DISCUSSION PAPER SERIES

IZA DP No. 7499

Exploitation Aversion: When Financial Incentives Fail to Motivate Agents

Jeffrey Carpenter
David Dolifka

July 2013

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Exploitation Aversion: When Financial Incentives Fail to Motivate Agents

Jeffrey Carpenter
Middlebury College
and IZA

David Dolifka
Middlebury College

Discussion Paper No. 7499
July 2013

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7499
July 2013

ABSTRACT

Exploitation Aversion: When Financial Incentives Fail to Motivate Agents

Empirical studies of the principal-agent relationship find that extrinsic incentives work in many instances, linking rewards to performance increases effort, but that they can also backfire, reducing effort. Intrinsic motivation, the internal drive to work to master a skill or to improve one's self image, is thought to be the key to whether incentives work or not. If the incentives crowd-out intrinsic motivation, and the effect is large enough, the net motivational effect on effort will be negative. We posit that an aversion to being exploited, i.e. being used instrumentally for the benefit of another, is one facet of intrinsic motivation, triggered by the combination of high-powered incentives and egoistic principal intent, that can cause incentives to fail. Using an experiment that provides the material circumstances necessary for exploitation to occur, we find that agent compliance is significantly lower for exploitative principals who use high-powered incentives and have a financial interest to do so, compared to neutral principals who use the same contracts but do not benefit from them. To corroborate our interpretation of the results we show that a surveyed "exploitation aversion" scale moderates this effect. Exploitation averse participants are less likely to comply with the incentives than exploitation tolerant participants when the principal signals an exploitative intent, but they are no less likely to comply with the same incentives when the principal is neutral. Our results have implications for the design and implementation of incentive structures within organizations.

NON-TECHNICAL SUMMARY

We study whether an aversion to being exploited, i.e. being used instrumentally for the benefit of another, is triggered by the combination of high-powered incentives and egoistic manager intent, which can cause incentives to backfire. Using an experiment, we find that worker compliance is lower for exploitative managers who use high-powered incentives and have a financial interest to do so, compared to neutral managers who use the same contracts but do not benefit from them. Our results have implications for the design and implementation of incentive structures within organizations.

JEL Classification: C92, J33, M52, M55

Keywords: financial incentives, intrinsic motivation, crowding, exploitation, experiment

Corresponding author:

Jeffrey Carpenter
Department of Economics
Middlebury College
Middlebury, VT 05753
USA
E-mail: jpc@middlebury.edu

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 mission of your organization. There is also empirical evidence that suggests that high-powered incentives 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 et al., 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 volunteers, Carpenter and Myers (2010) show that financial incentive have no effect on the labor supply of volunteer firefighters who are "image-concerned" 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 et al., (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?

In light of the puzzling empirical findings on the effectiveness of high-powered incentives, new theories have evolved, many of which have been influenced by the work in psychology of Deci and Ryan (1985). While economists have traditionally focused on "extrinsic" motivation, i.e., agents working to achieve some outcome (e.g., a financial reward), many psychologists have endeavored to understand "intrinsic" motivation, the internal drive to work to master a skill or to improve one's self image. In many of these new theories (e.g., Bénabou and Tirole, 2003) financial incentives motivate agent effort through the extrinsic element of motivation as predicted by personnel economics but may crowd-out intrinsic motivation. In the case of volunteers, for example, extrinsic rewards might reduce the intrinsic pride one takes in serving the public good and, as a result, the net effect of financial incentives on motivation may be negative.

