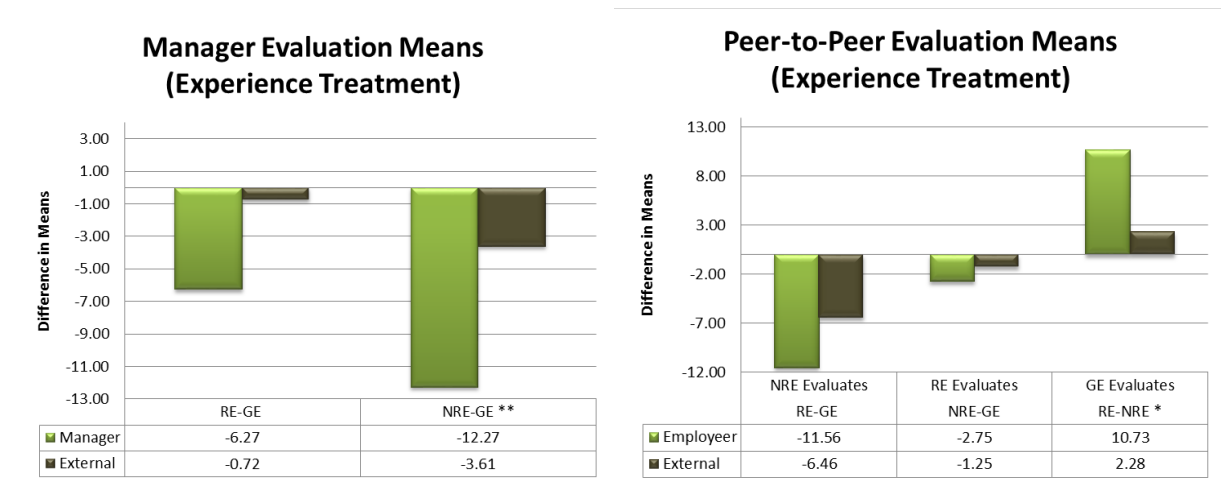
Figures F1 & F2



Figures F3 & F4



Figures F5 & F6



Figures F7 & F8

