Punishment Incentives in Principal-Agent Dynamics: Insights from a Public Goods Game Experiment

Sandro Casal[1], Päivi Maijanen[2], Luigi Mittone[1]* and Azzurra Morreale[2]

Abstract

This paper experimentally investigates the impact of punishment on agents' behavior in a principal-agent framework. The study focuses on agency problems that arise from conflicting incentive structures between principals (managers) and agents (employees). We aim to determine whether a punishment mechanism can reduce these agency problems and align agents' actions with the principal's objectives. In our experimental setup, managers, acting as principals, can use punishment to influence employees' (agents') efforts and decisions. The results indicate that punishment does affect employees' choices, leading them to select projects with higher returns for the manager. However, the punishment mechanism does not fully achieve its intended effect, as managers cannot consistently influence the level of employee contributions.

JEL Classification: C90, D82, H41

Keywords

agency theory — financial incentives — punishment — laboratory experiment

[1] *University of Trento, Italy*

[2] *LUT University, Business School, Finland*

***Corresponding author**: luigi.mittone@unitn.it

The authors are presented alphabetically, with each contributing equally to this research.

Introduction

Agency theory provides a comprehensive framework for understanding the complexities of principal-agent relationships, which are pivotal in organizational contexts. The theory primarily seeks to predict and explain the behavior of rational actors within these relationships, focusing on the inherent conflicts of interest that arise when the goals or preferences of the agent diverge from those of the principal (Wright et al., 2001). These conflicts, often called agency dilemmas, can lead to inefficiencies and potential losses for both parties involved (McCoy & Flesher, 1998).

Agency problems are pervasive across various organizational settings and manifest in multiple forms (Jensen & Meckling, 1976). The literature on agency theory has traditionally concentrated on the conflicts between shareholders (principals) and executives (agents) within firms (e.g., Agrawal & Knoeber, 1996; DeFusco, Johnson & Zorn, 1990; Hermalin & Weisbach, 1991; Jensen & Meckling, 1976; Garen, 1994). However, research has also extended to other relationships, such as those between suppliers (e.g., Zsidisin & Ellram, 2003; Lassar & Kerr, 1996; Camuffo, Furlan & Rettore, 2007; Manatsa & McLaren, 2008; Whipple & Roh, 2010), managers and employees (Van Puyvelde, Caers, Bois & Jegers, 2013), outsourcing arrangements (Logan, 2000; Bahli & Rivard, 2003), and financial management (Trautmann & Zia, 2015; Casal et al., 2019).

The ramifications of agency problems are significant, often resulting in inefficiencies, motivational issues, or organizational failures, as exemplified by historical cases like the Enron scandal (Arnold & Lange, 2004). Consequently, scholars have been actively engaged in analyzing these relationships and devising strategies to manage them effectively.

Financial incentives have emerged as a critical tool in addressing the misalignment of interests between principals and agents (Eisenhardt, 1989). Jensen and Meckling (1976) suggest that well-structured incentives can mitigate agency problems by promoting desirable behaviors among agents. However, the effectiveness of monetary incentives, particularly those incorporating punitive elements, remains contentious. While some studies highlight the positive impact of monetary incentives on performance (e.g., Jenkins, Mitra, Gupta, & Shaw, 1998; Stajkovic & Luthans, 2001), others suggest that such incentives might undermine intrinsic motivation (Kohn, 1993; Deci, Ryan, & Koestner, 1999).

A crucial evolution in the theory involves the concept of bounded self-interest. This suggests that contrary to what agency theory traditionally assumes (i.e., that actors are self-interested and rational), they are not solely driven by wealth maximization but also by other non-monetary considerations, such as fairness and reciprocity (Bosse & Philips, 2016). This shift calls for a deeper understanding of human behavior in the principal-agent relationship, where perceptions of fairness and reciprocity play a significant role in decision-making.

Experimental methodologies have proven invaluable in exploring these dynamics (e.g., Espín et al., 2017). By recreating real-world scenarios in controlled environments, such experiments provide empirical insights that complement theoretical frameworks. A prominent example is the public goods game (PGG), which models collective action problems and allows for studying incentive mechanisms like cooperation and defection (Guala, 2005). This experimental approach offers a robust method for examining the principal-agent relationship.

Initially commissioned by the Italian Ministry of Public Administration, this experimental study contributes to the ongoing discussion on the effectiveness of punitive monetary incentives in the agency dilemma by exploring how these incentives influence the principal-agent relationship in the public sector. Based on a PGG, the experiment aims to study a specific kind of organizational condition initially suggested by the Ministry. The Ministry's approach aimed to incentivize senior management to achieve a specific strategic goal. However, this goal was misaligned with employees' personal objectives. Furthermore, budget constraints prevented senior managers from offering positive incentives to motivate employees. Consequently, the only available method to influence employee behavior was the use of penalties.

An example of a situation of misalignment between managers and employees is when the employees decide to prioritize personal objectives - such as job security, fair compensation, and reasonable workload goals - that do not necessarily align with the Ministry's strategic agenda. This conflict results in employees fulfilling their own performance goals without necessarily advancing the Ministry's strategic goals, leading to missed opportunities and failure to meet desired outcomes (Ayers, 2015). Another typical example of this kind of conflict is when employees in the public sector must frequently change roles or become disengaged. When this situation arises, employees prioritize short-term tasks, creating friction between personal job objectives and ministry-wide strategic goals.

We experimentally mimicked the situation just described by introducing a strategic interaction between principals (the top managers) and agents (the employees) regarding the use of the employees' efforts in supporting different projects. Specifically, this study examines how managers utilize punishment incentives to influence project selection and employees' efforts, yielding insights into their effectiveness in driving organizational outcomes.

Our findings reveal that punishment incentives affect employees' project selection choices, aligning them with projects that offer higher returns to the manager. However, managers did not achieve their desired level of influence on employees' contributions to the projects.

This paper is structured as follows. In Section 2, we provide a review of relevant theoretical frameworks and empirical literature. Section 3 outlines our research design and methodology, detailing the experimental approach and hypotheses. Section 4 presents our results, and Section 5 concludes with a summary of key insights, limitations, and suggestions for future research.

Theoretical Background

Agency theory investigates the complexities of the principal-agent relationship, aiming to predict rational behavior within these dynamics (Wright et al., 2001). In this relationship, the agent acts on behalf of the principal, who delegates decision-making authority. Conflicts arise when the agent's goals or preferences differ from those of the principal, creating agency problems that can lead to significant costs for the organization (McCoy & Flesher, 1998). To address these issues, agency theory emphasizes the role of incentives in aligning the interests of agents and principals (Eisenhardt, 1989).

Eisenhardt (1989) identifies several factors contributing to agency problems: divergent goals, difficulties in verifying agent actions, and differences in risk preferences. Due to their concentrated employment ties, risk-averse agents may opt for safer choices, potentially leading to opportunity costs for principals seeking higher returns (Wiseman & Gomez-Mejia, 1998; Chari et al., 2019). This conflict is evident in scenarios where short-term gains for the agent might compromise long-term benefits for the principal, reflecting the tension between immediate rewards and sustainable outcomes.

Information asymmetry between principals and agents exacerbates agency problems through moral hazard and adverse selection (Shapiro, 2005). Effective incentives and monitoring mechanisms are crucial for aligning agent behaviors with the principal's goals and addressing associated agency costs such as monitoring expenditures, bonding costs, and residual losses (Jensen & Meckling, 1976; Cuevas-Rodriguez, Gomez-Mejia, & Wiseman, 2012). Research highlights a positive relationship between incentives and performance, suggesting that well-structured incentives can enhance productivity (Roth & O'Donnell, 1996; Anderhub et al., 2002). However, other studies indicate the complexity of this relationship, influenced by various factors such as task nature and individual differences, resulting in only a weak correlation between monetary compensation and performance (Jensen & Murphy, 1990; Garen, 1994). Concerns about the potential crowding-out effect of incentives on intrinsic motivation, where external rewards may diminish individuals' internal drive, underscore the need for nuanced incentive design (Frey & Jegen, 2001).

