Journal of Economic Psychology 61 (2017) 213–224

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

Exploitation aversion: When financial incentives fail to motivate agents ☆

Jeffrey Carpenter [a,b,*], David Dolifka [a]

[a] Departments of Economics, Middlebury College, United States
[b] IZA, Germany

ARTICLE INFO

Article history:
Received 7 October 2016
Received in revised form 16 February 2017
Accepted 27 April 2017
Available online 4 May 2017

JEL classification:
C92
J33
M52
M55

Keywords:
Financial incentives
Motivation
Crowding
Power
Exploitation
Experiment

ABSTRACT

Studies of the principal-agent relationship find that monetary incentives work in many instances but that they can also backfire. Various mechanisms for this failure have been examined (e.g., intrinsic motivation, image concerns). We posit that an aversion to being exploited, i.e., being used instrumentally for another's benefit, can also cause incentives to fail. Using an experiment we find that compliance is lower for exploitative principals compared to neutral ones despite using the same contracts. To corroborate our results we show that surveyed "exploitation aversion" mediates this effect. Our results have implications for the design and implementation of incentives within organizations.

1. Introduction

Economists routinely advise principals to use financial incentives to motivate their agents. The basic rationale is compelling. If possible, make rewards contingent on agent performance and you should be able to align the interests of the agent with the goals of your organization. There is also empirical evidence that suggests that high-powered incentives that link pay to performance work. One of the most influential of these studies is Lazear (2000) who finds that after the Safelite Glass Corporation switched from using low-powered incentives (hourly wages) to high-powered ones (a piece rate) the average output per worker increased substantially. Embracing experimental methods to better identify the pure causal effects of the incentives (as separated from any sorting), a number of recent studies have confirmed the effectiveness of financial incentives both in the lab (e.g., Anderhub, Gaechter, & Koenigstein, 2002) and the field (e.g., Shearer, 2004). The problem, however, is that financial incentives do not always work as intended, and sometimes they actually appear to backfire. Considering vol-

☆ We thank the editor, two thoughtful referees, Eric Gong, Peter Matthews and Sandra Polania Reyes for their helpful comments. We also acknowledge the financial support of Middlebury College.

* Corresponding author at: Department of Economics, Middlebury College, United States.
E-mail addresses: jpc@middlebury.edu (J. Carpenter), ddolifka@middlebury.edu (D. Dolifka).

http://dx.doi.org/10.1016/j.joep.2017.04.006
0167-4870/

unteers, Carpenter and Myers (2010) show that financial incentive have no effect on the labor supply of many volunteer firefighters and Mellstrom and Johannesson (2008) show that paying people to donate blood actually reduces their willingness to do so, especially for women. In a more traditional principal-agent setting Gneezy and Rustichini (2000) find that paying donation solicitors modest compensation reduces their performance compared to those who are unpaid and Ariely, Gneezy, Loewenstein, and Mazr (2009) and Ariely, Bracha, and Meier (2009) find a similar result at high levels of compensation. Given, the contracts that are offered across all these studies are relatively similar, it is puzzling that sometimes they increase effort, sometimes they have no effect, and sometimes they actually reduce, or crowd-out effort. Because of this variation in outcomes, it is no longer clear what advice a principal should heed and so it is critically important to identify the circumstances that cause financial incentives to backfire.

A closer look at the literature suggests that financial incentives can fail for a variety of reasons (Bowles & Polanía-Reyes, 2012; Gneezy, Meier, & Rey-Biel, 2011). One of the most studied reasons is that incentives might crowd out "intrinsic motivation," the internal drive to work to master a skill or to improve one's self concept (Deci & Ryan, 1985). In this framework, extrinsic (financial) incentives can reframe an interaction from one in which effort is required based on moral reasoning to one in which effort becomes a choice (e.g., Titmuss, 1970 or Cardenas, Stranlund, & Willis, 2000) or they can adversely affect an agent's sense of autonomy (e.g., Lepper, Greene, & Nisbett, 1973 or Falk & Kosfeld, 2006). Agents concerned with their public appearance or self image might also react adversely to the implementation of financial incentives (Bénabou & Tirole, 2006). In the case of volunteers, for example, extrinsic rewards might reduce the pride one takes in serving the public or tarnish, to some extent, one's reputation as an altruist (Ariely, Gneezy et al., 2009; Ariely, Bracha et al., 2009; Carpenter & Myers, 2010). Financial incentives might also provide information to the agent on the principal's assessment of their ability or the extent to which the principal trusts the agent to do a good job (Fehr & Rockenbach, 2003). If the principal is providing financial incentives because she does not think the worker is very talented or trustworthy, the agent might, again, react poorly.

We conjecture that another reason why financial incentives might backfire is that, through their choice of incentives, principals may signal selfish intentions that can reduce motivation. The sort of intentions we have in mind for the principal have a long tradition in the social sciences and the history of economic thought. Specifically, we examine whether choosing incentives to exploit an agent will cause the agent to reconsider compliance. To be precise, in our experiment we operationalize a very specific notion of exploitation in the workplace, one that works through agent perceptions of a principal's motives to affect motivation. As a result, we focus as much on intentions as outcomes. Like Feinberg (1988), who states exploitation grows upon a "morally unsavory" desire and Buchanan (1985) who refers to it as "merely instrumental" we define exploitation as the utilization of another to achieve one's own ends. Whether facilitated by status or leverage, whether gains and losses are distributed fairly or unfairly, whether the intentions are malicious or only selfish, exploitation for the purposes of our study involves the instrumental use of agent capabilities by a principal to advance his or her own goals.

To examine the potentially subtle issue of exploitative intentions experimentally, we designed a new experiment with three unique features. First, we formulated an underlying game structure that provided the material conditions necessary for exploitation. In our game, principals could choose contracts that would force agents to expend more effort than is socially optimal. Second, it was in the extrinsic interests of the agents to comply with these potentially exploitative contracts (i.e., they resulted in Nash equilibria). This feature guaranteed that if compliance did not occur, it was for intrinsic reasons. Third, we created two principal treatments to separate neutral and exploitative intent. Rather than comparing a condition in which a human chooses a contract to one in which a randomizing mechanism determines the contract parameters, as is common in the related literature (e.g. Falk & Kosfeld, 2006), it was important that humans chose in both our conditions so that we control for the basic effect of human "agency". Had we done it the standard way, the treatments would differ in both the accountability of humans versus machines (as managers) making a choice and the ability of the manager to exploit. Specifically, in one case, the neutral one, contracts may satisfy the material conditions for exploitation but agents cannot attribute exploitative intent to the principal. In the second case, the contracts may again be materially exploitative but this time the agents should infer the intention to exploit.

