Two experiments on the interaction of information, incentives and motivation

Anton Suvorov (HSE)

HSE Summer School in Experimental Economics 2019

July 2019

Incentives

- "Today, for many economists, economics is to a large extent a matter of incentives: incentives to work hard, to produce good quality products, to study, to invest, to save, etc."

J.-J. Laffont and D. Martimort ,
"The Theory of Incentives", 2002

Incentives

- Incentives are ubiquitous: rewards and punishments, promotions and demotions, piece-rates and stock options, monetary prizes and symbolic rewards are used to motivate students and professors, artists and scientists, taxi-drivers and top executives...
- Of course, no hope to give any comprehensive treatment in a single lecture; necessarily a highly selective and somewhat eclectic talk, driven largely by personal research agenda.

Incentives

- Conventional contract theory focuses largely on the agency problem studied within the principal-agent framework. Instruments are typically rewards/punishments contingent on observable and verifiable outcomes (e.g., piece rates for workers or sharecropping contracts in agriculture). However, also
  - career concerns (Holmstrom, 1982/1999; Dewatripont, Jewitt and Tirole, 1999; etc.);
  - subjective evaluation and feedback provision, implicit contracts (e.g., MacLeod, 2003; Levin, 2003; Fuchs, 2007; Suvorov and van de Ven, 2011; etc.);
  - leadership in organizations (Hermalin, 1998; Bolton, Brunnermeier and Veldkamp, 2012; etc.)
  - etc.

Why do incentives work?

- They explicitly or implicitly make the agent's material payoffs contingent on performance (which, itself, depends on effort).
- They provide information to the agent (e.g., feedback or leader's actions); information then changes the agent's incentives.
- They affect the link between the agent's actions and the public beliefs about the agent's characteristics (ability, altruism, etc.).
- In this talk mostly focus on the informational aspects or the interplay between the direct and informational effects.
- Also, we will pay much attention to *intrinsic motivation*.

Intrinsic motivation

- We can speak of intrinsic motivation when people do things without obvious external reasons.
- Deci and Ryan (2017) "Intrinsically motivated behaviors are those that are performed out of interest and for which the primary “reward” is the spontaneous feelings of effectance and enjoyment that accompany the behaviors."

Extrinsic motivation

- Deci and Ryan (2017): "Intrinsic motivation contrasts with extrinsic motivation, represented by behaviors that are instrumental for some separable consequence such as an external reward or social approval, avoidance of punishment, or the attainment of a valued outcome."
- Kreps (1997): "what is called intrinsic motivation may be (at least in part) the worker's response to fuzzy extrinsic motivators, such as fear of discharge, censure by fellow employees, or even the desire for coworkers' esteem".

Extrinsic incentives and intrinsic motivation

- Psychologists have been concerned that extrinsic incentives may have perverse effects ("crowd out") intrinsic motivation.
- Economists since the late 1990s joined the debate. For instance, Frey (1997) "Not Just For The Money" cites some empirical evidence and has a reduced-form model of motivation crowding out.
- Understanding mechanisms underlying potential crowding-out effects has been a challenge:
  - extrinsic incentives often work very well;
  - without understanding the mechanisms one cannot identify when and why negative effects can occur;
  - therefore, no scope for normative judgement and policy advice.

Hidden costs of rewards

Explanations

Extrinsic incentives "crowd out" *intrinsic motivation* (Deci and Ryan, 1985)

- Undermine *self-determination*
- Reduce *perceived competence*

Thus, both controlling and information effects of rewards are important

Hidden costs of rewards

Explanations

Overjustification effect:

- Cognitive dissonance theory (Festinger, 1957): "why are they paying me if the task is interesting and I am good at it?"
- Self-perception theory (Bem, 1967): "if I have been once paid for doing it, probably the thing should be boring..."

Also insufficient justification:

- E.g., Tom Sawyer and the fence story

Bénabou and Tirole (2003) explore these ideas in a game-theoretic framework

Benabou and Tirole (2003)

- Principal-agent model. Key assumptions:
  - The agent is uncertain about his ability, task difficulty, costs or benefits of implementing the task (young inexperienced employee, student, child...).
  - The principal, in contrast, is informed (experienced manager, professor, parent...).
  - The principal is uncertain about the agent's self-confidence (or degree of optimism about costs and benefits), i.e. two-sided asymmetric info.
  - The principal may offer performance-contingent bonuses.

Benabou and Tirole (2003)

- In equilibrium:
  - Bonuses weakly decrease in the agent's ability.
  - Thus, a higher bonus is bad news about ability.
  - Yet, a high bonus is effective short-term reinforcer (otherwise, would not be given).
  - In the long run, bonuses "reduce liking" of the task.

Benabou and Tirole (2003)

- Robustness and extensions:
  - It does not really matter whether uncertainty is about the agent's ability (suitability for the task), his costs or benefits.
  - The principal's decision to delegate control rights to the agent also transmits info; under reasonable conditions delegation is good news about ability.
  - The principal's decision to help an agent may bring bad news (if the principal's involvement substitutes the agent's effort); if the principal's and the agent's efforts are complements, help is good news.

Suvorov (2003)

- What happens in Benabou and Tirole's model when the relationship is repeated?
  - Ratchet effect for the agent (does not want to appear enthusiastic) and for the principal (concerned about creating "addiction" by offering a reward).
  - Learning effect for the agent.
- If learning effect is not very strong, similar monotonicity results obtain; otherwise, the principal promises high bonuses to the strong agent to promote learning.

Motivation

Bremzen, Khokhlova, Suvorov and van de Ven (2014): experimental test of Benabou and Tirole (2003) model.

- Before: extensive experimental literature in psychology and economics on "hidden costs" and "crowding out", but no paper that would fit well the assumptions of the model.

