



Loss averse agents and lenient supervisors in performance appraisal[☆]



Lucia Marchegiani^a, Tommaso Reggiani^{b,*}, Matteo Rizzolli^c

^a Department of Management, University of Rome 3, via Silvio D'Amico 77, 00145 Rome, Italy

^b School of Business and Economics, LUMSA University-Rome (Palermo campus) & IZA, via Filippo Parlato 65, 90145 Palermo, Italy

^c Department of Economics, LUMSA University-Rome, via Pompeo Magno 22, 00191 Rome, Italy

ARTICLE INFO

Article history:

Received 30 April 2015

Received in revised form 26 July 2016

Accepted 27 July 2016

Available online 17 August 2016

JEL classification:

C91

M50

J50

Keywords:

Performance appraisal

Type I and Type II errors

Leniency bias

Severity bias

Economic experiment

Loss aversion

Reference-dependent preferences

ABSTRACT

A consistent empirical literature shows that in many organizations supervisors systematically overrate their employees' performance. Such leniency bias is at odds with the standard principal-agent model and has been explained with causes that range from social interactions to fairness concerns and to collusive behavior between the supervisor and the agent. We show that the principal-agent model, extended to consider loss-aversion and reference-dependent preferences, predicts that the leniency bias is comparatively less detrimental to effort provision than the severity bias. We test this prediction with a laboratory experiment where we demonstrate that failing to reward deserving agents is significantly more detrimental than rewarding undeserving agents. This offers a novel explanation as to why supervisors tend to be lenient in their appraisals.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Supervisors¹ routinely evaluate agents' performance without directly observing their efforts. Consequently, evaluation errors inevitably arise, generally undermining agents' incentives. These errors take two forms: (i) a supervisor (she) may assess low performance when in fact the agent (he) is duly exerting effort and thus she does not reward a deserving agent

[☆] We would like to thank Gani Aldashev, Steffen Altmann, Benito Arrunada, Jeff Butler, Susan Cartwright, Anna Grandori, Fabrizio Galeotti, Andrea Ichino, Patrick Kampkötter, Michael Neugart, Matteo Ploner, Dirk Sliwka, Gari Walkowitz, Peter Werner, Giorgio Zanarone, the Associate Editor and the anonymous referee for their helpful comments. For the same reason we would also like to thank the participants at seminars at University of Bergen, University of Bologna, Free University of Bozen-Bolzano, University of Cologne, CUNEF (Madrid), and conference participants Academy of Management 2012 (Boston), EALE 2013 (Turin). 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 would also like to 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 "University of Cologne – Design & Behavior: Economic Engineering of Firms and Markets" (FOR 1371).

* Corresponding author.

E-mail addresses: lmarchegiani@uniroma3.it (L. Marchegiani), t.reggiani@lumsa.it (T. Reggiani), m.rizzolli@lumsa.it (M. Rizzolli).

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

(this is defined as a Type I error²); or (ii) a supervisor may observe high performance when in fact the agent is not exerting effort and therefore she may reward an undeserving agent (this is a Type II error). Systematic biases in performance appraisal usually emerge in two forms: *leniency bias* occurs when the supervisor assesses high performance “too often”, while *severity bias* occurs when the supervisor assesses low performance “too often”. A well-established and consistent empirical literature shows that in many organizations supervisors have systematic leniency biases (Prendergast, 1999). 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; Vafai, 2010; Thiele, 2013, and more cited below). These papers always focus on the principal’s side. In this paper we instead focus on the marginal impact of the two errors on agents’ incentives to exert effort, and we explore how severe and lenient appraisal compare to one another in undermining agents’ performance. To our knowledge there is no experiment that compares how agents behave under each of the two biases. We show that agents with reference-dependent preferences are comparatively more motivated under leniency bias. This theoretical result is supported by the lab evidence we provide. 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 explanations.

In the paper we extend a standard version of the principal-agent model where severity and leniency biases are stylized. Under standard assumptions concerning risk aversion and separability of utility and effort, the model predicts that 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. Following the path set by some other recent theory papers (Daido and Itoh, 2007; Herweg et al., 2010; Armantier and Boly, 2015; Daido and Murooka, 2016), we show that, under reference-dependent preferences à-la Koszegi and Rabin (2006), leniency bias is comparatively less detrimental to effort provision.

By means of a lab experiment we then discriminate among the predictions of the two models. In our laboratory experiment, subjects carry out an effort task that is initially paid piece-rate. In the following phase they must declare whether they would accept each of three hypothetical contracts: the *fair* contract has no appraisal error while the other two contracts imply either a *severe* or a *lenient* bias in appraisal. One of the three contracts is then randomly picked and subjects who accepted the contract can then carry out the task. Our main finding shows that failing to reward a deserving agent under a severe contract is significantly more detrimental to effort provision than rewarding an undeserving agent under a lenient contract, as the model with reference-dependent preferences predicts. We also demonstrate that regret aversion predicts the same behavioral pattern.

2. Literature review

Several streams of literature, including “Personnel Economics” (Lazear, 1999), “Agency Theory” (Hölmstrom, 1979; Aron and Olivella, 1994; Prendergast, 1999; Maestri, 2012), and “Organizational Studies” (Steers et al., 2004), deal with errors in performance appraisal. Most organizations rely on subjective performance appraisal in order to motivate their employees (Prendergast and Topel, 1993; Prendergast, 1999; MacLeod, 2003; Kambe, 2006; Maestri, 2012). These measures, used alone (Bull, 1987; MacLeod and Malcomson, 1989; Levin, 2003) or in combination with objective measures (Schmidt and Schnitzer, 1995; Pearce and Stacchetti, 1998; Bol and Smith, 2011), are nevertheless prone to errors. Among the most important types of errors classified by the literature on subjective appraisal there is *leniency bias* and *severity bias*.³ These errors reduce the scope of appraisal because they restrict the range of useful measures of performance, and thus weaken the incentive (MacLeod, 2003). A number of papers look at the origins of these appraisal errors with a particular focus on leniency bias. Such errors can be generated by unconscious cognitive and behavioral biases in the observing, elaborating, or recalling of rater performance information or in the process of generating the appraisal rating (Prendergast, 2002) or by feelings such as empathy and affection (Cardy and Dobbins, 1986; Varma et al., 1996) and manipulation (Higgins et al., 2003). The overconfident beliefs of an agent may also cause misalignment between the agent’s self-assessment and the supervisor’s performance appraisal (Maestri, 2012; Sautmann, 2013). Furthermore, a supervisor may find it convenient to provide lenient evaluations either because she colludes with the agent (See Tirole, 1986; Prendergast and Topel, 1996; Strausz, 1997; Vafai, 2010; Thiele, 2013; De Chiara and Livio, 2015), because of social interactions (Judge and Ferris, 1993; Grund and Przemek, 2012; Giebe and Guertler, 2012), or because of the desire to compensate for some uncontrollable stochastic effects that may undermine the rater’s evaluations (Bol and Smith, 2011). All these potential explanations for leniency bias are excluded by our design.

Empirical evidence of the existence of leniency bias has been long provided (e.g. Kingsbury, 1922; Thorndike, 1949; Landy and Farr, 1980). For instance Schoorman (1988) reports that supervisors who were involved in the hiring decision and

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 may thus 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).

³ Other rater’s errors are: (i) *central tendency* error derives from the 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 rater has affinity with the rater; (v) *recency error* happens when recent performance is given too much weight as opposed to early performance within a given time interval, and on the opposite (vi) *first impression error* when early performance is given too much weight as opposed to more recent performance within a given time interval (See Thomas and Meeke, 2010 on classification of rater’s errors. See Rabin and Schrag, 1999 specifically on first impression bias).

who viewed the applicant as favorable, subsequently tended to give more favorable performance ratings. Bretz et al. (1992) show that while most organizations' evaluation systems use five level scales to differentiate employee performance, very often only the higher three of such levels are used, with two-thirds of employees in any of the organizations' workforce that have been analyzed rated falling in the top two performance levels. Moers (2005) finds that, on average, performance ratings expressed on the subjective dimension are higher than performance ratings on a benchmark objective dimension. This result thus corroborates the prediction that supervisors provide more lenient ratings when subjectively assessing performance. Bol (2011) finds evidence of leniency bias in subjectively performed assessments within a large bank in the Netherlands. Bol's work studies the impact of such rating biases on workers' incentives and shows that leniency bias may increase employees' performance to a certain extent. This finding goes against the standard prediction that managers' performance evaluation biases are detrimental for compensation contracting (Landy and Farr, 1980; Rynes et al., 2005).