If intrinsic motivation is potentially the key to predicting when high-powered incentives will succeed or fail, then it is clearly worth unpacking this broad concept both to identify the dimensions of intrinsic motivation that are activated by financial incentives and to determine whether these dimensions complement the incentives, thus adding to the intended effect on motivation or substitute for them, potentially diminishing motivation. Considering the existing literature, Bowles and Polanía-Reyes (2012) have identified a number of ways that financial incentives might reduce intrinsic motivation. Extrinsic incentives can reframe an interaction from one in which effort is required based on moral reasoning to one in which effort becomes a choice because the incentives highlight a possible tradeoff that previously seemed unthinkable (e.g., Cardenas, Stranlund and Willis, 2000), they can adversely affect an agent's sense of autonomy (Falk and Kosfeld, 2006), and they can provide information about the principal who has chosen the incentives. Our study is designed to examine this last dimension. More particularly, we conjecture that, through their choice of incentives, principals may signal selfish intentions that can reduce intrinsic motivation, causing clear incentives to backfire.

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 more carefully consider 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 crowd out intrinsic 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. In one case, the neutral one, contracts may satisfy the material conditions for exploitation but agents can not

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 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 specifications. In addition, we show that a survey instrument designed to measure "exploitation aversion" moderates 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. 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. 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. The contracts we allowed were also simple and easy to understand but, importantly, they were the choice of the principal. 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, $e_i \in [1, 2, ...10]$. 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 Figure 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 optimal 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 Figure 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 below 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 Figure 1 occurs where effort is equal to five, though this is more obvious in Figure 2(a) 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.

e	1	2	3	4	5	6	7	8	9	10
$c(e)$	5	10	25	45	70	120	180	250	330	420
$c'(e)$	5	10	15	20	25	50	60	70	80	90

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 (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. Figure 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 $e = 10$. Hence, agents actually have a material incentive to shirk when $e_{min} = 10$, but rather than changing the marginal cost so that this payoff was small and positive, we decided it would be more interesting to leave it as another check on whether participants understood the game.

Returning to Figure 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?

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 affect the principal's payoff.

In the *neutral* condition, principals were simply paid a flat rate of $\pi_p^{Neutral} = 200$ per period. 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 because of it, 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 data 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 Figure 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. 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. The agents did not know what values were under consideration nor did they know the number of choices from which the the principal could choose (a complete set of experimental instructions are presented in the appendix). Principal contracts were then transmitted to the teams of agents using the z-Tree programming environment (Fischbacher, 2007) who then saw the appropriate table (again, Figure 2(b) shows the table for $e_{min} = 8$) and chose whether to comply with the contract (and contribute at least e_{min}) or not. Once the twenty periods were over, the participants completed a post-experiment survey while the experimenters calculated the earnings for each participant.

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. Figure 3 plots mean effort choices in the first block by period and treatment. As one can clearly see, our participants quickly 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). Figure 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 and then we dig into the details of agent contract compliance.

The choices of the principals, separated by treatment, are illustrated in Figure 4. The most important aspect of the figure is that there is full support for the distribution of e_{min} choices in each treatment. Nearly full support, that is. 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 Figure 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 ($p = 0.02$).

Starting with the most general analysis of compliance, in Figure 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 Figure 4) must account for some of the difference in compliance. With this in mind, in Figure 6 we plot average compliance by treatment and e_{min} choice. Confirming again that agents understood their incentives, every contract with $e_{min} \leq 5$ is complied with. Indeed, all contracts such that $e_{min} \leq 7$ are complied with, despite $e_{min} = 6, 7$ being potentially exploitative. The obvious reason for for full compliance up to $e_{min} = 7$ is that the symmetric payoff up to $e = 7$, while less than the social optimal, is still larger than what would be received at the, now focal, Nash equilibrium. In terms of treatment differences, Figure 6 clearly shows that the compliance differential is occurring exclusively at higher levels of e_{min}, a result that one would expect is 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. One possible explanation lies with the asymmetric information surrounding the choice of e_{min}. Perhaps agents feared that principals would retaliate with additional harsh contracts if they did not comply. That said, this fear must have been less motivating when agents felt exploited because the differential persists.

To be more formal about our analysis of the treatment differential, in Table 1 we present linear probability estimates of compliance. In the first column, we reproduce what was seen in Figure 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.

5 Robustness

Using Table 2, 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 1, the estimates are identical.