Hidden costs of rewards

Classical experiments in psychology

Deci (1971):

- Students work on an interesting task: solve puzzles
- Performance-contingent monetary rewards introduced for the experimental group, no rewards for the control group
- Then, rewards are withdrawn. Experimental group shows then lower engagement in the task compared to the control group

Lepper et al. (1973):

- Field experiment with 4-year-olds
- Draw with "magic markers" under 3 conditions: "no reward", "promised reward", "unexpected reward".
- Promised rewards reduce subsequent interest in playing with the markers, unexpected rewards have no effect

Related literature

Experiments in psychology

Large experimental literature in psychology on "hidden costs of rewards".

- Dozens of papers since Deci (1971)...
- Controversial results (e.g., meta analysis by Deci et al. (1999) and critical view by Eisenberger et al. (1999)), but delayed negative effects often found
- As Lepper et al. (1999) put it:

> *"...the relevant issue for further research was not whether rewards have negative, or positive effects "in general" but rather when and why these different effects might occur."*

But rewards are administered by the experimenter with unclear objectives $\Longrightarrow$ there is no direct test of the information effects

Related literature

Experiments in economics

- Growing experimental literature in economics on "crowding out". Mostly, on direct negative effect.
  - Gneezy & Rustichini (2000b): effect of small rewards ("pay enough or don't pay at all")
  - Frey and Oberholzer-Gee (1997): willingness to accept a nuclear facility (survey)
  - Fehr, Gächter and Kirchsteiger (1997), Fehr and Gähter (2001),...: experimental labor markets
  - Falk and Kosfeld (2006): monitoring signals distrust and reduces effort
  - Ariely et al. (2009): high rewards make people choke under pressure
  - also Galbiati et al. (2010) (on rewards and sanctions), Charness et al. (2010) (autonomy), Dickinson and Villeval (2004) (monitoring), Gneezy & Rustichini (2000a),...

Related literature

Theories in economics

- Frey (1997): reduced form "crowding out" theory
- Bénabou and Tirole (2003): informational content of rewards
- Bénabou and Tirole (2006): rewards undermine social signaling
- Seabright (2009): social signalling + matching
- Sliwka (2007): rewards signal about social norms
- Ellingsen and Johannesson (2008): rewards signal the principal's character

Contribution

Explicit test of attribution mechanism: do people infer bad news from a high bonus?

Controlled experiment, the structure of the game and stakes of participants are common knowledge to both players (up to explicitly introduced asymmetric information about task difficulty).

Main challenge: separate *informational effect* from the direct incentive effect and fairness-based considerations $\Longrightarrow$ key feature of the experiment: the agent works on two projects, a joint project and an independent own project.

Contribution

To our knowledge this is the first paper to address "hidden costs of rewards" in an experiment, where
- principals are participants with clear objectives.
- information effects directly singled out

Our results support all major implications of BT model.
- rewards contain "bad news"
- this is correctly perceived by the agents
- these "hidden costs" coexist with immediate positive effects

The model

The model is a simplified variant of Bénabou and Tirole (2003)

Two risk-neutral players: a principal (she) and an agent (he).

The agent works on a task: he chooses high effort ($e = 1$, cost c) or low effort ($e = 0$, no cost).

The task is equally likely to be easy (cost is low, $c = c_L$) or difficult (cost is high, $c = c_H$).

The model

Information structure

- Informed-principal model
- The agent does not know the task's difficulty, but observes a private informative signal $s \in \{s_H, s_L\}$ that is correct with probability $r > 1/2$.
- The principal knows with certainty the project's difficulty, but does not know the agent's "self-confidence" (his signal s).

Timing

- The principal observes the cost level and then specifies the bonus $b \geq 0$ for hard work and fixed wage w
- The agent receives the private signal and observes the bonus, and then decides on the effort level

Payoffs

The principal gets default payoff P_0 plus additional payoff ΔP if the agent works hard:

$$W = P_0 + e(\Delta P - b) - w$$

The agent gets default payoff A_0 plus additional payoff ΔA if he works hard:

$$V = A_0 + e(\Delta A - c + b) + w$$

Equilibrium features

Bénabou and Tirole show that in any Perfect Bayesian equilibrium:

1) rewards (bonuses) are positive immediate reinforcers: a higher bonus increases the probability of high effort;

2) rewards (bonuses) are informative and convey *bad news*: the principal gives (weakly) higher bonus if the task is difficult.

Intuition: when the task is easy, agents are more self-confident on average, so it is cheaper for a principal not to give additional extrinsic incentives. A high bonus signals *objectively motivated lack of trust* (difficult task, but also weak agent).

Additional assumptions and features

Under full information, the agent is intrinsically motivated to exert effort on the easy task and, but prefers to shirk on the difficult one:

$$c_L < \Delta A < c_H.$$

The principal can offer only two possible bonuses, $b \in \{0, \bar{b}\}$.

If offered a high bonus $\bar{b}$, the agent should work given any beliefs (i.e. $\bar{b} + \Delta A > c_H$).

Parametrization

Costs: $c \in \{15, 45\}$, with equal probability.

Payoffs: $A_0 = 25, P_0 = 10, \Delta A = \Delta P = 30$.

The principal determines the bonus for the agent, $b \in \{0, 20\}$, and an upfront fixed wage, $w \in \{0, 5, 10\}$.

The agent observes the specified bonus and wage, receives a private signal about costs (correct 75% of the time, i.e. $r = 3/4$). The signal is "good" or "bad" reflecting low and high costs respectively.

He then decides to exert effort or not: $e \in \{0, 1\}$.

Equilibrium

We restrict attention to Perfect Bayesian equilibria satisfying D1 (Cho and Kreps, 1987), assuming risk-neutral players.