Non-monetary incentives, such as reciprocity, fairness, and social preferences, also play significant roles in the principal-agent interaction. Research by Fehr and Gächter (2000b) and by Anderhub et al. (2002) indicate that these factors can sometimes outweigh purely self-interested behaviors, leading to higher levels of cooperation and performance. Monitoring and incentive-alignment systems help mitigate opportunistic behavior, further supporting the importance of considering both monetary and non-monetary incentives. Our study focuses exclusively on using punishment as an incentive mechanism. Experimental evidence shows that punishment can effectively promote cooperation and deter free-riding behaviors (Fehr & Gächter, 2000a), and its design, including its cost and perceived fairness, is crucial in shaping behavior (Balliet et al., 2011).

Despite its potential to influence behavior, punishment-based incentives may also create an atmosphere of threat that undermines long-term motivation. Fehr and Gächter (1998) found that punishment often induces a negative psychological response, where employees feel controlled rather than motivated, leading them to contribute less.

The interplay between short-term and long-term incentives is another critical aspect of agency theory. While short-term incentives can drive immediate performance improvements, they may also encourage short-sighted behavior that undermines long-term organizational goals. Research by Kaplan and Norton (2001) on the Balanced Scorecard approach underscores the importance of aligning incentives with long-term strategic objectives to ensure sustainable performance. In our experimental design, we incorporate a long-term investment component that benefits the entire organization, aligning with long-term incentives. Structuring incentives to encourage employees to prioritize these long-term benefits over immediate personal gains is a common challenge in the principal-agent relationship. This approach aims to balance immediate performance with sustainable organizational growth.

This study contributes significantly to the existing literature by examining the impact of punishment as an incentive mechanism in a controlled experimental setting that simulates real-world organizational conflicts of interest. It explores how negative incentives can influence decision-making processes in collective action scenarios, offering empirical evidence on the effectiveness of punishment in promoting cooperative behavior. Additionally, this study enriches our understanding of principal-agent dynamics by analyzing the interplay between short-term and long-term incentives and addressing the gap in the literature regarding the exclusive use of punishment.

Methodology

Experimental design

The experimental design is structured to replicate the conflicts of interest that can arise within a medium/large organization, namely those between managers and employees and between different operational units (such as departments, divisions, etc.) within the same organization. In our setting, 5 groups, each consisting of 4 members, are formed. In each group, one participant out of the four assumes the role of manager while the other three take on the role of employees; these groups represent a unit of the organization, which - in the experiment - is composed of all the groups participating in the same experimental session.

At the beginning of each period, each member receives 20 tokens, and only the employees must decide how to allocate them, knowing that these tokens can be used in two projects: Project A and Project B.[1] Once the project in which the employee wants to invest their tokens is chosen, they must decide how many tokens to invest (from 0 to 20); the first choice (between Project A and Project B) is therefore mandatory, but it is still possible not to invest any tokens in the chosen Project. However, investing tokens in a project that was not selected in the previous step is impossible.

The two projects have different returns for managers and employees. Specifically:

- **Project A**
  - Manager: receives a number of tokens equal to 60% of the total tokens invested in this project by the employees of their group.
  - Employees: only those who have chosen to invest their tokens in Project A receive 40% of the total tokens invested in this project by the employees of their group.
- **Project B**
  - Manager: does not receive any tokens from this project.
  - Employees: only those who have chosen to invest their tokens in Project B receive 40% of the total tokens invested in this project by the employees of their group.

However, investing tokens in Project B indirectly funds the creation of a third project, Project C, with returns equally distributed among all experiment participants. Project C aims to replicate the broader social welfare benefits generated by the employees' efforts. It is assumed that this will enhance the reputation of the entire organization and benefit all its members, including both employees and managers. This means that by financing a specific Project B, not only will the employees within the group benefit, but also all members of the other participating groups involved in each experimental session, including the managers. Essentially, all members of the organization benefit from the returns of Project C. Specifically, each Project C receives indirect financing: each of the 20 participants in the session (regardless of role, group, and choices made while playing the role of employees) receives an amount equal to 16% of the total sum of the tokens invested in the five projects B.[2] Summarizing, Project C gives the following returns:

- **Project C**
  - Managers: each one receives from Project C an amount equal to 16% of the total tokens invested in the five Projects B.
  - Employees: each one receives from Project C an amount equal to 16% of the total tokens invested in the five Projects B.

[1] The tokens are intended to mimic the effort expended by the employee in their respective project.

[2] An alternative way to interpret the individual return from Project C is to view it as 40% of the sum of five amounts, each representing the return earned by employees from their respective Project B within the five groups of the experimental session. Notably, these amounts correspond to 40% of the total invested by the same employees in their respective Project B.

Moreover, introducing Project C helps justify the conflict between employees' and managers' goals in a manner consistent with the intrinsic logic of public goods. In fact, managers also benefit from Project C, albeit to a lesser extent compared to their primary focus, Project A. In addition, the financing mechanism of Project C allows the introduction of a more complex interaction among the participants, bridging all the groups in a unique, common organizational good.

Therefore, the payoffs resulting from the investment choices are as follows:

i. Manager: $20+0.6\cdot A+0.16\cdot C$

ii. Employee who has invested in Project A: $20-c+0.4\cdot A+0.16\cdot C$

iii. Employee who has invested in Project B: $20-c+0.4\cdot B+0.16\cdot C$

where:
A: represents the total tokens invested by the group in Project A
B: represents the total tokens invested by the group in Project B
C: represents the total tokens invested by all groups in their respective Project B
c: represents the tokens invested by the employee in the project they selected.

As anticipated, this setup simulates a scenario that reflects the intrinsic conflict of interest between managers and employees. Project A is explicitly designed to attract the interests of managers; conversely, Project B is structured to contribute more to the organization's efficiency and long-term returns. Indeed, Project B, through the financing mechanism used to feed Project C, offers immediate zero compensation for the manager but ensures a future return for the entire organization. Since it generates a greater return for the whole organization, Project B allows employees to obtain a higher final compensation than they would simply by pursuing Project A, which, as mentioned, explicitly advantages only the manager.

At the end of each round, each participant receives feedback regarding the choices made by the employees of their group (choice between Project A and B, and number of tokens invested).

Groups, roles, and interaction rules described above do not change over the course of the 25 rounds of the experimental session. At the end of each round, the manager can punish the employees of their group for their choices by assigning up to a total of 10 penalty points to the three employees.[3]

There is a difference between the first 5 rounds and the subsequent 20 since this punishment has no payoff consequences in the first 5 rounds. However, in the subsequent 20 rounds, the employee's payoff is reduced by an amount equal to 10% penalty points received. In contrast, the manager's payoff is reduced by an amount that varies depending on the points assigned. Table 1 summarizes the cost of punishment for the manager:

Table 1. Costs for punishments

points	0	1	2	3	4	5	6	7	8	9	10
cost	0	1	2	4	6	9	12	16	20	25	30

Behavioral Predictions

The hypotheses of this study are derived from a synthesis of agency theory and public goods game studies, aiming to elucidate the dynamics of decision-making within a principal-agent framework.

Without punishment, self-interested employees are predicted to favor Project B, which offers a higher payoff. Drawing from agency theory, which posits that agents prioritize outcomes that maximize their individual utility (Eisenhardt, 1989), participants playing the role of employees are expected to opt for the project yielding the most significant personal benefit. Given the opportunity to contribute to either of two projects, with Project B offering the potential for higher profits, participants are anticipated to select Project B over Project A when punitive measures are merely hypothetical.

