Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen - Bolzano

Fakultät für Wirtschaftswissenschaften
Facoltà di Economia
School of Economics and Management

BEMPS –
Bozen Economics & Management
Paper Series

NO 01 / 2013

Severity vs. Leniency Bias In Performance Appraisal: Experimental Evidence

Lucia Marchegiani, Tommaso Reggiani, Matteo Rizzolli

Severity vs. Leniency Bias in Performance Appraisal: Experimental evidence[*]

Lucia Marchegiani[†] Tommaso Reggiani[‡] Matteo Rizzolli[§]

February 5, 2013

Abstract

Performance appraisal can be biased in two main ways: *lenient* supervisors assign predominantly high evaluations (thus rewarding also undeserving agents who have exerted no effort) while *severe* supervisors assign predominantly low evaluations (thus failing to reward deserving agents who have exerted effort). The principal-agent model with moral hazard predicts that both biases will be equally detrimental to effort provision. We test this prediction with a laboratory experiment and we show that failing to reward deserving agents is significantly more detrimental than rewarding undeserving agents. This finding is compatible with empirical evidence on real world supervisors being preponderantly biased towards lenient appraisals. We discuss our result in the light of alternative economic theories of behavior. Our result brings interesting implications for strategic human resource management and personnel economics and contributes to the debate about incentives and organizational performance.

Keywords: Agency theory, Performance appraisal, Type I and Type II errors, Leniency bias, Severity bias, Economic experiment. **JEL code**: C91, M50, J50.

[*] We thank Gani Aldashev, Steffen Altmann, Benito Arrunada, Jeff Butler, Susan Cartwright, Gary Charness, Anna Grandori, Fabrizio Galeotti, Andrea Ichino, Michael Neugart, Matteo Ploner, Dirk Sliwka and Giorgio Zanarone for helpful comments. For the same reason we also thank participants of seminars at CUNEF (Madrid), at the University of Bergen, at the Free University of Bozen-Bolzano, at the University of Cologne and conference participants of EURAM2011 (Tallinn), Young Economists Meeting (Bologna), ESA 2011 (Chicago), Academy of Management 2012 (Boston), EALE 2012 (Stockholm). We thank J. Abeler and M. Bigoni for help with the z-Tree code and the Einaudi Institute of Economics and Finance for access to their lab. We acknowledge the generous financial support of the School of Economics and Management at the Free University of Bozen-Bolzano. Tommaso Reggiani acknowledges the support of the Deutsche Forschungsgemeinschaft through the research group "Design & Behavior - Economic Engineering of Firms and Markets" (FOR 1371).

[†] University of Rome 3 (lmarchegiani@uniroma3.it)

[‡] University of Cologne & IZA (tommaso.reggiani@uni-koeln.de)

[§] Free University of Bozen (matteo.rizzolli@unibz.it)

1 Introduction

In many situations supervisors[1] evaluate agents' performance without directly observing their efforts. Evaluation errors inevitably arise and they generally undermine agents' incentives. These errors take two forms: i) a supervisor (she) may assess low performance when in fact the agent (he) is duly exerting high effort and thus she does not reward a deserving agent (this is defined as a Type I error[2]); and ii) a supervisor may observe high performance when in fact the agent is not exerting high effort and therefore she may reward the undeserving agent (this is a Type II error). In this paper we study both theoretically and experimentally the marginal impact of the two errors on agents' incentives to perform. If systematic biases in performance evaluation emerge they mainly take the following two forms: *leniency bias* occurs when the supervisor assesses high performance "too often" while s*everity bias* occurs when the supervisor assesses low performance "too often". The goal of the paper is to explore how severe and lenient appraisal compare to one another in inducing agents' performance.

Several streams of literature, including personnel economics (Lazear, 1999), agency theory (Hölmstrom, 1979a; Aron and Olivella, 1994; Prendergast, 1999; Maestri, 2012), human resources management and organizational studies (Steers, Mowday, and Shapiro, 2004), deal with errors in performance appraisal. There exists an empirical literature showing that supervisors required to assess employees' performance subjectively have systematic leniency biases (Bretz, Milkovich, and Read, 1992; Prendergast and Topel, 1993; Jawahar and Williams, 1997; Prendergast, 1999; Moers, 2005; Berger, Harbring, and Sliwka, 2012). Many authors have extended the principal-agent model in order to provide a theoretical explanation for this consistent empirical evidence (See Tirole, 1986; Prendergast and Topel, 1996; Strausz, 1997; Vafaï, 2010; Thiele, 2011, and more cited below). To our knowledge there is no study, either theoretical or empirical, confronting how agents behave under each of the two biases. Our initial conjecture was that the supervisor's bias might be simply compensating for an agent's bias. If agents are more sensitive to Type I errors than to Type II errors it might be optimal for supervisors to be lenient regardless of any other possible additional explanation. The research question of this

[1] We will use the synonyms *supervisor*, *rater*, and *principal* interchangeably.

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

paper is thus the following: does a supervisor's *severity* bias induce less effort provision by the agent than a corresponding *leniency* bias?

To address this question we use a standard principal-agent model where both Type I and Type II errors in performance appraisal are considered and where severity and leniency biases are stylized. The main theoretical prediction delivered by the model states that both error types should be treated equally as they both jeopardize the agent's effort performance by the same token. Therefore, leniency and severity biases should be equally detrimental to the agent's effort provision as long as the sum of the two errors is kept constant. To test this main theoretical prediction, we devised a real effort laboratory experiment where subjects' performance determines their final payment and appraisal is subject to evaluation errors. Our main finding shows that there is a substantial gulf between the theoretical predictions and the empirical laboratory evidence. In particular, failing to reward a deserving agent (Type I error) is significantly more detrimental to effort provision than rewarding an undeserving agent (Type II error). Or, in other terms, a severity bias is more detrimental than a leniency bias.

The paper is organized as follows: Section Two provides a review of the related literature. Section Three introduces a parsimonius principal-agent model where both Type I and Type II errors are considered. Section Four describes both the experimental design and the procedures adopted to test the theoretical predictions delivered by the model. Data analysis is discussed in Section Five. In Section Six, we discuss the experimental findings in light of some behavioral theories. Section Seven concludes. However, before proceeding to the literature review, and in order to prove the importance of this research question, let us illustrate several stylized situations in which the choice between a severity bias and a leniency bias becomes relevant and informs the way in which supervision, evaluation and adjudication are organized and carried out.

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

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

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

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

All these situations point to a common problem which is the research object of the paper: the supervisor's activity is prone to both Type I and Type II errors and both are detrimental to the agent's performance. In all cases the supervisor controls the error trade-off and therefore it becomes crucial to assess how bad one error is compared to the other. In other words it is crucial to understand whether leniency bias is more demotivating than severity bias or vice versa.

2 Objective and Subjective Appraisal and the Lenient Supervisor's Puzzle

How do errors arise in a supervisor's assessment of an agent's effort? It is useful to distinguish between i) objective errors, ii) unintended subjective errors and iii) intentional subjective errors.

Objective errors are modeled by the standard principal-agent model with moral hazard. In a principal-agent relation, when only an objective evaluation of performance is feasible (output is observable), evaluation errors arise because i) the agent's effort is non-observable and ii) the agent's effort provision and observable performance are stochastically related. In this model the more accurate the performance evaluation, the lower the incentive constraint (Laffont and Martimort (2002); see more details in Section 3 on the model).

However, organizations where performance can be evaluated objectively are rare (Gibbons, 1998; Prendergast, 1999) and the "fascination with an 'objective' criterion, [where] individuals seek to establish simple, quantifiable standards against which to measure and reward performance" leads many pay-for-performance schemes to establish severely distorting incentives (Kerr, 1975). The problem of "rewarding A while hoping for B" (Kerr, 1975; Baker, Gibbons, and Murphy, 1994) can only be partly mitigated by adding more indirect information on the agent's effort (Hölmstrom, 1979b) as very often the problem lies in defining exactly what the principal's objective is (Baker, 1992; Jensen and Meckling, 1976).

Most organizations, then, rely on **subjective performance** appraisal in order to motivate their employees (Prendergast and Topel, 1993; Prendergast, 1999; MacLeod, 2003; Kambe, 2006; Maestri, 2012). Compared to firms' owners, managers usually have private information concerning the overall performance of their subordinates (MacLeod, 2003; Thiele, 2011). Subjective performance measures can be used alone (Bull, 1987; MacLeod and Malcomson, 1989; Levin, 2003) or in combination with objective measures (Schmidt and Schnitzer, 1995; Pearce and Stacchetti, 1998). In any case, subjective performance evaluation is also prone to errors. The two most important errors classified by the literature on subjective appraisal are *leniency bias* and *severity bias*[3]. These errors reduce the scope of appraisal because they restrict the

[3] Other rater's errors are: (i)*central tendency* error derives from a propensity to avoid assigning extreme values; (ii) *halo effect* refers to a rater's judgment on one scale influencing ratings on other scales; (iii) *contamination errors* affect the construct validity of ratings by relying on irrelevant information; (iv) *similar-to-me error* occurs when ratings are influenced because the ratee has affinity with the rater; (v) *recency errors* happens

range of useful measures of performance, and thus weaken the incentive (MacLeod, 2003).

Within the realm of subjective errors, Kane (1994) distinguishes between i) unintended (he uses the term *nonvolitional)* subjective errors and (ii) intentional *(volitional)* subjective errors. Along the same lines Prendergast (2002) distinguishes between i) biases based on personal feelings and ii) biases based on personal returns. Among the first group there are the errors arising from unconscious cognitive and behavioral biases in the observing, elaborating, or recalling of ratee performance information or in the process of generating the appraisal rating. Feelings such as empathy and affection also play an important role (Cardy and Dobbins, 1986; Varma, Denisi, and Peters, 1996) and the rater's assessment can also be manipulated by the ratee (Higgins, Judge, and Ferris, 2003). Also agent's overconfident beliefs may cause misalignment between agent's self-assessment and supervisor's performance appraisal (Maestri, 2012; Sautmann, forthcoming).

