האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

INCENTIVE REVERSAL

by

EYAL WINTER

Discussion Paper # 505 **February 2009**

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Incentive Reversal

Eyal Winter[*†]

June 18, 2008

Abstract

By incentive reversal we refer to situations in which an increase of rewards for all agents results in fewer agents exerting effort. We show that externalities among peers may give rise to such intriguing situations even when all agents are fully rational. We provide a necessary and sufficient condition on the organizational technology in order for it to be susceptible to incentive reversal. The condition implies that some degree of complementarity is enough to allow incentive reversal.
Keywords: Incentives, Peer Effects, Team Production, Externalities

1 Introduction

The effect of rewards on performance is central to almost any debate on the optimal functioning of organizations. Rewards, if contingent on performance, can be used to boost agents' incentives to exert effort. Much of the principal-agent literature and the more recent literature on contract theory is based on this principle. The objective of this paper is to demonstrate that in a team environment rewards may affect performance in a non-monotonic way. Put differently, a promise to reward agents more generously in the case of success may paradoxically reduce agents' incentives to exert effort. We will provide this argument in a framework of fully rational agents, without building on any psychological argument, or making any behavioral assumption[1]. The argument we provide builds on the externalities among agents and uses a very simple moral hazard model in which agents' effort decisions are mapped into a probability of success for the joint project. We show that increasing rewards for all agents can result in the shirking of the set of agents who exert effort. In fact, the effect can be so dramatic that under low rewards all agents exert effort while under higher rewards almost all of them shirk. An increase in reward can change an agent's strategic consideration by making an effort decision become a dominant strategy.

[*]Center for the Study of Rationality and the Economics Department, The Hebrew University of Jerusalem, Jerusalem 91904 mseyal@pluto.huji.ac.il

[†]I wish to thank the editor and a referee for their constructive suggestions. This paper was supported by the Fritz Thyssen Stiftung.

[1]Incentive reversal driven by psychology has been shown empirically and experimentaly by Gneezy and Rustichini (2000) as well as Fehr and Falk (2002). See also Benabu and Tirole (2003) for a theoretical model explaining the phenomenon.

Suppose that under the low set of incentives agent i finds it rational to exert effort only when he observes some of his peers doing so, but under the high set of incentives he would choose to exert effort regardless of his peers' actions. In such a scenario the increase of incentives on part of agent i may rationally decrease the willingness to exert effort on the part of those whom he observes. If effort becomes a dominant strategy for agent i those whom he observe lose the endogenous incentives imposed by agent i. Consequently, the set of players exerting effort in equilibrium can shrink dramatically. The main result of this paper will characterize the technologies susceptible to incentive reversal. Specifically, we show that any technology that has increasing returns to scale for some range of production (i.e., is not concave on the entire range of production) is prone to yield reverse incentives. Furthermore, we show that if the technology has increasing returns to scale on the entire range of production, i.e., when it is supermodular, then incentive reversal can take a dramatic form in which as a result of reward increase all but one agent move from effort exertion to shirking.For the relationship between supermodularity and complementarity see Milgrom and Roberts (1990), Topkis (1998) and Segal (2003). The broader implication of our results and the intuition they provide is that in organizational environments in which peers have some information about each other's effort and where workers deal with tasks that have some degree of complementarity one should be cautious in setting up incentives. Raising rewards, which naively seems a helpful mean to boost incentives, may have a reverse effect. [2]

While we phrase our results in terms of incentives in organizations, their implications reach beyond this framework. Incentive reversal of the sort we describe here can arise in other economic environments. A fund-raiser who elicits donations for a cause should be cautious in his/her campaign. Suppose that donors are approached sequentially and that the cause requires a certain threshold of funds (making the fund-raising technology satisfy complementarity). Boosting the attractiveness of the cause in a way that would make it a dominant strategy for late movers to donate may make early movers reluctant to chip in their contributions. In fact, a phnomenon similar in spirit to the idea of incentive reversal often takes place in fund raising campaigns. A donation which is contingent on a matching from a different donor often allows fundraisers to raise more money. A similar phenomenon can arise with a design of environmental incentives for pollution abatement. Raising fines in a way that would make abatement a dominant strategy for some may discourage others to follow suit. A third environment in which our results may be relevant is the one of the Presidential election. Because of different time zones across the US the Presidential election is practically carried out sequentially, with voters at the west coast being early movers, and those in the east being late movers. If a party's campaign spending has been excessively high in some late moving states it may create an incentive problem for voters in early moving states. Voters' turnout in early moving states may be substantially lower as voters, and party activists may anticipate that turnout in the east will be sufficiently high (due to high campaign spending) and will be less motivated to show up or encourage

[2]Some papers in behavioral economics document empirical evidence of a very different type of incentive reversal. One prominent example is Gneezy and Rustichini's (2000) evidence on daycare centers. When daycare centers in Israel introduced small fines for parents who failed to pick up their kids on time, the overall effort towards on-time pickup declined. Roughly, the fine was perceived by parents as a convenient substitute for the shame and embarrassement linked with a late pickup.

other supporters of their party to do so[3].

This paper is part of a large literature on incentives in organizations using principal-multiagent models, much of which stems from Holmstrom's (1982) seminal paper. Papers such as Itoh (1991), Baliga and Sjoestrom (1998), Che and Yoo (2001), and Winter (2007) discuss the design of optimal incentives in teams and the way they are affected by the underlying environment (such as information among peers and prospects of collusion). Che and Yoo (2001) have also pointed to the role of implicit incentives among peers, which has a central role also in our context. To the best of my knowledge, however, the possibility of incentive reversal has not been mentioned in this literature, largely because this literature is concerned with *optimal* incentive mechanisms, while demonstrating incentive reversal requires a comparison of two mechanisms at least one of which is suboptimal.