Our results are clear and robust. Like the existing literature, the use of high-powered financial incentives in our experiment backfires sometimes, however, we are able to "adjust the carburetion" to increase or decrease compliance. Principals who choose contracts that exploit agents (i.e., cause them to choose higher than efficient effort levels) see a lower level of compliance only when the exploitative contract choice is accompanied by an exploitative intent. Neutral principals, using the same incentives benefit from higher levels of compliance than those whose own material incentives signal exploitative intent to the agents. The compliance difference is approximately ten percent and it is robust to the inclusion of various demographic controls and different econometric specifications. In addition, we show that a survey instrument designed to measure exploitation aversion mediates the compliance differential across treatments, confirming that agents are rejecting contracts because they perceive them as exploitative.

We proceed by describing the details of our experiment. We then present, in Section 3, an overview of our participants and their experimental choices. In Section 4 we analyze the determinants of contract compliance and in Section 5 we examine the robustness of our results. We discuss related work in the final section before concluding with a few suggestions for future research.

2. Study design

We designed an experiment to test whether strong financial incentives might backfire (reducing compliance) when agents perceive them as exploitative. Although exploitation can occur outside the workplace, we couch our study in the employment context, specifically, the principal-agent setting. To comply with this setting there must be some conflict of interest between the manager and the worker out of which exploitation can arise. Further, workers may often feel some sense of being controlled by the manager as in the intrinsic motivation literature (Deci & Ryan, 1985). To focus our attention on exploitation we need to maintain a sense of managerial control in our treatments - in other words, ours is the converse of the Falk and Kosfeld (2006) study.

Our definition suggests that for agents to feel exploited, they must not only feel manipulated, the manipulation has to be the result of the principal's choice. In other words, principal agency, and the resulting culpability were also important design considerations. This meant that we needed real people to occupy the managerial role in all conditions otherwise we would not be able to disentangle the effect of exploitative intent from the human agency (or lack there of) of the manager. At the same time we had to design an experiment that would control for other possible differences that could affect agent choices. Specifically, we need to compare agents who faced the same material incentives but the contracts dictating the incentive came from one of two sources: a neutral principal or an exploitative one.

In the end, we decided to create as subtle a manipulation as possible. This choice, however, necessitated that the rest of the experiment be very straightforward. With respect to the underlying incentive structure, this meant that we sought to create a principal-agent game that was transparent and could be easily understood by novice players, once exposed. On top of this structure we allowed principals to implement financial incentives in the form of a contract to provide a minimum amount of effort. The contracts we allowed were also simple and easy to understand but, importantly, they were the choice of the (human) principal in all cases so that any effect of being controlled by the principal would be common to both treatments. We now describe the experiment in detail.

The underlying principal-agent game that we created is a hybrid of two standards in the literature: the team production game (known in a different context as the voluntary contribution mechanism) and the gift exchange game. Consider agents who work in teams of size n and have effort endowments of $e = 10$. Individual agents choose integer effort levels, $e_i \in [1, 2, \ldots 10]$. The team's contributed efforts are then aggregated and multiplied by a productivity parameter, β, to create material benefits $\beta \sum e_i$ that are shared equally among team members. Effort, however, is costly to contribute. Specifically, the cost of effort $c(e)$ is increasing and convex. Subtracting the cost of effort from the material benefits results in the following payoff for the i^{th} agent.

$$\pi_i = \frac{\beta \sum e_i}{n} - c(e_i)$$

As illustrated in Fig. 1, this structure leads to both an interior Nash equilibrium choice of effort and an interior social optimum. Taking the derivative of π_i with respect to e_i yields the equilibrium condition $\frac{\beta}{n} = c'(e)$ while the social optimum occurs when the marginals are taken after summing the individual agent payoffs (i.e., where $\beta = c'(e)$). The benefit of the internal Nash equilibrium is that it allows us to separate equilibrium play from simply contributing nothing, regardless of the incentives, two outcomes that are confounded in the standard linear team production experiment. In other words, this structure gives us a bit more information on whether our participants understand the incentives. More importantly, however, the concomitant interior social optimum is at the core of our design. Although financial incentives that result in team effort choices between the Nash level and the social optimum will actually be helpful for the team members, those that cause agents to choose effort levels beyond the social optimum will hurt them. This creates the material conditions necessary for exploitation. If the principal has an incentive, along with the will, to extract efforts beyond the social optimum, workers should feel exploited.

Notice in Fig. 1 that the marginal cost of effort is monotonically increasing but only piecewise linear. This is the result of our experiment-specific choice of $c(e)$ presented in Table 1 and was done purposefully to make the incentives around the social optimum as clear as possible. For this specification of $c(e)$ and β set equal to 40, the Nash equilibrium in teams of four agents occurs where e^* equals two. The social optimum in Fig. 1 occurs where effort is equal to five, though this is more obvious in Fig. 2(a), a screen shot which illustrates how the game was summarized for the participants. Using this table participants could first estimate how much effort they thought that the other three agents in their team would contribute, on average, and then consider their own effort choices.

After everyone played an initial ten rounds of the baseline game to experience the incentives of the interaction, principals were assigned to each team and they implement a version of a forcing contract on the team of agents. The archetypal forcing contract (Holmstrom, 1982) sets a minimum output that must be achieved by the team as a whole because individual efforts are either unobserved, not verifiable or otherwise non-contractible. If the team fails to make the target, they receive only a low penalty wage instead of their share of the benefits. The advantage, for us, of this sort of contracting is that the incentives couldn't be clearer. However, to make things even simpler we removed any complications that might arise from participants trying to coordinate on various equilibria. This was done by allowing principals to implement the contracts at the level of the individual agent. They could set a minimum required effort (an e_{min}) for each worker in the team, though in each period they set just one e_{min} for all the members of the team, again to keep things simple. If the agent complied with the forcing contract