Among the second group there is the intentional distorting of appraisals in order to serve the supervisor's goals. For instance, if the principal is the residual claimant on the agents' production and the agents' pay is based on her subjective appraisal, she may underreport the performance of her subordinates in order to save costs. This would amount to an intentional severity error. On the other hand, many raters are not residual claimants but are themselves part of hierarchies and therefore their utility functions may deviate from the principal's objectives. In particular a supervisor may find it convenient to provide lenient evaluations because she colludes with the agent (See Tirole, 1986; Prendergast and Topel, 1996; Strausz, 1997; Vafaï, 2010; Thiele, 2011) or because of more complex motivation (Judge and Ferris, 1993; Grund and Przemeck, 2012; Giebe and Guertler, 2012). Most of the experiments in personnel economics and social psychology try to reproduce in the lab the conditions under which leniency bias happens to be more common than severity bias (Berger, Harbring, and Sliwka, 2012). Several studies show that supervisors are leniency-biased if they are asked to provide direct personal feedback to the agents (Klimoski and Inks, 1990; Fisher, 1979). In general, supervisors tend to be lenient either because their incentives are correlated to the agents' performances (Ilgen, Mitchell, and Fredrickson, 1981), or else because they believe that they evaluations will be

when recent performance is given too much weight as opposed to early performance within a given time interval and on the opposite (vi) *first impression error* when early performance is given too much weight as opposed to more recent performance within a given time interval (See Thomas and Meeke 2010 on classification of rater's errors. See Rabin and Schrag 1999 specifically on first impression bias).

used against the interests of the ratees (Villanova, Bernardin, Dahmus, and Sims, 1993; Kane, Bernardin, Villanova, and Peyrefitte, 1995). The focus of this literature is on the rater's behavior. This paper complements this stream of literature inasmuch as it focuses on agents' behavior (reaction) when exposed to leniency bias vis-à-vis severity bias. Evaluation accuracy is also central to Law & Economics studies of adjudicative procedures (Kaplow, 1994; Grechenig, Nicklisch, and Thoni, 2010; Rizzolli and Saraceno, Forthcoming; Kaplow, 2011; Rizzolli and Stanca, 2012) and to social psychology studies of judges' and witnesses' behavior in trial (Cutler and Penrod, 1989; Hendry, Shaffer, and Peacock, 1989).

The use of lab experiments to test theoretical predictions provides several important advantages (Falk and Heckman, 2009; Charness and Kuhn, 2010) in comparison with observational datasets that are typically used in labor/personnel economics or managerial case studies; above all the opportunity to control for all the crucial variables of the economic environment and the possibility of varying *ad hoc* the precise variables of interest (Falk and Fehr, 2003; Falk and Gächter, 2008)[4]. For this reason, experimental analyses are increasingly found in management oriented publications, especially in the field of human resources management (see among others Nieken, 2010; Gürtler and Harbring, 2010; Schnedler and Vadovic, 2011).

3 The Model

We consider a model with one supervisor and one agent where the supervisor cannot contract the agent's effort and only final output is observable. This is a standard principal-agent model with objective errors and moral hazard. However, its implications and predictions about the agent's behavior are also valid for the case of subjective errors. Finally, we borrow the nomenclature of leniency and severity biases from the subjective performance appraisal literature and we implement it into the principal-agent setting.

Let e be a measure of effort. The agent's choice is binary (no effort, effort) and we normalize no effort as 0 and effort as 1: $e\, in\, \{0.1\}$. Effort implies disutility for the agent. We define this disutility as a generic function of the level of effort $g(e)$ that can be normalized to $g(0) = g_0 = 0$ and $g(1) = g_1 = g$. The agent is an expected utility maximizer and has a utility function $V = v(w) - g(e)$ where w is the wage and can take the following two values w_0, w_r (w_0 is the

[4] On the other hand the external validity of lab findings can be questioned (Gneezy and List, 2006).

baseline wage and w_r is the rewarding wage). The utility function is separable in monetary utility and disutility of effort following the usual assumptions of concavity for the former and convexity for the latter. Note that $v(w_r) - v(w_0) = \Delta w$ is the net utility of the reward. Performance is interpreted in terms of the project's observable output. Output has a stochastic component and the performance level $\tilde{q}$ can only take two values $\{\underline{q}, \overline{q}\}$. We normalize $\underline{q} = 0$ and assume $\overline{q}$ to be positive. Another way of interpreting this is that the principal fixes a performance target equal to $\overline{q}$ and considers as zero any performance below that level.

Effort influences performance in the following way: $Pr(\tilde{q} = \overline{q} \mid e = 0) = \beta$ and $Pr(\tilde{q} = \overline{q} \mid e = 1) = 1 - \alpha$ with $1 - \alpha > \beta$. The assumption that $1 - \alpha > \beta$ implies that effort increases performance in the sense of *first-order stochastic dominance* since $Pr(\tilde{q} \leq q^* \mid e)$ is decreasing with e for any given performance q^*. Moreover, note that i) the probability that performance is zero when effort is high is smaller than the probability that performance is zero when effort is zero $\big(Pr(\tilde{q} \leq \underline{q} \mid e = 1) = \alpha < 1 - \beta = Pr(\tilde{q} \leq \underline{q} \mid e = 0)\big)$ and ii) the probability that performance is no higher than $\overline{q}$ is equal to 1 both when effort is 0 and when it is 1 $(Pr(\tilde{q} \leq \overline{q} \mid e = 1) = 1 = Pr(\tilde{q} \leq \overline{q} \mid e = 0))$. These two properties tells us that the principal prefers the stochastic distribution of performance when the agent exerts the positive effort level $e = 1$ as long as her utility function $u(.)$ is increasing in performance. Indeed, the principal's payoff when the agent exerts positive effort is $(1 - \alpha)u(\bar{q}) + \alpha u(\underline{q})$ and it is larger than the payoff when the agent exerts no effort $\beta u(\bar{q}) + (1 - \beta)u(\underline{q})$ as long as $u(.)$ is increasing. Note in fact that $(1 - \alpha)u(\bar{q}) + \alpha u(\underline{q}) = \beta u(\bar{q}) + (1 - \beta)u(\underline{q}) + (1 - \alpha - \beta)\big(u(\bar{q}) - u(\underline{q})\big)$ (Laffont and Martimort, 2002, p. 149).

For the sake of our argument we can ignore the participation constraint on the principal. We focus instead on the agent's constraints. Given the stochastic nature of performance and the non-observability of effort, the principal can only offer a contract where the agent's compensation is a function of the random output $\tilde{q}$. In other words the supervisor decides the performance target/level of observed output $\overline{q}$ which triggers the rewarding wage w_r. Note that the supervisor may commit the following evaluation errors:

Type I error: with probability α, the agent that exerts effort $(e = 1)$ does not meet the performance target $\overline{q}$ and thus he does not receive his due reward w_r

Type II error: with probability β, the agent that exerts zero effort nevertheless meets the performance target $\overline{q}$ and thus he is undeservedly rewarded with w_r.[5]

The agent's *participation constraint* is defined by the agent's reservation utility that corresponds to his best outside option. Therefore it can be written as $\alpha v(w_0) + (1-\alpha)v(w_r) - g \geq \widehat{u}$. This ensures that if the agent exerts effort, it yields at least his outside opportunity utility level. If we assume that the outside option of our experimental subjects, once sitting in the lab, is the expected wage when exerting no effort, then $\widehat{u} = (1-\beta)v(w_0) + \beta v(w_r)$ and the participation constraint coincides with the incentive constraint. The agent's *incentive constraint* instead shows the conditions under which the agent exerts effort:

$$\alpha v(w_0) + (1-\alpha)v(w_r) - g \geq (1-\beta)v(w_0) + \beta v(w_r) \tag{1}$$

which leads to the following cutoff value that defines the minimum incentive needed to induce the subject to exert effort $\hat{g} = (v(w_r) - v(w_0))\,(1-\alpha-\beta)$.

The incentive constraint tells us that the agent will exert effort as long as the cost of effort is smaller than the net reward of performance discounted by the probabilities of both Type I and Type II errors. Note also that on one hand the larger the probability of β (being rewarded undeservingly), the larger the returns from not exerting effort. On the other hand, however, the larger the probability of α (not being rewarded when deserving it), the smaller the returns of exerting effort. Note that the sum $(1-\alpha-\beta)$ defines the *accuracy* of the performance appraisal. Accuracy can be kept constant with very different error trade-offs as long as $\alpha_{low} + \beta_{high} = \alpha_{high} + \beta_{low}$. We will exploit this implication of the model to characterize the lenient and severe biases as our treatment conditions in the experiment.

Accordingly to this analysis, it is possible to draw a clear main theoretical prediction [1] to be tested. In addition to that, the assessment of two complementary predictions [2 and 3] is important in order to check the robustness of the main test.

[5] The derivation of the probabilities of errors from the definition of $\overline{q}$ is outside the scope of the present work. However, it is intuitive to say that the sum of errors $(\alpha + \beta)$ is minimized for some intermediate levels of $\bar{q}$. This is because when the performance target is set very low it is very easy to meet the target both with high effort and with none. Therefore with low $\overline{q}$ we have no Type I errors (i.e. not meeting the target when exerting effort) and many Type II errors (i.e. meeting the target when exerting no effort). Therefore there is little incentive for the agent to invest effort. The more the performance target increases, the smaller the probability of Type II error becomes and thus switching to effort becomes convenient. At some intermediate level of $\overline{q}$ we have a few Type II errors and a few Type I errors. Finally, when the performance target becomes extremely high, the probability of Type II errors (i.e. meeting the target when exerting no effort) becomes virtually nil but at the same time the probability of Type I error (i.e. not meeting the target when exerting effort) is very high and therefore there is little incentive to exert effort.