We start in Section 2 by demonstrating incentive reversal with the simplest possible example. We use a two-agent example and show how a 15% increase in rewards for both agents shifts the equilibrium in the organization from full effort to 50% effort. In Section 3 we set up a simple model of moral hazard similar to ones used in Winter (2004) and Winter (2006) and define the notion of incentive reversal. Unlike the current paper both Winter (2004) and Winter (2006) deal with optimal mechanisms. In Winter (2004) the strategic environment is one of simultaneous effort decisions (no information among peers) and it is shown that full implementation of effort requires the principal to discriminate among her agents. The result is used to explain the role of hierarchies in organizations without authority. Winter (2006) introduces an environment of sequential production, and the paper deals with various optimal design questions, such as the allocation of agents (depending on their skills) to different production slots, as well as the allocation of tasks across different stages of the game.

In Section 4 we provide necessary and sufficient conditions for incentive reversal in the two-agent case, which will be used as an intermediary result towards the general case. Section 5 provides a characterization of incentive reversal for arbitrary number of agents, and shows that its form is made severe with increasing returns to scale. We conclude in Section 6. All proofs except for the proof of Proposition 1 are relegated to the Appendix.

2 Example

Two agents form a team to manage a joint project with each of them in charge of a different task. Each agent can either shirk or exert effort. If an agent exerts effort he performs his task successfully with certainty. If he shirks his task succeeds with probability $\alpha < 1$. The common cost of effort is c. The joint project will succeed if and only if both tasks end successfully. The principal who can neither monitor agents' effort nor the outcome of individual tasks offers the agents rewards that are contingent only on the project's outcome. Specifically, if the project suceeds, agent 1 gets v_1 and agent 2 gets v_2, and they both get zero if the project fails.

[3]The above examples seem even more relevant in view of recent laboratory results on incentive reversal. Klor, Kube, Winter and Zultan (2008) designed 2-person and 3-person effort games which are susceptible to incentive reversal. Both cases have shown sharp evidence for incentive reversal. Moreover, as predicted by the theory incentive reversal emerged only in a sequential environment and not in a simultaneous one.

Assume now that agents move sequentially. Agent 1 acts first and agent 2 observes agent $1's$ effort decision (but not the outcome of his task) and makes his own effort decision. We wish to raise the following question: Is it possible that higher rewards for both agents will generate less effort in equilibrium?

Let us set $\alpha = 0.9$, $c = 1$, and assume first that $v_1 = 5.5$ and $v_2 = 11$. It is easy to verify that under these rewards both agents exert effort in the unique (subgame-perfect) equilibrium of the game. Player 2's optimal strategy is to exert effort if and only if player 1 exerts effort (which follows from the fact that $v_2 - c > \alpha v_2$, and $\alpha v_2 - c < \alpha^2 v_2$). Player 1's optimal strategy is therefore to exert effort (since $v_1 - c > \alpha^2 v_1$).

Suppose now that the principal raises the rewards of both agents by 15%, yielding $v_1^* = 6.33$ and $v_2^* = 12.66$. It is now a dominant strategy for agent 2 to exert effort (since $\alpha^2 v_2^* < \alpha v_2^* - c$). But now the first agent, who realizes that the second will invest no matter what agent 1 is doing, loses his incentive to exert effort: $\alpha v_1^* > v_1^* - c$. Thus, the unique equilibrium of the game yields only player 2 investing. Hence, the principal spent more money and got less effort.

In the sequel of the paper we will characterize the technologies under which such reverse incentives can happen.

3 The Model

The organizational project involves a set N of n identical agents who collectively manage a project. The project involves a sequential production. Each agent in his turn has to decide whether to exert effort in the performance of his tasks or not. We denote by $d_i = 1$ agent $i's$ decision to exert effort and by $d_i = 0$ his decision to shirk. When an agent is making his effort decision he is informed about the effort decisions of all his peers who acted earlier. The cost of effort is c and is constant across all players. The technology of the organization maps a profile of effort decisions into a probability of the project's success. We denote by $p(s)$ the probability that the project succeeds if exactly s agents exert effort and $n - s$ shirk. The technology is assumed to be increasing, i.e., if $s_1 > s_2$ then $p(s_1) > p(s_2)$.

The principal who cannot monitor the agents for their effort but knows only the project's outcome sets up a mechanism $v = (v_1, ..., v_n)$ by which agent i receives the payoff v_i if the project succeeds and zero otherwise. [4] For a given mechanism v players are facing a perfect information game.

Denoting by T_i the set of agents preceding agent i in the order of moves, we can specify the game formally as follows. The strategy for player i is a function σ_i: $2^{T_i} \longrightarrow \{0, 1\}$ specifying to each player whether to exert effort or to shirk as a function of the information he possesses on other agents' decisions. For every strategy profile $\sigma = (\sigma_1, ..., \sigma_n)$ we denote by $E(\sigma)$ the set of agents who exert effort under the profiles σ. Finally, the payoff for player i under $\sigma = (\sigma_1, ..., \sigma_n)$ is given by $f_i(\sigma) = v_i p(E(\sigma)) - c$ if $i \in E(\sigma)$ and $f_i(\sigma) = v_i p(E(\sigma))$ if $i \notin E(\sigma)$. Generically, the extensive form game described above has a unique subgame-perfect equilibrium. For the non-generic case we assume that indifferences are resolved in favor of exerting effort. We denote by $E(v)$ the set of agents who exert

[4] Zero payment in case of failure is a standard assumption of limited liability.

effort in equilibrium under the reward vector v.