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 Figure 4), they will also differ in the history of play. To account for this we added the lag of e_{min} to our estimating equation. In the second column of Table 2 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 have some corroboration of this conclusion. 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.

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 questions loaded negatively suggesting that our intuition about how to classify the responses was correct. To keep things simple, we used the factor scores to create an exploitation aversion indicator, splitting the full sample of agents at the median. 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 Figure 7, 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 2, that surveyed exploitation aversion moderates 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 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. One shortcoming of this literature, however, is that intrinsic motivation is an amorphous, "catch-all," category that is likely to have a number of different dimensions. The purpose of our study is to begin to pull apart the various aspects of intrinsic motivation and identify the ones that attenuate the effectiveness 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 this 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 dircetly 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 intrinsic motivation. In our experiment all agents are exposed to the same contracts, regardless of treatment and so, given we find an effect, it can't just be the material outcomes that matter. In fact, the only difference between our 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 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. Our 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. In Bartling, Fehr and Schmidt (2012), the experimenters examine whether employment contracts are viable when principals face moral hazard about how to complete 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.

Next steps in this area of research might include running a version of the experiment using the real effort paradigm. In our experience, intrinsic motivation is very strong in real effort experiments, so strong that treatment effects are usually put to a very strict test (e.g., van Dijk et al., 2001). It would be interesting to see if the effects of exploitation aversion survive when intrinsic motivation is particularly salient. 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.

7 References

Anderhub, Vital; Simon Gaechter and Manfred Koenigstein. 2002. "Efficient Contracting and Fair Play in a Simple Principal-Agent Experiment." Experimental Economics, 5(1), 5-27.

Ariely, Dan; Uri Gneezy; George Loewenstein and Nina Mazr. 2009. "Large Stakes and Big Mistakes." Review of Economic Studies, 76(2), 451-69.

Bartling, Bjorn; Ernst Fehr and Klaus Schmidt. 2012. "Use and Abuse of Authority: A Behavioral Foundation of the Employment Relation," Department of Economics, University of Zurich Working Paper.

Bénabou, Roland and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." Review of Economic Studies, 70(3), 489-520.

Bowles, Samuel and Sandra Polanía-Reyes. 2012. "Economic Incentives and Social Preferences: Substitutes or Complements?" Journal of Economic Literature, 50(2), 368-425.

Buchanan, Allen. 1985. Ethics, Efficiency, and the Market. Totowa, N.J.: Rowman and Allanheld.

Cardenas, J.C.; J. Stranlund and C. Willis. 2000. "Local Environmental Control and Institutional Crowding-Out." World Development, 28(10), 1719-33.

Carpenter, Jeffrey and Caitlin Knowles Myers. 2010. "Why Volunteer? Evidence on the Role of Altruism, Reputation and Incentives." Journal of Public Economics, 94(11-12), 911-20.

Deci, Edward and Richard Ryan. 1985. Intrinsic Motivation and Self-Determination in Human Behavior. New York: Plenum.

Falk, Armin and Michael Kosfeld. 2006. "The Hidden Cost of Control." American Economic Review, 96(5), 1611-30.

Feinberg, Joel. 1988. Harmless Wrongdoing: Moral Limits of the Criminal Law. Oxford: Oxford University Press.

Fischbacher, Urs. 2007. "Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments." Experimental Economics, 10(2), 171-78.

Gneezy, Uri and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." Quarterly Journal of Economics, 115(3), 791-810.

Holmstrom, Bengt. 1982. "Moral Hazard in Teams." Bell Journal of Economics, 13, 324-40.

Lazear, Edward. 2000. "Performance Pay and Productivity." American Economic Review, 90(5), 1346-61.

Mellstrom, Carl and Magnus Johannesson. 2008. "Crowding out in Blood Donation: Was Titmuss Right?" Journal of the European Economic Association, 6(4), 845-63.