[1] Main Prediction: Neglecting due rewards (severity - α) is equally detrimental to agents' effort provision than rewarding undeserving agents (leniency - β).

[2] Sub-Prediction : Neglecting due rewards (severity - α) decreases agents' effort provision with respect to perfect appraisal.

[3] Sub-Prediction : Rewarding undeserving agents (leniency - β) decreases agents' effort provision with respect to perfect appraisal.

3.1 Model Parameters for the Experimental Treatments

To test the behavioral implications of the model, we devise three treatments (*fair*, *severe* and *lenient*) in two different configurations (*high* and *low*) described in Table 1. The first variation concerns the accuracy of appraisal: under the *fair* treatment accuracy is maximal, while accuracy is heavily biased by Type I errors under the *severe* treatments and by Type II errors under the *lenient* treatments and yet accuracy is the same under both *severe* and *lenient* treatments. The second variation concerns the initial endowment. In the *low* configuration $w_0 = 0$ while in the *high* configuration $w_0 =$ €5.28. The net reward paid to performing agents is in both cases $w_r - w_0 =$ €6.60.[6]

Table 1: Table of Treatment Parameters.

Conf.	**Treatment**	**Type I**	**Type II**	**Accuracy**	**Baseline**	**Reward**	**Incentive constraint**
		α	β	$1-\alpha-\beta$	w_0	w_r	$((w_r) - v(w_0))\,(1-\alpha-\beta)$
Low	$T0_L$ **- Fair**	0	0	1	€0	€6.60	$v($€$6.60)$
	$T1_L$ **- Severe**	4/5	0	1/5	€0	€6.60	$\frac{1}{5}v($€$6.60)$
	$T2_L$ **- Lenient**	0	4/5	1/5	€0	€6.60	$\frac{1}{5}v($€$6.60)$
High	$T0_H$ **- Fair**	0	0	1	€5.28	€11.88	$v($€$11.88) - v($€$5.28)$
	$T1_H$ **- Severe**	4/5	0	1/5	€5.28	€11.88	$\frac{1}{5}\left(v(\text{€}11.88) - v(\text{€}5.28)\right)$
	$T2_H$ **- Lenient**	0	4/5	1/5	€5.28	€11.88	$\frac{1}{5}\left(v(\text{€}11.88) - v(\text{€}5.28)\right)$

Fair treatment ($T0_L$ and $T0_H$). There are no evaluation errors ($\alpha, \beta = 0$) and thus the expected returns for the agents are $v(w_r)$ and $v(w_0)$ for exerting effort ($e = 1$) and no effort ($e = 0$) respectively. The incentive constraints are thus simply the utility differences between the rewarding and baseline wages ($v(w_r) - v(w_0)$) in each treatment. This treatment is "fair" in the sense that the agent gets what he deserves.

[6] The two configurations - *low* and *high* - stylize different payment structures often found in real world situations. In the *low* configurations the entire wage corresponds to the performance reward and thus any evaluation error affects the assignment of the whole salary. On the other hand, in the *high* configurations only a certain - albeit still large - amount of the salary depends on the interaction between the performance and the evaluation error.

Severe treatment ($T1_L$ and $T1_H$). In T1 there are no Type II errors ($\beta = 0$) but the probability of Type I error is significant ($\alpha = 0.8$).[7] Given the high number of Type I errors, the net returns from exerting effort are small but still positive (€1.32). On the other hand, the returns from not exerting effort are zero. The incentive constraint is smaller than in the *fair* treatment and equal to $\frac{1}{5}v(€6.60)$ and to $\frac{1}{5}\left(v(€11.88) - v(€6.60)\right)$ for $T1_L$ and $T1_H$ respectively. Note that the incentive constraints are the same if utility is linear. This treatment is "severe" in the sense that the deserving agent very often does not get what he deserves.

Lenient treatment ($T2_L$ and $T2_H$). In T2 there is a significant probability of Type II error ($\beta = 0.8$) but there are no Type I errors ($\alpha = 0$). The returns from exerting effort are large (€6.60). On the other hand the expected returns from not exerting effort are also large (€5.28) and the difference is still €1.32. The incentive constraints are $\frac{1}{5}v(€6.60)$ for $T2_L$ and $\frac{1}{5}\left(v(€11.88) - v(€6.60)\right)$ in $T2_H$. Note that the incentive constraints are the same as for $T2_L$ and $T2_H$ respectively in the *severe* treatment. This treatment is "lenient" in the sense that the undeserving agent very often gets what he does not deserve.

Table 1 shows that, in each configuration, the incentive constraint is the same for the *severe* and *lenient* treatments. This property is exploited to test some predictions that can be derived from Equation 1. Before proceeding to test the theoretical predictions we illustrate the experimental protocol in the next paragraph.

4 The Experiment

The research question of the present work deals with a variable - evaluation errors - that is basically impossible to observe in the field because of the unobservability of effort and the stochastic relation between performance and effort. In the lab, instead, we can superimpose an exogenous probability of error in evaluating performance and at the same time we can perfectly observe effort. This ideally allows us to identify precisely the impact of errors on effort provision and thus on performance.

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

The experimental design was composed of three different phases: the preliminary **Phase I** was used to elicit individuals' risk attitudes via a standard incentivized choice of lotteries. Following Holt and Laury (2002) subjects were asked to carry out a standard series of lotteries (see Table 6 in the Appendix) to measure individual risk aversion. Outcomes of the lotteries were communicated to subjects only at the end of the experiment in order to prevent revenue effects.

In **Phase II** individual ability in the default task was measured. Following Abeler, Falk, Götte, and Huffman (2011), the real effort task consisted in counting the number of occurrences of the digit *1* in as many tables as possible, where each table was composed of 50 digits and, among these, the number of *1*s was randomly generated[8]. In order to elicit individual productivity, subjects were offered a pure piece-rate compensation scheme. They received €0.03 for each table correctly processed[9]. Furthermore both a countdown timer and a counter reporting the number of tables processed were provided. After 10 minutes subjects received summary statistics on the number of tables correctly processed, the number of tables incorrectly processed and the total amount of money generated in this phase. Let us define $q_{phase-II}$ as the number of tables correctly processed by each individual. This number is key to computing the individual performance target in the next phase. The average $q_{phase-II}$ was 45.9 with standard deviation of 11.5. The number of tables counted in the second phase, paid by the piece, represents a reliable measure of an individual's ability. In **Phase III** we imposed on each individual specific performance targets based on their piece-rate performance in Phase II. The intent of scaling the individual's specific target to his own ability was to roughly normalize the cost of effort for the task across subjects. This is because if, on one hand, individuals with different abilities counted in Phase II different numbers of tables, then on the other hand, by setting individual performance targets proportional to these numbers for Phase III, the costs of effort required to reach these targets should be roughly the same for each individual.

[8] This task has several advantages: it does not require any prior specific knowledge; performance is objective and easily measurable; and there is little room for learning effects. At the same time, the task is boring and pointless at least for most of the subjects and thus it can be claimed that the task entails a positive cost of effort. The task is also clearly artificial, and output does not provide intrinsic or extrinsic value to the experimenter. This should rule out any tendency for subjects to use effort provision during the experiment as a way to reciprocate for incentives provided by the experimenter or the possibility that subjects carry out the task for some intrinsic motivation.

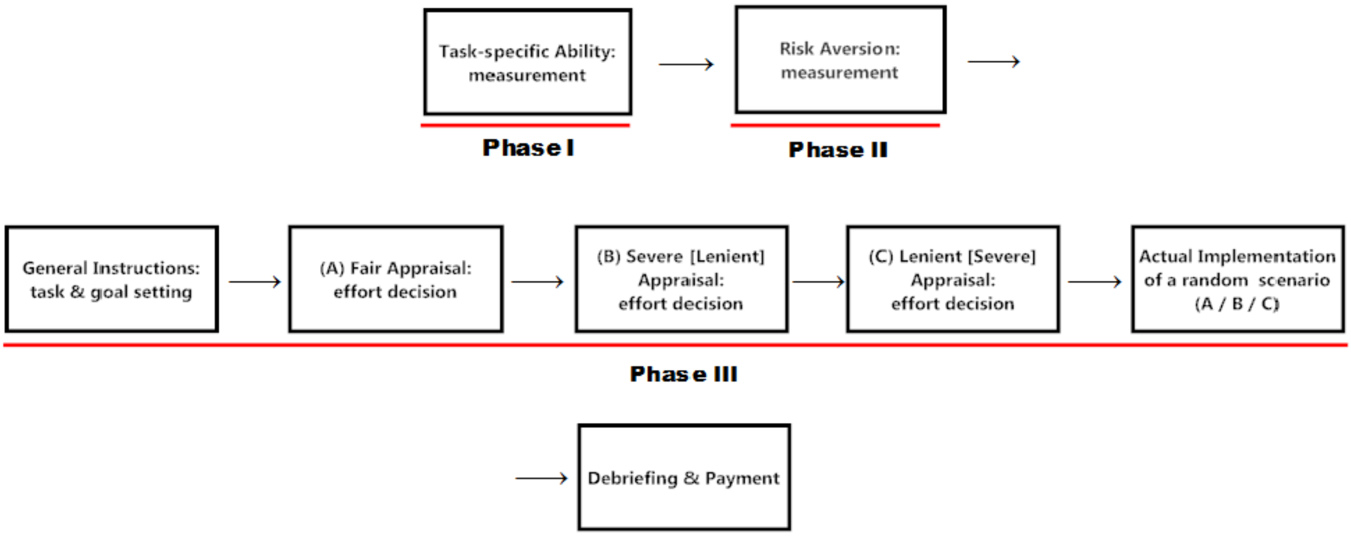
[9] The piece-rate was relatively low in order to provide both a positive incentive to exert effort in this phase and at the same time - considering the flow of the whole experimental protocol - to prevent distortions in the two different treatment configurations (high/low) adopted in Phase III.