By incentive reversal we refer to situations in which an increase of rewards for all the agents results with the shrinkage of the (equilibrium) set of investing agents. Formally, we say that the technology p is susceptible to reverse incentives if there exist two reward vectors v^1, v^2 such that $v^1 < v^2$ (coordinatewise) and yet $E(v^1) \not\supseteq E(v^2)$. If p is not susceptible to reverse incentives we will say that it is *immune* to reverse incentives.

Two properties of the technology p will play an important role in our analysis. We say that p satisfies increasing returns to scale (IRS) if $D(k) = p(k+1) - p(k)$ is increasing in k. We say that p satisfies decreasing returns to scale (DRS) if $D(k)$ is weakly decreasing in k. The properties of IRS and DRS correspond to the convexity and the concavity of the technology p respectively. The IRS condition represents situations with complementarity across agents' tasks since the higher an agent's marginal contribution, the more other agents contribute. In contrast, DRS represents substitution across tasks since effort becomes less effective the more agents contribute. If there are only two agents in the organization p must have either DRS or IRS. We examine this case in detail in the next section.

4 The Two-agent Case

The intriguing situation of incentive reversal that we demonstrated in Section 2 relied on the complementarity of the two tasks that form the project. In this section we show that complementarity is both a necessary and sufficient condition for incentive reversal in the two-agent case. We will later address the case of an arbitrary number of agents to characterize incentive reversal, and will use the analysis in this section as a step in the proof for the general case.

Proposition 1 : If $n = 2$, then p is susceptible to reverse incentives if and only if it has IRS.

We will start by showing that any technology which is susceptible to incentive reversal must be convex. This result has a flavor of reverse engineering. A typical result in economics takes the technology as part of the premise and the implication consists of a comparative statics result. In contrast, we assume a certain comparative statics result and deduce the properties of the technology.

Lemma 1: If for some $v^1 < v^2$ we have $E(v^1) \not\supseteq E(v^2)$ then the technology p has IRS.

In the proof of Lemma 1 we wil show that incentive reversal implies that under the low vector of incentives player 2's strategy must be to exert effort if and only if player 1 does so, which as we shall see implies complementarity.

Proof of Lemma 1: If $E(v^1) \not\supseteq E(v^2)$, then the vector of equilibrium actions under v^2 cannot be $(1,1)$ and is therefore either $(0,0), (0,1)$, or $(1,0)$. Consider the four strategies of player 2:

(a) Exert effort if and only if player 1 does so.

(b) Always exert effort (i.e., effort is a dominant strategy)

(c) Exert effort if and only if player 1 shirks.

(d) Never exert effort (i.e., shirking is a dominant strategy)

It is easy to verify that regardless of the technology (as long as it is monotonic increasing) player 1 has higher incentive to exert effort under (a) than under (b) or (d) (meaning that if player 1 exerts effort when facing a player 2 with a strategy (b), then for the same reward he will also exert effort when facing a player 2 with a strategy (a)). Likewise, he has a higher incentive to exert effort under (b) or (d) than under (c). This will play a role in our analysis later. We now distinguish cases:

Case 1: The equilibrium actions under v^1 are $(1,1)$. Since under v^1 both exerted effort it must be the case that player 2's best response to an effort by player 1 under v^1 is to exert effort. So this must also be the case under v^2. Hence, $(1,0)$ is not possible under v^2. Assume now that $E(v^2) = (0,0)$. Consider the best response of player 2 to effort by player 1 under v^2. If this best response is shirking, then shirking is a dominant strategy for player 2 under v^2. Hence it must also be a dominant strategy under v^1, but this contradicts $E(v^1) = (1,1)$. So the best response of player 2 to effort by player 1 under v^2 must be effort. So under v^2 player 2 exerts effort if and only if player 1 does so. Hence, player 2's strategy is (a) where player 1's incentive to exert effort is the highest and even more so since $v^2 > v^1$. This contradicts $E(v^2) = (0,0)$. Hence, $(0,1)$ is the only possible case. So under v^1 the equilibrium outcome is $(1,1)$ and under $v^2 > v^1$ it is $(0,1)$. Consider again player 2's strategies under both payoff vectors. Under v^1 his strategy must either be (a) or (b), while under v^2 it must either be (b) or (c). Moreover if it is (b) under v^1 it must also be (b) under v^2 since $v^2 > v^1$. But the latter contradicts that the equilibrium under v^2 is $(0,1)$. Hence player 2 exerts effort under v^1 if and only if player 1 does so. This implies the following two incentive constraints: $p(2)v_2^1 - c > p(1)v_2^1$ and $p(1)v_2^1 - c < p(0)v_2^1$ or $\frac{c}{p(2)-p(1)} < v_2^1 < \frac{c}{p(1)-p(0)}$, or $p(2) - p(1) > p(1) - p(0)$, which means IRS.

Case 2: The equilibrium actions under v^1 are $(0,1)$. In this case it must be that under v^2 the equilibrium actions are $(0,0)$. This is impossible: if under v^1 player 2's best response to player 1's choosing 0 was to choose 1, it should also be the case under v^2.