Finally, a small number of papers deal with leniency bias in a lab context. Most of the experiments in personnel economics and social psychology try to reproduce the conditions in the lab under which leniency bias happens to be more common than severity bias (Berger et al., 2013). 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 aligned with the agents' performance (Ilgen et al., 1981), or because they believe that their evaluations will be used against the interests of the ratees (Villanova et al., 1993; Kane et al., 1995). While the focus of this literature is only on the rater's behavior, this paper complements this stream of literature inasmuch as it focuses on agents' behavior when exposed to leniency bias or to severity bias.

Our theoretical contribution stands within a recent stream of papers that have extended the principal-agent model to account for reference-dependent preferences, following the seminal work of Koszegi and Rabin (2006, 2007). This literature has already produced new interesting insights. Herweg et al. (2010) show that, in presence of reference dependent preferences, the optimal contract is a binary payment scheme even when a rich performance measure is available. Along the same lines Daido et al. (2013) show that assigning the task to a single loss averse agent in all possible states can be optimal even when the principal can write a contingent contract at no cost. Armantier and Boly (2015) show both theoretically and experimentally that the performance of agents with reference dependent preference can be increased by the combined use of bonuses and penalties. Daido and Murooka (2016) show that, because of agent loss aversion, the principal can induce higher performance by stochastically compensating agents' low performance. Finally, Daido and Itoh (2007) show that agents with reference dependent preferences who possess high expectations about their performance can be induced to choose high effort with low-powered incentives; this explains the so called "Pygmalion" and "Galatea" effect. Our model instead explains *leniency bias* often found in the context of performance appraisal.

3. The model

We consider a model with one supervisor and one agent where the supervisor cannot contract the agent's effort and only the agent's final output is observable.⁴ This is a standard principal-agent model with objective errors and moral hazard. However, its implications for and predictions on the agent's behavior are also valid in the case of subjective errors. Finally, we borrow the nomenclature of leniency and severity biases from the subjective performance appraisal literature and implement it in 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: $e \in \{0, 1\}$. The agent's choice of action is private information and unobservable to the principal. Effort implies disutility for the agent. We define this disutility as a generic function of the level of effort $g(e)$, which can be normalized to $g(0) = g_0 = 0$ and $g(1) = g_1 = g$.

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 $\{q, \bar{q}\}$. We normalize $q = 0$ and assume \bar{q} to be positive. Another way of interpreting this is that the principal fixes a performance target equal to \bar{q} and considers any performance below that level as zero. Effort influences performance in the following way: $Pr(\tilde{q} = \bar{q} | e = 0) = \beta$ and $Pr(\tilde{q} = \bar{q} | e = 1) = 1 - \alpha$ with $1 - \alpha > \beta$. The assumption that $1 - \alpha > \beta$ implies that effort increases performance in the sense of *first-order stochastic dominance* since $Pr(\tilde{q} \leq q^* | e)$ decreases with e for any given performance q^* .

The agent receives a wage w that can take the following two values: the baseline wage w_0 whenever performance is q , and the rewarding wage w_r when performance is \bar{q} with $w_r \geq w_0 \geq 0$.⁵

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 \bar{q} which then triggers the rewarding wage w_r . Note that the supervisor may commit the following evaluation errors:

⁴ An alternative interpretation of the condition of unobservability of the variable "effort" is that this variable describes a complex good whose quality cannot be determined by a court; thus a contract cannot depend on the realization of this variable.

⁵ The agent is assumed to have limited liability.

Type I error: with probability α , the agent that exerts effort ($e = 1$) does not meet the performance target \bar{q} and thus does not receive his due reward w_r .
 Type II error: with probability β , the agent that exerts no effort ($e = 0$) meets the performance target \bar{q} and is thus undeservedly rewarded with w_r ⁶.

The sum $(1 - \alpha - \beta)$ defines the *accuracy* of the performance appraisal. More accuracy implies a smaller sum of the two errors, and therefore a larger ability of the principal to discriminate between the agent exerting effort and the one shirking. Note that accuracy can be kept constant with very different error trade-offs. There could be a *lenient* setting characterized by a certain $\alpha_{small}, \beta_{large}$ tradeoff that has the same level of accuracy as a *severe* setting characterized by a $\alpha_{large}, \beta_{small}$ tradeoff as long as $\alpha_{small} + \beta_{large} = \alpha_{large} + \beta_{small}$. We will exploit this implication of the model to characterize the lenient and severe biases as our treatment conditions in the experiment.

3.1. Principal's and agent's participation constraints

In this simple model the principal is always interested in harvesting high effort from the agent. This is because (i) the probability that performance is zero when effort is high is smaller than the probability that performance is zero when effort is zero – $\Pr(\bar{q} \leq q | e = 1) = \alpha < 1 - \beta = \Pr(\bar{q} \leq q | e = 0)$ – and also (ii) the probability that performance is no higher than \bar{q} is equal to 1 both when effort is 0 and when it is 1 – $\Pr(\bar{q} \leq \bar{q} | e = 1) = 1 = \Pr(\bar{q} \leq \bar{q} | e = 0)$. These two properties suggest 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 $u(\cdot)$ increases in performance. Indeed, the principal's payoff when the agent exerts positive effort is $(1 - \alpha)u(\bar{q}) + \alpha u(q)$ and it is larger than the payoff when the agent exerts no effort $\beta u(\bar{q}) + (1 - \beta)u(q)$ as long as $u(\cdot)$ is increasing.

Note that $(1 - \alpha)u(\bar{q}) + \alpha u(q) = \beta u(\bar{q}) + (1 - \beta)u(q) + (1 - \alpha - \beta)(u(\bar{q}) - u(q))$ (Laffont and Martimort, 2002, p. 149). To highlight our main results in a simple manner, and following Daido and Murooka (2016), we assume away the agent's participation constraint. This is possible because the agent's *participation constraint* is defined by the agent's reservation utility, which corresponds to his best outside option $U(e \equiv 1, w) \geq u$. This ensures that if the agent exerts effort, it yields at least his outside opportunity utility level. By design the outside option of subjects participating to the experiment, once sitting in the lab, is the expected wage when exerting no effort $U(e = 0, w)$. Therefore, for the purpose of our experiment, the participation constraint coincides with the incentive constraint.

3.2. Agent's utility with reference-dependent preferences

The key assumption of this principal-agent model is that agents have expectation-based reference-dependent preferences à-la Koszegi and Rabin (2006, 2007) so that their overall utility consists of intrinsic consumption utility, which depends on the wage ($v(w)$), on the psychological gain-loss utility $\mu(v(w)|\hat{w})$ and on the disutility of effort $g(e)$ (Daido and Itoh, 2007; Herweg et al., 2010; Armantier and Boly, 2015; Daido and Murooka, 2016). Following the literature, we assume separability in effort and income and across dimensions. If the agent actually exerts effort e and receives wage w , he thus obtains overall utility $U(w, e) = v(w) - g(e) + \mu(v(w)|\hat{w})$, where $\mu(\cdot)$ is a gain-loss function that satisfies the assumptions of Koszegi and Rabin (2006) (see Assumptions A0–A4, pg. 1139). Crucial to their theory is the definition of the reference point \hat{w} , which depends on expectations and is stochastic if the outcome is stochastic. Koszegi and Rabin (2006) model \hat{w} as a distribution of stochastic reference points drawn from a distribution, 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 Koszegi and Rabin (2007), preferences are assumed to be linear in probabilities and we only consider the coefficient of loss aversion λ . In this model, the agent has two consumption dimensions: effort cost and wage. We assume that the agent feels psychological gain-loss only in the wage dimension. This is because the agent's expected and actual effort choice coincide, and therefore there is therefore neither a gain nor a loss in the effort dimension.⁷ The gain-loss function is defined over income as

$$\mu(w|\hat{w}) = \begin{cases} \eta(v(w) - v(\hat{w})) & \text{if } v(w) - v(\hat{w}) \geq 0 \\ \eta\lambda(v(w) - v(\hat{w})) & \text{if } v(w) - v(\hat{w}) < 0 \end{cases}$$

⁶ The derivation of the probabilities of errors from the definition of \bar{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 effort and with none. Therefore with low \bar{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). Thus, there is little incentive for the agent to invest effort. The more the performance target increases, the smaller the probability of Type II errors becomes; thus, switching to effort becomes convenient. At some intermediate level of \bar{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 errors (i.e., not meeting the target when exerting effort) is very high and therefore there is little incentive to exert effort.