> H1: When the punishment has no payoff consequences, employees will choose project B, which gives them a higher payoff.

When the punishment mechanism has payoff consequences, it is hypothesized that most employees will shift their choice to Project A, which offers superior benefits to managers. Rooted in agency theory, this suggests that implementing punishment aligns with the interests of managers and employees (Jensen & Meckling, 1976). Consequently, employees should prioritize the project, yielding higher returns for managers and assuming an efficient incentive structure.

> H2: When the punishment mechanism has payoff consequences, most employees will choose project A.

The contribution levels of employees are expected to increase after the introduction of monetary punishment, based on prior incentive studies. For example, Fehr and List (2004) and Hannan, Hoffman, and Moser (2005) found that punishment positively impacts employee behavior. However, it must be noted that Fehr and List (2004) also observed that punishment was more efficient only as a threat, not when it was actually implemented. Hence, if punishment is used to address employees' decision-making, it may actually work against the manager's desires.

> H3: The agents will contribute more when a payoff-relevant punishment is introduced.

[3] The manager can assign penalty points because they disagree with the employee's choice of projects to invest and/or because they consider the employee's contribution too low.

Over the course of the experiment, it is hypothesized that there will be a gradual increase in free-riding behavior among employees, consistent with prior studies (Fehr & Gächter, 2000a; Andreoni, 1988). However, the impact of punitive measures by the managers is expected to diminish this effect, resulting in a lesser extent of free riding compared to scenarios without punitive measures.

> H4: Over the course of the game, a gradual increase in free-riding behavior among agents is anticipated, alongside a diminishing impact due to punitive measures.

Experimental Procedures

The experiment was conducted at the Cognitive and Experimental Economics Laboratory (CEEL) at the University of Trento (Italy). A total of 80 participants were recruited for 4 sessions (20 participants per session). Upon arrival at the laboratory, participants were randomly assigned to computer workstations and received instructions on the experiment's procedures. They were given a few minutes to read instructions,[4] which the experimenter read aloud. Before starting the experiment, subjects answered questions to confirm their understanding of the instructions.

At the end of the experiment, participants completed a questionnaire covering various aspects of the experiment, including their overall impression of the game, observations on other participants' behavior, and any messages they would have liked to convey to fellow participants (either managers or employees), if given the opportunity. On average, each participant earned €12,[5] including €3 of the show-up fee.

Results

In this section, we present the primary findings from the four experimental sessions, focusing on the selection and contributions to the project and the impact of introducing a punishment mechanism.

Figure 1 depicts the number of choices made for Project B across all rounds, highlighting a clear preference for Project B up to round 5. These initial rounds were characterized by the absence of payoff consequences resulting from the punishment points allocated by managers to those employees investing in project B.

According to the experimental design, Project B directly benefits employees and indirectly benefits the entire organization, whereas Project A primarily benefits managers. The high frequency of selecting Project B indicates that employees prioritize their direct benefits and the overall welfare of the organization over the incentives for managers, particularly during the initial five rounds when the punishment mechanism is not yet implemented.

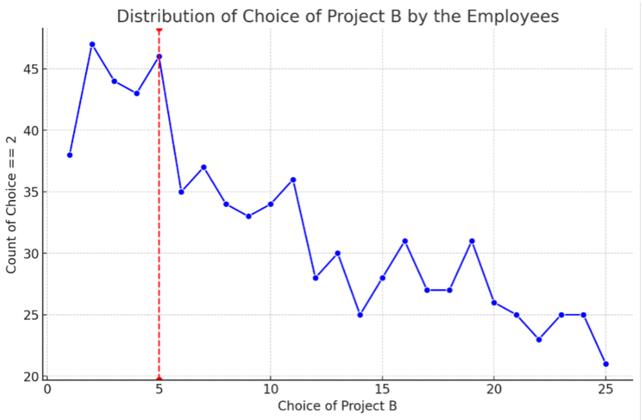

Figure 1. The selection of Project B by the employees

After round 5, there is a noticeable shift, with employees increasingly selecting Project A over Project B. Without actual punishment, employees tend to favor Project B over Project A, consistent with the expectation that self-interested and rational employees would choose Project B for better overall earnings. In the second phase (rounds 6-25), managers can use punishment to influence the employees' decision-making. We hypothesized that with the introduction of the punishment mechanism, most employees would select Project A, resulting in higher payoffs for managers. The data supports this hypothesis, showing an increase in the selection of Project A (Figure 1). While Project B was preferred in the absence of punishment, its selection declined significantly when employees faced the possibility of a payoff-relevant punishment from managers. This observation aligns with the concept that punishment can effectively guide employees' decision-making toward outcomes favored by managers. Agency theory suggests that agents will act in accordance with principals' desires when appropriate incentives are applied, a conclusion supported by prior empirical studies (Roth & O'Donnell, 1996; Anderhub et al., 2002).

For a deeper understanding of the role played by the payoff-relevant punishment on the employees' behavior, it is necessary to analyze the punishment points assigned to induce changes in project selection.

Figure 2 shows that starting from round 6, managers reduce the punishment points assigned to employees. This reversal in the managers' choices is because inflicting punishment points from round 6 onwards becomes costly for the managers. However, referring back to Figure 1, despite the decrease in punishment points imposed, it can be seen that the corresponding choices of employees shift from Project B to Project A. As previously mentioned, this effect is due to the introduction of payoff-relevant punishments. Returning to Figure 2, it becomes clear that the behaviors of managers and employees align over the rounds, with a parallel decrease in punishment points imposed and a corresponding increase in choices for Project A. Balliet et al. (2011) observed that

[4] Reported in Appendix A (translated from Italian).

[5] One round was randomly selected to determine the real earning of the subject.