In the unique PBE that satisfies D1, the principal never pays a fixed wage and offers no bonus when costs are low, and randomizes between no bonus $(p^* = 1/3)$ and a bonus $(1 - p^* = 2/3)$ when costs are high;

The agent exerts effort after a high bonus and/or after a good signal: and randomizes between effort $(q^* = 1/9)$ and no effort $(1 - q^* = 8/9)$ after no bonus in combination with a bad signal.

Identification

Our objective is to test whether or not agents infer bad news from a bonus.

A key element of the design is to introduce an independent, own project for the agent.

The agent makes an effort decision for the joint project, and for his own project. The principal has no stake in the agent's own project.

Identification

The two projects are identical in all respects except that the bonus and wage offered by the principal only applies to the joint project and not the own project.

Thus, if the agent infers information from the bonus, this also applies to the own project, but there is no incentive effect for the own project.

In equilibrium, after a high bonus the agent exerts no effort on the own project, and after a low bonus the agents exerts effort (good signal) or is indifferent (bad signal).

Identification

To test whether agents indeed react to information contained in the bonus, we also implemented a setting with uninformed principal

In equilibrium with an uninformed principal: the principal always offers a bonus and the agent always exerts effort in the joint project. In the own project, the agents exerts effort after a good signal and no effort after a bad signal.

Hypotheses

Hypothesis 1: An informed principal is more likely to offer a high bonus when she observes a high level of costs.

Hypotheses

Hypothesis 1: An informed principal is more likely to offer a high bonus when she observes a high level of costs.

Hypothesis 2: A higher bonus increases effort by the agent in the joint project.

Hypotheses

Hypothesis 1: An informed principal is more likely to offer a high bonus when she observes a high level of costs.

Hypothesis 2: A higher bonus increases effort by the agent in the joint project.

Hypothesis 3 a) With an informed principal, the agent infers bad news from a high bonus and consequently reduces effort in his own project;

Hypotheses

Hypothesis 1: An informed principal is more likely to offer a high bonus when she observes a high level of costs.

Hypothesis 2: A higher bonus increases effort by the agent in the joint project.

Hypothesis 3 a) With an informed principal, the agent infers bad news from a high bonus and consequently reduces effort in his own project; b) with an uninformed principal, the agent infers no information from the bonus, and effort in his own project is unaffected by the size of the bonus.

Experimental design

- See for instance, Charness and Kuhn (2011), for a discussion of tradeoffs an experimentalist is facing in the design of a principal-agent experiment.
- Participants are randomly matched in pairs
  - straightforward choice.
- Rematched every round
  - we wanted to avoid reputation-building, reciprocity (beyond a single round), etc.;
  - we guaranteed that there can be at most a single match between any two participants within streaks of 5 consecutive rounds.
- Stated effort task
  - lacks realism (?) but
  - gives control over cost function, "ability" and other important parameters.

Experimental design

- Labor market framing
  - purists may object: meaningful framing may invoke unobservable differences in interpretation;
  - on the other hand, framing in experiments matters (e.g., Liberman et al., 2004 on framing in prisoner's dilemma) and "neutral" framing may also affect behavior;
  - natural framing may fundamentally change reasoning (Chrostowski and Griggs, 1985: success rate inreased from 10% to 70% when the 4-cards problem (A, K, 4, 7) was reformulated as beverage (beer or soda) and age (22 or 16) problem;
  - Cooper and Kagel (2003): in a signaling game experiment natural framing (in a limit pricing game) accelerates learning, i.e. is a substitute for experience;
  - given the complexity of our game, we opted for the natural, labor market framing.

Experimental design

- All participants play in both informed and uninformed conditions
  - largely, for budgetary reasons
- and in both roles
  - to encourage to look at the problem from both sides and thus, potentially, accelerate learning;
  - in total they play 20 rounds in the informed condition and 12 rounds in the uninformed condition.
- Feedback: participants observe the cost of the project and all payoffs at the end of every round
  - again, to increase chances that subjects comprehend the strategic canvass of the game.

Experimental design

- Participants are paid for every round.
- Advantages of paying for a subset of rounds are that
  - stakes are larger;
  - no wealth effects;
  - no scope for "gambling for resurrection";
  - no (when single paid round) or smaller (when several, but few paid rounds) incentives to hedge the payoffs in different rounds.
- However, pay in every round is easier to comprehend and is unlikely to provoke strong negative consequences in our game.

Experimental design

- A total of 156 participants (8 sessions, subdivided in 12 independent groups)
  - this division on independent groups allowed to increase the number of independent group-level observations without exploding the number of participants (budgetary considerations).
- Order of conditions different between sessions
  - to control for order effects: we did not expect (and did not find) particular order effects here.
- In 8 out of 12 groups, we also elicited social preferences measures.
- Computerized using Z-tree (Fischbacher, 2007)

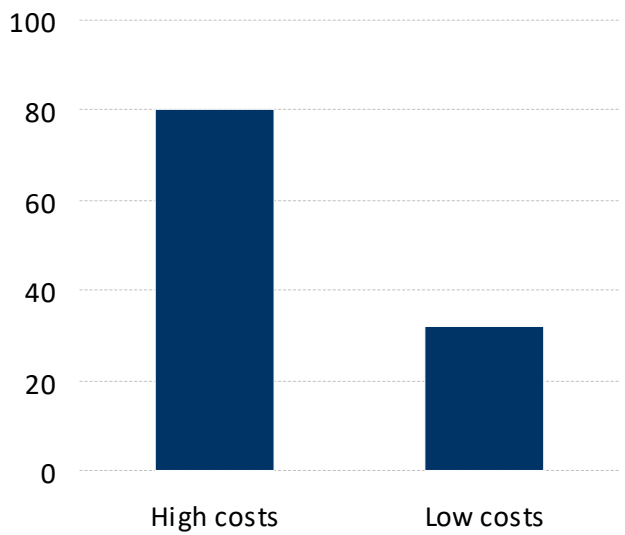
Experimental design

- Sessions lasted for about 90 minutes.
- Average earnings approximately $13.
- All sessions took place in Moscow among undergraduate and graduate students
  - first-year undergraduate students as subjects allowed to increase external validity compared to second-year NES master students; no qualitative difference in results.