⁷ On this choice we follow Herweg et al. (2010) and Armantier and Boly (2015). Daido and Murooka (2016) also model the disutility of effort with respect to a reference point and show how the deterministic nature of the choice of effort implies that the gain-loss effort dimension is canceled out.

where $\eta \geq 0$ represents the weight of the gain-loss payoff on the utility function, and $\lambda \geq 1$ is the coefficient of loss aversion.⁸ For convenience, and following the relevant literature, we assume from now on that $\eta = 1$ and linear utility⁹ so that $v(w_r) - v(w_0) = |w_r - w_0| = \Delta w$.

In order to derive the incentive constraint we need to set the equilibrium concept to be used. The choice-acclimating personal equilibrium (CPE) (Koszegi and Rabin, 2007) is defined as the stochastic choice that is best given that also determines the reference point. This equilibrium concept seems appropriate in our setting since the decision of the agent ($e = 0, 1$) is determined long before the outcome and payment occur, and hence the agent updates his belief about the action he took before the outcome is realized¹⁰. Because the agent knows that his belief will change on the basis of his chosen action before the outcome and payment occur, he takes this change into account when he decides what action to take (Daido and Murooka, 2016). Therefore, the action of each agent himself determines his reference point under CPE, and the utility when exerting effort is thus

$$U(e = 1, w) = \alpha [v(w_0) + (\alpha(v(w_0) - v(w_0)) + (1 - \alpha)\lambda(v(w_0) - v(w_r)))] + (1 - \alpha) [v(w_r) + (\alpha(v(w_r) - v(w_0)) + (1 - \alpha)(v(w_r) - v(w_r)))] - g$$

Similarly the utility when exerting no effort is

$$U(e = 0, w) = (1 - \beta) [v(w_0) + (\beta\lambda(v(w_0) - v(w_r)) + (1 - \beta)(v(w_0) - v(w_0)))] + \beta [v(w_r) + (\beta(v(w_r) - v(w_r)) + (1 - \beta)(v(w_r) - v(w_0)))]$$

The Incentive Constraint is satisfied whenever the returns of exerting effort loom larger than the returns of exerting no effort $U(w, e = 1) \geq U(w, e = 0)$ and thus, rearranging, the incentive constraint is satisfied whenever

$$\underbrace{(1 - \alpha - \beta) \Delta w}_{\text{Consumption IC}} + \underbrace{(1 - \lambda)(1 - \alpha - \beta)(\alpha - \beta)\Delta w}_{\text{gain-loss IC}} \geq g \tag{1}$$

The first part of the incentive constraint depends on the utility of the monetary net reward multiplied by the level of accuracy, while the second part of the incentive constraint crucially depends on our assumptions concerning loss aversion.

Let us first examine a case in which the agent suffers no loss aversion ($\lambda = 1$). The incentive constraint simplifies to $\Delta w (1 - \alpha - \beta) \geq g$ and shows 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 the one hand, the larger the probability of β (being rewarded undeservingly), the larger the returns from not exerting effort. On the other hand, the larger the probability of α (not being rewarded when deserving it), the smaller the returns of exerting effort.

Proposition 1. *With no loss aversion, neglecting due rewards to deserving agents (α) is equally detrimental to effort provision as rewarding undeserving agents (β).*

Notice that $\frac{\partial IC}{\partial \alpha} = \frac{\partial IC}{\partial \beta} = -(\Delta w)$. Type-I and type-II errors both increase the incentive constraint (in the sense that a larger net reward $w_r - w_0$ is needed in order to obtain the same amount of effort g) and they do so in the same way.

Notice that if at the margin both errors are equally detrimental to deterrence, then a *lenient* setting and a *severe* setting should harvest the same level of effort from agents as long as the level of accuracy remains the same.

3.3. Loss aversion

The degree of loss aversion is modeled through λ . Empirical estimations usually set $\lambda \simeq 2$ for the general population, although there is a certain degree of variance on individual loss aversion. Notice that once we set $\lambda > 1$ the symmetric prediction of Proposition 1 concerning the relative impact of the two errors no longer holds.

Proposition 2. *With loss aversion, neglecting due rewards to deserving agents (α) is more detrimental to effort provision than rewarding undeserving agents (β).*

⁸ Loss aversion is a well known behavioral trait and its impact on agent's choices has been widely studied since the seminal work of Kahneman and Tversky (Kahneman and Tversky (1979, 1984), Tversky and Kahneman (1991)). 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.

⁹ Linear utility can be easily assumed for small stakes such as those involved in our experiment (See Rabin, 2000).

¹⁰ The timing of the model is the following: (a) The principal sets a pair of errors (α, β) that define the accuracy of the appraisal, (b) The principal offers a wage payment scheme subject to the limited liability, (c) The agent chooses his action subject to his incentive constraint, (d) The performance level takes one of the two values $\{q, \bar{q}\}$ and the wage is paid accordingly.

Table 1

Table of treatment parameters.

Conf.	Treatment	Type I α	Type II β	Accuracy $1 - \alpha - \beta$	Baseline w_0	Reward w_r
Low	$T0_L$ – Fair	0	0	1	€0	€6.60
	$T1_L$ – Severe	4/5	0	1/5	€0	€6.60
	$T2_L$ – Lenient	0	4/5	1/5	€0	€6.60
High	$T0_H$ – Fair	0	0	1	€5.28	€11.88
	$T1_H$ – Severe	4/5	0	1/5	€5.28	€11.88
	$T2_H$ – Lenient	0	4/5	1/5	€5.28	€11.88

Note that $\frac{\partial IC}{\partial \alpha} = \Delta w [(1 - \lambda)(1 - 2\alpha) - 1]$ and $\frac{\partial IC}{\partial \beta} = \Delta w [(1 - \lambda)(2\beta - 1) - 1]$ and that $\frac{\partial IC}{\partial \alpha} > \frac{\partial IC}{\partial \beta}$ as long as $1 - \alpha - \beta > 0$, which is always the case because of the assumption of first order stochastic dominance.

Notice also that $\frac{\partial IC}{\partial \alpha} \leq 0$, at least for conventional values of $\lambda \simeq 2$. When $\alpha = 0$ then $\frac{\partial IC}{\partial \alpha} = -\lambda \Delta w$, while when $\alpha = 1$ then $\frac{\partial IC}{\partial \alpha} = (\lambda - 2)\Delta w$. Moreover $\frac{\partial^2 IC}{\partial \alpha^2} = -2(1 - \lambda)\Delta w$.

Corollary 3. Neglecting due rewards to deserving agents (α) is always detrimental to performance unless loss aversion is particularly high $\lambda > 2$ and α is close to 1

Finally, observe that $\frac{\partial IC}{\partial \beta} \leq 0$ for the same relevant parameters. And therefore when $\beta = 0$, then $\frac{\partial IC}{\partial \beta} = (\lambda - 2)\Delta w$ while when $\beta = 1$, then $\frac{\partial IC}{\partial \beta} = -\lambda \Delta w$. Moreover $\frac{\partial^2 IC}{\partial \beta^2} = 2(1 - \lambda)\Delta w$.

Corollary 4. Rewarding undeserving agents (β) is always detrimental to performance unless loss aversion is particularly high $\lambda > 2$ and β is close to 0

3.4. Model parameters for the experimental treatments

To test the behavioral implications of the model, we devise three contracts (*fair*, *severe* and *lenient*) in two different endowment configurations (*high* and *low*) described in Table 1. The first variation concerns the accuracy of appraisal: Under the *fair* contract, accuracy is maximal, while accuracy is heavily biased by Type I errors under the *severe* contract ($\alpha_{large}, \beta_{small}$) and by Type II errors under the *lenient* contract ($\alpha_{small}, \beta_{large}$). Accuracy is the same under both *severe* and *lenient* contracts. 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 $w_r - w_0 = €6.60$.¹¹

Fair contract ($T0_L$ and $T0_H$). There are no evaluation errors ($\alpha, \beta = 0$). This contract is “fair” in the sense that the agent gets what he deserves.

Severe contract ($T1_L$ and $T1_H$). In T1 there are no Type II errors ($\beta = 0$) but the probability of Type I error is significant ($\alpha = 0.8$).¹² 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. This contract is “severe” in the sense that the deserving agent very often does not get what he deserves.

Lenient contract ($T2_L$ 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. This contract is “lenient” in the sense that the undeserving agent very often gets what he does not deserve.

Table 2 computes the incentive constraint without loss aversion (column 3) and with the gain-loss component in column 4. Some theory predictions to be tested with the lab experiment can now be formulated.