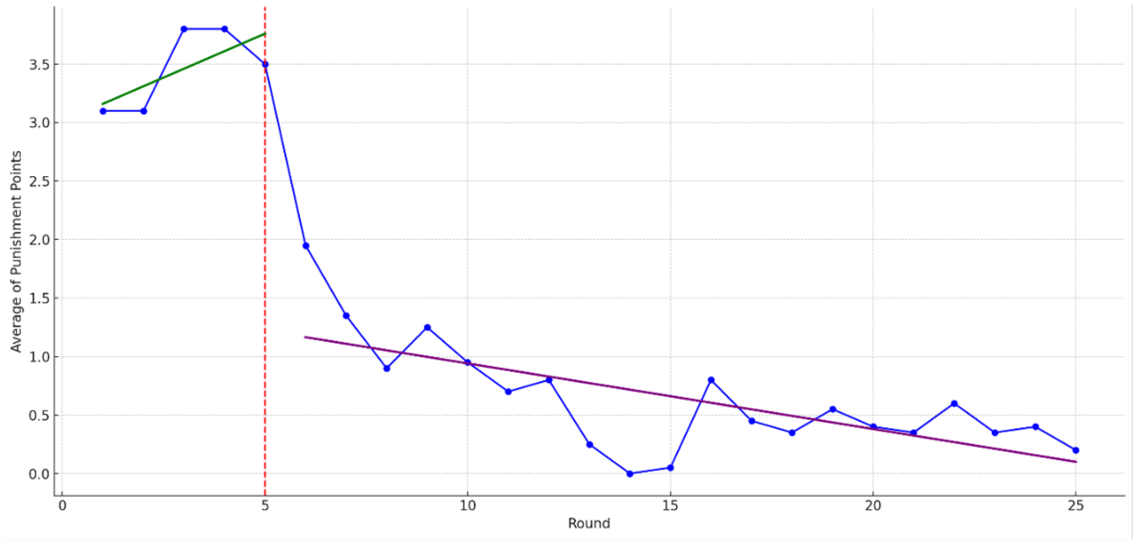

Figure 2. Punishment points assigned to change the project

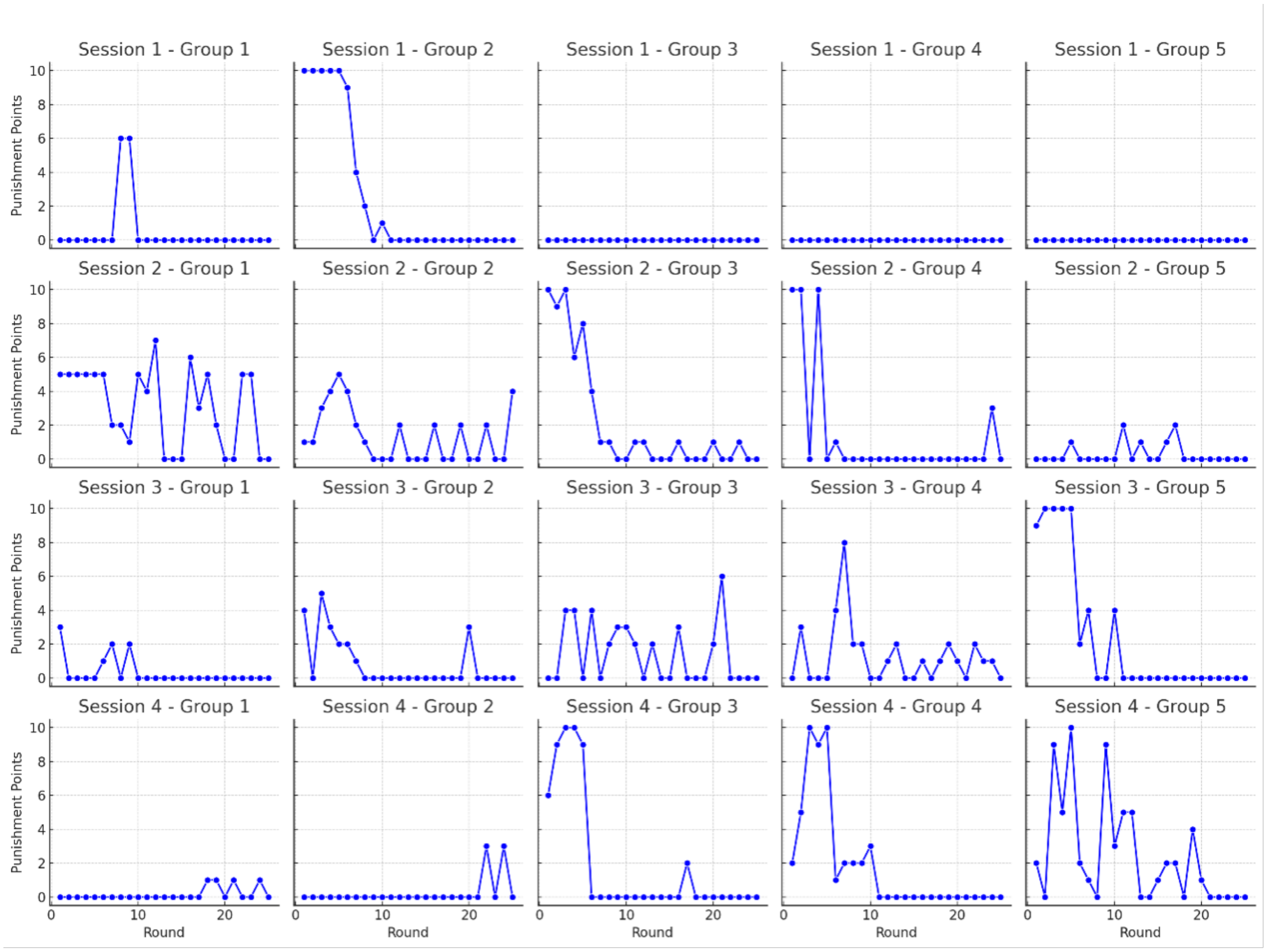

Figure 3. Punishment points assigned to change the project by each manager in each group

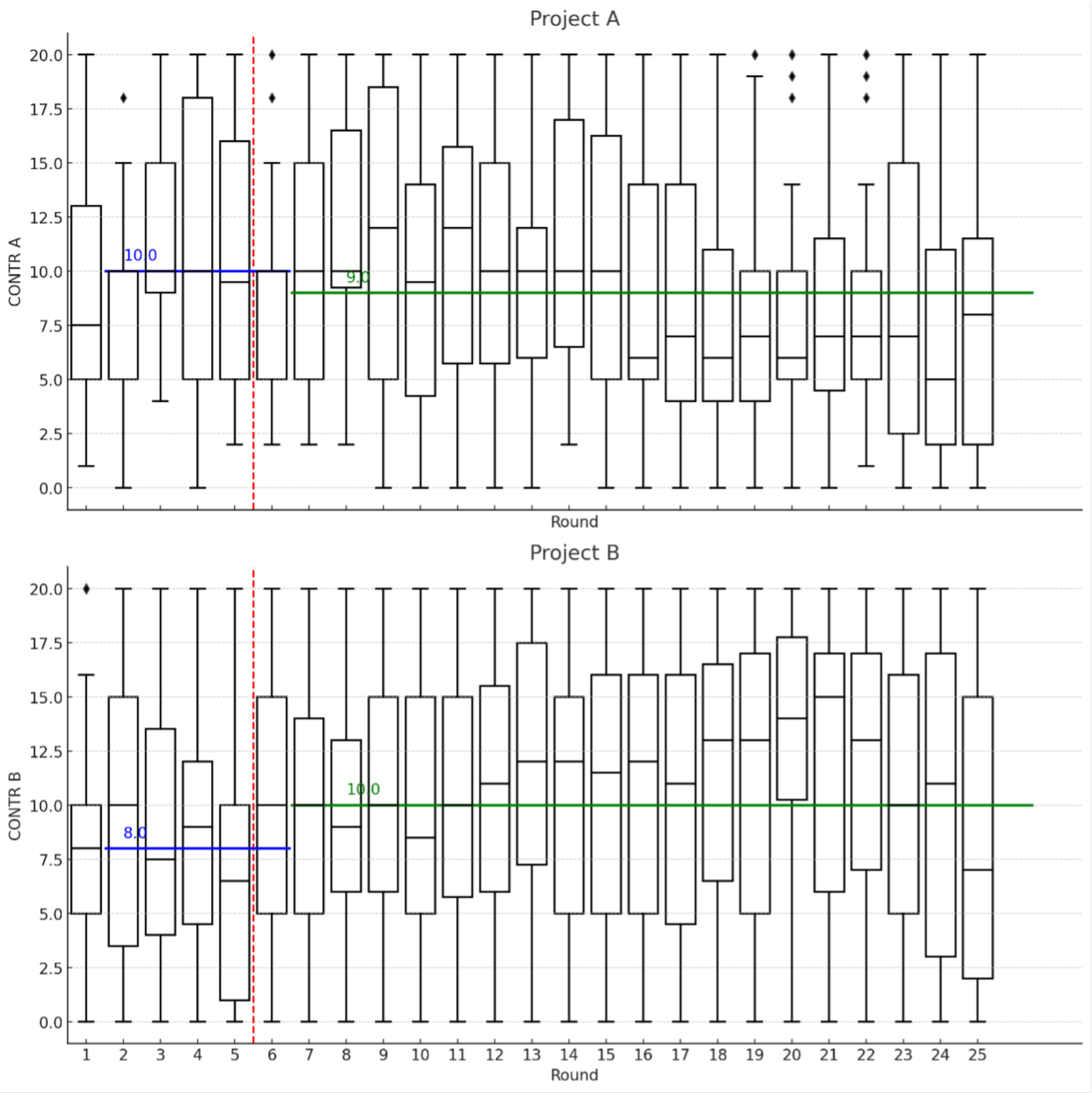

Figure 4. Contributions to project A and B

punishments in PGG are more effective when they impose costs on the punisher rather than being freely allocated. This pattern is evident in our experiment as well. The declining trend in the use of punishment (violet line) may be attributed to employees switching to Project A as the experiment progressed, rendering punishment unnecessary as it had already influenced behavior in earlier stages.