Case 3: The equilibrium actions under v^1 are $(1,0)$. Again in this case it must be that under v^2 the equilibrium actions are $(0,0)$. It must be the case that under v^1 player 2 strategy is either (c) or (d), whereas under v^2 it is either (a) or (d). But this means that player 1 has more incentive to exert effort under v^2 than under v^1. This contradics $E(v^2) = (0,0)$. and completes the proof of the Lemma. **Q.E.D.**

We now proceed with the converse of Lemma 1 namely,

Lemma 2: If p satisfies IRS, then it is susceptible to reverse incentives.

The proof of Lemma 2 is based on the intuition that we provided in the example of Section 2. In order to generate incentive reversal we design the payoffs in such a way that under the high scheme player 2 will have a dominant strategy to exert effort while under the low scheme player 2 exerts effort (in equilibrium) if and only if player 1 does so.

Proof of Lemma 2: Consider the following vectors of rewards for the two agents $v_1 = \frac{c}{p(2)-p(0)}$ and $v_2 = \frac{c}{p(2)-p(1)}$. Under this reward vector there exists a subgame-perfect

equilibrium in which both agents exert effort. The strategies of the players in this equilibrium are as follows. Player 1 exerts effort and player 2 exerts effort if and only if player 1 exerts effort. To verify that this is an equilibrium simply note that v_1 and v_2 satisfy $v_1 p(2) - c = v_1 p(0)$. Hence, player 1 is indifferent between exerting effort and shirking given the strategy of player 2. Furthermore, $v_2 p(2) - c = v_2 p(1)$, which means that player 2 is best responding to the action taken by player 1. Next, if player 1 is shirking, then player 2 is better off shirking as well because $v_2 p(1) - c < v_2 p(0)$, which follows from the fact that $p(2) - p(1) > p(1) - p(0)$ (the IRS condition). It is easily seen that for any $\varepsilon > 0$ arbitrarily small the equilibrium described above is the unique subgame-perfect equilibrium of the game given by the rewards $v_1^* = v_1 + \varepsilon$ and $v_2^* = v_2 + \varepsilon$ for $\varepsilon > 0$ small enough. (it is sufficient to take $\varepsilon < \frac{1}{2}(\frac{c}{p(1)-p(0)} - v_1)$, which is positive because of IRS), and define the following new vectors of rewards $\overline{v} = (\overline{v_1}, \overline{v_2})$ by $\overline{v_1} = v_1^* + \varepsilon$ and $\overline{v_2} = \frac{c}{p(1)-p(0)} + \varepsilon$. Note that $\overline{v_2} > v_2^*$ because of IRS. We can now compare the equilibria under the two mechanisms. Under v^* the unique equilibrium outcome is with both players exerting effort and hence $E(v^*) = \{1, 2\}$. Under $\overline{v}$, on the other hand, player 2's optimal action is to exert effort also when agent 1 is shirking. This is because $\overline{v_2} p(1) - c > \overline{v_2} p(0)$. Hence, exerting effort is a dominant strategy for player 2. But now player 1 will find it optimal to shirk under $\overline{v_1}$ because he is no longer threatened by the shirking of player 2. The unique equilibrium of the game under $\overline{v}$ has player 1 shirking and player 2 exerting effort. Hence, $E(\overline{v}) \subseteq \{2\} \subsetneq \{1, 2\} = E(v^*)$, which establishes the result. **Q.E.D.**

The Proof of Proposition 1: follows directly from Lemma 1 and Lemma 2. **Q.E.D.**

It is interesting to point out that if agents make their effort decisions simultaneously (rather than sequentially), then incentive reversal is not possible. In a simultaneous game there may be a multiplicity of equilibria. However, it is easy to verify that if a certain reward vector (v_1, v_2) admits a Nash equilibrium in which both agents exert effort, then any increase of rewards for both agents will sustain this equilibrium as well. Furthermore, if (v_1, v_2) sustains only an equilibrium in which one of the agents exerts effort, then a vector of increased rewards will either sustain an equilibrium in which one agent exerts effort (not necessarily the same agent) or an equilibrium in which both exert effort. This suggests that information about peers (which is prevalent in almost every team environment) is crucial for incentive reversal.

We now move to discuss the multiple-agent case.

5 The General Case

For an arbitrary number of agents the technology p may satisfy neither IRS nor DRS. There can be, therefore, two potential extensions for Proposition 1. It turns out, however, that incentive reversal can prevail under much weaker conditions than IRS. In fact, any technology that is not DRS is susceptible to incentive reversal.

Theorem 1: A technology p is immune to incentive reversal if and only if it has decreasing returns to scale.

Theorem 1 indicates that it is the immunity to reverse incentives which is exceptional, and not its susceptibility to reverse incentives. Only when the marginal contribution of effort is declining for the entire range of production can we guarantee that no incentive reversal is possible. If the technology is not concave there must be a range of production such that the marginal contributions are increasing in this range. This means that we can generate two schemes v^1 and v^2 such that under v^1 the equiibrium behavior for last player in this range is to exerts effort if and only if his predecessor did so, and under v^2 he has a dominant strategy to exert effort. This will alow us to generate incentive reversal using similar ideas as in Proposition 1. The proof of Theorem 1 is given in the Appendix via Proposition 2 and Proposition 3.

We conclude this section by showing that incentive reversal may take a rather dramatic form if the technology is one of increasing returns to scale. For such technologies an increase of rewards can paradoxically cause all but a single agent to move from exerting effort to shirking. This phenomenon is sustained by a domino effect in which one agent after another realizes that his own action will not affect that taken by each of his subsequent peers, which induces almost everyone to shirk.