Phase III lasted 40 minutes, four times the length of Phase II. The performance target was set equal to $4 \times 0.9 \times q_{phase-II}$[10]. We implemented all three treatments (*fair*, *severe* and *lenient*; see Table 1) in each session and we had two sessions for the *low* configuration and two sessions for the *high* configuration.

Table 2: Experimental Sessions

Session	Conf.	Treatment Order	Endowment	Reward	Subjects
i	*Low*	$T0_L, T1_L, T2_L$	€0	€6.60	23
ii	*Low*	$T0_L, T2_L, T1_L$	€0	€6.60	17
iii	*High*	$T0_H, T1_H, T2_H$	€5.28	€11.88	23
iv	*High*	$T0_H, T2_H, T1_H$	€5.28	€11.88	21

Figure 1: Flow of the experiment


Task-specific Ability: measurement
Risk Aversion: measurement
Phase I
Phase II
General Instructions: task & goal setting
(A) Fair Appraisal: effort decision
(B) Severe [Lenient] Appraisal: effort decision
(C) Lenient [Severe] Appraisal: effort decision
Actual Implementation of a random scenario (A / B / C)
Phase III
Debriefing & Payment

Absent any evaluation error (e.g. in T0) the accomplishment of the task was rewarded with €6.60[11] on top of the initial endowment that characterized the two alternative treatment configurations: €0 in the *low* endowment configuration, €5.28 in the *high* endowment configuration.

Each subject could choose between i) leaving the room with the baseline wage w_0 or ii) attempting to perform the task and gain w_r. They had to declare their choice (i or ii) for each of the three possible treatments. However, after the three choices, only one treatment

[10] Note that 4 is the ratio of the duration of Phase III (40 minutes) to the duration of Phase II (10 minutes). Moreover we thought the higher fatigue created by the longer task justified a 10% discount on the performance target and at the same time it signals that the performance target can be achieved by exerting a high but not extraordinarily high level of effort. This is confirmed by the data. All but one of the subjects who engaged in the task eventually matched the performance target.

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

was randomly selected and its parameters applied[12]. Subjects were informed about which treatment was actually randomly implemented only after they had stated their decisions for all three scenarios. If, for the implemented treatment, the subject had chosen i), then he immediately had to proceed to the questionnaire phase; if he had chosen ii) the task begun and the subject had to let 40 minutes pass before moving to the questionnaire phase.

Subjects made truthful choices for the three scenarios because the choices implied real consequences: if the subject chose A then he had to spend 40 minutes in the lab anyway before moving to the questionnaire and to the payment phase, and if he chose B then the real effort task was skipped entirely. Therefore the subjects had no reason to misrepresent their true preferences[13].

The three main treatments were deployed in two different configurations (*high* and *low*). This was to check whether different levels of initial endowment could play a role in determining systematically different perceptions of the evaluation errors (see also Footnote 6).

Between each of the phases, subjects had the opportunity to rest. Common instructions for the subsequent phase were read and described aloud while instructions concerning each single treatment were delivered on screen. Feedback information, on the outcomes of the lotteries in Phase I and on whether the supervisor-automaton had made an evaluation error in the implemented scenario, were provided at the end of the experimental session. Control questions for each of the different phases and treatments were administered through the computer. The experiment was programmed and conducted with z-Tree (Fischbacher, 2007) and sessions took place at the Einaudi Institute of Economics and Finance in Rome on April 6, April 8, April 14 and May 2, 2011. We ran a total of four sessions with 84 participants. Subjects were recruited online with ORSEE (Greiner, 2004). Wilcoxon rank-sum tests indicate that there were no significant differences in the socio-demographic characteristics of the subjects across sessions: mainly undergraduate students with very different backgrounds (humanities, medicine, hard sciences, social sciences). Average age was 22.47 (s.d. 2.16), females 40%, males 60%. Via

[12] The fair Treatment is always submitted first as it represents the benchmark case. The severe and lenient treatments are submitted after the fair in inverted order following table 2.

[13] This is thus a within-subject design as we are able to observe the variations of subjects' effort choices across the three treatments (*fair*, *severe* and *lenient*) and it implements the strategy method as the choices are elicited before one is randomly chosen and implemented. This procedure avoids income effects and also rules out any potential order effect of subjects' choice being influenced by previous decisions. Of course the choice of the within-subject design has its own drawbacks, which are well-known in the literature (Charness, Gneezy, and Kuhn, 2011).

the strategy method we elicited 84 observations for each treatment. Average payoff was about €10.21.

5 Experimental Results

Equation 1 of the model shows that the incentive constraint can be kept constant even with very different Type I/Type II error trade-offs as long as accuracy (the sum of errors) is kept constant. This property has been exploited in creating the *severe* and *lenient* treatments. We can use our treatments to test some predictions of the model by looking at the different shares $Z(Ti_j)$ of performing agents when exposed to the different treatments (i) and configurations (j).

The design envisaged that, for each treatment, individuals had to state whether they would leave the room with the baseline wage w_0 (effort= 0) or would attempt to perform the task and gain w_r (effort = 1). Performing agents are the ones who exerted effort and met the performance target. Those who chose to leave the room or failed the target are defined as non-performing agents. In this respect, we only focus on the "extensive margin" of the share $Z(Ti_j)$ of subjects willing to exert effort.[14]

Table 3: Summary of Results

Share of population	**Fair** - $T0$	**Severe** - $T1$	**Lenient** - $T2$
$Z(Ti_L)$ *Low configuration*	0.825	0.325	0.625
$Z(Ti_H)$ *High configuration*	0.909	0.545	0.705

Share of Switchers	**Fair vs. Severe**	**Fair vs. Lenient**	**Lenient vs. Severe**
Switchers *Low configuration*	0.5 ***	0.2 **	0.3 ***
Switchers *High configuration*	0.355 ***	0.204 ***	0.16 **

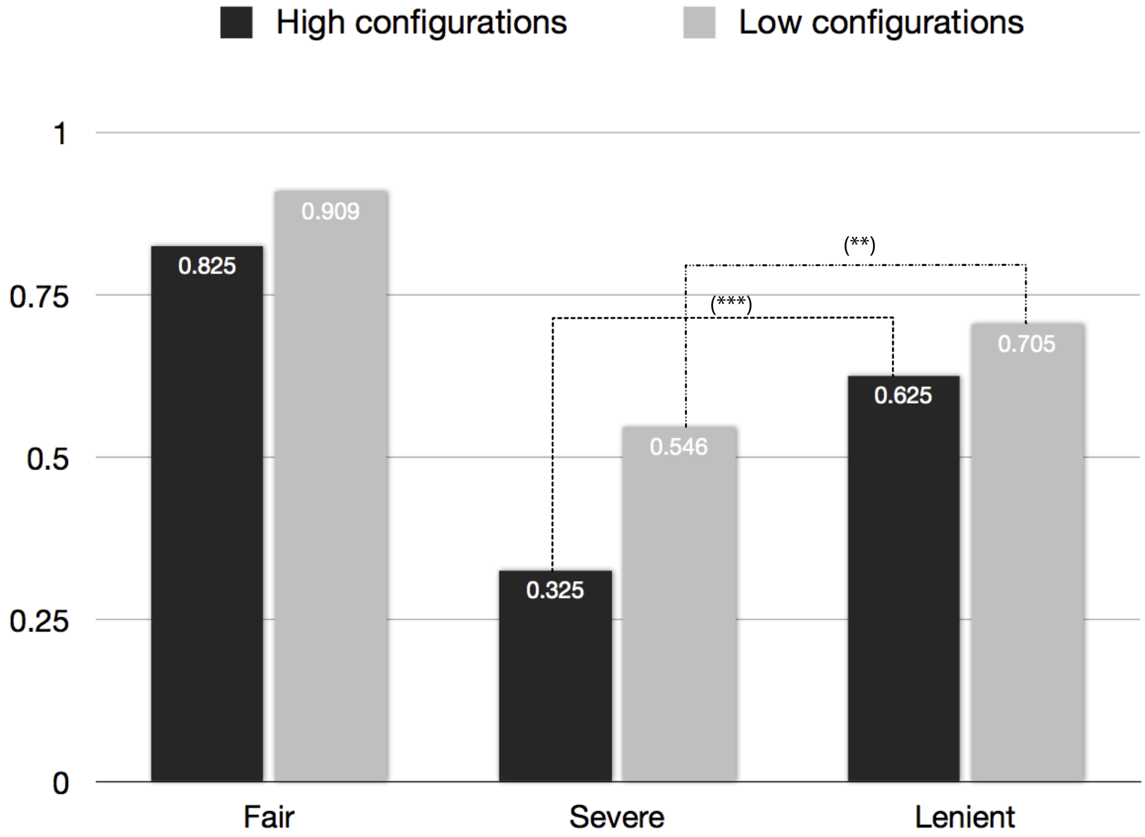
Significance levels (McNemar's test): $^{***}p < 0.01$,$^{**}p < 0.05$, $^{*}p < 0.10$

Result [1]. Neglecting due rewards (α) is more detrimental to agents' effort provision than rewarding undeserving agents (β).

The model predicts that Type I and Type II errors should be equally detrimental to effort

[14] Tuning the target to 90% of the maximal individual capacity ensures the feasibility of the goal and allows us to dispel uncertainty concerns related to the actual feasibility of the task. For this main reason - in this setting - the analysis of the effort "intensive margin" is redundant: in fact all but one of the subjects who chose to carry out the task eventually matched the performance target.