Prediction 1: Type-I errors (α) decrease performance with and without loss aversion. By comparing Fair and Severe treatments, one can immediately see that the incentive constraint decreases for any value of λ .

¹¹ This amount is paid for a task lasting 40 min and is thus proportional to an hourly wage of €10. The two configurations – *low* and *high* – stylize different simple bonus contracts 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.

¹² 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 world.

Table 2
Incentive constraints with loss aversion and without it.

Conf.	Treatment	IC with no loss aversion $(1 - \alpha - \beta)\Delta w$	IC with loss aversion $(1 - \alpha - \beta)\Delta w + (1 - \lambda)$ $[\alpha(1 - \alpha) - \beta(1 - \beta)]\Delta w$
Low	$T0_L$ – Fair	$v(\text{€}6.60)$	–
	$T1_L$ – Severe	$\frac{1}{5}v(\text{€}6.60)$	$\dots + (1 - \lambda)\frac{4}{25}v(\text{€}6.60)$
	$T2_L$ – Lenient	$\frac{1}{5}v(\text{€}6.60)$	$\dots + (\lambda - 1)\frac{4}{25}v(\text{€}6.60)$
High	$T0_H$ – Fair	$v(\text{€}11.88) - v(\text{€}5.28)$	–
	$T1_H$ – Severe	$\frac{1}{5}(v(\text{€}11.88) - v(\text{€}5.28))$	$\dots + (1 - \lambda)\frac{4}{25}v(\text{€}6.60)$
	$T2_H$ – Lenient	$\frac{1}{5}(v(\text{€}11.88) - v(\text{€}5.28))$	$\dots + (\lambda - 1)\frac{4}{25}v(\text{€}6.60)$

Prediction 2: Type-II errors (β) decrease performance without loss aversion and with loss-aversion if $\lambda < 6$. By comparing Fair and Lenient treatments, it can be observed that the incentive constraint decreases if $\lambda < 6$. Typical λ observed in field and lab experiments is between 1.5 and 3; therefore, prediction 2 holds for all reasonable values of λ .

Prediction 3: Under Loss Aversion type-I errors (α) decrease performance more than type-II errors (β). By comparing the incentive constraint under Severe and Lenient treatments, in each configuration, the incentive constraint is the same for the *severe* and *lenient* treatments if there is no loss aversion but diverges when there is loss aversion. In particular, the $IC_{Severe} < IC_{Lenient}$ if $\lambda < 2$

Before proceeding to test the theoretical predictions, in the following section we illustrate the experimental protocol.

4. Experimental protocol

The research question of the present work deals with a variable, appraisal errors, that is practically impossible to observe in the field because of the unobservability of effort and because of the stochastic relation between performance and effort. In the lab, however, we can both superimpose exogenous probabilities of errors and we can perfectly observe effort. This allows us to precisely identify the impact of errors on effort provision. A recurrent objection to our design concerns the imposition of exogenous error probabilities. This was an intentional choice made in order to precisely control the level of accuracy in comparing lenient and severe appraisals.¹³

The experimental protocol is divided into three different phases: Preliminary **Phase I** was used to elicit individuals' risk attitudes via a standard incentivized task (for example Bruhin et al., 2010; Abdellaoui et al., 2011; Vieider et al., 2015) that assesses certainty equivalents (CE) of risky prospects. Subjects were asked to choose between binary lotteries and different sure amounts (see Table 8 in the online material). Individuals typically prefer the lottery when sure amounts are low, and switch over to preferring the sure amount as the latter gets larger. The switching point is called the certainty equivalent (CE) of the lottery and it is a measure of the individual attitude towards risk taking.

Only one subject was paid for this task. The identity of the subject, the chosen lottery to be paid, and the outcome of the lottery were all randomly determined at the end of the experiment in order to prevent revenue effects during the main treatments.

In Phase II, individual ability in the default task was measured. Following Abeler et al. (2011), the real effort task consisted of 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 1s was randomly generated.¹⁴ In order to elicit individual productivity, subjects were offered a pure piece-rate compensation scheme. They received €0.03 for each table correctly processed.¹⁵ Furthermore, both a countdown timer and a counter reporting the number of tables processed were provided. Phase II lasted 10 min. Between Phase II and Phase III, subjects had the opportunity to rest. During this pause, 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.

¹³ An alternative design with very high and very low performance targets would also generate, respectively, a severe and a lenient contract. Albeit more realistic, we feel it would be virtually impossible with such a design to isolate and disentangle the effect of each of the two errors and thus assess the marginal impact of each type of error on individual performance. For an example of such a design see Brooks et al. (2014).

¹⁴ 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.

¹⁵ 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.

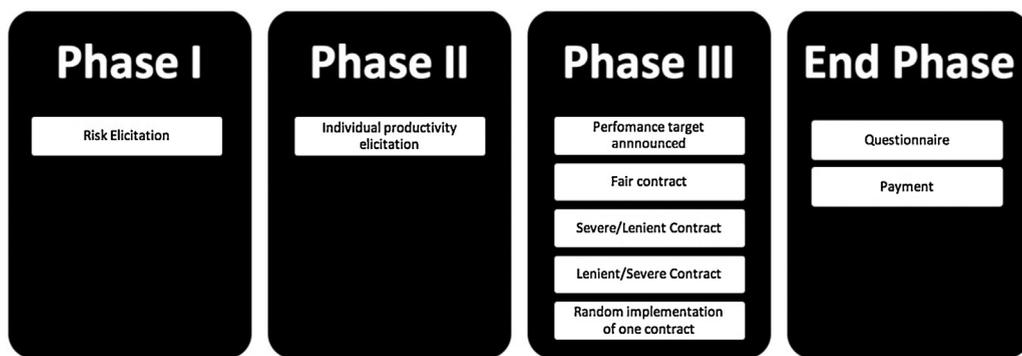


Fig. 1. Flow of the experiment.

Table 3
Experimental sessions.

Session	Conf.	Treatment order	Endowment	Reward	Subjects
i	Low	T_{0L}, T_{1L}, T_{2L}	€0	€6.60	23
ii	Low	T_{0L}, T_{2L}, T_{1L}	€0	€6.60	17
iii	High	T_{0H}, T_{1H}, T_{2H}	€5.28	€11.88	23
iv	High	T_{0H}, T_{2H}, T_{1H}	€5.28	€11.88	21

Let us define $q_{\text{phase-II}}$ as the number of tables correctly processed by each individual in 10 min in Phase II. The average $q_{\text{phase-II}}$ was 45.9 with a standard deviation of 11.5. The number of tables counted in the second phase, paid by the piece, represents a reliable measure of individual ability.

In Phase III, subjects were offered three hypothetical contracts (*fair*, *severe* and *lenient*)¹⁶. Instead of being paid piece-rate as in phase-II, now subjects were asked to reach a performance target of correctly counted tables within 40 min. For each of the three contracts, subjects had to decide:

- (i) either to reject the terms of the contract and thus not provide effort. In this case, they were allowed to leave the room with the baseline wage w_0 after having completed a long questionnaire.
- (ii) or to accept the terms and perform the contracted task that lasted 40 min. In this case, if they accomplished the task and completed the questionnaire, they could potentially gain w_r .

Performance target and individual productivity. In Phase III, each individual was requested to reach a specific performance target based on his or her piece-rate performance in Phase II. Phase III lasted 40 min and thus, four times the length of Phase II. The performance target was set equal to $0.9 \times \left(\frac{40 \text{ min}}{10 \text{ min}}\right) \times q_{\text{phase-II}}$ ¹⁷. Scaling the individual's specific target to his or her own ability roughly normalizes the cost of effort for the task across subjects. The performance target was announced only after Phase II had been carried out so there could be no ratchet-effect-like strategic behavior present in Phase II.

Elicitation of subjects' decisions. Subjects had to declare their choice (i or ii) for each of the three hypothetical contracts. However, after the three choices were made, only one contract for each subject was randomly selected and its parameters applied, so that in each session a certain number of subjects were assigned to each of the three contracts. Subjects were informed of which contract was actually implemented only after they had stated their decisions for all three contracts. If, for the implemented contract, the subject had chosen to reject the contract (ii), then he or she immediately had to proceed to the questionnaire phase; if he or she had chosen to accept the contract (i), the task began and the subject had to let 40 min pass before moving to the questionnaire phase. Thus this was a within-subject design, and it implemented the strategy method as the hypothetical choices were elicited before one was randomly chosen and applied.¹⁸

Order effects. Among the three contracts, the *fair* one was always submitted first, as it represents the benchmark case. These *severe* and *lenient* contracts were submitted after the *fair* one in alternate sequence to control for order effects (see column 3 of Table 3 for details). Both parametric and non-parametric tests (Appendix A.1, Table 6 and footnote 25) rule out any statistically significant difference in the decisions to accept the terms of the contract when these *severe* contract was presented first as opposed to when the *lenient* one was.