Results. Behavior of principals

In the informed condition (main treatment), the bonus is given 80% of the time if costs are high, and only 32% when costs are low. (difference significant at p=.002, Wilcoxon signed rank test, group as unit of obs.)

Figure 1: Bonuses



Bonus in the Main Treatment

	(1) all rounds	(2) All rounds	(3) rounds 1-10	(4) rounds 11-20
High costs	0.480***	0.502***	0.460***	0.560***
	(0.041)	(0.048)	(0.041)	(0.070)
Female		0.002	-0.043	0.070
		(0.077)	(0.080)	(0.100)
Altruist		-0.007	0.007	-0.028
		(0.053)	(0.060)	(0.075)
Trusting		-0.017	-0.033	0.011
		(0.068)	(0.057)	(0.084)
Fair		0.025	-0.015	0.085
		(0.058)	(0.056)	(0.077)
Reciprocal		0.043	0.046	0.043
		(0.059)	(0.054)	(0.081)
N Obs	1,461	1,001	547	454
N Subjects	156	110	110	110
N Groups	12	8	8	8
Pseudo R2	0.181	0.203	0.182	0.242

Probit estimates, reporting marginal effects. Robust s.e. clustered at the group level in parentheses. All specifications include the treatment order as a control variable.
***<0.01, ** p<0.05, * p<0.1

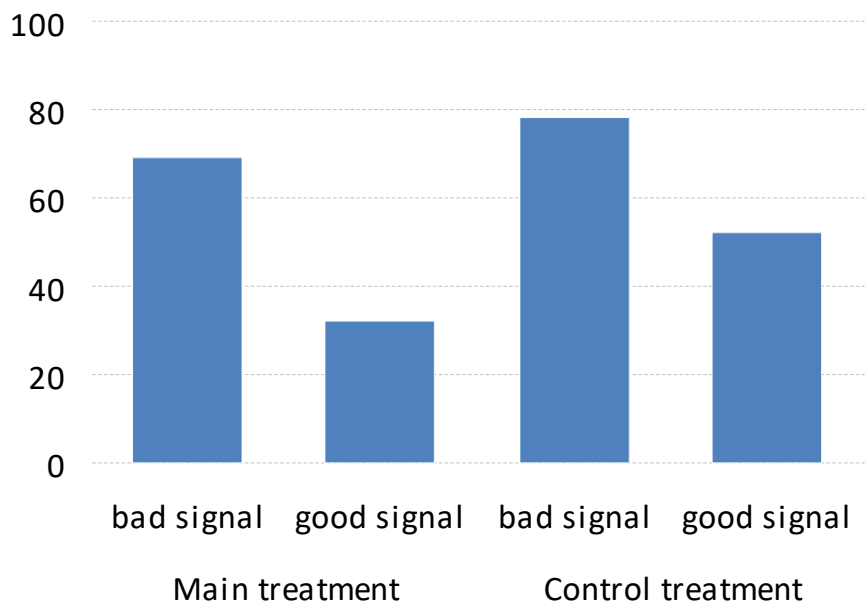
Results. Behavior of principals

Result 1: A bonus is very informative about the level of costs in the informed condition. High costs increase the likelihood of a bonus by around 50 percentage points.

▸ wage distribution

Results. Behavior of agents. Joint project in main treatment

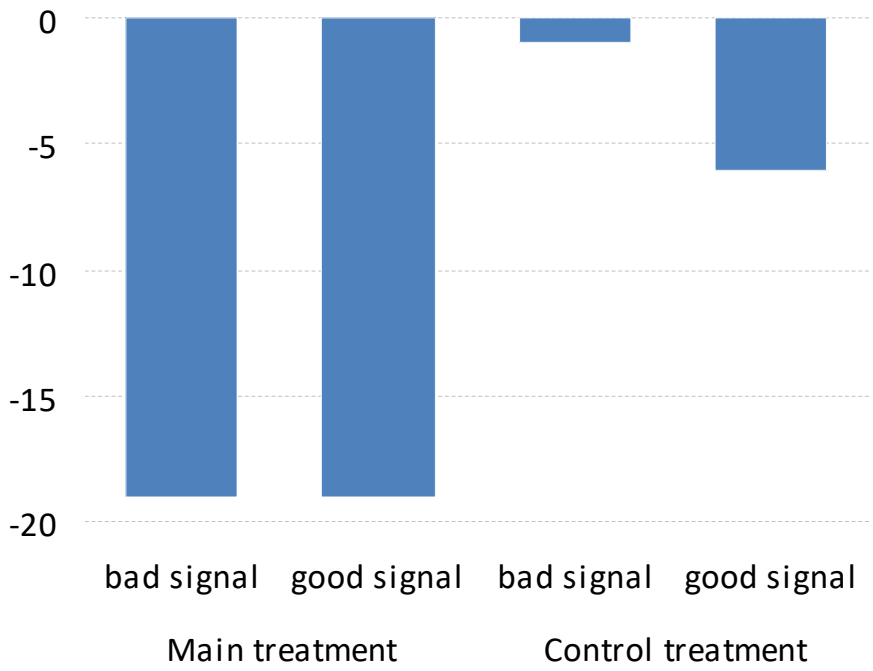
Difference in mean effort in the joint project



Result 2: With an informed principal, a bonus increases effort in the joint project

Results. Behavior of agents. Own project

Effort (on own project) significantly lower in the main treatment, but not in the control treatment.



(p values: .002/.002/.906/.267.