Figure 3 offers a detailed view of the points assigned by each manager across different sessions and groups.[6] As already anticipated in the discussion of Figure 2, managers allocated more punishment points during the initial phase when the punishment was costless and had no effect on the employees' behavior. In fact, since punishment points incurred no cost to the leader in the first phase, managers were expected to assign more punishment points during this period. A two-sample t-test revealed a statistically significant difference ($p < 0.001$) in the average punishment points assigned to change projects between rounds 1-5 and 6-25.

Regarding individual contributions to specific projects, Figure 4 compares the trends of average individual contributions to Project A and Project B. The average individual contributions of agents selecting Project A remained stable across rounds without (rounds 1-5) and with (rounds 6-25) payoff consequences for punishment. A Mann-Whitney U test confirmed that there was no statistically significant difference in contributions to Project A ($p = 0.228$) during rounds 1 to 5 compared to rounds 6 to 25. In contrast, the same test for Project B revealed a significant increase in contributions during rounds 6-25 compared to rounds 1-5 ($p < 0.001$).

[6] We observe variability in punishment points, with some managers assigning high points in certain rounds while others assign minimal or no points. Over time, these patterns suggest attempts to influence behavior.

Thus, introducing the payoff-relevant punishment mechanism increases contributions to Project B. Conversely, this effect is not observed for individual contributions to the joint Project A. This outcome necessitates interpretation alongside the preceding observations: although there is an increase in the number of employees inclined to select Project A during rounds 6-25, individual contributions to Project A counterintuitively decline. In addition, a Mann-Whitney U test applied to contributions during rounds 6-25 reveals that contributions to Project A were significantly lower than those to Project B during the same period ($p < 0.001$). This is to say that the effects induced by the payoff-relevant punishments are effective in determining the choices between A or B but don't impact the magnitude of contributions to A. This result is further visualized in Figure 5, showing the average individual contributions to the two projects during the game.

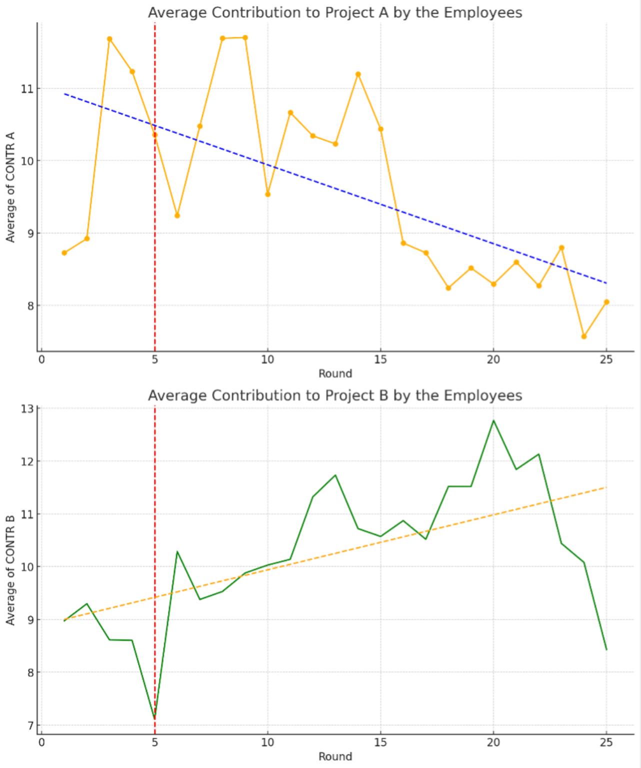

Figure 5. Average contribution to project A and B

Two potential explanations emerge. First, managers observing contributions to Project A within their group may fail to provide additional incentives to motivate employees to increase their contributions. Alternatively, management's punishments may counterproductively affect contributions to Project A. A speculative explanation for this behavior is that employees may feel justified in contributing less because they believe they are already aligning with the managers' objectives by choosing project A. Our experimental findings appear to support the latter explanation (see Figure 5).

In contrast, for the joint Project B, although there has been a decline over time in the number of employees selecting this project, individual contributions tend to increase. Analyzing the punishment points aimed at encouraging higher contributions (Figure 6), it becomes clear that more punishment points were issued during the first phase than in the second one.

Initially, the increased use of punishment points in the first phase likely contributed to higher contributions to Project A. A Mann-Whitney test revealed that contributions to Project A were significantly higher than contributions to Project B within the same rounds ($p = 0.039$). In the second phase, as contributions to Project A declined, the issuance of punishment points decreased. A two-sample t-test revealed a statistically significant difference in the average punishment points assigned to increase contributions between rounds 1-5 and 6-25 ($p < 0.001$).

Figure 7 offers a detailed view of the points assigned by each manager to increase contributions across different sessions and groups. Looking at Figure 7, one can observe variability, with a few managers assigning no points altogether and the majority assigning points at the beginning of the game, thereby confirming the average trend noted in Figure 6.

Conclusions and Discussion

This study investigated the efficacy of punishment as an incentive mechanism in the principal-agent interaction characterized by conflicting interests. Specifically, the research aimed to discern whether punishment could effectively motivate agents in scenarios exemplified by principal-agent dynamics, particularly within the context of public goods games involving multiple agents and a single principal.

Based on prior literature and agency theory, it was hypothesized that in the absence of punishment, self-interested agents would choose the project offering the higher payoff, namely Project B. This prediction has been confirmed. However, an analysis of contributions during rounds 1-5 revealed that employees opting for Project A contributed more than those selecting Project B. This outcome stems from the strategic design of the experiment. The lack of payoff-relevant punishment in the first phase – punishment which was only applied to employees deviating from the manager's preferred Project A – combined with the higher marginal returns associated with Project B, suggests that employees assumed free-riders would gravitate towards Project B. Consequently, investing in Project B during this phase posed a greater risk of financial loss due to the increased likelihood of free-riding.

Dynamics shifted entirely with the introduction of payoff-relevant punishment. As predicted, employees' choices increasingly converged toward Project A to avoid punishment from the manager. However, this shift was accompanied by a counterintuitive outcome: the introduction of payoff-relevant punishment led to a decrease in contributions to Project A, while contributions to Project B increased. This finding suggests that punishment was ineffective in fostering higher average contributions toward the intended target, Project A. One possible explanation is that employees may have interpreted