¹⁶ We did not frame the decision as “contracts” but – generically – as “situations.”

¹⁷ 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.

¹⁸ 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 et al., 2012).

Table 4
Summary of results.

Share of population	Fair – T0	Severe – T1	Lenient – T2
$Z(T_{iL})$ Low configuration	0.825	0.325	0.625
$Z(T_{iH})$ High configuration	0.909	0.545	0.705
Share of switchers ^a	Fair T0 vs. Severe T1	Fair T0 vs. Lenient T2	Lenient T2 vs. Severe T1
Switchers Low configuration	0.5***	0.2**	0.3***
Switchers High configuration	0.355***	0.204***	0.16**

Significance levels (exact McNemar's test): *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Bonferroni–Holm stepwise multiple-testing procedure is reported in [Appendix A.2](#).

^a Share of agents exerting high effort in the “one treatment” but exerting no effort in the “other treatment”.

Non-contamination of individuals' decisions. All subjects started Phase III at the same time and thus had to make their hypothetical choices simultaneously. Once they had all taken their three decisions, one contract was randomly selected for each subject. If a subject had chosen to reject one of the contracts (i), and that contract was randomly chosen, he or she had to first fill in a questionnaire that lasted several minutes, and only at that point could he or she leave the lab. It was only then that other subjects, observing him or her leaving the room could infer he decided (i) in at least one of the three contracts. This observation could not possibly contaminate ex-post their decision and their behavior, because their own choice had already been made several minutes before, and also because they did not know which contract the subject leaving the room had been randomly assigned to.¹⁹ It is also important to notice that only one subject among all those that had chosen (ii) eventually failed to reach the performance target. This corroborates the idea that observing others leaving the lab did not significantly affect their decision to perform.

Plausibility of the outside option. Subjects were aware that their choice (i or ii) for each of the three hypothetical contracts implied real consequences: If the subject chose (ii), then he or she would still be required to spend 40 min in the lab before moving to the questionnaire and to the payment phase, and if he or she chose (i), then the real effort task would be skipped entirely. Therefore the subjects had no reason to misrepresent their true preferences.

Participation and incentive constraints. In the model the participation constraint is assumed to be non-binding and the focus is on the incentive constraint. This is reflected in the experimental design. Subjects stated their hypothetical participation decisions under each of the three contracts; then, once one contract was chosen, they could proceed to carry out the task. On one hand, subjects took both a participation decision (accept or reject the contract) and also an incentive decision (exert enough effort to meet the target). In theory, these two decisions could lead to three different outcomes: (i) reject the contract, (ii-a) accept the contract and exert no or little effort, and (ii-b) accept the contract and exert enough effort to reach the target. On the other hand, given our experimental design, these three potential outcomes indeed collapse into only two possible payments, either w_0 or w_r . Under severe contracts doing (ii-a) leads to w_0 with certainty while under the lenient contract doing (ii-a) produce either w_0 or w_r if there is a type-II error. Under all contracts the final outcomes are either w_0 or w_r and thus the participation constraint is totally absorbed by the incentive constraint. The choice is ultimately between (i) rejecting the contract and (ii) accepting the contract and meeting the target. In fact only 1 out of 45 subjects who accepted the contract that was finally implemented (out of 84 observations) failed the target.

High and low configurations. The three main contracts 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 11). There were two low configuration sessions (40 subjects) and two high configuration sessions (44 subjects).

Common instructions for the ongoing phase were read and described aloud while instructions concerning each single contract were delivered individually 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 contract, were provided at the end of the experimental session. Control questions for each of the different phases and contracts 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 of 2011. A total of four sessions with 84 participants were carried out. Subjects were recruited online with ORSEE ([Greiner, 2015](#)). Nonparametric statistical checks 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). The

¹⁹ An issue with potential ex-ante contamination has been raised by one of the editors. Such contamination may happen because before choosing between (i) and (ii) and anticipating that other subjects in the lab, those *staying*, may observe the choice ex-post, subjects pondering the decision to *leave* may change their minds and stay instead. Our design cannot rule out this effect; however, we emphasize that *staying* subjects did not know which contract was randomly implemented for each *leaving* subject. By leaving the lab earlier, subjects were thus signaling that they had answered (i) to one, two or all three of the contracts without being able to specify which one. It is difficult to imagine that this very poorly informative signal could alter the choice between (i) and (ii) and, more important, that this alteration would be asymmetrical between lenient and severe contracts.

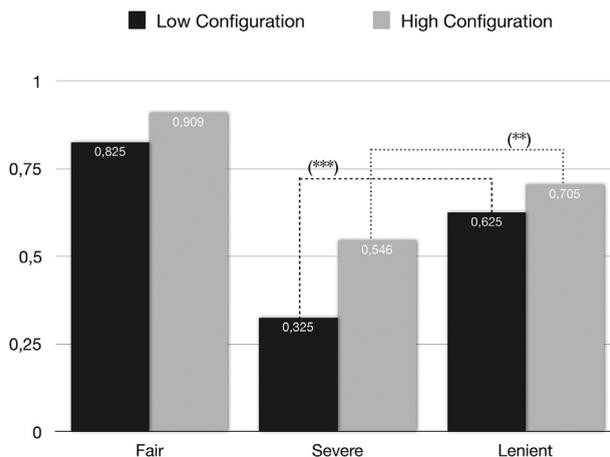


Fig. 2. Percentage of population exerting effort under *fair*, *severe* and *lenient* treatments in *low* and *high* configurations.

average age of participants was 23.3; 40% were females and 60% males. Using the strategy method, 84 observations for each contract were elicited. The average payoff was approximately €10.20.

5. Experimental results

In our experiment we focus on the extensive margin²⁰ by looking at the different shares $Z(T_{ij})$ of agents exerting effort when exposed to the different contracts (i) and configurations (j).

Result 1. 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 (defined as Z) in *fair* treatments ($\alpha = 0, \beta = 0$) with the same share in *severe* treatments ($\alpha = 0.8, \beta = 0$).

An inspection of Fig. 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 highly statistically significant according to a two-sided exact McNemar's test (p -value < 0.001) based on 40 independent paired observations. 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. Again, this negative effect is highly statistically significant at a 1% level (two-sided McNemar's test, p -value < 0.001 ; #40 independent paired observations).

Result 2. Rewarding undeserving agents (β) decreases agents' effort provision. We now compare the share of subjects exerting effort under *fair* treatments (with $\alpha = 0, \beta = 0$) with those under *lenient* treatments ($\beta = 0.8, \alpha = 0$). In the *low* configuration, the share of subjects exerting effort decreases from 0.825 (T_{0L}) to 0.625 under *lenient* treatment (T_{2L}). This negative effect is statistically significant at a 5% level (two-sided McNemar's test, p -value = 0.0386; #40 independent paired observations). In the *high* configuration, the percentage of subjects exerting effort drops from 0.909 (T_{0H}) to 0.705 in (T_{2H}). This negative effect is highly statistically significant at 1% level (two-sided McNemar's test, p -value = 0.0039; #44 independent paired observations).

Result 3. Neglecting due rewards to deserving agents (α) is more detrimental to agents' effort provision than rewarding undeserving agents (β). Testing Prediction 3, we compare the share of agents exerting effort under the *severe* and the *lenient* treatments. An inspection of Fig. 2 shows that the percentages of individuals exerting effort under the *lenient* and *severe* treatments are by no means equal. Under the *low* configuration treatments, the share of population exerting effort almost doubles from 0.325 (T_{1L}) to 0.625 (T_{2L}), moving from severe appraisal to lenient appraisal. The positive difference is highly statistically significant at a 1% level (two-sided McNemar's test, p -value = 0.0075, #40 independent paired observations). Under the *high* configuration treatments, the share of subjects exerting effort increases from 0.545 (T_{1H}) to 0.705 (T_{2H}). The positive effect is highly statistically significant at a 1% level, (two-sided McNemar's test, p -value = 0.0156, #44 independent paired observations).

In line with the predictions of our model, this experiment provides evidence that severity bias and leniency bias generate asymmetric effects on subjects' effort provision: the negative effect of the severity bias in T_{1L} and T_{1H} is substantially and significantly greater than the effect of the leniency bias in T_{2L} and T_{2H} , respectively.