Figure 2: Percentage of population exerting effort under *fair*, *severe* and *lenient* treatments in *low* and *high* configurations



provision. A given increase in the probability of β compensated by an equal decrease in the probability of α, and the converse, leaves the incentive constraint unchanged and therefore individuals should not change their behavior as long as *accuracy* is kept constant. In order to test this prediction we compare the share of agents exerting effort under the *severe* and the *lenient* treatments[15]. An inspection of Figure 2 shows that the percentages of individuals exerting effort under the *lenient* and *severe* treatments are by no means equal. In fact under the *low* configuration treatments the share of population exerting effort almost doubles from 0.325 ($T1_L$) to 0.625 ($T2_L$). The positive effect is highly statistically significant on the basis of a two-sided null hypothesis and 40 independent paired observations (McNemar's $\chi^2 = 8$, $p - value = 0.0047$). Under the *high* configuration treatments the share of subjects exerting effort increases from 0.545 ($T1_H$) to 0.705 ($T2_H$). The positive effect is highly statistically significant, on the basis of a two-sided null hypothesis and 44 independent paired observations (McNemar's $\chi^2 = 7$, $p - value = 0.0082$).

[15] We thus test the following: $H_0 : Z(T1_L) = Z(T2_L)$ - versus -$H_1 : Z(T1_L) < Z(T2_L)$ and $H_0 : Z(T1_H) = Z(T2_H)$ - versus -$H_1 : Z(T1_H) < Z(T2_H)$

Contrary to the predictions of the model, this experiment provides evidence that severity bias and leniency bias do not generate symmetric effects on subjects' effort provision: the negative effect of the severity bias in $T1_L$ and $T1_H$ is substantially and significantly greater than the effect of the leniency bias in $T2_L$ and $T2_H$ respectively,

Result [2]. Neglecting due rewards (α) decreases agents' effort provision.
In order to test whether neglecting due rewards decreases agents' effort provision, we contrast the share of performing agents (define as Z) in *fair* treatments ($\alpha = 0, \beta = 0$) with the same share in *severe* treatments ($\alpha = 0.8, \beta = 0$)[16].

An inspection of Figure 2 shows how sharply the percentage of population exerting effort drops between the *fair* and *severe* treatments. In the *low* configuration (dark bars) under *fair* treatment, 0.825 of subjects exert effort while the same share falls to only 0.325 under *severe* treatment. This negative effect is statistically significant on the basis of a two-sided null hypothesis and 40 independent paired observations (McNemar's[17] $\chi^2 = 20, p{-}value < 0.0001$). The results are similar in the *high* configuration: under *fair* treatment, 0.909 of subjects exert effort while only 0.545 of them exert effort under *severe* treatment. Also this drop is strongly statistically significant on the basis of a two-sided null hypothesis and 44 independent paired observations (McNemar's $\chi^2 = 16$, $p - value < 0.0001$).

Result [3]. Rewarding undeserving agents (β) decreases agents' effort provision.
In order to test whether rewarding undeserving agents decreases agents' effort provision, we compare the share of subjects exerting effort under *fair* treatments (with $\alpha = 0, \beta = 0$) with those under *lenient* treatments ($\beta = 0.8, \alpha = 0$).[18] In the *low* configuration the share of subjects exerting effort decreases from to 0.825 ($T0_L$) to 0.625 under *lenient* treatment ($T2_L$). This negative effect is statistically significant at the 5% level, on the basis of a two-sided null

[16] We thus test the following: $H_0 : Z(T0_L) = Z(T1_L)$ - versus -$H_1 : Z(T0_L) > Z(T1_L)$
and $H_0 : Z(T0_H) = Z(T1_H)$ - versus -$H_1 : Z(T0_H) > Z(T1_H)$

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

[18] We thus test the following: $H_0 : Z(T0_L) = Z(T2_L)$ - versus -$H_1 : Z(T0_L) > Z(T2_L)$
and $H_0 : Z(T0_H) = Z(T2_H)$ - versus -$H_1 : Z(T0_H) > Z(T2_H)$

hypothesis and 40 independent paired observations (McNemar's $\chi^2 = 5.33$, $p - value = 0.022$). In the *high* configuration the percentage of subjects exerting effort drops from 0.909 ($T0_H$) $T3$ to 0.705 in ($T2_H$). This negative effect is highly statistically significant, on the basis of a two-sided null hypothesis and 44 independent paired observations (McNemar's $\chi^2 = 9$, $p - value = 0.0027$).

Finally, an examination of Figure 2 suggests that the percentage of population exerting effort under each treatment qualitatively increases in the *high* endowment configuration compared with the *low* endowment configuration. Under the two *fair* treatments ($T0$ and $T3$) the share of agents exerting effort in the *low* endowment configuration is equal to 0.825 and 0.909 respectively. The difference between configurations[19] turns out not to be statistically significant at any conventional level ($p - value = 0.2567$). When comparing the same share under the *severe* treatments ($T1_L$ with 0.325 vs. $T1_H$ with 0.545) the difference between configurations turns out to be significant at the 5% level ($p - value = 0.0433$). Finally under the *lenient* treatments the difference between the share in $T2_L$ (0.652) and $T2_H$ (0.705) is not statistically significant at any conventional level ($p - value = 0.4426$). There seems to be some mild evidence that, if anything, both leniency and severity biases are more detrimental when they affect the whole wage assignment (*low* configuration) than when they affect only a portion of the salary (*high* configuration). The negative effect is more pronounced in the case of Type I error.

All the results produced by the non-parametric analysis are confirmed by complementary parametric analysis (non-linear models) reported in the Appendix 8.1.

6 Discussion

Subjects behave differently under the *severity* and *leniency* biases. This is the interesting puzzle that emerges from our experimental test. In the following paragraphs we rule out some potential explanations for the asymmetry and we discuss the main result in light of different economic theories of behavior.

First, let us consider social preferences such as inequity aversion, gift-exchange behavior

[19] To test these difference we have run a between-subjects analysis contrasting treatment outcomes by configuration with a Wilcoxon rank-sum test.

and reciprocity. These theories of behavior are often called on in explaining organizational behavior (see Sebald and Walzl, 2012 for a recent experiment on subject performance appraisal and reciprocity). However, our experimental design does not allow interaction between agents, thus inequality in outcomes arising from the agent's choices is ruled out by design. Moreover, our experiment does not comprise a real supervisor (our supervisor is represented by a passive automaton) to be thankful or resentful towards, thus hindering the explicative effect of reciprocity and gift-exchange.

6.1 Why It Cannot Be Risk Aversion

The experimental design is neutral to subjects' different levels of risk aversion, at least the risk aversion that comes from conventional decreasing marginal utility of income. To see why, consider the following table where standard generic concave utility functions with separable costs of effort are reported for both *low* and *high* configurations. Subjects decide whether to exert effort whenever the difference in utility (the last column of Table 4) is positive. Note that all four of the "biased" treatments ($T1_j$ and $T2_j$) have the same difference in expected utility.

Table 4: Expected Utility Functions

	Exp. Utility with Effort	**Exp. Utility with No Effort**	**Exp. Delta Utility**
$T0_L$	$v(€6.60) - g$	$v(€0)$	$v(€6.60) - v(€0) - g$
$T1_L$	$\frac{1}{5}v(€6.60) + \frac{4}{5}v(€0) - g$	$v(€0)$	$\frac{1}{5}\left(v(€6.60) - v(€0)\right) - g$
$T2_L$	$v(€6.60) - g$	$\frac{4}{5}v(€6.60) + \frac{1}{5}v(€0) - g$	
$T0_H$	$v(€11.88) - g$	$v(€5.28)$	$v(€11.88) - v(€5.28) - g$
$T1_H$	$\frac{1}{5}v(€11.88) + \frac{4}{5}v(€5.28) - g$	$v(€5.28)$	$\frac{1}{5}\left(v(€11.88) - v(€5.28)\right) - g$
$T2_H$	$v(€11.88) - g$	$\frac{4}{5}v(€11.88) + \frac{1}{5}v(€5.28)$	

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

Moreover, in order to control for risk aversion in the data, we also ran an incentivized Holt and Laury (2002) lottery test in Phase I. Correlations between the individual measure of risk aversion[20] and the choice of exerting effort for both treatments $T1$ and $T2$ are very weak and statistically not significant (Spearman's $\rho - correlation = 0.032$, $p - value = 0.77$ in $T1$ and

[20] In terms of the switching point from the risky to the safe option in Table 6

$= 0.062$, $p - value = 0.57$ in $T2$).

A final consideration concerns the difference in expected utility between *high* and *low* configurations. Under standard decreasing marginal returns of income we have that $v(€6.60) \geq v(€11.88) - v(€5.28)$ and $\frac{1}{5}v(€6.60) \geq \frac{1}{5}\left(v(€11.88) - v(€5.28)\right)$. Therefore for each treatment (*fair*, *severe*, *lenient*), the agent's marginal utility of the reward payment should be higher in the *low* endowment configuration than in the *high* endowment one. This should imply that the incentive constraints in the *high* configuration treatments are no lower than the ones in the *low* configuration treatments. Our qualitative results in this respect go against such a hypothesis as we observed - if anything - more subjects exerting effort in the *high* configuration treatments than in the *low* ones. This may suggest that the agent's utility is linear in this range of values.[21]

6.2 Loss Aversion, Disappointment Aversion and Reference Points

Another likely candidate explanation for the asymmetry in behavior observed in the *severe* vs. *lenient* treatments is loss aversion. Loss aversion implies that the disutility suffered when losing a certain monetary amount relative to a reference point is larger than the utility enjoyed when gaining the same certain monetary amount relative to the same reference point (Kahneman and Tversky, 1979, 1984). Key for understanding what role risk aversion may play in our experiment is to understand where the reference point is set.

Early models of loss aversion set the reference point simply on the wealth owned at the status quo (Kahneman, Knetsch, and Thaler, 1990; Benartzi and Thaler, 1995; Genesove and Mayer, 2001). If we use the status quo in the context of our experiment, then every subject participating in our experiment never suffers any loss. This is because there is no situation in which the subject has to pay the experimenter and the net monetary returns of exerting effort are positive under all three treatments (€6.60 with certainty for $T0$ and the expectation of €1.32 for $T1$ and $T2$). More advanced models set the reference points elsewhere: for instance at the *lagged status quo* (Thaler and Johnson, 1990; Gomes, 2005).