Proposition 4: If p has increasing returns to scale, then there exist two reward vectors $v^1$and v^2 with $v^2 > v^1$ such that under v^1 all agents exert effort in eqilibrium, while under v^2 only one agent does so.

It is instructive, at this stage, to discuss the relation between the principal's optimal mechanism and incentive reversal. By the optimal mechanism we refer to the vector of rewards that induces all players to exert effort (in subgame perfect equilibrium[5]) and does so at the minimal total cost for the principal. Winter (2006) shows that under increasing returns to scale technology p the optimal mechanism (implementing effort by all players) is given by $v = (\frac{c}{p(n)-p(0)}, \frac{c}{p(n)-p(1)}, ..., \frac{c}{p(n)-p(n-1)})$. With this mechanism the principal never falls into the trap of incentive reversal. It is easy to verify that under this mechanism each player i is indifferent between shirking and exerting effort provided that all his predecessors exerted effort. However, if one or more of these players shirk player i strictly prefers to shirk as well. Hence, the implicit incentive to exert effort generated by the sequential structure of the game allows the principal to reduce the explicit incentive yielding the optimal mechanism. We have seen in Proposition 1 (as well as in Theorem 1) that the implicit incentive described above plays a central role in the presence of incentive reversal. If the principal raises the payoff of the player to the extent that choosing effort becomes a dominant strategy it will remove the implicit incentive on part of the second to last player who will choose to shirk. This shirking will induce all earlier players to shirk as well.

Consider now a decerasing returns to scale technology p. The optimal mechanism for such a technology is given by $v' = (\frac{c}{p(n)-p(n-1)}, \frac{c}{p(n)-p(n-1)}, ..., \frac{c}{p(n)-p(n-1)})$. (see Winter (2006)) and each player is exerting effort in a dominant strategy. Hence, implicit incentives play no role for such a technology, which is the reason why the technology is not susceptible to incentive reversal.

[5]Assuming that indifferences are resolved in favor of exerting effort.

6 Discussion

Externalities among peers in team environments can give rise to intriguing situations of incentive reversal in which all agents are promised higher rewards while the set of agents exerting effort in equilibrium shrinks. We show that such situations can arise without any behavioral assumption, and characterize the technological environment that breeds them. Our basic assumption is that agents respond rationally to the effort decisions taken by their peers while taking into account the organizational technology. Recently, Gould and Winter (2007) provided evidence to this effect using a database on professional baseball players. The empirical analysis shows that a player's performance increases with the performance of players with whom his production relationship is one of complementarity. On the other hand players' performance decline with the performance of their substitutes. For example, a player's batting average significantly increases with the batting performance of his batting peers, but decreases with the quality of the team's pitching. These findings jointly with our results on incentive reversal offer a potential implication (that of course should be judged only in view of other incentive considerations): Bonuses based only on the team performance (like those offered during playoffs) should not be excessively high for batters who appear late in the order as it may quash the incentives of batters who appear earlier in the order.

In this paper we have assumed that agents move sequentially and that the resulting game is one of perfect information. While this assumption may seem stringent, it is merely simplifying. Incentive reversal can arise in a partially sequential environment. Assume for example that player 2 observes the effort decision of player 1, and the rest of the players are ignorant of each other's effort. In this almost simultaneous framework incentive reversal can be shown to occur using the same argument as in our two-person example.

We have mentioned that under incentive reversal the high vector of incentives cannot be optimal because the principal would be better off shifting to the low vector. However, our results are relevant to the design of optimal mechanisms in three respects. Firstly, in real-life organizations it is reasonable to assume that the optimal mechanism is reached through a series of improvements over existing mechanisms. Our observation about incentive reversal says that if the principal detects that the number of agents exerting effort is too low, the way to induce more effort is not necessarily by increasing rewards. It may be possible to boost effort by reducing rewards. Secondly, incentive reversal is relevant to optimal mechanisms with principals who face external constraints. Suppose, for example, that incentive reversal occurs for the reward vectors v^1 and v^2 with $v^1 < v^2$and the principal is constrained to offer at least v^2 to his agents because of the non-competitiveness of the market. Suppose that v^1 is the optimal mechanism under the unconstrained problem. Incentive reversal says that in the constrained problem, rather than offering the contract in which the entire payoff v2 is contingent on the success of the project the principal may be better off using a different contract in which he promises a payment of $v^2 - v^1$ regardless of the project's outcome in addition to a payment of v^1, which is contingent on its success. Thirdly, another (related) form of incentive reversal may arise when rewards are fixed but the cost of effort varies. Consider an organization that operates under an optimal mechanism that induces all agents to exert effort. Suppose that at some stage the cost of effort drops for all agents (for example, because experience has improved their

skill). If the technology is not concave it may happen that the same mechanism will now induce a smaller set of agents to exert effort. This can also happen when effort becomes a dominant strategy for late movers killing the incentives of earlier movers to exert effort. Such incentive reversal can take place in the short run of an organization operating under the optimal incentive mechanism until the mechanism is amended to take into account the new vector of the cost of effort.