▸ effort own informed ▸ effort own uninformed

Results. Effort in the main treatment

Effort in the Main Treatment

	(1)	(2)	(3)	(4)	(5)	(6)
	joint project		**own project**			
	all rounds	all rounds	all rounds	all rounds	rounds 1-10	rounds 11-20
Bonus	0.636***	0.655***	-0.230***	-0.195***	-0.062	-0.342***
	(0.036)	(0.054)	(0.059)	(0.069)	(0.094)	(0.073)
Good signal	0.304***	0.316***	0.498***	0.480***	0.511***	0.467***
	(0.033)	(0.056)	(0.064)	(0.090)	(0.102)	(0.092)
Bonus X Good signal	-0.367***	-0.358***	0.008	-0.062	-0.125	-0.013
	(0.045)	(0.064)	(0.049)	(0.064)	(0.091)	(0.089)
Wage 5	0.017	0.047	0.086***	0.070*	0.057	0.100
	(0.044)	(0.047)	(0.031)	(0.038)	(0.054)	(0.062)
Wage 10	0.037	0.093*	-0.060	-0.071	-0.020	-0.113
	(0.042)	(0.050)	(0.075)	(0.076)	(0.102)	(0.076)
Female		-0.049		0.106***	0.144***	0.060
		(0.050)		(0.032)	(0.040)	(0.081)
Altruist		0.053**		-0.042	-0.104**	0.042
		(0.023)		(0.036)	(0.050)	(0.062)
Trusting		0.051		-0.010	-0.062**	0.067
		(0.033)		(0.028)	(0.032)	(0.059)
Fair		-0.116***		0.073	0.049	0.122
		(0.037)		(0.069)	(0.073)	(0.079)
Reciprocal		0.003		-0.060*	-0.039	-0.088**
		(0.043)		(0.032)	(0.048)	(0.036)
Observations	1,467	1,007	1,467	1,007	553	454
Pseudo R2	0.267	0.264	0.246	0.220	0.189	0.287

Probit estimates, marginal effects. Robust standard errors clustered at the group level.

Results. Effort in control treatment

Effort in Control Treatment

	(1)	(2)	(3)	(4)
	joint project		own project	
Bonus	0.846***	0.865***	-0.026	-0.048
	(0.029)	(0.042)	(0.071)	(0.090)
Good signal	0.558***	0.610***	0.732***	0.689***
	(0.076)	(0.108)	(0.082)	(0.104)
Bonus X good signal	-0.361***	-0.389***	-0.053	-0.034
	(0.077)	(0.106)	(0.063)	(0.077)
Wage5	-0.213***	-0.257***	0.062	-0.024
	(0.051)	(0.054)	(0.063)	(0.057)
Wage10	0.190***	0.108***	-0.079**	-0.088**
	(0.052)	(0.039)	(0.034)	(0.041)
Female		-0.008		0.006
		(0.105)		(0.070)
Altruist		-0.003		-0.029
		(0.085)		(0.072)
Trusting		-0.001		-0.029
		(0.042)		(0.074)
Fair		-0.080		0.072
		(0.065)		(0.081)
Reciprocal		0.061		0.042
		(0.050)		(0.079)
Number of observations	936	660	936	660
Pseudo R-squared	0.435	0.467	0.376	0.342

Probit estimates, marginal effects. Robust s.e. clustered at the group level.

Results. Behavior of agents

Result 3: Agents correctly infer bad news from a bonus in the informed condition, which leads them to reduce effort in the own project. Effort in the own project is not substantially affected by the size of the bonus in the control treatment.

Results. Behavior of agents. Additional results

- Interaction effects *Bonus* x *Goodsignal* have the signs that correspond to the model's predictions.
- Learning: Inference of bad news is particularly strong in later rounds.
- No systematic gender effects.

Results. Behavior of agents

Figure 4: change in effort in the own project over rounds after high bonus.

Results. Role of social preferences

We also find some evidence of crowding out of motivation for small rewards: mean effort is lower after a fixed wage of 5 than after no positive upfront fixed wage.

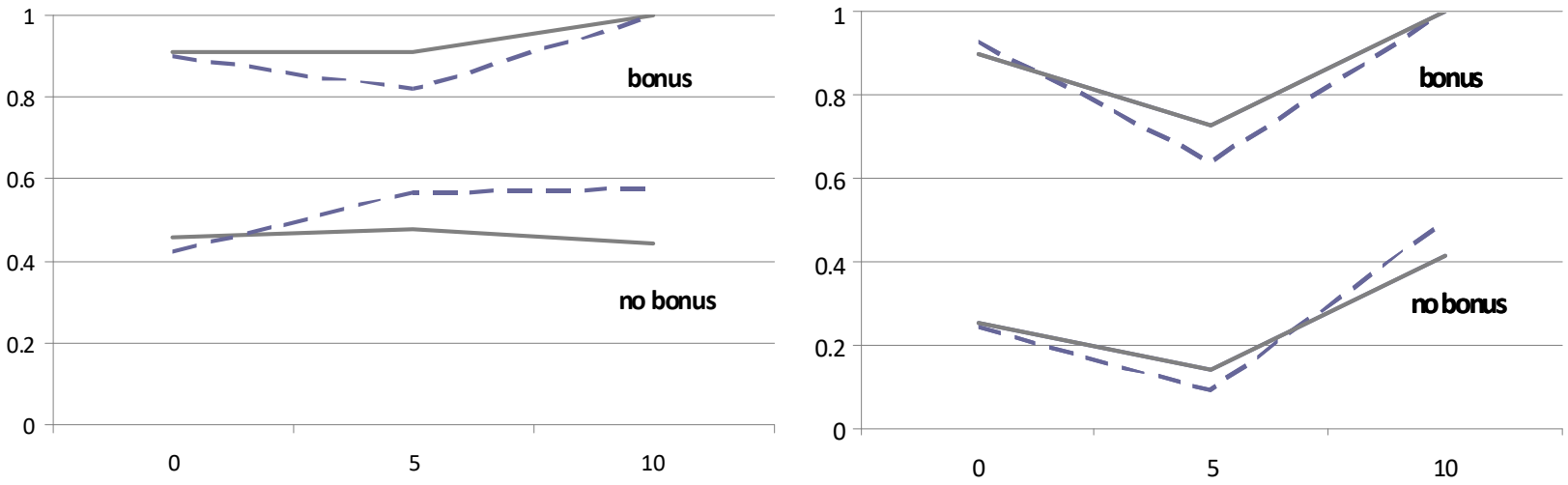
This resembles the result of Gneezy and Rustichini (2000) but with a fixed wage rather than a piece rate.