Shearer, Bruce. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment." Review of Economic Studies, 71(2), 513-34.

van Dijk, Frans; Joep Sonnemans and Frans van Winden. 2001. "Incentive Systems in a

Real Effort Experiment." European Economic Review, 45, 187-214.

8 Appendix - Experiment instructions

Instructions to first-stage team production game

Introduction

Thank you for participating in our study today. In this experiment, you will be asked to make choices affecting both your own monetary payoffs and the payoffs of other individuals in the room. All of your decisions will remain strictly confidential. To assure confidentiality and anonymity, we request that you do not speak with other participants at any point during this session. Please direct all questions to the lab assistant.

You will automatically receive $5.00 for just showing up and you will have the opportunity to make additional earnings depending on your choices in the experiment. You will earn Experimental Monetary Units (EMUs), which will be converted to dollars at the end of the session. The exchange rate will be 100 EMUs = $1.00. Given this exchange rate, experimental earnings for a participant with 1,800 EMUs, for example, would be $23.00 ($5 show up reward + $18 in EMU conversion). You will be paid in cash, rounding to the nearest dollar, immediately following the conclusion of today's experiment. Funding for this study has been provided by Middlebury College.

We will now proceed to the instructions for Round 1. Please read along with the lab assistant and raise your hand if you have any questions. There will be a short quiz after these instructions. The purpose of the quiz is solely to test your understanding of the game; your performance will have no effect on your earnings. Furthermore, to ensure your complete understanding of the game, we will provide solutions to the quiz questions.

Round 1 – Ten Periods – Computer-Based Game.

In this first round, you will be playing ten periods of a team production game. At the beginning of the round, you will be randomly assigned to a team with three other participants. Your team will not change during this round. To clarify, you will be matched with the same three people for each of the ten periods in Round 1.

To begin each period, you will be given an endowment of 10 units of effort. You may choose to contribute anywhere from 1 to 10 of your units of effort to the team. The output of the team will depend on the total amount contributed by all its members. Specifically, the total output of the team will be the sum of the efforts provided by the workers multiplied by a productivity factor of 40, or

$$\text{Team Output} = 40\times(e_1+e_2+e_3+e_4)$$

Where e_1 is the effort provided by the first worker, e_2 is the second worker's effort, and so on.

As compensation, each member of the team will get an equal share of this output (i.e., $\frac{1}{4}$ of it), regardless of how much effort that team member provided. This means that, regardless of how much effort you provide, you will earn

$$\text{Individual Earnings} = \tfrac{1}{4} \times \text{Team Output} = [40\times(e_1+e_2+e_3+e_4)]/4$$

which, for simplicity, just equals $10\times(e_1+e_2+e_3+e_4)$ because the 4s cancel. For example, worker 1 provides e_1 and receives a payout of $10\times(e_1+e_2+e_3+e_4)$.

Providing effort, however, will be costly and your final profit for each period will take this cost into account. Your cost of providing effort, called c(e), increases as you work "harder." The relationship between your effort and the cost of that effort is summarized in the following table.

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

All costs are calculated in EMUs. For example, it only costs 45 EMUs to contribute 4 units of effort but it costs 250 EMUs to contribute 8 units of effort.

Your final profit for each period will be the difference between your individual earnings and the cost associated with your individual level of effort. This is summarized as

$$\text{Profit} = \text{Earnings} - \text{Cost of Effort or Profit for Worker 1} = \{10\times(e_1+e_2+e_3+e_4)\} - c(e_1)$$

Here are some examples:

Example 1: Suppose all four workers in a group provide 5 units of effort. Because they all do the same thing, they will all receive the same payoff. The total effort will be 20 units and the earnings of each worker will be 10×20=200 EMUs. From the table above, the cost of the effort provided for each worker will be 70 EMUs and so each worker receives a profit of 130 EMUs for the period.