While our model and results are based on full rationality, our analysis offers an insight into a different type of incentive reversal that is behaviorally based. There is considerable empirical and experimental evidence on psychological peer effects that shows that workers are typically reluctant to exert effort when they observe their peers shirking (see, e.g., Fischbacher (2002), Gaechter and Fehr (2001)). This reluctance may in fact be quite effective in sustaining a high level of effort within teams, because it serves as an implicit threat against shirking. In such teams an increase in rewards may quash this implicit threat. Some agents may find it attractive enough to exert effort even when observing their peers shirking, which in turn may encourage these peers to shirk. This will give rise to an incentive reversal quite similar in spirit to the one described in this paper. Indeed, it is one that may arise for any technology and for a wider range of initial reward vectors.

7 Appendix

Proposition 2: If p does not have DRS, then it is susceptible to incentive reversal.

Proof: We will construct two reward vectors v^1 and v^2 such that $v^1 < v^2$ and the set of players exerting effort under v^2 is a strict subset of the set of players investing under v^1. If p does not have DRS, then there exists some $k < n$ such that

$$p(k+1) - p(k) > p(k) - p(k-1) * *$$

Let v^* satisfy $v^* > \frac{c}{\min_j [p(j)-p(j-1)]}$ and note that for a player who is promised v^* exerting effort is a dominant strategy. Consider now the following rewards vector:

$$v^1 = \begin{cases} v^* \text{ for } j > n-k+1 \\ 0 \text{ for } j < n-k \\ \frac{c}{p(k+1)-p(k)} \text{ for } j = n-k+1 \\ \frac{c}{p(k+1)-p(k-1)} \text{ for } j = n-k \end{cases}$$

For this reward vector the unique equilibrium of the game is for players $j > n-k+1$ to exert effort, for players $j < n-k$ to shirk, and for players $j = n-k$ and $j = n-k+1$ to exert effort as well. Furthermore, we argue that player $n-k+1$ exerts effort if and only if player $n-k$ exerts effort. This follows from the inequality ** and from the fact that for $j = n-k+1$ we have $v_j^1 p(k+1) - c \geq v_j^1 p(k)$ (which means that it is optimal for j to invest if $n-k$ invests) and $v_j^1 p(k) - c < v_j^1 p(k-1)$ (which means that j is better off

shirking if $n-k$ shirks). Consider now a new reward vector given by

$$v^2 = \begin{cases} v^* \text{ for } j > n-k+1 \\ 0 \text{ for } j < n-k \\ v^* \text{ for } j = n-k+1 \\ \frac{c}{p(k+1)-p(k-1)} \text{ for } j = n-k \end{cases}$$

It is now a dominant strategy for player $n-k+1$ to exert effort. Hence it is optimal for player $n-k$ to shirk because for $j = n-k$ we have $v_j^2 p(k+1) - c < v_j^2 p(k-1)$ (recall that if $n-k$ shirks $n-k+1$ shirks as well). The set of players exerting effort in equilibrium is now $\{j : j \geq n-k+1\}$ and it is a strict subset of the set of players investing under v^1, which is $\{j : j \geq n-k\}$. We finally note that the equilibrium outcome will not change if we replace v^2 with $v^2 + \varepsilon = (v_1^2 + \varepsilon, ..., v_n^2 + \varepsilon)$, which establishes the result. **Q.E.D.**

Proposition 3: If p satisfies DRS, then it is immune to incentive reversal.

For the proof of Proposition 3 we need the following two lemmas:

Lemma 3: Let p have DRS. Consider any decision node of player i after players $1, 2, , ..., i-1$ have already acted. Denote

$S_i^1 = \{j > i; d_j = 1 \text{ at the subgame where } d_i = 1\}$ and $S_i^0 = \{j > i; d_j = 1 \text{ at the subgame where } d_i = 0\}$, and let $s_i^k = \#S_i^k$, where $k = 0, 1$. Then we have $s_i^1 \leq s_i^0$.

Proof: We prove the claim by (backward) induction. The argument for $i = n-1$ is straightforward and it follows from the fact that if player n exerts effort when player $n-1$ exerts effort then he will do so also when player $n-1$ shirks, which follows from the concavity of the technology. We now assume the statement is true for all players $j \geq i$ and consider player $i-1$. Let $d_i(k)$ be the action of player i after player $i-1$ chooses k, where $k = 0, 1$, and let $S_i^1(k), S_i^0(k)$ be the sets defined above in the subgame where player $i-1$ chose k, where $k = 0, 1$.

We will distinguish the following cases:

Case 1: $d_i(1) = 1$ and $d_i(0) = 0$. Consider first the subgame in which $i-1$ chooses 1. If $s_i^1(1) < s_i^0(1)$, then it cannot be the case that i invests at this node (because by shirking he will generate at least as many investing agents and will save the cost of effort). Furthermore, by the induction hypothesis, $\#S_i^1(1) \leq \#S_i^0(1)$. Hence, $s_i^1(1) = s_i^0(1)$. Consider now the subgame in which $i-1$ chooses 0. Because of the symmetry of the technology, the equilibrium continuation at each player's decision node depends only on the number of agents who exerted effort prior to that stage and does not depend on who they are. This implies that $s_i^0(1) = s_i^1(0)$. By the induction hypothesis we again have $s_i^1(0) \leq s_i^0(0)$. If $s_i^1(0) = s_i^0(0)$ then we have $s_i^1(1) = s_i^0(1) = s_i^1(0) = s_i^0(0)$. But if this is the case it must be optimal for i to exert effort after $i-1$ chooses 0 if it was optimal for him to do so after $i-1$ chose 1. This follows from the property of decreasing returns to scale of p. Specifically, $i's$ marginal contribution to the project's success is greater after $i-1$ chooses 0. So it must be that $s_i^1(0) < s_i^0(0)$, and thus $s_i^0(0) > s_i^1(0) = s_i^0(1) = s_i^1(1)$. Note now that $s_{i-1}^1 = s_i^1(1) + 1$ and $s_{i-1}^0 = s_i^0(0)$ and hence because $s_i^0(0) > s_i^1(1)$ we have $s_{i-1}^0 \geq s_{i-1}^1$, which is what we need.