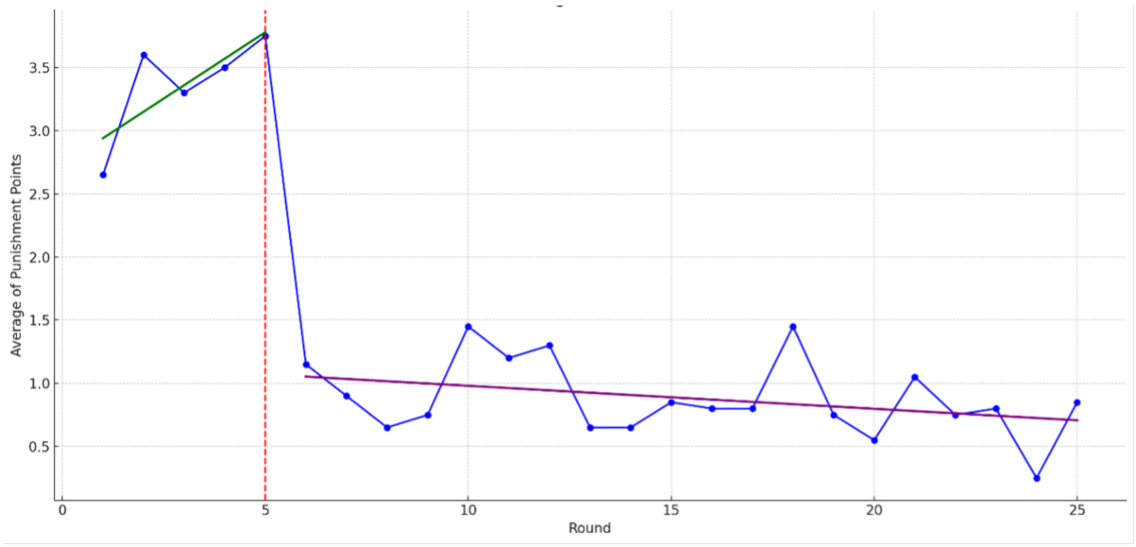

Figure 6. Punishment points assigned to increase contributions

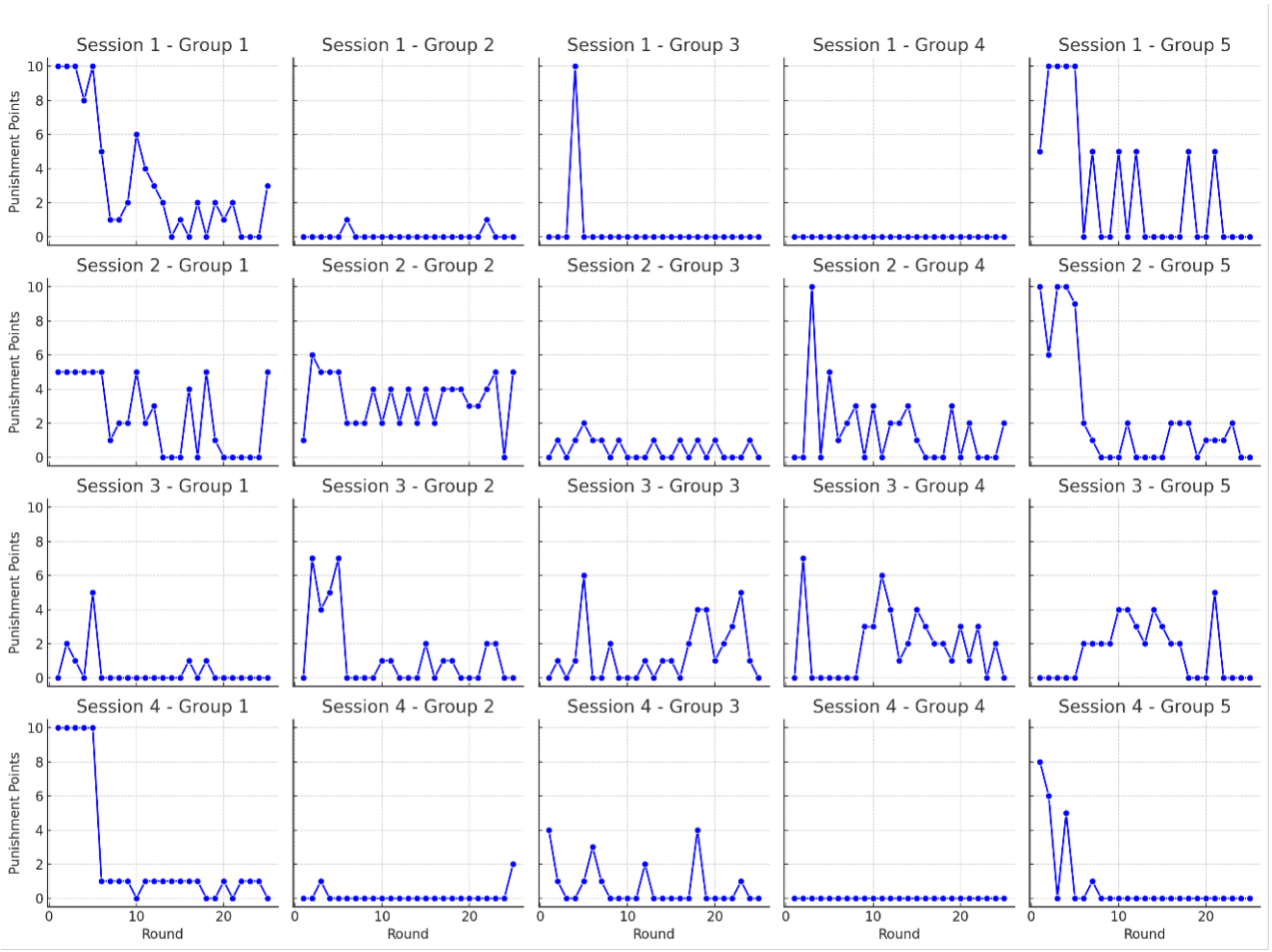

Figure 7. Punishment points assigned to increase contributions by each manager in each group

the punishment as solely enforcing the selection of Project A over Project B, rather than as a strategy to encourage greater contributions overall. Furthermore, prior literature suggests that contributions would stabilize or slightly increase over time in the presence of payoff-relevant punishments. Contrary to findings from many public goods game (PGG) studies (e.g., Fehr and Gächter, 1998; Fehr & Gächter, 2000a; Andreoni, 1988), contributions to Project A declined over time despite the implementation of payoff-relevant punishments. Similarly, and again deviating from most of the literature, contributions to Project B not only failed to decline but actually increased over time, even though free riders in Project B were not subject to the risk of punishment.

As previously suggested, a speculative explanation for this phenomenon is that employees may have felt coerced into selecting Project A, which could have negatively impacted their willingness to contribute. For instance, Kohn (1988) has argued that incentives may create a sense of being controlled and thus have a negative impact on people. Coherently with this interpretation, an opposite trend is observed in Project B, where no punishment was introduced. As the game progressed, employees increased their contributions to Project B, recognizing its potential as an investment in the organization's future. Project B's ability to yield broader societal benefits enhances the collective return on investment, motivating employees to contribute more.

An alternative explanation for our findings is that the institutional framework used in the experiment to implement punishments lacked the sophistication needed for managers to exert a fine-tuned incentivization of the employees' behavior. While the experimental design allowed managers to inform employees about the reasons for the punishment—either selecting the 'wrong' project or contributing insufficiently—this mechanism was likely not robust enough to meaningfully influence and foster fine-grained behavioral changes among the employees.

Our findings reveal a notable dichotomy in the effects of punishment-based incentives. On the one hand, these incentives significantly influenced employees' project choices, leading them to select options that provided higher returns for the manager. On the other hand, the incentives did not achieve the same level of control over employee contributions to these projects. This suggests that while punishment incentives effectively shaped *which projects* employees chose, they were far less effective in influencing *how much effort* employees put into those projects. This dichotomy highlights a limitation of punishment-based mechanisms. Although they can guide certain behaviors, such as aligning project selection with managerial goals, they fail to drive the desired level of effort or engagement. This finding suggests that punishment may work as a short-term corrective measure but fails to encourage the sustained effort required for long-term organizational success. Our experimental results, however, do not allow us to isolate the reasons for this disparity clearly. After fulfilling the manager's desired project selection, one promising interpretation is that employees felt that their behavior was "fair enough" and thus psychologically justified reducing their effort. By agreeing to follow the manager's goals, they may have perceived that they had met their obligations, losing self-motivation to go beyond the minimum required effort. This interpretation aligns with research on fairness in organizations (Andreoni, 1988; Fischbacher et al., 2001), suggesting that employees may feel less compelled to exert additional effort when they believe they have met a fairness threshold.