²⁰ 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 intensive marginal effort is redundant; in fact, all but one of the subjects who chose to carry out the task eventually matched the performance target.

Table 5
Expected utility functions.

	Exp. utility with effort	Exp. utility with no effort	Expected Δ utility
$T0_L$	$v(\in 6.60) - g$	$v(\in 0)$	$v(\in 6.60) - v(\in 0) - g$
$T1_L$	$\frac{1}{5}v(\in 6.60) + \frac{4}{5}v(\in 0) - g$	$v(\in 0)$	$\frac{1}{5}[v(\in 6.60) - v(\in 0)] - g$
$T2_L$	$v(\in 6.60) - g$	$\frac{2}{5}v(\in 6.60) + \frac{1}{5}v(\in 0)$	$\frac{1}{5}[v(\in 6.60) - v(\in 0)] - g$
$T0_H$	$v(\in 11.88) - g$	$v(\in 5.28)$	$v(\in 11.88) - v(\in 5.28) - g$
$T1_H$	$\frac{1}{5}v(\in 11.88) + \frac{4}{5}v(\in 5.28) - g$	$v(\in 5.28)$	$\frac{1}{5}[v(\in 11.88) - v(\in 5.28)] - g$
$T2_H$	$v(\in 11.88) - g$	$\frac{4}{5}v(\in 11.88) + \frac{1}{5}v(\in 5.28)$	$\frac{1}{5}[v(\in 11.88) - v(\in 5.28)] - g$

Finally, an examination of Fig. 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 *fair* contracts, the shares of agents exerting effort in the *low* and *high* endowment configurations are equal to 0.825 and 0.909, respectively. The difference between configurations²¹ was not statistically significant at any conventional level (proportion test, p -value = 0.254). When comparing the same share under these severe treatments, $T1_L$ with 0.325 vs. $T1_H$ with 0.545, the difference between configurations is statistically significant at a 5% level (p -value = 0.042).

Under the *lenient* treatments, the difference between the share in $T2_L$ with 0.652 and $T2_H$ with 0.705 is not statistically significant at any conventional level (p -value = 0.439). 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 a Type I error (*severity bias*).

All statistical results delivered by the non-parametric analysis are confirmed by (i) complementary regression analysis (OLS – Linear Probability Model) and (ii) Bonferroni-Holm stepwise multiple-testing procedure (Westfall et al., 2010; List et al., 2016) to control for the family wise error rate (FWER). Both robustness checks are reported in Appendix A.1.

6. Discussion

Subjects behave differently under the *severity* and *leniency* biases. This was a key prediction from our principal-agent model with reference-dependent preferences, and one that was confirmed by our experimental test. In the following sections we discuss this main result in light of different economic theories of behavior.

6.1. The role of risk aversion

The experimental design is neutral to subjects' different levels of risk aversion, or at least the kind of risk aversion that comes from conventional decreasing marginal utility of income. The reason for this can be seen in 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 5) is positive. Note that both contracts have the same difference in expected utility in each configuration as $EU(\Delta T1_L) = EU(\Delta T2_L)$ and $EU(\Delta T1_H) = EU(\Delta T2_H)$.

Whether the attempt to rule out risk aversion by construction can be considered successful depends crucially on the acceptance of the separability of the utility functions in monetary utility and effort (see Laffont and Martimort, 2002 – p. 149) and whether we focus on standard risk aversion derived by the diminishing marginal utility of income. On the other hand, Rabin (2000) points out that diminishing marginal utility of income can hardly explain non trivial modest-scale risk-aversion as the one typically observed in lab experiments. If this is the case, our attempt to neutralize expected utility risk aversion by design becomes irrelevant and –furthermore– risk attitudes arising from other behavioral traits could have informed our results. This is the reason why we have implemented a popular risk measurement task to elicit the certainty equivalent (CE) of a lottery (See Bruhin et al., 2010; Abdellaoui et al., 2011; Sutter et al., 2013) in Phase I. Correlations between the CE and the choice of exerting effort for both treatments $T1$ and $T2$ are very weak and statistically not significant (Spearman's ρ -correlation = 0.078, p -value = 0.48 in $T1$ and = 0.091, p -value = 0.40 in $T2$) suggesting that the measure of risk aversion elicited by our task does not explain our main treatment effects.²²

²¹ To test these differences we ran a between-subjects analysis contrasting treatment outcomes by configuration with a proportion test.

²² A word of caution should be added to this conclusion. The empirical measurement of risk preferences is the subject of an intense ongoing debate (See among others (Harrison and Rutstrom, 2008; Charness et al., 2013; Crosetto and Filippin, 2015)). Following Rabin (2000), Koszegi and Rabin (2007) propose their reference dependence model of utility as a better candidate to explain observed modest-scale risk aversion in the lab. One may thus conjecture that, if the same reference dependence utility drives subjects' choice in the effort task and subjects' choices in the lottery task, some correlation in the data on effort choice and on risk choice may surface. However, it does not. Our speculation about why this is not the case boils down to mainly two reasons: (i) Reference Dependence might not be the driving force behind our experimental effect (in the next paragraph we provide another candidate explanation for our asymmetric result) and/or; (ii) the risk elicitation task we chose does not adequately harvest risk-aversion determined by reference dependent preferences.

A final consideration concerns the difference in expected utility between *high* and *low* configurations. Under standard decreasing marginal utility of income we have that $v(\in 6.60) \geq v(\in 11.88) - v(\in 5.28)$. For each contract (*fair, severe, lenient*), the agent's marginal utility of the reward payment should therefore be higher in the *low* endowment configuration than in the *high* endowment configuration. 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.²³

6.2. Regret aversion

Regret aversion (Loomes and Sugden, 1982; Bell, 1983; Hayashi, 2008) is another candidate explanation for our asymmetric result and it is not captured by the reference-dependent model.²⁴ When a decision results in a bad outcome, a subject may feel disappointed when he or she expected a better result, and may feel regret when he or she 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 $p=4/5$ the subject exerts effort and obtains the same outcome as if he had not exerted effort). Conversely, 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 or she eventually does not get paid. Thus, anticipated regret might offer an alternative explanation for the observed pattern.

6.3. Social preferences, organizational justice and self-image

Different streams of research in management, psychology and economics all point to the fact that employees' incentives and motivation are greatly affected by the perceived fairness of performance appraisal and related compensation plans (Akerlof and Yellen, 1990; Cropanzano et al., 2007; Folger, 2001). Organizational justice theories and recent behavioral research in economics claim that employees care about the absolute compensation level as much as how they compare with other fellow workers and how fair they perceive the procedure to be. If an employee perceives his rating to be unfair, then he or she will react negatively towards the compensation system and will not be motivated by it (Cohen-Charash and Spector, 2001; Greenberg, 1987).

Social preferences such as inequity aversion (Grund and Przemeczek, 2012), gift-exchange behavior (Akerlof, 1982) and reciprocity (Sebold and Walzl, 2014) are often called on in explaining organizational behavior that deviates from the standard principal-agent predictions. 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 include a real supervisor (our supervisor is represented by a passive automaton) to be grateful or resentful towards, thus hindering the explicative effect of reciprocity and gift-exchange. Concerns for procedural justice may somehow explain the asymmetric behavior we observe. In our experiment, 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* contracts go against the subject's interests while the *lenient* contracts favor them. This consideration could provide a plausible explanation for the higher negative response observed under *T1* than under *T2*.

Finally, behavioral research has long shown how individuals are affected by self-serving biases and tend to overestimate their own performance and contribution (Thornton, 1980; Harris and Schaubroeck, 1988). Therefore, self-appraisal of performance is typically lenient and employees affected by self-serving bias may perceive a non-inflated rating as unfair. Supervisors thus optimally adjust for such a bias by inflating their own scores.

7. Conclusion

We have considered both severity and leniency biases and have shown how they both undermine the impact of performance appraisal on the agent's incentive constraint. We have first shown that the standard principal-agent model predicts that both biases should impact performance symmetrically. However, once one considers loss aversion and reference-dependent preferences, the principal-agent model predicts that leniency bias is less detrimental to effort provision. This theoretical result is very much in line with consistent empirical evidence of the prevalence of leniency bias in real world organizations using performance appraisals. We then present an experiment to test this prediction. The experiment finds

²³ Gift-exchange theory (Akerlof, 1982) could represent a plausible candidate explanation for this result. As long as agents get a fair basic wage (*high* endowment configuration) they are more willing to exert effort. In particular, this consideration is corroborated by the fact that more agents exert effort under T_{0H} than T_{0L} .