Case 2: $d_i(0) = d_i(1) = 1$. In this case using the same arguments as above (i.e., induction hypothesis plus symmetry) we have $s_i^1(0) = s_i^0(1) \geq s_i^1(1)$. Furthermore, $s_{i-1}^0 = s_i^1(0) + 1 \geq s_i^1(1) + 1 = s_{i-1}^1$ and we are done.

Case 3: $d_i(0) = d_i(1) = 0$. We now have $s_i^0(1) = s_i^1(0) \leq s_i^0(0)$. Furthermore, $s_{i-1}^0 = s_i^0(0) \geq s_i^0(1) = s_{i-1}^1$ as needed. Finally,

Case 4: $d_i(0) = 1$ and $d_i(1) = 0$. By symmetry we have $s_i^1(0) = s_i^0(1)$. Furthermore, $s_{i-1}^0 = s_i^1(0) + 1 > s_i^0(1) = s_{i-1}^1$, and we are done. **Q.E.D.**

Lemma 4: Let p be a technology with DRS and let $v^1 < v^2$ be two reward vectors. Denote by s^1 and s^2 the number of players who exert effort in equilibrium under the vectors v^1 and v^2, respectively, when p is the prevailing technology. Then we have $s^1 \leq s^2$.

Proof: We prove the result by induction on the number of players n. For $n = 2$ the result follows directly from Proposition 1. Suppose now that the statement is true for any game with $n \leq k$ and consider now a game with $k+1$ players. Let G^1 and G^2 be the games with $k+1$ players with reward vectors v^1 and v^2 respectively. We denote by s^1 and s^2 the number of players exerting effort in the equilibrium of the two games respectively. We will show that $s^1 \leq s^2$. We next denote by d_1^1 and d_1^2 the equilibrium actions of player 1 in the games G^1 and G^2 respectively. We note that following the action of player 1 we enter a subgame that involves k players with a technology $p^* = p$ if player 1 chooses 0 and $p^*(s) = p(s+1)$ if player 1 chooses 1. In both cases the technology remains concave. We denote by G_0^1 and G_1^1 the subgames of G^1 following a choice of $d_1^1 = 0$ and $d_1^1 = 1$ by player 1 respectively. Likewise G_0^2 and G_1^2 are the subgames of G^2 following a choice of $d_1^2 = 0$ and $d_1^2 = 1$ by player 1 respectively. Finally, s_0^1 , s_1^1, s_0^2, and s_1^2 denote the number of players exerting effort in the equilibrium of the subgames G_0^1, G_1^1, G_0^2, and G_1^2 respectively. We now distinguish the following cases:

Case 1: $d_1^1 = 1$ and $d_1^2 = 0$. By the Lemma 3 $s_0^1 \geq s_1^1$. If $s_0^1 > s_1^1$ it cannot be optimal for player 1 to exert effort in G^1. Hence, we must have $s_0^1 = s_1^1$. Suppose first that $s_0^2 = s_1^2$, then it must be that $s_0^2 = s_1^2 > s_0^1 = s_1^1$. Otherwise player 1 would choose to exert effort in G^2 where his reward is at least as high as in G^1. This follows from the concavity of p as player 1's marginal contribution to s_1^2 is greater than it is to s_1^1. But if $s_0^2 > s_1^1$, then combined with the action of player 1 we have $s^1 \leq s^2$, which is what we need. We now assume that $s_0^2 > s_1^2$. By the induction hypothesis, $s_1^2 \geq s_1^1$. Hence we have $s_0^2 > s_1^1$, and combined with player 1's action we again have $s^1 \leq s^2$, as needed.

Case 2: $d_1^1 = 0$ and $d_1^2 = 1$. By Lemma 1, $s_0^2 \geq s_1^2$. Since $d_1^2 = 1$ we must have $s_0^2 = s_1^2$. Otherwise, if $s_0^2 > s_1^2$, player 1 is better off not exerting effort at G^2. By the induction hypothesis we have $s_0^1 \leq s_0^2 = s_1^2$. Furthermore, $s^1 = s_0^1 \leq s_0^2 = s^2 - 1 < s^2$.

Case 3: $d_1^1 = d_1^2$. By the induction hypothesis we have $s_1^2 \geq s_1^1$ and $s_0^2 \geq s_0^1$. If $d_1^1 = d_1^2 = 0$ we have $s^2 = s_0^2 \geq s_0^1 = s^1$ and if $d_1^1 = d_1^2 = 1$ we have $s^2 = s_1^2 \geq s_1^1 = s^1$ as needed. **Q.E.D.**

The Proof of Proposition 3 follows directly from Lemmas 3 and 4. **Q.E.D.**

The Proof of Theorem 1: follows directly from Propositions 2 and 3. **Q.E.D.**

The **Proof of Proposition 4:** We define $v^1 = (\frac{c}{p(n)-p(0)}, \frac{c}{p(n)-p(1)}, ..., \frac{c}{p(n)-p(n-1)})$. We argue that under v^1 the unique equilibrium yields all players investing.[6] Specifically, the strategy profile is for player 1 to invest and for all other players to invest if and only if all preceding players have invested. To verify this note first that $p(n)v_n^1 - c = p(n-1)v_n^1$

[6]Notethat this vector is the one with minimal total rewards among those incentivizing all agents to exert effort (see Winter (2006)).

and so by our tie-breaking rule player n invests if all preceding players invested as well. If, on the other hand, a set of preceding players of size k chose to shirk, then the incentive constraint faced by player n is given by $p(n-k)v_n^1 - c < p(n-k-1)v_n^1$, which follows from the fact that $p(n)-p(n-1) > p(n-k)-p(n-k-1)$, which in turn follows from the property of IRS. Hence, n will not invest if at least one of his preceding players shirked.