Disappointment aversion models set the reference point on the certainty equivalent value of the lottery. Disappointment arises when a lottery outcome turns out to be bad relative to the

[21] Gift-exchange theory (Akerlof 1982) could represent a plausible candidate explanation for this behavioral outcome. As long as agents get a fair basic wage (*high* endowment configuration) they are more frequently willing to perform the task. In particular, this consideration is corroborated by the fact that more agents exert effort under $T0_H$ than $T0_L$.

reference point and therefore agents may experience disappointment at not having had a better outcome. This emotion of disappointment can worsen the disutility that the outcome produces directly. Similarly, relatively good outcomes of the lottery can produce the pleasing emotion of "elation" over and above the utility that the positive lottery outcome produces directly (Elster, 1998). A disappointment-averse subject is thus one who derives more disutility from disappointment than utility from elation. This aversion reduces for him the certainty equivalent value of the lottery. Disappointment and elation are defined by contrasting the lottery outcome with the mean of the lottery (Bell, 1985; Loomes and Sugden, 1986; Gul, 1991). Disappointment aversion, however, does not seem to explain the asymmetry we observed in our experiment. In fact, following Loomes and Sugden (1986) it must be noted that the basic utility of the lottery outcome is affected by whether it ends up being above or below the certainty equivalent value of the lottery $\bar{w}$. Disappointment and elation are represented by a single differentiable real-valued function $D(.)$ which assigns a decrement or increment of utility to every possible value $w_{0,r} - \bar{w}$. The utility $v(.)$ is assumed to be linear at least for the relevant interval and therefore we simplify the notation with $v(w_{0,r}) = w_{0,r}$. The expected modified utility is thus $w_{0,r} + D(w_{0,r} - \bar{w})$. Each subject tries to anticipate any disappointment or elation, and chooses so as to maximize expected modified utility. As in Loomes and Sugden (1986), $D(w_{0,r} - \bar{w})$ is such that when $D(.) = 0$ the subject behaves exactly as maximizing expected utility. Moreover $D(w_{0,r} - \bar{w})$ is convex for all positive values of $(w_{0,r} - \bar{w})$ and concave for all negative values. The intuition here is that the intensity of disappointment and of elation increases at the margin. Finally, disappointment is assumed to have the same intensity as elation, so $D(w_{0,r} - \bar{w}) = -D(\bar{w} - w_{0,r})$. This last assumption somehow departs from the loss aversion framework. Following Loomes and Sugden (1986), the new incentive constraint (see Equation 1) is thus as follows:

$$\begin{aligned}\alpha\left[w_0 + D(w_0 - \bar{w})\right] + (1-\alpha)\left[w_r + D(w_r - \bar{w})\right] - g \\ \geq (1-\beta)\left[w_0 + D(w_0 - \bar{w})\right] + \beta\left[w_r + D(w_r - \bar{w})\right] \quad (2)\end{aligned}$$

In the *severe* treatment the subject takes part in the lottery if he chooses to exert effort and in this case $\bar{w} = \frac{1}{5}$€6.60 = €1.32. If he chooses to exert no effort the outcome is certain and therefore no disappointment or elation is possible. Conversely, in the *lenient* treatment, the subject takes part in the lottery if he chooses to exert no effort. In this case $\bar{w} =$ €5.28. In the case where he exerts effort the outcome is certain. Now the cutoff value for the incentive constraint can be computed and is the same for both $T1$ and $T2$: $\hat{g}_{T1} = \hat{g}_{T2} =$ €$1.32 + \frac{1}{5}D($€$5.28) - \frac{4}{5}D($€$1.32)$. Therefore, disappointment aversion does not predict the asymmetry in behavior observed in our experiment.

Kőszegi and Rabin (2006, 2007) maintain from the disappointment aversion literature that the reference point depends on expectations. However, instead of fixing the reference point at the mean of the lottery they model $\bar{w}$ as a distribution of stochastic reference points drawn from a distribution G (in our case the distribution is binomial), assuming that how a person feels about gaining or losing with respect to a reference point depends on the changes in consumption utility associated with such gains or losses. Following Kőszegi and Rabin (2007) preferences are assumed to be linear in probabilities and we only consider the coefficient of loss aversion λ. The incentive constraint with Koszegi and Rabin preferences is thus as follows:

$$\begin{aligned}
&\alpha \left[w_0 + (\alpha\lambda(w_0 - w_0) + (1-\alpha)\lambda(w_0 - w_r))\right] + \\
&\qquad (1-\alpha)\left[w_r + (\alpha(w_r - w_0) + (1-\alpha)(w_r - w_r))\right] - g \\
&\qquad \geq (1-\beta)\left[w_0 + (\beta\lambda(w_0 - w_r) + (1-\beta)\lambda(w_0 - w_0))\right] + \\
&\qquad\qquad \beta\left[w_r + (\beta(w_r - w_r) + (1-\beta)(w_r - w_0))\right] \quad (3)
\end{aligned}$$

We now plug our experimental parameters for T1 and T2 into the above incentive constraint. In the *severe* treatment the above condition simplifies to $0.8\,(-0.2\lambda w_r) + 0.2\,(w_r + 0.8w_r) \geq -g$ and therefore the incentive constraint is satisfied for $\hat{g}_{T1} \leq$ €$(2.376 - 1.056\lambda)$. In the *lenient* treatment instead the incentive constraint is $w_r - g \geq 0.2\,(-0.8\lambda w_r) + 0.8w_r + 0.8\,(0.2w_r)$ and the cutoff value is thus $\hat{g}_{T2} \leq$ €$(0.264 + 1.056\lambda)$. Note that for $\lambda = 1$ the two cutoff values are the same. However, when the coefficient of loss aversion $\lambda > 1$ then $\hat{g}_{T1} < \hat{g}_{T2}$. A large body of literature following Kahneman and Tversky (1979); Tversky and Kahneman (1992) estimates the coefficient of loss aversion to be $\lambda \geq 2$ and therefore the Koszegi and Rabin model correctly predicts the asymmetric result of our experiment[22].

6.3 Regret Aversion

Regret aversion (Loomes and Sugden, 1982; Bell, 1983; Hayashi, 2008) is another candidate explanation for our asymmetric result. Regret and disappointment are two different emotions that affect decision making. Both are negative emotions encountered when facing risky decisions, and both arise in the context of a mental comparison between the outcome actually obtained and an outcome that might have been (Zeelenberg, Van Dijk, Manstead, and van der Pligt, 2000). When a decision results in a bad outcome, a subject may feel disappointed when he had expected a better result, and the same subject may feel regret when he realizes that the outcome would have been better had he chosen differently (Van Dijk and Zeelenberg, 2002). Note that in our *severe* treatment the choice of exerting effort could be conducive to regret (with a probability of $\frac{4}{5}$ that the subject exerts effort and obtains the same outcome as if he had not exerted effort). Instead the choice of not exerting effort is regret-free (note that no information was provided on what could have happened if the subject had exerted effort). By contrast, in the *lenient* treatment, the choice to exert effort is regret-free as no counterfactual is provided (no feedback on whether the subject would have been paid for doing nothing). Instead, by choosing not to exert effort, the subject may feel regret if he eventually does not get paid. Therefore anticipated regret might seem to generate the observed pattern.

6.4 Organizational Justice and Procedural Fairness

A final thought goes to the literature on procedural fairness, which has been addressed in both economics and management. The procedural fairness literature claims that subjects do not primarily focus on the "ex-post distributive fairness" (Kagel and Roth, 1995) realized between agents but rather on the "ex-ante procedural fairness" (Bolton, Brandts, and Ockenfels, 2005) of the generated outcome. In management studies, procedural fairness represents a core concept of Organizational Justice. In this context, it has been proved that employees' satisfaction depends on final fair distribution of the rewards as well as on fair procedures in determining the distribution, and very often the procedure used to implement a decision plays a larger role

[22] Note that with $\lambda > 2.25$ the cutoff parameter in T1 is actually negative.

than the actual outcome of the decision *per se* in affecting subjects' welfare. In this respect, justice is intended as the worker's subjective feeling about the fairness of the procedures towards the self rather than for other workers or the supervisor. In our experiment, *severe* ($T1$) and *lenient* ($T2$) treatments represent two symmetrical cases of imperfect compensation procedures. While severity bias exposes the worker to potential harm, leniency bias puts the supervisor in the weaker role. Although both biases are procedurally unfair, the *severe* treatments go against the subject's interests while the *lenient* treatments favor them.

This consideration could provide a plausible explanation of the higher negative response observed under $T1$ than under $T2$.

7 Implications and Conclusion

A Type I error is the failure to reward effort. A Type II error is the erroneous rewarding of no effort. Together, they are a pervasive phenomenon in every evaluation procedure. This is true within the firm, in subjective performance appraisal and also when performance can be measured objectively, but it is only stochastically related to effort. The implications of the present work apply also to evaluation procedures outside the firm, such as judicial procedures and educational assessment. The combination of the two types of error leads to two main situations: when Type I errors are predominant, there is severity bias, while when Type II errors are predominant, there is leniency bias. Evaluation biases undermine the impact of performance appraisal on the agent's incentive constraint. In the paper we first show that the standard principal agent model predicts that both biases should impact performance symmetrically. We then present an experiment to test this prediction. The experiment finds strong support for the existence of an asymmetric impact of evaluation errors on agents' willingness to exert effort. In particular we show that an agent exposed to evaluation errors is more sensitive to Type I errors. This result is not predicted by the model even when we consider: a) risk aversion within the expected utility framework; b) loss aversion with the reference point at the status quo; or c) disappointment aversion. However, Koszegi and Rabin's (2006; 2007; 2009) model of preferences with stochastic reference points and loss aversion, as well as regret aversion theory, predicts the difference in effort provision observed in our *severe* and *lenient* treatments. Although the experimental method has limited external validity, this particular result may have direct practical implications in real-world contexts. Since intangibles are increasingly important in business organizations and knowledge-intensive jobs are difficult to assess, errors in evaluating employees' performance may well be a relevant phenomenon. Our research suggests that when a perfect assessment of employees' effort provision is not viable, it may be wise for the supervisor to be cautious when neglecting rewards and - in general - have a pro-employee bias in conducting her assessment, as this may well be beneficial for employees' motivation and effort provision in the longer term. The take-home message from the experiment is the following: if one must err, better err on the lenient side. So, it is better for the quality control manager to choose a relatively high number of defective products as a benchmark to assign rewards; better for the board to choose achievable goals in order to incentivize its CEO; better for the firm to pick the lenient manager for the HR office and better for the teacher to show the easy test instead of the impossible one.