Example 2: Now suppose that workers 1, 2 and 3 continue to provide 5 units of effort but worker 4 chooses to provide just 2 units. Here the total effort for the team is 17 units and each worker earns 10×17=170 EMUs. The first three workers, who provided 5 units of effort each, receive profits for the period of 170-70=100 EMUs and the fourth worker, who provided 2 units of effort, earns 170-10=160 EMUs.

Example 3: Lastly, let's assume all four workers now provided 2 units of effort. Here team output is 8 units and each worker earns 10×8=80 EMUs. They all provide 2 units of effort, which costs them each 10 EMUs, and so they all earn 80-10=70 EMUs in profit for the period.

These are just examples. Your choices are completely up to you and can change from period to period. To simplify things as much as possible, we are providing a table that does

all of these calculations for you. It shows you how much profit you will earn for each possible effort level and for the possible average effort levels of the other three members of your team. This table will also appear on your computer screen each period to help you in your effort choices.

Are there any questions?

We will now begin with the quiz, and then go immediately into the first round of the game. Remember, you may not speak to any other participants or ask questions once the experiment has started.

Additional Instructions to Second Stage Forcing Contract Game

Round 2 – Ten Periods – Computer-Based Game.

There will be a short quiz prior to beginning Round 2. Like before, your responses to the quiz questions will not alter your earnings. In this round, the experiment will proceed very much like in round one, except we will alter the team structure to incorporate a "boss" to oversee each team of workers. Bosses will be chosen at random such that there is one boss for every four workers. Participants will maintain the same role for all ten periods of this round. As in Round 1, the groups will remain the same for all 10 periods. These groups will not be the same groups as in the first round.

It may be helpful to think of each period as two separate stages.

Stage 1

During stage 1, the boss will choose and enforce a minimum required level of effort, "e_{min}". This minimum required effort will be a number between 1 and 10. At the beginning of each period, bosses will select e_{min} from the options provided by the computer program.

Stage 2

Workers will be notified of the required e_{min} for the team once it has been chosen. They will then have to decide whether or not to contribute a level of effort at or above e_{min}. If workers provide an effort level at or above e_{min}, their profits will be calculated in the same way as during the first round (i.e., earnings minus cost of effort). However, if workers contribute less than e_{min}, their profits will be set to zero. Each period, the computer program will display the relevant payoff table that accounts for the boss' choice of e_{min} to aid in the decision-making process.

[Exploitative Principal] Bosses will be paid depending on their choice of e_{min}. Specifically, the boss earns $20 \times e_{min}$. For example, a boss would earn 80 EMUs if s/he selected an e_{min} of 4 and would earn 200 EMUs if s/he selected an e_{min} of 10. The boss's pay does not depend on the effort provided by the team but it does depend on her/his choice of e_{min}.

[Neutral Principal] Bosses will be paid a fixed wage of 200 EMUs per period. Boss pay does not depend on the effort provided by the team or on her/his choice of e_{min}.

Are there any questions?

You may now begin Round 2 of this experiment.