²⁴ Earlier reference-dependent models were used to model disappointment aversion (Loomes and Sugden, 1986). 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 et al., 2000).

Table 6
OLS linear probability model: probability of exerting high effort – marginal effects.

Variable	Outcome: EFFORT
Severity bias (T1)	-0.428***
Leniency bias (T2)	-0.202***
High configuration	0.154**
Risk aversion	0.017
Ordering	-0.015
Male	-0.122*
Constant (T0)	0.754***
Obs.	252

Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. Clusters at individual level.

strong support for the existence of an asymmetric impact of evaluation errors on agents' willingness to exert effort. The source of error in our experiment was completely exogenous and generated by the program algorithm. An interesting potential extension of the present work may envisage a principal/subject committing the appraisal error. This may exacerbate the asymmetric effect of errors even further.

Although the experimental method has limited external validity, with this experiment we are able to provide clear causal evidence about the basic underlying mechanism that characterizes reactions to performance appraisal errors. Since intangibles are increasingly important in business organizations, and knowledge-intensive jobs are difficult to assess, errors in evaluating employees' performance may very 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 implications of the present work may also apply to evaluation procedures outside the firm, such as judicial procedures and educational assessment. The take-home message from the experiment is the following: If one must err, better to err on the lenient side.

Appendix A.

A.1. Regression analysis

We run a Linear Probability Model (with conservative standard errors clustered at the subject level) to provide further evidence on 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 the goal realized. We analyze how the probability of exerting effort is influenced by the types of bias that subjects are exposed to, namely Severity (T1) vs. Leniency (T2) bias (treatment dummies), different endowment levels (configuration dummy), individual risk aversion (discrete variable capturing the switching-point in the Phase I "risk elicitation lotteries table", Table 8) and ordering effects in submitting reverse series of scenarios during the experiment (dummy variable, =1 if sequence T0, T1, T2; =0 if sequence T0, T2, T1).

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 (42 percentage points vs. 20 percentage points.) and statistically different compared to the detrimental effect generated by the Type II error (Wald test p -value = 0.001).

With the type of bias kept constant, the higher endowment configuration leads to a mild increase in the probability of observing an effort action.

Subjects' risk preferences do not affect the action choice. Also the sequential order in which scenarios are submitted to the participants do not statistically affect the behaviour.²⁵ We only find a weak gender effect suggesting that males are on average less willing to provide effort.

A.2. Bonferroni–Holm stepwise multiple-testing procedure

Following Westfall et al. (2010), in order to control for the family wise error rate (FWER) that could arise due to multiple-testing, a Bonferroni–Holm stepwise multiple-testing procedure (Holm, 1979) is performed (Table 7) considering the exact McNemar's p -values of the multiple tests. Fixing a conventional $\alpha = 0.05$, the rejection of the null hypothesis does hold for all multiple McNemar's tests.

²⁵ The same null result is confirmed also performing two-samples tests of proportions by "ordering", both for T1 (p -value = 0.908) and T2 (p -value = 0.757).

Table 7Bonferroni–Holm stepwise multiple-testing procedure for $\alpha = 0.05$.

Comparison	Actual p -value	Rank	Criterion	Reject H_0
T_{0L} vs. T_{1L}	0.0000	6	0.0083	Yes
T_{0H} vs. T_{1H}	0.0000	5	0.01	Yes
T_{0H} vs. T_{2H}	0.0039	4	0.0125	Yes
T_{2L} vs. T_{1L}	0.0075	3	0.0167	Yes
T_{2H} vs. T_{1H}	0.0156	2	0.025	Yes
T_{0L} vs. T_{2L}	0.0386	1	0.05	Yes

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jebo.2016.07.022>.

References

- Abdellaoui, M., Baillon, A., Placido, L., Wakker, P.P., 2011. The rich domain of uncertainty: source functions and their experimental implementation. *Am. Econ. Rev.* 101 (2), 695–723.
- Abeler, J., Falk, A., Götte, L., Huffman, D., 2011. Reference points and effort provision. *Am. Econ. Rev.* 101 (2), 470–492.
- Akerlof, G.A., 1982. Labor contracts as partial gift exchange. *Q. J. Econ.* 97 (4), 543–569.
- Akerlof, G.A., Yellen, J.L., 1990. The fair wage–effort hypothesis and unemployment. *Q. J. Econ.* 105 (2), 255–283.
- Armanter, O., Boly, A., 2015. Framing of incentives and effort provision. *Int. Econ. Rev.* 56 (3), 917–938.
- Aron, D.J., Olivella, P., 1994. Bonus and penalty schemes as equilibrium incentive devices, with application to manufacturing systems. *J. Law Econ. Organ.* 10 (1), 1–34.
- Bell, D., 1983. Risk premiums for decision regret. *Manag. Sci.* 29 (10), 1156–1166.
- Berger, J., Harbring, C., Sliwka, D., 2013. Performance appraisals and the impact of forced distribution: an experimental investigation. *Manag. Sci.* 59 (1), 54–68.
- Bol, J.C., 2011. The determinants and performance effects of managers' performance evaluation biases. *Acc. Rev.* 86 (5), 1549–1575 (20.01.14).
- Bol, J.C., Smith, S.D., 2011. Spillover effects in subjective performance evaluation: bias and the asymmetric influence of controllability. *Acc. Rev.* 86 (4), 1213–1230 (20.01.14).
- Bretz, R., Milkovich, G., Read, W., 1992. The current state of performance appraisal research and practice: concerns, directions, and implications. *J. Manag.* 18 (2), 321–352.
- Brooks, R.R., Stremtizer, A., Tontrup, S., 2014. Stretch it but don't break it: The hidden risk of contract framing. In: *UCLA School of Law, Law-Econ Research Paper No. 13–22*.
- Bruhlin, A., Fehr-Duda, H., Epper, T., 2010. Risk and rationality: uncovering heterogeneity in probability distortion. *Econometrica* 78 (4), 1375–1412.
- Bull, C., 1987. The existence of self-enforcing implicit contracts. *Q. J. Econ.* 102 (1), 147–159.
- Cardy, R., Dobbins, G., 1986. Affect and appraisal accuracy: liking as an integral dimension in evaluating performance. *J. Appl. Psychol.* 71 (4), 672.
- Charness, G., Gneezy, U., Imas, A., 2013. Experimental methods: eliciting risk preferences. *J. Econ. Behav. Organ.* 87, 43–51.
- Charness, G., Gneezy, U., Kuhn, M., 2012. Experimental methods: between-subject and within-subject design. *J. Econ. Behav. Organ.* 81 (1), 1–8.
- Cohen-Charash, Y., Spector, P.E., 2001. The role of justice in organizations: a meta-analysis. *Organ. Behav. Hum. Decis. Process.* 86 (2), 278–321.
- Cropanzano, R., Bowen, D., Gilliland, S., 2007. The management of organizational justice. *Acad. Manag. Perspect.* 21 (4), 34–48.
- Crossetto, P., Filippin, A., 2015. A theoretical and experimental appraisal of four risk elicitation methods. *Exp. Econ.* 1, 1–29.
- Daido, K., Itoh, H., 2007, September. The Pygmalion and Galatea Effects: An Agency Model With Reference-Dependent Preferences and Applications to Self-Fulfilling Prophecy. Discussion Paper Series 35. School of Economics, Kwansai Gakuin University.
- Daido, K., Morita, K., Murooka, T., Ogawa, H., 2013. Task assignment under agent loss aversion. *Econ. Lett.* 121 (1), 35–38.
- Daido, K., Murooka, T., 2016. Team incentives and reference-dependent preferences. *J. Econ. Manag. Strateg.*, <http://dx.doi.org/10.1111/jems.12166>.
- De Chiara, A., Livio, L., 2015. The threat of corruption and the optimal supervisory task. In: *ECARES Working Papers (R&R on JEBO)*.
- Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. *Exp. Econ.* 10 (2), 171–178.
- Fisher, C., 1979. Transmission of positive and negative feedback to subordinates: a laboratory investigation. *J. Appl. Psychol.* 64 (5), 533.
- Folger, R., 2001. Fairness as deonance. *Theor. Cult. Perspect. Organ. Justice*, 3–33.
- Giebe, T., Guertler, O., 2012. Optimal contracts for lenient supervisors. *J. Econ. Behav. Organ.* 81 (2), 403–420.
- Greenberg, J., 1987. A taxonomy of organizational justice theories. *Acad. Manag. Rev.* 12 (1), 9–22.
- Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with orsee. *J. Econ. Sci. Assoc.* 1 (1), 114–125.
- Grund, C., Przemek, J., 2012. Subjective performance appraisal and inequality aversion. *Appl. Econ.* 44 (17), 2149–2155.
- Harris, M.M., Schaubroeck, J., 1988. A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Pers. Psychol.* 41 (1), 43–62.
- Harrison, G.W., Rutstrom, E.E., 2008. Risk aversion in the laboratory. *Risk Aversion Exp.* 12, 41.
- Hayashi, T., 2008. Regret aversion and opportunity dependence. *J. Econ. Theory* 139 (1), 242–268.
- Herweg, F., Muller, D., Weinschenk, P., 2010. Binary payment schemes: moral hazard and loss aversion. *Am. Econ. Rev.* 100 (5), 2451–2477.
- Higgins, C., Judge, T., Ferris, G., 2003. Influence tactics and work outcomes: a meta-analysis. *J. Organ. Behav.* 24 (1), 89–106.
- Holm, S., 1979. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, 65–70.
- Hölmstrom, B., 1979. Moral hazard and observability. *Bell J. Econ.* 10 (1), 74–91.
- Ilgen, D.R., Mitchell, T.R., Fredrickson, J.W., 1981. Poor performers: supervisors' and subordinates' responses. *Organ. Behav. Hum. Perform.* 27 (3), 386–410.
- Judge, T., Ferris, G., 1993. Social context of performance evaluation decisions. *Acad. Manag. J.* 36 (1), 80–105.
- Kahneman, D., Tversky, A., 1984. Choices, values, and frames. *Am. Psychol.* 39 (4), 341–350.
- Kahneman, D.J., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. *Econometrica* 47 (2), 263–292.
- Kambe, S., 2006. Subjective evaluation in agency contracts. *Jpn. Econ. Rev.* 57 (1), 121–140.
- Kane, J., Bernardin, H., Villanova, P., Peyrefitte, J., 1995. Stability of rater leniency: three studies. *Acad. Manag. J.* 38 (4), 1036–1051.
- Kingsbury, F.A., 1922. Analyzing ratings and training raters. *J. Pers. Res.* 1, 377–383.
- Klimoski, R., Inks, L., 1990. Accountability forces in performance appraisal. *Organ. Behav. Hum. Decis. Process.* 45 (2), 194–208.
- Kozegi, B., Rabin, M., 2006. A model of reference-dependent preferences. *Q. J. Econ.* 121 (4), 1133–1165.
- Kozegi, B., Rabin, M., 2007. Reference-dependent risk attitudes. *Am. Econ. Rev.* 97 (4), 1047–1073.
- Laffont, J., Martimort, D., 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton Univ Pr.
- Landy, F.J., Farr, J.L., 1980. Performance rating. *Psychol. Bull.* 87 (1), 72.
- Lazear, E., 1999. *Personnel Economics: Past Lessons and Future Directions*.