Assume now by induction that all players $k+1, k+2, ..., n$ are using the strategy specified above and consider player k. If all players acting prior to player k invested, then player k is facing the following incentive constraint: $p(n)v_k^1 - c \geq p(k-1)v_k^1$ and k will exert effort. If some set of players preceding k of size r shirked, then we have $p(k-r)v_k^1 - c < p(k-r-1)v_k^1$ (again because of IRS), and player k will choose to shirk as well. This establishes that under v^1 all players exert effort in equilibrium.

We next define v^2 as follows: $v_j^2 = v_j^1$ for $j = 1, 2, ..., n-1$ and $v_n^2 = \frac{c}{p(1)-p(0)}$. We first argue that under v^2 it is a dominant strategy for player n to exert effort. Indeed, because of IRS, if player $j's$ best response is to exert effort when k other players are exerting effort, then it is also his best response when $r > k$ players exert effort. Furthermore, because $v_n^2 p(1) - c \geq p(0)$, it is optimal for player n to invest even when no one else does so. Hence, investing is a dominant strategy for n. Consider now the decision of player $n-1$ at the subgame where all preceding players invest. If player $n-1$ invests, his expected payoff is $v_{n-1}^2 p(n) - c$; if he shirks it is $v_{n-1}^2 p(n-1)$. So $n-1$ will invest only if $v_{n-1}^2 \geq \frac{c}{p(n)-p(n-1)}$. But since p is increasing we have $p(n)-p(n-2) > p(n)-p(n-1)$ and hence $v_{n-1}^2 < \frac{c}{p(n)-p(n-1)}$ and player $n-1$ must shirk. Furthermore, if some of the players acting before $n-1$ chose to shirk, then player $n-1's$ incentive to shirk is even greater because when $k > 1$ we have $v_{n-1}^2 p(n-k) - c < v_{n-1}^2 p(n-k-1)$ if and only if $\frac{c}{p(n-k)-p(n-k-1)}$, which holds since $p(n-k) - p(n-k-1) < p(n) - p(n-1) < p(n) - p(n-2)$ (where the first inequality follows from IRS). We thus obtain that player $n-1$ shirks regardless of the action taken by the earlier players. Using backward induction we can now obtain that all players $j < n$ shirk regardless of the action taken by their predecessors and only player n invests. We finally note again that the analysis of the equilibrium will not change if instead of v^2 we take $v^2 + \varepsilon$ for sufficiently small ε. **Q.E.D.**

References

[1] Baliga, S. and T. Sjostrom (1998) "Decentralization and Collusion," *Journal of Economic Theory* 83, 162-232.

[2] Benabou R.and J. Tirole (2003) "Intrinsic and Extrinsic Motivation" *Review of Economic Studies* 70, 3, 489-520.

[3] Che, Y.K. and Yoo, S.W. (2001) "Optimal Incentives for Teams," *American Economic Review*, 91, 525-541.

[4] E. Fehr and A. Falk (2002) "Psychological foundations of incentives." *European Economic Review 46,* 687-724.

[5] Fischbacher, U., S. Gaechter and E. Fehr (2001) "Are People Conditionally Cooperative? Evidence from a Public Good Experiment., *Economic Letter 71,* 397-404.

[6] Gneezy, U. and A. Rustichini (2000) "Pay Enough or Don't Pay At All," *Quarterly Journal of Economics* 791-810.

[7] Gould, E. and E. Winter (2007) "Interactions Between Workers and the Technology of Production: Evidence from Professional Baseball." Forthcoming in *The Review of Economics and Statistics.*

[8] Holmstrom, B. (1982) "Moral Hazard in Teams," *Bell Journal of Economics* 13, 324-340.

[9] Itoh, H. (1991) "Incentives to Help Multi-Agent Situations," *Econometrica* 59, 611-636.

[10] Klor, E., S. Kube, E. Winter and R. Zultan (2008) "An Experimental Evidence for Incentive Reversal," Mimeo. The Economics Department. The Hebrew University of Jerusalem.

[11] Milgrom,P. and J. Roberts, (1990) "Rationalizability, learning, and equilibrium in games with strategic complementarities," *Econometrica* 58 1255- 1277.

[12] Topkis, D. Supermodularity and Complementarity, Princeton University Press, Princeton, NJ, 1998.

[13] Segal, I. (2003) "Collusion, Exclusion, and Inclusion in Random-Order Bargaining," *Review of Economic Studies* 70, 439-460.

[14] Winter, E. (2004) "Incentives and Discrimination," *American Economic Review* 94, 764-773.

[15] Winter, E. (2006) "Optimal Incentives for Sequential Production Processes," *Rand Journal of Economics.,* 37 (2), 376-390

[16] Winter, E. (2007) "Transparency among Peers and Incentives," Mimeo, The Center for the Study of Rationality, The Hebrew University ofJerusalem.