8 Appendix

8.1 Parametric Analysis

We run a probit model to provide further evidence concerning the statistical significance as well as the economic meaning of the asymmetric effects highlighted through the non-parametric analysis reported in Section 5. The outcome variable is binary by design. It takes value 0 in the case of no provision of effort and value 1 when effort is actually exerted and goal realized. We analyze how the probability of exerting effort is influenced by the types of error that subjects are exposed namely Type I vs. Type II error (treatment dummies), different endowment levels (configuration dammy), individual risk aversion (discrete variable) and ordering effects in submitting reverse series of scenarios during the experiment (dummy variable).

As highlighted by the non-parametric analysis, even though both severity and leniency biases have a sizable and highly significant negative effect with respect to the fair case of unbiased appraisal, the relative effect of the Type I error is twice as large (0.41 vs. 0.22). With the type of error kept constant, the higher endowment configuration leads to a mild increase in the probability of observing an effort action.

Subjects' risk preferences and the sequential order in which scenarios are submitted to the participants do not affect the action choice.

Table 5: Probit model, marginal effect reported

Variable	Outcome: EFFORT§
Type I	-0.412***
Type II	-0.224***
High conf.	0.129**
Risk	-0.019
order	-0.001
obs.	254

§baseline No-error = 0.883 (constant)

Significance levels: $^{***}p < 0.01$,$^{**}p < 0.05$, $^{*}p < 0.10$

8.2 Translation of the Instructions

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

# **SITUATION – A –** $(T0_{High})$ # #

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

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

# **CONTROL QUESTIONS**

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

# **EFFORT CHOICE**

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

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

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

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

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

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

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

# **CONTROL QUESTIONS**

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

# **EFFORT CHOICE**

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

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

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

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

#   **SITUATION – C –**   ($T2_{High}$)   # #

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

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

# CONTROL QUESTIONS

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

# EFFORT CHOICE

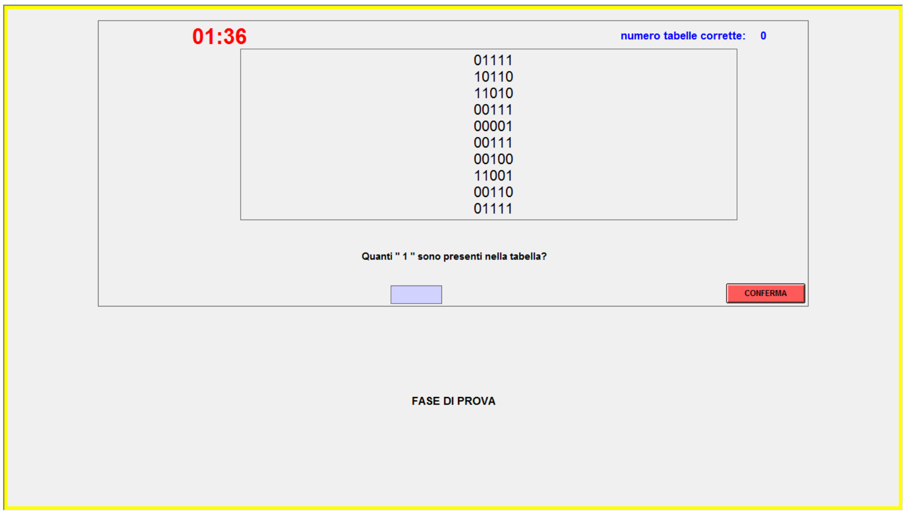
If Phase III of the experiment corresponds to SITUATION C as just described, will you perform the task or will you skip the task? Remember that:

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

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

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

8.3 Screenshots



01:36

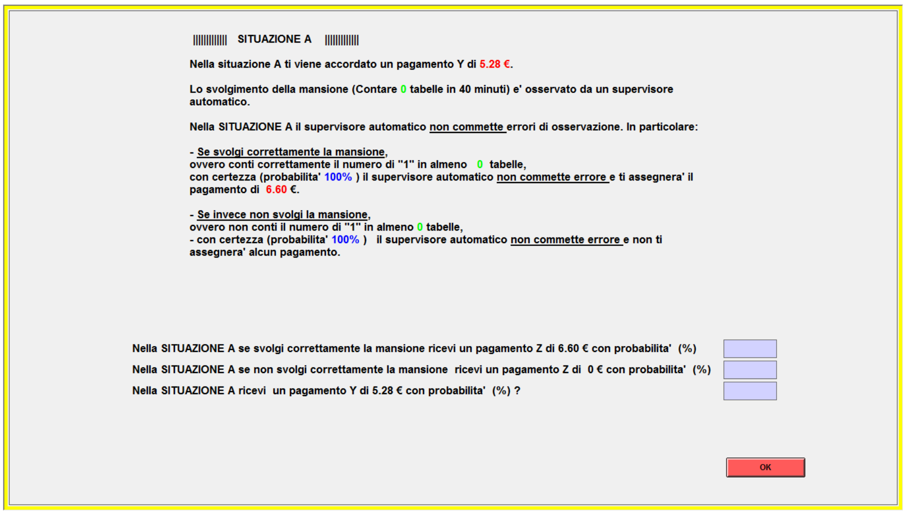
numero tabelle corrette: 0

```
01111
10110
11010
00111
00001
00111
00100
11001
00110
01111
```

Quanti " 1 " sono presenti nella tabella?

FASE DI PROVA

Figure 3: Screenshot of the Real Effort Task



|||||||||||| SITUAZIONE A ||||||||||||

Nella situazione A ti viene accordato un pagamento Y di 5.28 €.

Lo svolgimento della mansione (Contare 0 tabelle in 40 minuti) e' osservato da un supervisore automatico.

Nella SITUAZIONE A il supervisore automatico non commette errori di osservazione. In particolare:

- Se svolgi correttamente la mansione,
ovvero conti correttamente il numero di "1" in almeno 0 tabelle,
con certezza (probabilita' 100%) il supervisore automatico non commette errore e ti assegnera' il pagamento di 6.60 €.

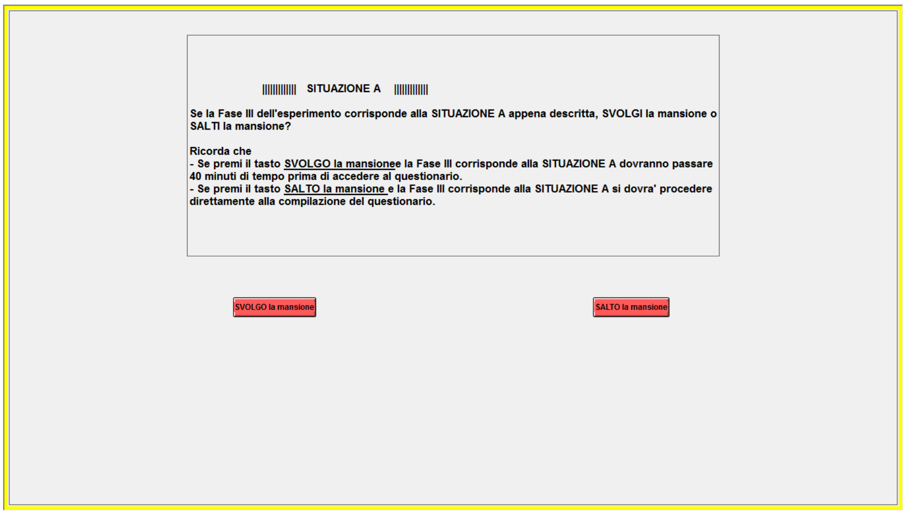
- Se invece non svolgi la mansione,
ovvero non conti il numero di "1" in almeno 0 tabelle,
- con certezza (probabilita' 100%) il supervisore automatico non commette errore e non ti assegnera' alcun pagamento.

Nella SITUAZIONE A se svolgi correttamente la mansione ricevi un pagamento Z di 6.60 € con probabilita' (%)

Nella SITUAZIONE A se non svolgi correttamente la mansione ricevi un pagamento Z di 0 € con probabilita' (%)

Nella SITUAZIONE A ricevi un pagamento Y di 5.28 € con probabilita' (%) ?

Figure 4: Screenshot of the Situation Presentation and Control Questions



|||||||||||| SITUAZIONE A ||||||||||||

Se la Fase III dell'esperimento corrisponde alla SITUAZIONE A appena descritta, SVOLGI la mansione o SALTI la mansione?

Ricorda che
- Se premi il tasto SVOLGO la mansionee la Fase III corrisponde alla SITUAZIONE A dovranno passare 40 minuti di tempo prima di accedere al questionario.
- Se premi il tasto SALTO la mansione e la Fase III corrisponde alla SITUAZIONE A si dovra' procedere direttamente alla compilazione del questionario.

Figure 5: Screenshot of the Effort Choice Phase

Table 6: **Risk Elicitation Lotteries**

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

References

Abeler, J., A. Falk, L. Götte, and D. Huffman (2011): "Reference points and effort provision," *The American Economic Review*, 101(2), 470–492.

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

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

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

Baker, G., R. Gibbons, and K. J. Murphy (1994): "Subjective Performance Measures in Optimal Incentive Contracts," *The Quarterly Journal of Economics*, 109(4), pp. 1125–1156.