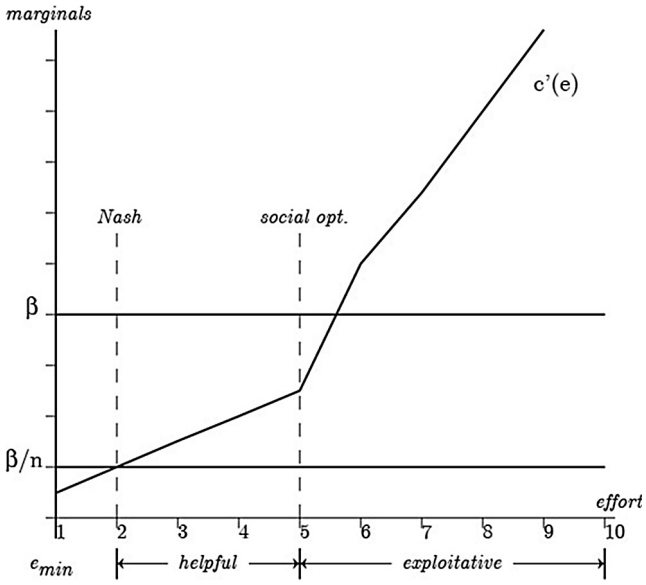

Fig. 1. Experimental design.

Table 1
The cost of effort.

e	1	2	3	4	5	6	7	8	9	10
$c(e)$	5	10	25	45	70	120	180	250	330	420

		Your Choice of Effort									
	e	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	35	40	35	25	10	-30	-80	-140	-210	-290
	2	65	70	65	55	40	0	-50	-110	-180	-260
Expected	**3**	95	100	95	85	70	30	-20	-80	-150	-230
Avg. of	**4**	125	130	125	115	100	60	10	-50	-120	-200
Other	**5**	155	160	155	145	130	90	40	-20	-90	-170
Players	**6**	185	190	185	175	160	120	70	10	-60	-140
	7	215	220	215	205	190	150	100	40	-30	-110
	8	245	250	245	235	220	180	130	70	0	-80
	9	275	280	275	265	250	210	160	100	30	-50
	10	305	310	305	295	280	240	190	130	60	-20

Fig. 2(a). Agent payoff table in the initial ten rounds (no forcing).

(i.e., chose an effort level at or above e_{min}), she received her share of the proceeds created by the team (minus her effort cost), as before. If she did not comply, if she contributed an effort level less than e_{min}, she received a penalty payoff set to zero for the period. The screen shot in Fig. 2(b) illustrates how the payoff table is transformed when e_{min} is set to eight. In equilibrium, agents should comply with all forcing contracts stipulating e_{min} between two and nine because the alternative is to receive nothing. As it turned out our payoff function generated a payoff of -20 when everyone chose $e = 10$ which presents an interesting case that allows us to examine whether agents shirk more often when they feel exploited even though everyone should shirk. That is, some players may naturally feel compelled to do as instructed despite it not being in their narrow material interests and the question is whether being exploited attenuates this compunction.

Returning to Fig. 1 we see the core of the design. Setting e_{min} between its lower bound of 2 and 5 will actually help a team of Nash players because their payoffs will increase. Therefore, workers should be happy to comply with any e_{min} in this range. However, values of e_{min} that are greater than 5 might be exploitative, depending on the incentives of the principal and the intentions signaled by those incentives and the principal's choices of e_{min}. The question is whether agents who feel exploited will be less likely to comply with these contracts, despite the material incentives, than a control group who should not feel exploited. That is, can exploitation aversion explain some of the instances in which high-powered incentives backfire?

		Your Choice of Effort									
	e	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
	1	0	0	0	0	0	0	0	-140	-210	-290
	2	0	0	0	0	0	0	0	-110	-180	-260
Expected	**3**	0	0	0	0	0	0	0	-80	-150	-230
Avg. of	**4**	0	0	0	0	0	0	0	-50	-120	-200
Other	**5**	0	0	0	0	0	0	0	-20	-90	-170
Players	**6**	0	0	0	0	0	0	0	10	-60	-140
	7	0	0	0	0	0	0	0	40	-30	-110
	8	0	0	0	0	0	0	0	70	0	-80
	9	0	0	0	0	0	0	0	100	30	-50
	10	0	0	0	0	0	0	0	130	60	-20

Fig. 2(b). Agent payoff table with forcing and $e_{min} = 8$.

To manipulate whether agents should feel exploited or not we ran two treatments that differed only in how the principals were compensated. In the *exploitative* condition, principals were paid according to their choices of e_{min}. Specifically, principals in the exploitation condition received a payment of $\pi_p^{Exploit} = 20 \times e_{min}$. Clearly, larger values of e_{min} were better for these principals. Originally we planned to use a more intuitive payment structure for exploitative principals, the product of a constant and team total effort, but we decided that this might introduce a confound. If the boss is compensated based on team output and a worker decides to shirk on the contract, she might be doing it because she is averse to exploitation as we hypothesize, but she might also do it because she is inequality averse and she wants to lower the boss' payoff. With the payoff scheme we implemented, workers cannot act to reduce the principal's payoff, a standard aspect of manager compensation schemes in the field.

In the *neutral* condition, principals were simply paid a flat rate of $\pi_p^{Neutral} = 200$ per period, a payment system consistent with managers earning fixed salaries in many naturally occurring firms. Here because principal compensation did not depend on e_{min}, the link between intentions, exploitation and compliance is severed and agents play under the same financial incentives but should not feel exploited. While this explains why we chose to compensate neutral principals with a flat payment, the level of 200 was set so that if there was any residual inequality aversion it could only work against our hypothesis (and dampen our estimates of the effect of exploitation aversion). The most principals in the exploitation condition could earn was 200 by setting $e_{min} = 10$ so if agents were inequality averse and they shirked on contracts to protest, they should be more likely to shirk in the neutral condition than in the exploitation condition.

We ran four sessions, two for each treatment with a total of 80 participants (exactly 20 per session). Each session lasted about an hour and participants earned an average of $22.78, including a $5 show-up payment. Because we used "partners" matching, we generated 640 effort choices from 16 independent groups.

There were twenty periods split into two blocks of ten during each session. Participant earnings were the sum of the earnings that they accumulated over all twenty periods. In the first ten periods of a session all twenty participants played the simple principal-agent game summarized in Fig. 2(a). The first block was intended to familiarize all the participants with the incentives of the game, in particular where the Nash equilibrium was and where the social optimum was so that when they played the second block it would be clear how some forcing contracts could be helpful while others might be exploitative.