9 Figures and tables

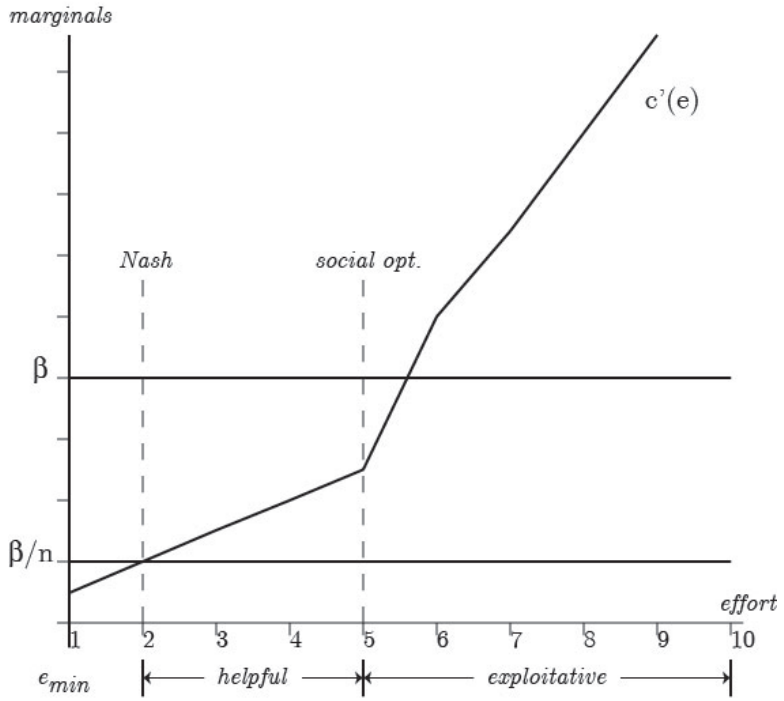

Figure 1. Experimental Design.

		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

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

		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

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

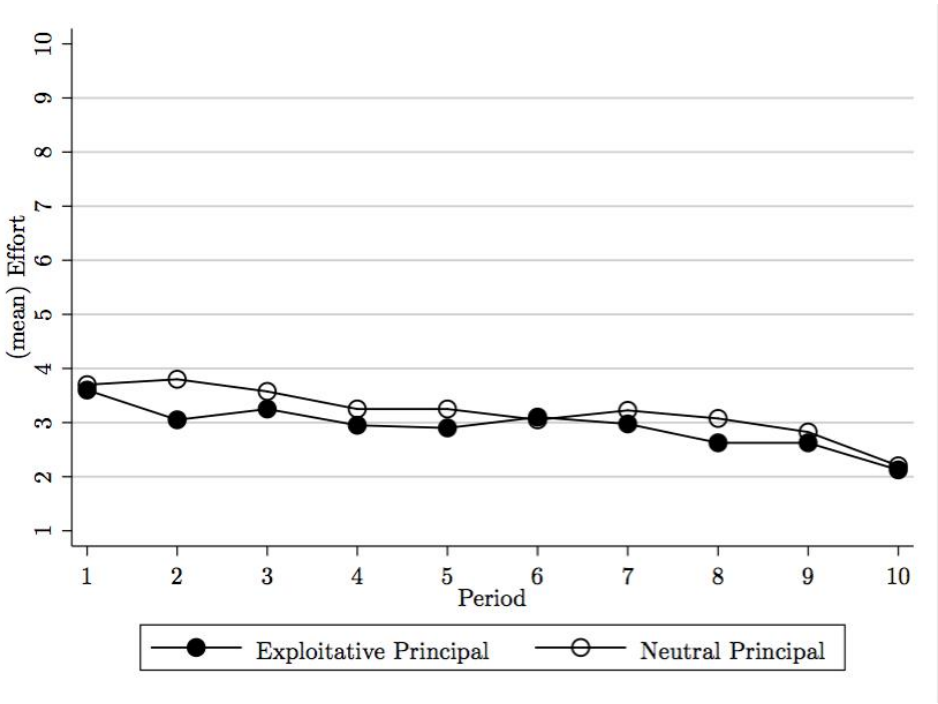

Figure 3. Effort choices in the initial ten periods without forcing (by treatment).