Effect only shows up in the control treatment.

However, not enough observations for reliable nonparametric tests (positive fixed wage relatively rare).

Results. Role of social preferences

Figure 5: mean effort in the joint project by wage level
Left panel: Main Treatment; right panel: Control Treatment.



Solid line for all participants; dashed line for reciprocal participants.

Discussion

First direct test of information effects with clear stakes of principal.

We find evidence of hidden costs of rewards: despite being *positive short-term reinforcers*, they bring *bad news* and *crowd out motivation*. Robust across specifications.

Possible extensions for future work:

- Real effort task
- Test if informed principals internalize the bad news effect of rewards
- Repeated relationship (as in Suvorov, 2003)
- Non-contractible, discretionary rewards (good news in Suvorov and van de Ven, 2009)

Pay for performance often works!

- Previous discussion might suggest that negative effects of pay for performance are dominant. This is wrong, of course; costs and benefits depend on the context and details of implementation.
- Lazear (2000) reports a natural experiment: piece rates have been introduced in Safelite, the largest US installer of automobile glass.
- Efficiency gains were substantial:
  - The average increase in productivity of 44%.
  - A half of this gain is due to higher worker efforts; another half – to the change in the composition of the workforce.
  - This gained is shared between the firm and the employees (about 10% increase in worker pay).

- Image motivation is an important driver of behavior. It leads to
  - emergence of endogenous social norms (Bernheim, 1994);
  - helps to explain the prevalence of equal division of surplus often observed in experiments (Andreoni and Bernheim, 2009);
  - increases charitable giving (Ariely, Bracha and Meier, 2009; DellaVign, List and Malmendier, 2012).
- Benabou and Tirole (2006) provide a unified framework in which prosocial behavior is determined by intrinsic motivation, extrinsic incentives and reputational (image) concerns.

Soraperra, Suvorov, van de Ven and Villeval (2019)

- One of the key quests in behavioral economics is the study of drivers of prosocial behavior.
- There are several key classes of explanations for prosocial behavior:
  - other-regarding preferences (altruism, inequity aversion, etc.);
  - warm glow;
  - reciprocity;
  - image concerns;
  - extrinsic incentives.
- In this paper we focus on image concerns and show (theoretically and experimentally) that they may have perverse effects in some circumstances.

This paper

This paper


CHEAP
RED
WINE


wine-searcher.com
1999
Henri Jayer
VOSNE-ROMANÉE
CROS-PARANTOUX
Henri Jayer
© File Photo

This paper

- Our goal is to examine if people are willing to make Pareto-damaging choices to preserve a good image.

This paper

- Our goal is to examine if people are willing to make Pareto-damaging choices to preserve a good image.
- And, if so, whether this happens for intrinsic or instrumental reasons.

This paper

- Our goal is to examine if people are willing to make Pareto-damaging choices to preserve a good image.
- And, if so, whether this happens for intrinsic or instrumental reasons.

Literature on Image concerns

- Bernheim (1994), Benabou and Tirole (2006), Andreoni and Bernheim (2009), etc: social image, status play important role in prosocial behavior.
- Benabou and Tirole (2006)
  - Image concerns are one of the key drivers of prosocial behavior: pro-social behavior allows people to signal that they are good.
  - Extrinsic incentives may mute the positive effect of image concerns.
- Similar effects in Janssen and Mendys-Kamphorst (2004), Seabright (2009), Andreoni and Bernheim (2009).
- Ariely, Bracha and Meier (2009) demonstrate the negative interaction of image concerns and extrinsic incentives in a lab experiment.

Image concerns: field experiments

- DellaVigna, List and Malmandier (2012) show that social pressure is an important determinant of charitable giving.
- Karing (2018) and Karing and Naguib (2018) show that social signaling opportunities help in promoting healthy behavior with substantial positive externalities: child vaccination in Sierra Leone and deworming in Kenia respectively.
- Funk (2010) shows that an introduction of voting by mail significantly reduced voting in small communities.
- DellaVigna et al. (2017) find that people who abstained from voting try to avoid answering surveys on this issue; the prospect of answering such a survey known before the voting may significantly increase turnout.
- **In these theories and empirical work image concerns push individuals to behave prosocially.**

Bad reputation

- Morris (2001)
  - In a repeated relationship even an advisor with preferences perfectly aligned with DM's may give a biased advice to improve reputation and differentiate himself from the opportunistic one ("political correctness").
- Ely and Välimäki (2003): reputational concerns of a long-term player may lead to a complete loss of surplus ("bad reputation").
- Grosskopf and Sarin (2010) test experimentally the "bad reputation" theory and fail to find support for it.
- Chung and Harbaugh (2017): build a model where transparency of the expert's incentives may affect informativeness of communication; they also show experimentally that "political correctness" deception by unbiased experts indeed emerges. They do not consider

Credence goods provision

- Big literature (surveyed in Dulleck and Kerschbamer, 2006). For the empirics see, in particular, Beck et al. (2014), Dulleck et al. (2011), Balafoutas et al. (2013).
- Main focus on overprovision. However, Schneider (2012) reports frequent undertreatment, but out of neglect rather than for strategic reasons; he finds no effect of reputational concerns.
- We have not found papers on underprovision for image-seeking reasons in the credence goods context.