- Levin, J., 2003. Relational incentive contracts. *Am. Econ. Rev.* 93 (3), 835–857.
- List, J.A., Shaikh, A.M., Xu, Y., 2016. Multiple hypothesis testing in experimental economics. NBER Working Paper – 21875.
- Loomes, G., Sugden, R., 1982. Regret theory: an alternative theory of rational choice under uncertainty. *Econ. J.* 92 (368), 805–824.
- Loomes, G., Sugden, R., 1986. Disappointment and dynamic consistency in choice under uncertainty. *Rev. Econ. Stud.* 53 (2), 271–282.
- MacLeod, W., Malcomson, J., 1989. Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica*, 447–480.
- MacLeod, W.B., 2003. Optimal contracting with subjective evaluation. *Am. Econ. Rev.* 93 (1), 216–240.
- Maestri, L., 2012. Bonus payments versus efficiency wages in the repeated principal-agent model with subjective evaluations. *Am. Econ. J.: Microecon.* 4 (3), 34–56.
- Moers, F., 2005. Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Acc. Organ. Soc.* 30 (1), 67–80.
- Pearce, D., Stacchetti, E., 1998. The interaction of implicit and explicit contracts in repeated agency. *Games Econ. Behav.* 23 (1), 75–96.
- Prendergast, C., 1999. The provision of incentives in firms. *J. Econ. Lit.* 37 (1), 7–63.
- Prendergast, C., 2002. Uncertainty and incentives. *J. Labor Econ.* 20 (2; PART 2), 115–137.
- Prendergast, C., Topel, R., 1993. Discretion and bias in performance evaluation. *Eur. Econ. Rev.* 37 (2–3), 355–365.
- Prendergast, C., Topel, R.H., 1996. Favoritism in organizations. *J. Polit. Econ.* 104 (5), 958–978.
- Rabin, M., 2000. Risk aversion and expected-utility theory: a calibration theorem. *Econometrica* 68 (5), 1281–1292.
- Rabin, M., Schrag, J.L., 1999. First impressions matter: a model of confirmatory bias. *Q. J. Econ.* 114 (1), 37–82.
- Rynes, S.L., Gerhart, B., Parks, L., 2005. Personnel psychology: performance evaluation and pay for performance. *Annu. Rev. Psychol.* 56, 571–600.
- Sautmann, A., 2013. Contracts for agents with biased beliefs: some theory and an experiment. *Am. Econ. J.: Microecon.* 5 (3), 124–156.
- Schmidt, K., Schnitzer, M., 1995. The interaction of explicit and implicit contracts. *Econ. Lett.* 48 (2), 193–199.
- Schoorman, F.D., 1988. Escalation bias in performance appraisals: an unintended consequence of supervisor participation in hiring decisions. *J. Appl. Psychol.* 73 (1), 58.
- Sebald, A., Walzl, M., 2014. Subjective performance evaluations and reciprocity in principal-agent relations. *Scand. J. Econ.* 116 (2), 570–590.
- Steers, R., Mowday, R., Shapiro, D., 2004. Introduction to special topic forum: The future of work motivation theory. *Acad. Manag. Rev.* 29 (3), 379–387.
- Strausz, R., 1997. Collusion and renegotiation in a principal-supervisor-agent relationship. *Scand. J. Econ.* 99 (4), 497–518.
- Sutter, M., Kocher, M.G., Glätzle-Rützler, D., Trautmann, S.T., 2013. Impatience and uncertainty: experimental decisions predict adolescents' field behavior. *Am. Econ. Rev.* 103 (1), 510–531.
- Thiele, V., 2013. Subjective performance evaluations, collusion, and organizational design. *J. Law Econ. Organ.* 29 (1), 35–59.
- Thomas, J., Meeke, H., 2010. Rater error. In: Corsini Encyclopedia of Psychology.
- Thorndike, R.L., 1949. *Personnel Selection; Test and Measurement Techniques*. Wiley, Oxford, England.
- Thornton, G.C., 1980. Psychometric properties of self-appraisals of job performance. *Pers. Psychol.* 33 (2), 263–271.
- Tirole, J., 1986. Hierarchies and bureaucracies: on the role of collusion in organizations. *J. Law Econ. Organ.* 2, 181.
- Tversky, A., Kahneman, D., 1991. Loss aversion in riskless choice: a reference-dependent model. *Q. J. Econ.* 106 (4), 1039–1061.
- Vafai, K., 2010. Opportunism in organizations. *J. Law Econ. Organ.* 26 (1), 158–181.
- Van Dijk, W., Zeelenberg, M., 2002. Investigating the appraisal patterns of regret and disappointment. *Mot. Emot.* 26 (4), 321–331.
- Varma, A., Denisi, A., Peters, L., 1996. Interpersonal affect and performance appraisal: a field study. *Pers. Psychol.* 49 (2), 341–360.
- Vieider, F.M., Chmura, T., Fisher, T., Kusakawa, T., Martinsson, P., Thompson, F.M., Sunday, A., 2015. Within-versus between-country differences in risk attitudes: implications for cultural comparisons. *Theory Decis.* 78 (2), 209–218.
- Villanova, P., Bernardin, H., Dahmus, S., Sims, R., 1993. Rater leniency and performance appraisal discomfort. *Educ. Psychol. Meas.* 53 (3), 789–799.
- Westfall, P.H., Troendle, J.F., Pennello, G., 2010. Multiple mcnemar tests. *Biometrics* 66 (4), 1185–1191.
- Zeelenberg, M., Van Dijk, W., Manstead, A., van der Pligt, J., 2000. On bad decisions and disconfirmed expectancies: the psychology of regret and disappointment. *Cogn. Emot.* 14 (4), 521–541.