At the beginning of the second block in each session one of the five teams of four from the first block was dissolved at random and the four members were randomly assigned to be the principal of one of the other four remaining teams. The period began by each principal choosing an e_{min} from a set of possible values that changed from one period to the next. Rather than allowing the principals to pick any e_{min} between two and ten, we realized that neutral principals had no incentive to set high values of e_{min} but exploitative ones did. To make sure we could compare agent compliance across treatments for each value of e_{min} we had to restrict the principal choices. In each period principals chose between a low value and a high value for e_{min}. The two possible values for each period were determined before the experiment began and the sequence was the same for each principal treatment and all four sessions.[1] However, information about the set of possible values of e_{min} was asymmetric. Principals saw the two values each period but agents only knew that the principals were choosing from a set. To preserve control and provide agents in both treatments with exactly the same information, the agents did not know what values were under consideration nor did they know the number of choices from which the principal could choose (a complete set of experimental instructions are presented in Appendix A). Principal contracts were transmitted to the teams of agents using the z-Tree programming environment (Fischbacher, 2007) who then saw the appropriate table (again, Fig. 2(b) shows the table for $e_{min} = 8$) and chose whether to comply with the contract (by choosing an effort level of e_{min} or more) or not. Once the twenty periods were over, the participants completed a post-experiment survey while the experimenters calculated the earnings for each participant.

[1] For example, in period 5 managers chose between $e_{min} = 3$ or $e_{min} = 5$ and in period 8 they chose between 7 and 9.

3. Data preliminaries

The mean age of our participants (all Middlebury College students) was 19.78 years, 54% were male, 63% reported being caucasian, 31% were social science majors and 62% reported having a grade point average above 3.25 (based on a 4 point scale). None of these characteristics differed significantly between the two principal compensation treatments ($p > 0.10$ in each case) so based on these observables, it appears that we achieved randomization to treatment.

Before turning to our main results – an analysis of agent compliance – we first want to see if there is any evidence that the first block of ten periods was useful in helping our participants learn the incentives of the underlying game. Fig. 3 plots mean effort choices in the first block by period and treatment. As one can clearly see, our participants learned to play the Nash equilibrium. Average effort choices start near 4 in the first period but are almost exactly 2, on average, by the end of ten periods. T-tests suggest that mean effort choices do not differ from 2 in either treatment during the final period of the first block ($p = 0.22$ for the neutral treatment and $p = 0.43$ for the exploitative treatment). Fig. 3 also suggests that there is no treatment difference in the time paths of this learning process. Random effects (at the agent level) estimates of effort choices including all ten periods confirm this: the p-value on the treatment coefficient is 0.20.

In sum, the first block of the experiment seems to have served its intended purpose. In both treatments, by round ten most participants (66% to be precise) are playing the Nash equilibrium and achieving relatively low payoffs compared to the social optimum.

4. Contract compliance

In this section we present our main results. The experiment was designed with one specific question in mind: are agents less likely to comply with high-powered financial incentives when they perceive that these incentives are being used by the principal to exploit them? To address this question we begin by cataloging the forcing contract choices of the principals to make sure all contracts were offered in the two treatments and then we dig into the details of agent contract compliance.

Although not the focus of our study, the choices of the principals, separated by treatment, are illustrated in Fig. 4. The most important aspect of the figure is that there is nearly full support for the distribution of e_{min} choices in each treatment. The lowest value of e_{min}, 2, was never chosen in the exploitative principal treatment, but other than this omission, principals in both treatments chose every e_{min} value a number of times. As a result, we are able to make an apples-to-apples comparison of agent compliance: controlling for the contract is there less compliance in the exploitative treatment than in the neutral treatment?

Before we leave Fig. 4, however, it is also interesting to note that the principals seemed to understand their incentives. As one can also see, the distribution of e_{min} choices is shifted to the right in the exploitative principal treatment compared to the neutral principal treatment. In other words, exploitative principals were sensitive to the fact that they earned more at higher values of e_{min}. On average (and from the same set of choices), the exploitative principals chose $e_{min} = 7$, neutral principals chose $e_{min} = 6.15$ and the difference is significant ($t = 2.31$, $p = 0.02$).[2] If managers chose forcing contracts at random, given the options they were presented, the mean observed choice would be 6.4. If instead they always picked the contract that was closest to the social optimum, the average would have been 6.1. Considering additional t-tests, both of these benchmarks are within the 95% confidence interval of the observed neutral managers' mean choice, but not within the bounds of the exploitative managers' choice.

Starting with the most general analysis of compliance, in Fig. 5 we pool across all contracts and all periods to test if there is any compliance difference by treatment. Overall, compliance is high, 84%, which is not too surprising given any contract stipulating $2 < e_{min} < 6$ helps the agents. At the same time, a treatment difference does emerge. Neutral principals enjoy a higher rate of compliance (91%) than exploitative principals (77%) and the 14% difference is highly significant ($p < 0.01$). Further the difference grows to 17% when we limit the sample to only those contracts satisfying the necessary material conditions for exploitation (i.e., $e_{min} > 5$).

Clearly, the differences in e_{min} choices made by the principals (seen in Fig. 4) must account for some of the difference in compliance. With this in mind, in Fig. 6 we plot average compliance by treatment and e_{min} choice. Confirming again that agents understood their incentives, every contract with $e_{min} \leqslant 5$ is complied with. Indeed, all contracts such that $e_{min} \leqslant 7$ are complied with, despite $e_{min} = 6,7$ being potentially exploitative. The obvious reason for full compliance up to $e_{min} = 7$ is that the symmetric payoff up to $e = 7$, while less than the social optimum, is still larger than what would be received at the, now focal, Nash equilibrium. In terms of treatment differences, Fig. 6 clearly shows that the compliance differential is occurring exclusively at higher levels of e_{min}, a result that one would expect if agents were averse to being exploited. The difference in compliance is negligible at $e_{min} = 9$, but it is substantial, close to 20%, at both $e_{min} = 8$ and $e_{min} = 10$. Further, it is the case that many fewer participants comply with $e_{min} = 10$, as expected. However, compliance does not fall to zero. As mentioned above, participants might feel a natural compunction to adhere to the contracts they are given, or perhaps agents feared that principals would retaliate with additional harsh contracts if they did not comply. That said, this compunction or fear must have been less motivating when agents felt exploited because the differential persists.

[2] This difference persists in regressions that control for data interdependences using clustering or random effects (both at the level of the principal). It is also the case that the mean e_{min} in both treatments is significantly larger than the social optimum level of 5 according to t-tests ($p < 0.01$ in both cases).