Model: Buyer

- Needs a repair of type $s \in \{0, 1\}$; $s = 1$ corresponds to the major repair, $s = 0$ to the minor repair.
- It is common knowledge that a minor repair is needed with probability $q > 1/2$, a major one with probability $1 - q$.
- The buyer does not know which repair is needed.
- The buyer's utility is

$$U_B = a(1 - (r - s)^2).$$

- Type of repair $r \in \{0, 1\}$ is chosen by the seller; $r = 1$ corresponds to the major repair and $r = 0$ to the minor one.
- $a > 0$ is a parameter (one can set $a = 1$).

Model: Seller

- Knows the type of repair the buyer needs s; chooses r.
- The seller's utility is

$$U_S = br + \theta(1 - (r - s)^2) + \alpha\eta\hat{\theta}_r.$$

- θ reflects a combination of intrinsic motivation (professional pride) and altruism towards the buyer.
- η reflects the sensitivity of a particular seller to image concerns.
- It is common knowledge that θ and η are independent and distributed unformly on $[0, 1]$, but only the seller knows the true values.
- $\hat{\theta}_r$ is the buyer's equilibrium belief about the seller's intrinsic motivation θ conditional on type of repair r she receives.
- $b > 0$ and $\alpha > 0$ are commonly known parameters.

The seller's incentives

- When the seller knows a minor repair is needed, he tells the truth if

$$\theta + \alpha\eta\hat{\theta}_0 \geq b + \alpha\eta\hat{\theta}_1. \tag{1}$$

- For each θ denote by $\underline{\eta}(\theta)$ the value of η that makes the seller indifferent between recommending a major and a minor repair when a minor repair is needed (given equilibrium beliefs):

$$\underline{\eta}(\theta) = \frac{b - \theta}{\alpha(\hat{\theta}_0 - \hat{\theta}_1)}. \tag{2}$$

The seller's incentives

- When the seller knows a major repair is needed, he tells the truth if

$$b + \theta + \alpha\eta\hat{\theta}_1 \geq \alpha\eta\hat{\theta}_0. \tag{3}$$

- For each θ denote by $\overline{\eta}(\theta)$ the value of η that makes the seller indifferent between recommending a major and a minor repair when a major repair is needed (given equilibrium beliefs):

$$\overline{\eta}(\theta) = \frac{b + \theta}{\alpha(\hat{\theta}_0 - \hat{\theta}_1)}. \tag{4}$$

A minor repair secures better image

Lemma

In any perfect Bayesian equilibrium the seller gets a strictly better social image when he provides a minor repair: $\hat{\theta}_0 > \hat{\theta}_1$.

Proof.

Assume that $\hat{\theta}_0 = \hat{\theta}_1$ in equilibrium. Then, (1) and (3) imply that sellers with $\theta \geq b$ provide the type of repair that the buyer needs and sellers with $\theta < b$ always do a major repair. Then, clearly $\hat{\theta}_0 > \hat{\theta}_1$ contradicting our assumption. □

A minor repair secures better image

Proof.

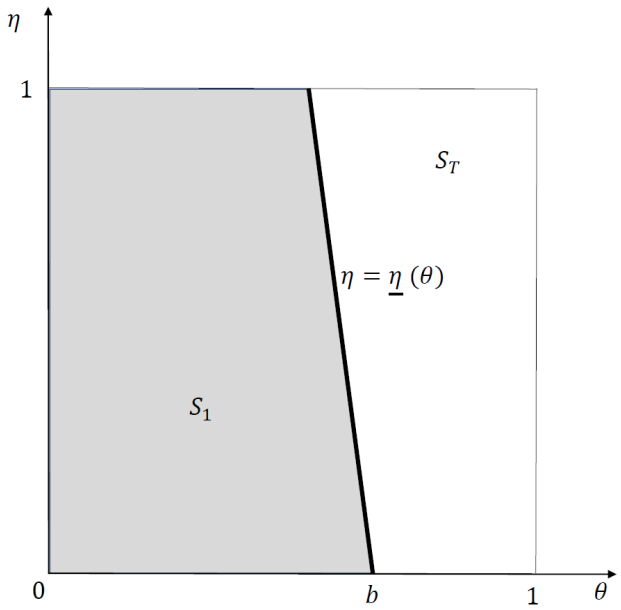
If $\hat{\theta}_1 > \hat{\theta}_0$ in equilibrium, then the incentive constraints imply that the types of seller with $\eta \leq \frac{\theta - b}{\alpha(\hat{\theta}_1 - \hat{\theta}_0)}$ (which is a set with a positive mass under our assumptions) provide the required type of repair, and the other types of seller always choose a major repair.

Then, using our assumption that θ and η are independent and uniformly distributed, it can be easily checked that this would imply $\hat{\theta}_1 < \hat{\theta}_0$ in contradiction to the assumption. □

Remark. Without assumption on independent and uniform distributions Lemma may not hold (counterexample in the paper). Uniform is unlikely to be crucial, though.