Our research contributes to the ongoing scholarly debate on agency theory and incentive mechanisms by providing experimental evidence on the effectiveness of punishment incentives in a public good game setting. We offer insights into the behavioral dynamics of agency relationships and highlight the nuanced interplay between incentives and decision-making processes. Our findings indicate that, while punishment may drive short-term goals, it is insufficient for fostering the long-term commitment and effort needed to achieve strategic organizational goals. In this regard, a more balanced approach that incorporates rewards or positive outcomes, such as praise, bonuses, or recognition, and punishment may be required to maintain immediate and long-term employee engagement.

Regarding the limitations of our study, the sample size of eighty participants restricts the generalizability of the findings, warranting caution in drawing definitive conclusions. Moreover, participants' status as students may influence decision-making dynamics, potentially diverging from real-world managerial scenarios. Future research avenues could explore larger datasets to enhance the robustness of findings and investigate the impact of rewards as an alternative incentive mechanism. Furthermore, scholars could investigate the role of intrinsic motivators such as reciprocity and fairness in shaping agent decisions within agency contexts, thus enriching our understanding of incentive structures beyond traditional extrinsic rewards.

Acknowledgments

We gratefully acknowledge the *Ministry of Public Administration* (*Ministero della Pubblica Amministrazione*) in Italy for funding the experiments. Additionally, we thank Viola Saredi for her valuable support in discussing the original experimental project and in collaborating on the data collection and to Ronja Sandström for her contribution to the literature review.

References

Agrawal, A., & Knoeber, C. (1996). Firm Performance and Mechanisms to Control Agency Problems between Managers and Shareholders. *Journal of Financial and Quantitative Analysis*, 31*(3), 377-397.

Ayers, R. S. (2015). Aligning Individual and Organizational Performance: Goal Alignment in Federal Government Agency Performance Appraisal Programs. *Public Personnel Management*, 44(2), 169-191.

Andreoni, J. (1988). Why free ride? Strategies and Learning in Public Goods Experiments. *Journal of Public Economics*, 37(1), 291-304.

Anderhub, Gächter, & Königstein. (2002). Efficient Contracting and Fair Play in a Simple Principal-Agent Experiment. *Experimental Economics*, 5 (1), 5–27.

Arnold, B., & Lange, P. (2004). Enron: an examination of agency problems. *Critical Perspectives on Accounting*, 15(6–7), 751-765.

Bahli, B., & Rivard, S. (2003). The Information Technology Outsourcing Risk: A Transaction Cost and Agency Theory-Based Perspective. *Journal of Information Technology*, 18*(3), 211–221.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological bulletin*, 137(4), 594.

Bosse, B., & Phillips, R. (2016). Agency theory and bounded self-interest. *Academy of Management Review*, 41*(2), 276–297.

Camuffo, A., Furlan, A., & Rettore, E. (2007). Risk sharing in supplier relations: an agency model for the Italian air-conditioning industry. *Strategic Management Journal*, 28 (12), 1257-66.

Casal, S., Ploner, M., & Sproten, A. N. (2019). Fostering the best execution regime: An experiment about pecuniary sanctions and accountability in fiduciary money management. *Economic Inquiry*, 57(1), 600-616.

Chari, M. D., David, P., Duru, A., & Zhao, Y. (2019). Bowman's risk-return paradox: An agency theory perspective. *Journal of Business Research*, 95, 357-375.

Cuevas-Rodriguez, G., Gomez-Mejia, L., & Wiseman, R. (2012). Has Agency Theory Run its Course?: Making the Theory more Flexible to Inform the Management of Reward Systems. *Corporate Governance: An International Review*, 20(6), 526-546.

Deci, E., Ryan, R., & Koestner, R. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627-668.

DeFusco, R., Johnson, R., & Zorn, T. (1990). The Effect of Executive Stock Option Plans on Stockholders and Bondholders. *The Journal of Finance*, 45(2), 617–627.

Eisenhardt, K. (1989). Agency Theory: An Assessment and Review. *The Academy of Management Review*, 14(1), 57-74.

Espín, A., Reyes-Pereira, F., & Ciria, L. (2017). Organizations should know their people: A behavioral economics approach. *Journal of Behavioral Economics for Policy*, 1(S), 41-48.

Fehr, E., & Gächter, S. (1998). Reciprocity and Economics: The Economic Implications of Homo Reciprocans. *European Economic Review*, 42(3-5), 845-859.

Fehr, E., & Gächter, S. (2000a). Cooperation and Punishment in Public Goods Experiments. *American Economic Review*, 90 (4), 980-994.

Fehr, E., & Gächter, S. (2000b). Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives*, 14*(3), 159-18.

Fehr, E., & List, J.A. (2004). The Hidden Costs and Returns of Incentives - Trust and Trustworthiness among CEOs. *Journal of the European Economic Association*, 2(5), 743-77.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71(3), 397-404.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys*, 15(5), 589-611.

Garen, J. (1994). Executive Compensation and Principal-Agent Theory. *Journal of Political Economy*, 102(6), 1175–1199.

Guala, F. (2005). *The methodology of experimental economics*. Cambridge University Press.

Hannan, R. L., Hoffman, V. B., & Moser, D. V. (2005). "Bonus versus penalty: does contract frame affect employee effort?". In *Experimental Business Research: Economic and Managerial Perspectives VOLUME II* (pp. 151-169). Springer US.

Hermalin, B., & Weisbach, M. (1991). The Effects of Board Composition and Direct Incentives on Firm Performance. *Financial Management,* 20(4), 101-112.

Jenkins, G.D., Mitra, A., Gupta, N., & Shaw, J.D. (1998). Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research. *Journal of Applied Psychology*, 83(5), 777-787.

Jensen, M., & Meckling W. (1976). Theory of the firm: Managerial behavior, agency costs, and ownership. *Journal of Financial Economics*, 3(4), 305-360.

Jensen, M., & Murphy, K. (1990). Performance Pay and Top-Management Incentives. *Journal of Political Economy*, 98(2), 225-264.

Kaplan, R. S., & Norton, D. P. (2001). Transforming the balanced scorecard from performance measurement to strategic management: Part 1. *Accounting horizons*, 15(1), 87-104.

Kohn, A. (1988). Incentives Can Be Bad for Business. *INC, pp.93-94.

Kohn, A. (1993). Why incentive plans cannot work. *Harvard Business Review*, 71(5), 54-63.

Lassar, W., & Kerr, J. (1996). Strategy and Control in Supplier-Distributor relationships: An Agency Perspective. *Strategic Management Journal*, 17(8), 613-632.

Logan, M. (2000). Using agency theory to design successful outsourcing relationships. *International Journal of Logistics Management*, 11(2), 21-32.

Manatsa, P., & McLaren, T. (2008). Information sharing in a supply chain: using agency theory to guide the design of incentives. *Supply Chain Forum: International Journal*, 9(1), 18-26.

Mittone, L. (2006). Dynamic behaviour in tax evasion: An experimental approach. *The Journal of Socio-Economics*, 35(5), 813-835.

McCoy, T., & Flesher, D. (1998). A case of an early 1900s principal-agent relationship in the Mississippi lumber industry. *Accounting, Business & Financial History,* 8(1), 13-31.

Roth, K., & O'Donnell, S. (1996). Foreign subsidiary compensation strategy: An agency theory perspective. *Academy of Management Journal*, 39(3), 678-703.

Shapiro, S. (2005). Agency Theory. *Annual Review of Sociology*, 3(1), 263-284.