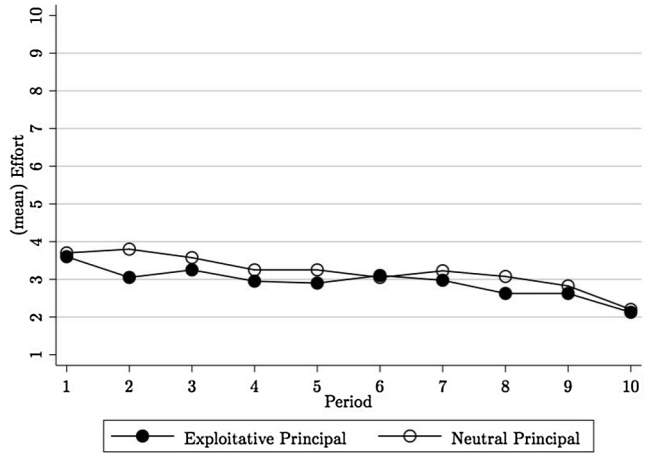

Fig. 3. Effort choices in the initial ten periods without forcing (by treatment).

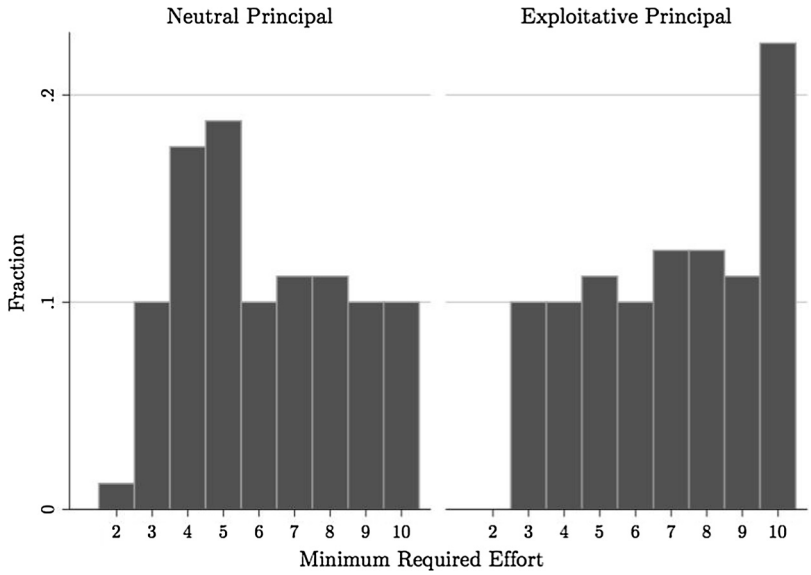

Fig. 4. Principal choices of forcing levels, e_{min}, (by treatment).

To be more formal about our analysis of the treatment differential, in Table 2 we present linear probability estimates of compliance. In the first column, we reproduce what was seen in Fig. 5. Pooling periods and contracts, agents of an exploitative principal comply 14.4% less ($p < 0.01$). In column (2) we restrict the sample to only those contracts that could be deemed exploitative, and the differential increases, as expected, to 16.8% ($p < 0.01$). In addition to restricting the sample, in column (3) we control for contract choice (and the difference in contracts chosen between treatments) and see that the differential does shrink to 9.7% but that it is still highly significant ($p < 0.01$). We also confirm that compliance is lower for contracts with larger values of e_{min}. Finally, in column (4) we add the observables that we collected (age, sex, major, race and GPA) and given the balance in our samples, it is not too surprising that they are orthogonal to the treatment effect (in addition none of the point estimates on the demographics are significant). When exploitation is perceived, our results suggest that high-powered financial incentives can backfire. Our best estimate suggests that the subtle perception of exploitation alone accounts for a compliance differential that is approximately ten percent.

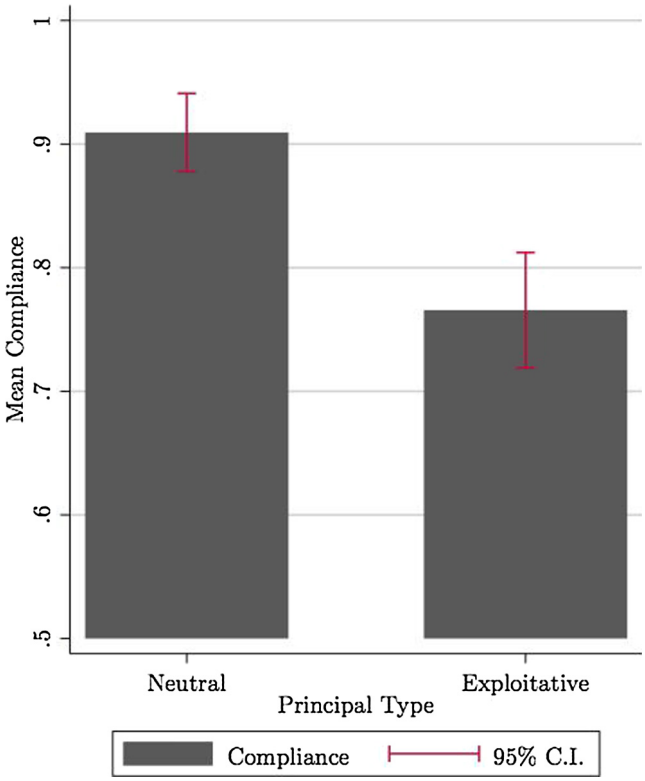

Fig. 5. Overall forcing contract compliance (by treatment).

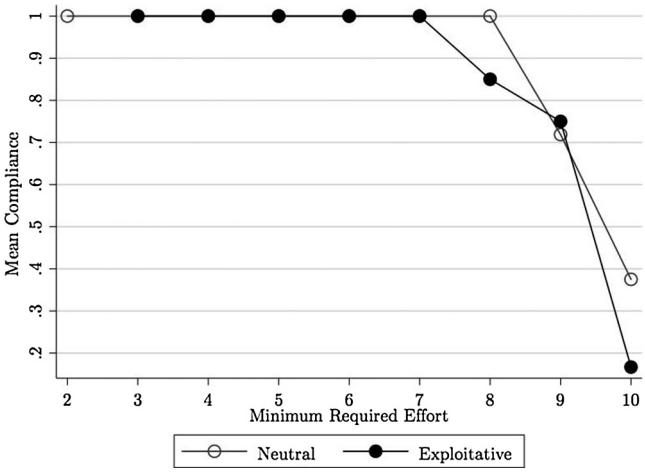

Fig. 6. Contract compliance by forcing level, e_{min}, (and treatment).