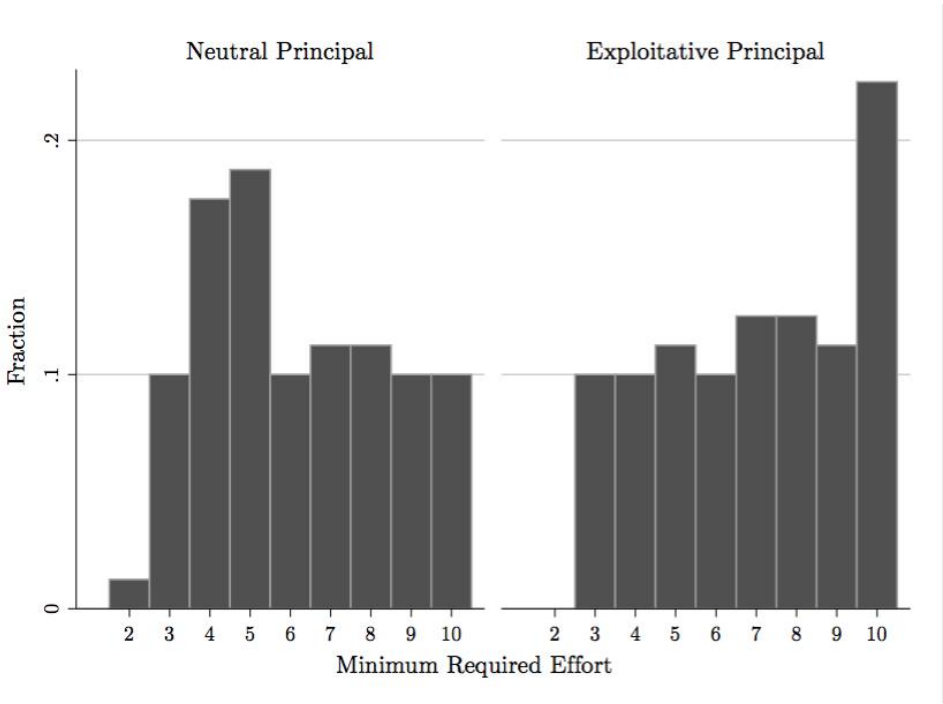

Figure 4. Principal choices of forcing levels, e_{min}, (by treatment).

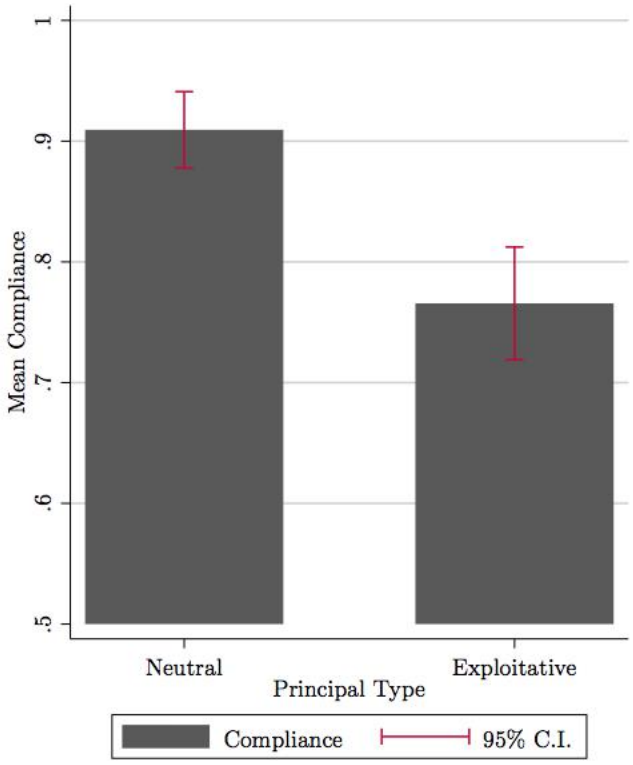

Figure 5. Overall forcing contract compliance (by treatment).