Bell, D. (1983): "Risk premiums for decision regret," *Management Science*, 29(10), 1156–1166.

——— (1985): "Disappointment in decision making under uncertainty," *Operations research*, 33(1), 1–27.

Benartzi, S., and R. Thaler (1995): "Myopic Loss Aversion and the Equity Premium Puzzle," *The Quarterly Journal of Economics*, 110(1), 73–92.

Berger, J., C. Harbring, and D. Sliwka (2012): "Performance appraisals and the impact of forced distribution: An experimental investigation," *Management Science*, Online first.

Bolton, G. E., J. Brandts, and A. Ockenfels (2005): "Fair Procedures: Evidence from Games Involving Lotteries," *The Economic Journal*, 115(506), 1054–1076.

Bretz, R., G. Milkovich, and W. Read (1992): "The current state of performance appraisal research and practice: Concerns, directions, and implications," *Journal of Management*, 18(2), 321–352.

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

Caplan, A., D. Aadland, and A. Macharia (2010): “Estimating Hypothetical Bias in Economically Emergent Africa: A Generic Public Good Experiment,” *Agricultural and Resource Economics Review*, 39(2), 344–358.

Cardy, R., and G. Dobbins (1986): “Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance.,” *Journal of Applied Psychology; Journal of Applied Psychology*, 71(4), 672.

Charness, G., U. Gneezy, and M. Kuhn (2011): “Experimental methods: Between-subject and within-subject design,” *Journal of Economic Behavior & Organization*.

Charness, G., and P. Kuhn (2010): “Lab labor: What can labor economists learn from the lab?,” *NBER Working Paper*.

Cutler, B. L., and S. D. Penrod (1989): “Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence.,” *Journal of Applied Psychology*, 74(4), 650 – 652.

Elster, J. (1998): “Emotions and Economic Theory,” *Journal of Economic Literature*, 36(1), 47–74.

Enderer, F., and G. Manso (2009): “Is Pay-For-Performance Detrimental to Innovation?,” *UC Berkeley: Department of Economics, UCB.*

Falk, A., and E. Fehr (2003): “Why labour market experiments?,” *Labour Economics*, 10(4), 399 – 406, Special Issue on Labour Market Experiments.

Falk, A., and S. Gächter (2008): “Experimental Labour Economics,” *The New Palgrave Dictionary of Economics, Basingstoke: Palgrave Macmillan.*

Falk, A., and J. J. Heckman (2009): “Lab Experiments Are a Major Source of Knowledge in the Social Sciences,” *Science*, 326(5952), 535–538.

Fehr, E., U. Fischbacher, et al. (2003): “The nature of human altruism,” *Nature*, 425(6960), 785–791.

Fischbacher, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10(2), 171–178.

Fisher, C. (1979): "Transmission of positive and negative feedback to subordinates: A laboratory investigation.," *Journal of Applied Psychology*, 64(5), 533.

Genesove, D., and C. Mayer (2001): "Loss aversion and seller behavior: Evidence from the housing market," *The Quarterly Journal of Economics*, 116(4), 1233–1260.

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

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

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

Gomes, F. (2005): "Portfolio Choice and Trading Volume with Loss-Averse Investors*," *The Journal of Business*, 78(2), 675–706.

Grechenig, K., A. Nicklisch, and C. Thoni (2010): "Punishment despite reasonable doubt - A public goods experiment with sanctions under uncertainty," *Journal of Empirical Legal Studies*, 7(4), 847–867.

Greiner, B. (2004): *An online recruitment system for economic experiments*, Forschung und wissenschaftliches Rechnen. GWDG Bericht 63, Göttingen, 2003 edn.

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

Gul, F. (1991): "A theory of disappointment aversion," *Econometrica: Journal of the Econometric Society*, pp. 667–686.

Gürtler, O., and C. Harbring (2010): "Feedback in Tournaments under Commitment Problems: Experimental Evidence," *Journal of Economics & Management Strategy*, 19(3), 771–810.

Hayashi, T. (2008): “Regret aversion and opportunity dependence,” *Journal of Economic Theory*, 139(1), 242–268.

Hendry, S. H., D. R. Shaffer, and D. Peacock (1989): “On testifying in one’s own behalf: Interactive effects of evidential strength and defendant’s testimonial demeanor on mock jurors’ decisions.,” *Journal of Applied Psychology*, 74(4), 539 – 545.

Higgins, C., T. Judge, and G. Ferris (2003): “Influence tactics and work outcomes: a meta-analysis,” *Journal of Organizational Behavior*, 24(1), 89–106.

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

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

Holt, C., and S. Laury (2002): “Risk Aversion and Incentive Effects,” *The American Economic Review*, 92(5), 1644–1655.

Ilgen, D. R., T. R. Mitchell, and J. W. Fredrickson (1981): “Poor performers: Supervisors’ and subordinates’ responses,” *Organizational Behavior and Human Performance*, 27(3), 386 – 410.

Jawahar, I., and C. Williams (1997): “Where all the children are above average: The performance appraisal purpose effect,” *Personnel Psychology*, 50(4), 905–925.

Jensen, M., and W. Meckling (1976): “Theory of the firm: Managerial behavior, agency costs and ownership structure,” *Journal of financial economics*, 3(4), 305–360.

Judge, T., and G. Ferris (1993): “Social context of performance evaluation decisions,” *Academy of Management Journal*, 36(1), 80–105.

Kagel, J. H., and A. E. Roth (1995): *The Handbook of Experimental Economics*. Princeton University Press.

Kahneman, D., J. Knetsch, and R. Thaler (1990): “Experimental tests of the endowment effect and the Coase theorem,” *Journal of political Economy*, pp. 1325–1348.

Kahneman, D., and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47(2), 263–292.

——— (1984): "Choices, values, and frames," *American Psychologist*, 39(4), 341–350.

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

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

Kane, J., H. Bernardin, P. Villanova, and J. Peyrefitte (1995): "Stability of rater leniency: Three studies.," *Academy of Management Journal*, 38(4), 1036–1051.

Kaplow, L. (1994): "The Value of Accuracy in Adjudication: An Economic Analysis," *Journal of Legal Studies*, 23(1), 307–401.

——— (2011): "Optimal Proof Burdens, Deterrence, and the Chilling of Desirable Behavior," *American Economic Review*, 101, 277–280.

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

Klimoski, R., and L. Inks (1990): "Accountability Forces in Performance Appraisal," *Organizational Behavior and Human Decision Processes*, 45(2), 194–208.

Kőszegi, B., and M. Rabin (2006): "A Model of Reference-Dependent Preferences," *The Quarterly Journal of Economics*, 121(4), 1133–1165.

——— (2007): "Reference-dependent risk attitudes," *The American Economic Review*, 97(4), 1047–1073.

——— (2009): "Reference-dependent consumption plans," *The American Economic Review*, 99(3), 909–936.

Laffont, J., and D. Martimort (2002): *The theory of incentives: the principal-agent model.* Princeton Univ Pr.

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

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

LOOMES, G., AND R. SUGDEN (1982): "Regret theory: An alternative theory of rational choice under uncertainty," *The Economic Journal*, 92(368), 805–824.

——— (1986): "Disappointment and dynamic consistency in choice under uncertainty," *The Review of Economic Studies*, 53(2), 271–282.

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

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

MAESTRI, L. (2012): "Bonus Payments versus Efficiency Wages in the Repeated Principal-Agent Model with Subjective Evaluations," *American Economic Journal: Microeconomics*, 4(3), 34–56.

MOERS, F. (2005): "Discretion and bias in performance evaluation: the impact of diversity and subjectivity," *Accounting, Organizations and Society*, 30(1), 67 – 80.

NIEKEN, P. (2010): "On the Choice of Risk and Effort in Tournaments - Experimental Evidence," *Journal of Economics & Management Strategy*, 19(3), 811–840.

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

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

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

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

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

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

RIZZOLLI, M., AND M. SARACENO (Forthcoming): “Better that ten guilty persons escape: punishment costs explain the standard of evidence,” *Public Choice*, pp. 1–17, 10.1007/s11127-011-9867-y.

RIZZOLLI, M., AND L. STANCA (2012): “Judicial errors and deterrence: theory and experimental evidence,” *Journal of Law and Economics*, Forthcoming.

SAUTMANN, A. (forthcoming): “Contracts for agents with biased beliefs: Some theory and an experiment,” *American Economic Journal: Microeconomics*.

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

SCHNEDLER, W., AND . J. O. E. . M. S. .-. VADOVIC, RADOVAN. . "LEGITIMACY OF CONTROL (2011): “Legitimacy of Control,” *Journal of Economics & Management Strategy*, 20(4), 985–1009.

SEBALD, A., AND M. WALZL (2012): “Subjective performance evaluations and reciprocity in principal-agent relations,” Discussion paper, Working Papers in Economics and Statistics - 2012-15 - University of Innsbruck.

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

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

THALER, R., AND E. JOHNSON (1990): “Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice,” *Management science*, 36(6), 643–660.

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

Thomas, J., and H. Meeke (2010): "Rater error," *Corsini Encyclopedia of Psychology.*

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

Tversky, A., and D. Kahneman (1992): "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and Uncertainty*, 5(4), 297–323.

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

Van Dijk, W., and M. Zeelenberg (2002): "Investigating the appraisal patterns of regret and disappointment," *Motivation and Emotion*, 26(4), 321–331.

Varma, A., A. Denisi, and L. Peters (1996): "Interpersonal affect and performance appraisal: A field study," *Personnel Psychology*, 49(2), 341–360.

Villanova, P., H. Bernardin, S. Dahmus, and R. Sims (1993): "Rater leniency and performance appraisal discomfort," *Educational and psychological measurement*, 53(3), 789–799.

Zeelenberg, M., W. Van Dijk, A. Manstead, and J. van der Pligt (2000): "On bad decisions and disconfirmed expectancies: The psychology of regret and disappointment," *Cognition & Emotion*, 14(4), 521–541.