5. Robustness

Using Table 3, we examine the robustness of our compliance results. First, in column (1) we acknowledge the panel nature of our data and estimate a linear probability model that includes agent-level random effects. As is apparent by comparing the results in column (1) to those in the last column of Table 2, the estimates are identical.[3]

The second thing that concerned us was that because the treatments differ to some extent in the contracts to which the agents were exposed (as we saw in Fig. 4), they will also differ in the history of play. To account for this we added the lag of

[3] The panel results described in this section (i.e., Table 3) are unchanged if we include multilevel effects at both the group and individual levels.

Table 2
Compliance regressions.

	(1)	(2)	(3)	(4)
Exploitative Principal (I)	−0.144***	−0.168***	−0.097***	−0.099***
	(0.029)	(0.043)	(0.035)	(0.036)
e_{min}			−0.193***	−0.194***
			(0.012)	(0.012)
Sample	full	$e_{min} > 5$	$e_{min} > 5$	$e_{min} > 5$
Controls	No	No	No	Yes
N	640	388	388	381
Adj. R^2	0.04	0.03	0.42	0.43

Notes: Dependent variable is a contract compliance indicator; e_{min} is the forcing contract choice; Linear probability estimates; (robust standard errors); * $p < 0.10$, ** $p < 0.05$, controls include age, sex, major, race and GPA.
*** $p < 0.01$.

Table 3
Robustness checks.

	(1)	(2)	(3)	(4)
Exploitative Principal (I)	−0.099***	−0.172***	−0.022	−0.096**
	(0.034)	(0.035)	(0.033)	(0.040)
e_{min}	−0.194***	−0.210***	−0.195***	−0.210***
	(0.011)	(0.011)	(0.011)	(0.011)
Lagged e_{min}		0.012*		0.013*
		(0.007)		(0.007)
Exploitation aversion (I)			0.004	0.017
			(0.037)	(0.045)
E.A. × Exploitative principal			−0.141**	−0.140**
			(0.058)	(0.063)
Sample	$e_{min} > 5$	$e_{min} > 5$	$e_{min} > 5$	$e_{min} > 5$
Controls	Yes	Yes	Yes	Yes
Agent random effects	Yes	Yes	Yes	Yes
N	381	318	381	318
R^2 (overall)	0.43	0.45	0.44	0.46

Notes: Dependent variable is a contract compliance indicator; e_{min} is the forcing contract choice; Linear probability estimates; (robust standard errors); controls include age, sex, major, race and GPA.
* $p < 0.10$.
** $p < 0.05$.
*** $p < 0.01$.

e_{min} to our estimating equation. In the second column of Table 3 we find that controlling for the lag of e_{min} actually increases the treatment point estimate substantially (to 17.2%, $p < 0.01$) because compliance tends to rise slightly, not fall, after being exposed to a relatively harsh contract (and there are more harsh contracts in the exploitative principal treatment).

Up to this point we have set up the material conditions for exploitation in our game and we have given one set of principals the incentive to exploit their agents. Given the construction of our neutral control, we are confident that the only thing that can be driving the difference in compliance that we have documented is an aversion to being exploited on the part of the agents. However, it would be nice to corroborate this conclusion with additional, more direct, evidence. Anticipating this, in our survey we asked players to respond to the following four statements (based on the Stanford Encyclopedia of Philosophy) to try to capture our participants' attitudes towards exploitation (the responses were gathered using a 5-point Likert scale). Q1: *If A willingly agrees to a transaction with B, this can't possibly be exploitation.* Q2: *If A and B both benefit from a transaction, this can't possibly be exploitation.* Q3: *If an unexpected blizzard hits tomorrow, the owner of the hardware store in town has every right to start charging double for snow shovels.* Q4: *The fairness of a transaction can be evaluated solely by comparing the gains of each involved party*. Considering our definition of exploitation, one based on intentions as much as outcomes, we classify affirmations of these statements as being tolerant of exploitation and rejections as being exploitation averse.

We reverse coded the exploitation question responses and added the resulting four scores to get a preliminary sense of the distribution of exploitation aversion preferences. The resulting sums are shown in Fig. 7 in which larger values indicate a greater aversion. To summarize our exploitation aversion scale, we conducted a factor analysis and the results were encouraging. Not only was the Eigenvalue on the first factor larger than one (the standard cutoff) indicating a strong common thread through the responses, all the (reverse-coded) questions loaded positively suggesting that our intuition about how to classify the responses was correct. Because the factor loadings were not the same across the four questions, we used the factor scores, instead of the preference sums used to generate Fig. 7, to create an exploitation aversion indicator, splitting the full sample of agents at the median. To examine the demographic determinants of our exploitation aversion scale, we

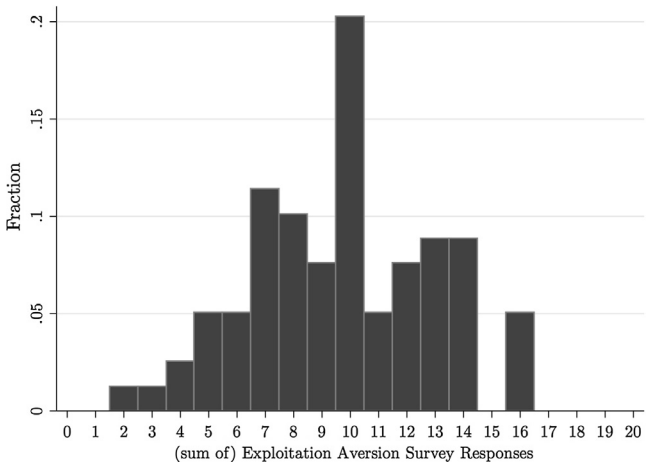

Fig. 7. The distribution of agent exploitation aversion survey responses.

regressed the indicator on the demographics we collected for the agents. While older agents were a bit more averse to exploitation, as were male agents, white agents and those agents with higher GPAs, the only robustly significant finding is that social science majors were less exploitation averse. Finally, and perhaps most importantly for the analysis that follows, we tested to see if our measure of exploitation aversion differed by treatment and it did not ($p = 0.77$). In other words, there is no evidence that participating in the exploitative principal treatment made agents more exploitation averse.