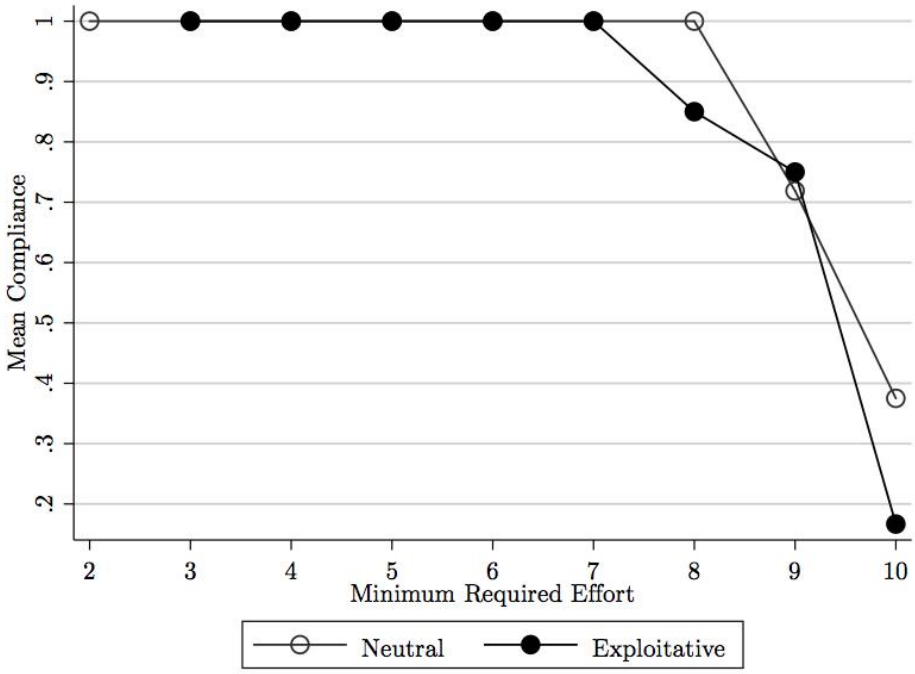

Figure 6. Contract compliance by forcing level, e_{min}, (and treatment).

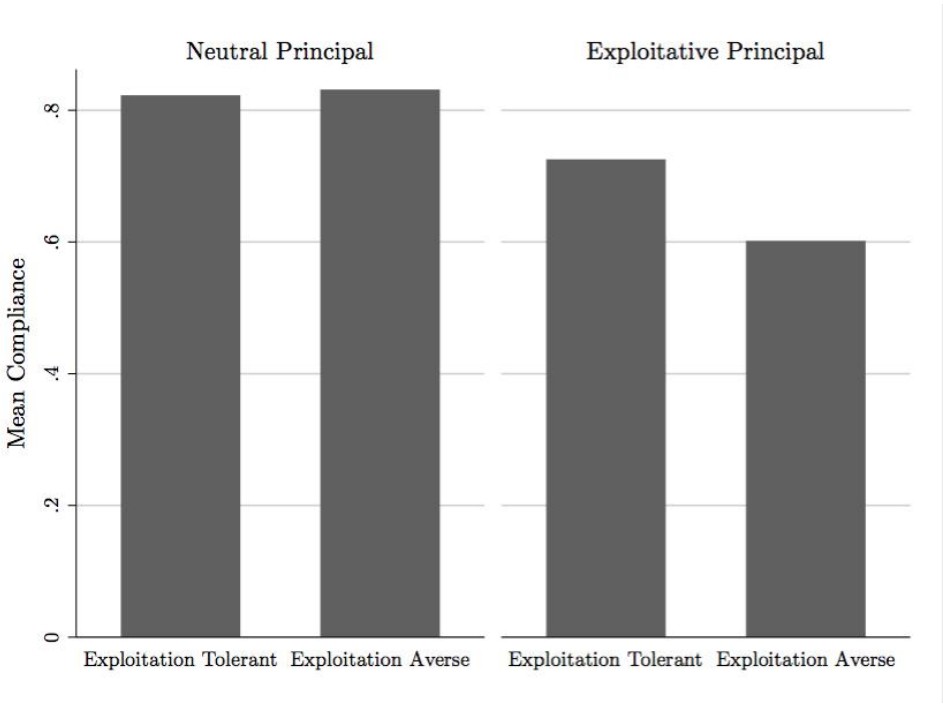

Figure 7. Contract compliance and surveyed exploitation aversion (by treatment).

Table 1: 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); $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$; controls include age, sex, major, race and GPA.

Table 2: 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); *$p < 0.10$, **$p < 0.05$, ***$p < 0.01$; controls include age, sex, major, race and GPA.