Stajkovic, A., & Luthans, F. (2001). Differential Effects of Incentive Motivators on Work Performance. *Academy of Management Journal*, 4(3), 580-590.

Trautmann, S., and B. Zia (2015). *Household Finance* in World Development Report 2015: Mind, Society, and Behavior. Washington, DC: The World Bank.

Van Puyvelde, S., Caers, R., Bois, C., & Jegers, M. (2013). Agency Problems between Managers and Employees in Nonprofit Organizations: A Discrete Choice Experiment. *Nonprofit Management and Leadership,* 24*(1), 63-85.

Wiseman, R., & Gomez-Mejia, L. (1998). A Behavioral Agency Model of Managerial Risk Taking. *The Academy of Management Review,* 23(1), 133-153.

Wright, P., Mukherji, A., & Kroll, M. (2001). A reexamination of agency theory assumptions: extensions and extrapolations. *Journal of Socio-Economics,* 30(5), 413–429.

Whipple, J., & Roh, J. (2010). Agency theory and quality fade in buyer-supplier relationships. *The International Journal of Logistics Management*, 21(3), 338-52.

Zsidisin, G., & Ellram, L. (2003). An agency theory investigation of supply risk management. *Journal of Supply Chain Management*, 39(3), 15-27.

Appendix

A: Experiment Instructions

Welcome,

Thank you for being here. You are about to participate in an experiment on economic decisions. Just for arriving on time, you will receive €3 at the end of the experiment.

You will soon receive the necessary instructions. Please read them carefully, as your additional earnings will depend on the decisions you make. If you have any doubts, you can ask a staff member by raising your hand.

Please do not talk to the other participants. If you disturb your colleagues or use the computer for activities not directly related to the experiment, you will be automatically excluded from the experiment and any compensation. You can trust that everything that happens during the experiment will be in line with the rules presented in the following instructions.

During the rounds of the experiment, we will use Experimental Currency Units (ECU): 1 ECU corresponds to €0.65.

Before the first round begins, the software will randomly create groups of 4 participants. The composition of these groups will remain the same throughout the entire experiment, which consists of 25 periods. Thus, your group will be made up of you and three other participants, whose identities you do not know. At the start of the experiment, you will also be randomly assigned a role: each group will consist of **one member** with the role of **Participant 1** and **three members** with the role of **Participant 2**.

The experiment consists of two consecutive parts. You will now receive the instructions for the first part. You will be given a few minutes to read the instructions, which will then be read aloud. Before the experiment begins, you will be asked to answer some simple questions to ensure you understand the instructions.

At the end of the first part, you will receive the instructions for the second part. Finally, there will be a brief questionnaire, after which you will be informed of your earnings.

PART 1

This first part will take place over 5 periods, each consisting of two decision-making phases. In each period, your initial endowment is 20 ECU.

First decision-making phase

If you are assigned the role of Participant 2, you will only participate in the first phase of each period. You will be asked to decide how to use your ECU endowment. There are two different common projects (Project A and Project B) to which you can contribute: you will select one of the two projects and, without knowing which project the other Participants 2 have chosen, you will decide how many ECU to allocate to the development of the project you selected.

If you are assigned the role of Participant 1, you will participate in both decision-making phases. First, you will be asked to select one of the two projects and hypothetically decide your contribution. Like the other members of your group, you will receive the initial endowment of 20 ECU, but you will not actually be required to contribute to the projects.

If you have the role of Participant 2, in each period, your earnings are determined based on the total contributions to the project you have chosen. In other words, if you select Project A, the sum of the ECU invested by all Participant 2 members of your group in Project A will be multiplied by a factor of 0.4. The result is your share, which you will receive as a redistribution of the value created by the common Project A. This share will be the same for all Participant 2 members in your group who selected Project A. You will only receive this share if you have selected Project A in that period.

Similarly, the sum of the ECU invested by all Participant 2 members of your group in Project B will be multiplied by the same factor of 0.4. Again, the result is your share, which you will receive as a redistribution of the value created by the common project B, but only if you selected B in that period.

If you have the role of Participant 1, in each round, your earnings are determined based on the total contributions of all Participant 2 members to Project A. Specifically, your share from Project A is calculated by multiplying the sum of the ECU invested by Participant 2 members of your group in Project A by a factor of 0.6. However, you will not receive any earnings from the contributions to Project B.

Regarding Project B, note that this will be used to develop an additional project, (Project C), which is common to all groups (not just yours).

Independently of role and of the project selected by the participant, each participant receives an equal share of 16% of the total ECU invested by all groups in their respective Project B (if you prefer, this part can be seen as 40% of the total returns obtained by the participants in the role of Participant 2 who invested in their own Project B).

Notice that, in order for Project C to be developed, at least one participant (with role 2) must contribute to his/her respective Project B.

At the end of each period, you will be informed about the projects selected by the Participant 2 members of your group and about your earnings, which will be calculated as follows based on your role:

i. Manager: $20+0.6\cdot A+0.16\cdot C$

ii. Employee who has invested in Project A: $20-c+0.4\cdot A+0.16\cdot C$

iii. Employee who has invested in Project B: $20-c+0.4\cdot B+0.16\cdot C$

where:
A: represents the total tokens invested by the group in Project A
B: represents the total tokens invested by the group in Project B
C: represents the total tokens invested by all groups in their respective Project B
c: represents the tokens invested by the employee in the project they selected.

Second Decision-Making Phase

If you have the role of Participant 1, in the second phase, you can observe how much each Participant 2 in your group contributed to which common project in the first phase. At this point, you are asked to decide, hypothetically, whether to reduce or keep unchanged the earnings that each Participant 2 obtained in that period during the first phase.

You can allocate up to a maximum of 10 points: each point reduces the earnings of the Participant 2 to whom it is assigned by 10%. For each point assigned, you can provide a reason for your decision. Therefore, points could be assigned to one or more Participant 2s because you wished they had selected a different common project and/or because you wanted them to contribute more. You can assign points to the same Participant 2 for both reasons if you wanted that Participant 2 not only to choose a different project but also to increase their contribution. Indeed, the higher the contributions of the Participant 2s in your group to Project A, the higher your earnings will be in each period.

Since these are hypothetical choices, your decisions will not affect the earnings of the Participant 2 members of your group, and your decisions will not be communicated to them. Therefore, in each of the 5 periods, your earnings and those of the Participant 2s will remain the same as determined during the first decision-making phase.

PART 2

This part of the experiment lasts for 20 periods, with each period divided into two phases. As in PART 1, in each period, your initial endowment is 20 UMS.

Regardless of your role, the first decision-making phase remains unchanged.

For the second decision-making phase, you will only participate if you have the role of Participant 1: you will observe the decisions made by the Participant 2s in your group and will decide, **no loner hypothetically**, whether to reduce or keep unchanged the earnings that each Participant 2 obtained in that period.

In this second part of the experiment, Participant 1 will have a real effect on the earnings of the Participant 2s, and her decisions will be communicated to the Participant 2s at the end of each period.

Notice that assigning points incurs a cost, which depends on the number of points assigned, as shown in the following table:

points	0	1	2	3	4	5	6	7	8	9	10
cost	0	1	2	4	6	9	12	16	20	25	30

Participant 2s will be informed of the points assigned to them and the reasons for these decisions: specifically, the number of points for choosing a different project and the number of points for having a too low contribution.

At the end of the second phase, the earnings are determined as follows:

- Participant 1: Earnings from Phase 1 - Cost of Assigned Points
- Participant 2: Earnings from Phase 1 - 0.10 × Points Received × Earnings from Phase 1

At the end of the final round (and after completing a brief questionnaire), the software will randomly select one round from the 25 rounds. The result you obtained in the selected round will represent your compensation for the experiment. The total amount of ECU earned will be converted into euros and paid in addition to the €3 participation fee.