If exploitation aversion is at the heart of the compliance differential that we have estimated, it should be the case that participants who are categorized as exploitation averse based on our survey should be less likely to comply with contracts stipulating $e_{min} > 5$ when the principal is exploitative than those who are categorized as exploitation tolerant. Further, the two types should not comply at different rates when the principal is neutral. As one can see in Fig. 8, this is what we find. For the neutral principals, compliance does not differ by surveyed exploitation aversion ($p = 0.88$) but the rate of compliance is 12% lower for exploitation averse agents when the principal is exploitative ($p = 0.05$).

We can also see, in the last two columns of Table 3, that surveyed exploitation aversion mediates the effect of the treatment. Much of the variation previously attributed to the exploitative principal treatment indicator is now being absorbed by the surveyed exploitation aversion of agents in this treatment (i.e., the interaction term). The estimated effect of exploitation aversion under a neutral principal in column (3) is very close to zero, 0.004 ($p = 0.92$) but for an exploitative principal the effect is $-0.141 + 0.004$ or -0.137 which is highly significant ($p < 0.01$). Considering column (4) which also controls for the lagged contract, including surveyed exploitation aversion again reduces the coefficient on the treatment indicator but this time not to zero. That said, the estimates of exploitation aversion in the two treatments are unchanged. To a great extent these results confirm that compliance is lower in the exploitative principal treatment because the agents felt exploited.

6. Discussion

There are a number of empirical studies showing how high-powered financial incentives can sometimes backfire. The leading hypothesis is that in some situations the financial incentives crowd out intrinsic motivation. The purpose of our study is to highlight another reason why agents might resist the use of standard economic tools like pay for performance. Our hypothesis, that agent motivation can be crowded out when financial incentives appear exploitative, has received relatively little attention in the literature despite being a very old concept in the history of economic thought. Given the lack of previous work to guide our choices, we chose to start with a simple and clean experiment to see if we could first, create the material conditions necessary for exploitation to matter in the lab and second, remove any confounds so that we could directly test for any effect of exploitation aversion.

Our results suggest that just the intention to exploit signaled by the use of incentives may be enough to reduce compliance. To control for the material consequences of the different possible forcing contracts that a principal could offer, our experiment was designed so that agents in both treatments would be exposed to all the possible contracts. Hence, the only difference between our comparison agents in the two treatments is how the principal is compensated and so it must be that knowledge of this is sufficient to signal intent and trigger exploitation aversion in our agents.

Our estimates of the effect of exploitation aversion on contract compliance, after a number of robustness checks, tend to be in the neighborhood of 10%. While this effect is not small, there is some reason to think that we might be measuring the lower bound. First, as mentioned before, our manipulation is subtle. Agents must not only recognize that they are being exploited (i.e., understand the material incentives of the game), they must correctly interpret the intentions of the principal conveyed through the contract and these intentions must trigger an unease, one substantial enough to cause the agents to

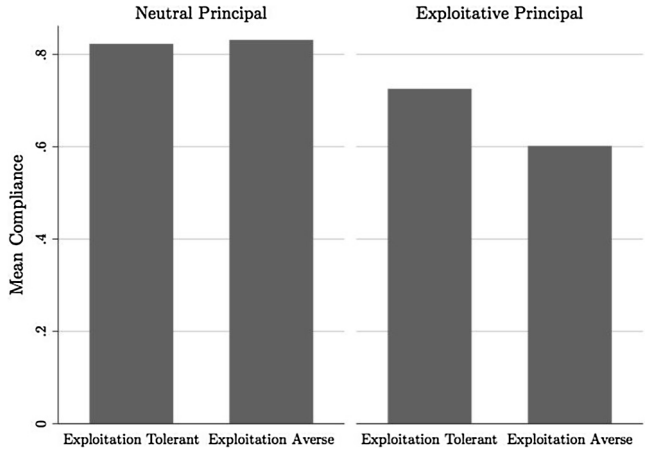

Fig. 8. Contract compliance and surveyed exploitation aversion (by treatment).

act contrary to their material incentives. If any of these features were more prominent or salient, we suspect contract compliance would fall even further. Second, because we worried that inequality aversion might also cause agents to reject contracts, we chose to make sure that it would always be a larger motivator in our neutral treatment. As a result, the difference in contract compliance might be lower that it would be otherwise.

Our results dovetail nicely with other related work in the economics literature. The experiment is similar in design to Falk and Kosfeld (2006) who find that principals who try to control their agents by explicitly restricting their choice sets (similar in effect to the forcing contracts we use) do worse than those who simply trust their agents to do the right thing (i.e., the extrinsic incentives backfire). This is a very nice demonstration of Deci and Ryan's (1985) self-determination theory, the dimension of intrinsic motivation which Bowles and Polanía-Reyes (2012) refer to as "control aversion." Notice, however, that control aversion cannot explain our results. The forcing contracts, and therefore the levels of control exerted by our principals, are the same across treatments. This same type of argument also eliminates explanations based on outcome-oriented models of negative reciprocity (Fehr & Schmidt, 1999) because the material consequences of a principal implementing a harsh contract are also the same across or treatments. The difference in compliance seen between treatments in our experiment also cannot be explained by "betrayal aversion" (Bohnet, Greig, Herrmann, & Zeckhauser, 2008) which has been developed in a different context, that of social risk - situations in which trusting by a first-mover may or may not pay off. In our setting agents move second so trust and this sort of uncertainty play no role.

The related behavior seen in Blount (1995) and more recently in Charness (2004) may, however, be interpreted as instances of exploitation aversion. In Blount's ultimatum games, second-movers announce lower minimally acceptable offers (i.e., they are more likely to sacrifice and comply with a given offer of less than half the pie) when offers came from a random device than when they were made by another participant who might be interpreted as exploiting the second-mover. In the gift exchange experiment conducted by Charness, when principals intentionally offer a low wage agents rarely respond with high effort, they feel exploited. However, when the low wage comes as the result of a random draw they are more willing to sacrifice to benefit the principal. Perhaps even more explicitly, Schnedler and Vadovic (2011) conduct an extension of the Falk and Kosfeld experiment, including a treatment in which robotic agents are used to increase the perceived "legitimacy" of the control exerted by the principal. The authors find increased legitimacy does affect the provision of effort, a result that suggests that the aversion to exploitation that we observe could be moderated by the perceived legitimacy of the controlling principal. Specifically, because it was random which role participants occupied in our experiment, one could argue that the control used by principals is fair and legitimate to some degree. This is yet another rationale for why what we observe might constitute a lower bound estimate.

Given our forcing contract can be seen as a type of contingent or punishing contract, the work of Fehr and Rockenbach (2003) or the extension of this work by Houser, Xiao, McCabe, and Smith (2008) on how sanctions and punishment can fail to yield the intended results may also be related to exploitation aversion. In particular, in Fehr and Rockenbach's experiment an "investor" makes a transfer to a "trustee" who can improve the payoff of both players and the trustee decides whether to honor this trust with a back-transfer, made at a cost. In a second sanctioning treatment, the investor chooses to either inflict a fine on trustees that do not send enough back or not. Comparing the three conditions, trustees spitefully send back the least when a fine is looming and send the most when the investor had the option of fining, but chose not to. Given the design, it is hard to know the extent to which the difference in trustee behavior is due to control aversion like in Falk and Kosfeld (2006); however, it might also be consistent with trustees feeling exploited. In their follow-up study, Houser et al. (2008) replicate the basic results of Fehr and Rockenbach but also include a treatment in which any punishment is assigned to the trustee exogenously by a random device. This case is interesting because it allows the authors to assess the effect of the punishment being intended by the investor. Contrary to many of the other studies discussed and our own results, Houser et al. do not find

an effect of intentions. It is important to keep in mind, however, the differences in design. For example, we purposefully avoided random devices in favor of a design in which incentives are chosen by human actors in all the conditions and our is a comparison between principals with different financial incentives, only some of whom are incentivized to exploit the agent.

Lastly, in Bartling, Fehr, and Schmidt (2013), the experimenters examine whether employment contracts are viable when principals face moral hazard about how to react to contracts that are state-dependent. The main result is that principal moral hazard does make it harder for employment contracts to survive but what is interesting with respect to our results is the moral dilemma faced by their principals. Bosses can pick one option that is efficient and shares the resulting surplus with the agent or they can pick an inefficient option that benefits the boss disproportionately. In other words, the principals in this experiment can exploit the agents. Though not the purpose of their experiment, the fact that employment contracts vanish to some extent in this setting due to workers resisting exploitation validates our more direct results.

In the end we see this experiment as a first step, a "proof of concept." Given our results indicate people are sensitive to being exploited, there are a number of interesting and important next steps. First, like the reciprocity literature developed through the 1990s, it will be important to develop a more formal theoretical framework that clarifies any differences and similarities between exploitation aversion and other similar aversions like those derived from intention-based models of reciprocity (e.g., Charness & Rabin, 2002 or Falk & Fischbacher, 2006). Second, one might run a version of the experiment using the real effort paradigm. In our experience, intrinsic motivation is often very strong in real effort experiments, so strong that treatment effects are usually put to a very strict test (e.g., van van Dijk, Sonnemans, & van Winden, 2001). It would be interesting to see if the effects of exploitation aversion survive when intrinsic motivation is particularly salient. Lastly, another potentially fruitful line of research might be to further develop an exploitation aversion scale that could be used with other related experiments, particularly ones in which motivational crowd-out has previously been observed.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joep.2017.04.006.

References

Anderhub, V., Gaechter, S., & Koenigstein, M. (2002). Efficient contracting and fair play in a simple principal-agent experiment. *Experimental Economics, 5*(1), 5–27.

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review, 99*(1), 544–555.

Ariely, D., Gneezy, U., Loewenstein, G., & Mazr, N. (2009). Large stakes and big mistakes. *Review of Economic Studies, 76*(2), 451–469.

Bartling, B., Fehr, E., & Schmidt, K. (2013). Use and abuse of authority: a behavioral foundation of the employment relation. *Department of Economics, Journal of the European Economic Association, 4*, 711–742.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review, 96*, 1652–1678.

Blount, S. (1995). When social outcomes aren't fair: The effect of causal attribution on preferences. *Organizational Behavior & Human Decision Processes, 63*(2), 131–144.

Bohnet, I., Greig, F., Herrmann, B., & Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey and the United States. *American Economic Review, 98*(1), 294–310.

Bowles, S., & Polanía-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature, 50*(2), 368–425.

Buchanan, A. (1985). *Ethics, efficiency, and the market*. Totowa, NJ: Rowman and Allanheld.

Cardenas, J. C., Stranlund, J., & Willis, C. (2000). Local environmental control and institutional crowding-out. *World Development, 28*(10), 1719–1733.

Carpenter, J., & Myers, C. K. (2010). Why volunteer? Evidence on the role of altruism, reputation and incentives. *Journal of Public Economics, 94*(11–12), 911–920.

Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics, 22*(3), 665–688.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 117*(3), 817–869.

Deci, E., & Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior, 54*(2), 293–315.

Falk, A., & Kosfeld, M. (2006). The hidden cost of control. *American Economic Review, 96*(5), 1611–1630.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature, 422*(6928), 137–140.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 769–816.

Feinberg, J. (1988). *Harmless wrongdoing: Moral limits of the criminal law*. Oxford: Oxford University Press.

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

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives, 25*(4), 191–210.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *Quarterly Journal of Economics, 115*(3), 791–810.

Holmstrom, B. (1982). Moral hazard in teams. *Bell Journal of Economics, 13*, 324–340.

Houser, D., Xiao, E., McCabe, K., & Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior, 62*(2), 509–532.

Lazear, E. (2000). Performance pay and productivity. *American Economic Review, 90*(5), 1346–1361.

Lepper, M., Greene, D., & Nisbett, R. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the overjustification hypothesis. *Journal of Personality and Social Psychology, 28*(1), 129–137.

Mellstrom, C., & Johannesson, M. (2008). Crowding out in blood donation: Was Titmuss right? *Journal of the European Economic Association, 6*(4), 845–863.

Schnedler, W., & Vadovic, R. (2011). Legitimacy of control. *Journal of Economics and Management Strategy, 20*(4), 985–1009.

Shearer, B. (2004). Piece rates, fixed wages and incentives: Evidence from a field experiment. *Review of Economic Studies, 71*(2), 513–534.

Titmuss, R. (1970). *The gift relationship*. London: Allen and Unwin.

van Dijk, F., Sonnemans, J., & van Winden, F. (2001). Incentive systems in a real effort experiment. *European Economic Review, 45*, 187–214.