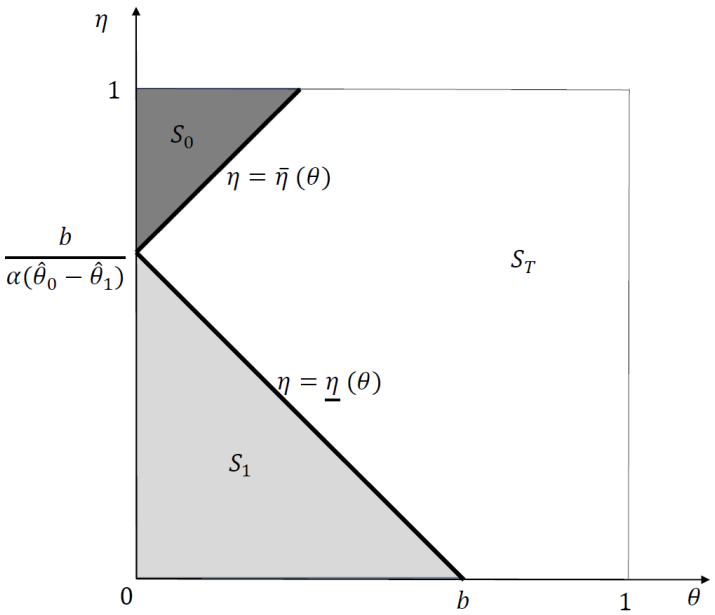
Equilibrium configurations

- Equilibrium with the low weight of image concerns ($\alpha < \overline{\alpha}$):



Equilibrium configurations

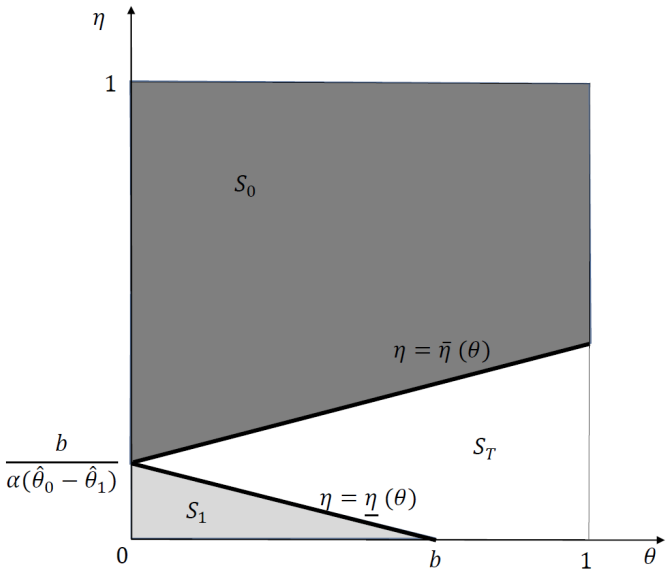
- Equilibrium with the high weight of image concerns ($\alpha > \bar{\alpha}$):



- Now there is another kind of seller opportunism: undertreatment by sellers with low intrinsic motivation and strong image concerns.

Equilibrium configurations

- Equilibrium with the high weight of image concerns ($\alpha > \bar{\alpha}$):



- Qualitatively similar to the previous case.

Equilibrium configurations

Theorem (Proposition)

There is always a PBE in the game. There exists threshold value $\bar{\alpha}$ such that

1. *if the weight of image concerns is weak enough ($\alpha < \bar{\alpha}$), in equilibrium sellers with $\eta < \underline{\eta}(\theta)$ offer a major repair regardless of the buyer's needs (area S_1 on Figure 1) and sellers with $\eta > \underline{\eta}(\theta)$ choose the kind of repair according to the buyer's true needs (area S_T on Figure 1);*
2. *if the weight of image concerns is strong enough ($\alpha \geq \bar{\alpha}$), in equilibrium sellers with $\eta < \underline{\eta}(\theta)$ offer a major repair regardless of the buyer's needs (area S_1 on Figures 2 and 3); sellers with $\underline{\eta}(\theta) \leq \eta \leq \overline{\eta}(\theta)$ choose the kind of repair according to the buyer's true needs (area S_T on Figures 2 and 3); sellers with $\eta > \overline{\eta}(\theta)$ offer a minor repair regardless of the buyer's needs*

Experimental design

- Participants randomly divided into sellers and buyers (roles are fixed).
- The buyer needs a minor repair with probability $\frac{5}{6}$ and a major repair with probability $\frac{1}{6}$.

Table 1: Payoffs

State of the world	Seller's choice	
	Minor repair	Major repair
Minor repair needed (prob. = 5/6)	(6,10)	(7, 5)
Major repair needed (prob. = 1/6)	(6, 5)	(7,10)

Notes: payoffs are denoted as (seller, buyer). The state of the world is determined by the roll of a die.

Terminology

- We say that a seller *makes a pro-social choice* if he provides a minor repair when the buyer needs it.
- We say that a seller *makes a Pareto-damaging choice* if he provides a minor repair when the buyer needs a major one.

Four treatments

- PRIVATE: buyer and seller remain anonymous; essentially, a dictator game.
- SOCIAL WEAK: buyer and seller can identify each other, but the buyer remains passive.
- SOCIAL STRONG: same as SOCIAL WEAK, but the buyer rates seller on scale 1 to 10 and in the end the seller publicly announces how many times he chose a major repair.
- REWARD: after each game the buyer divides 15 experimental currency units (ECUs) between herself and the seller.

Table 2: Overview of treatments

Treatment	Anonymity	Public announcement of seller's decisions	Buyer's task
PRIVATE	Yes	No	Passive
SOCIAL WEAK	No	No	Passive
SOCIAL STRONG	No	Yes	Rate seller
REWARD	No	No	Allocate money

Experimental procedures

- 8 sessions run in 2017 and 2019 in CREED, Amsterdam, as classroom experiments.
- 176 subjects (51% female).
- Subjects randomly rematched after every round.
- In 2017: 5 rounds in SOCIAL WEAK followed by 5 rounds in REWARD followed by 1 round in PRIVATE.
- In 2019: 3 rounds in SOCIAL WEAK followed by 5 rounds in SOCIAL STRONG followed by 1 round in PRIVATE.

Experimental procedures

- Strategy method was used (simpler logistics, more observations).
- In all treatments except PRIVATE subjects indicated if a matched participant was a friend or an acquaintance (highly reciprocal answers).
- Paper and pencil implementation.
- Sessions lasted 75-90 minutes.
- Average earnings 7.30 euros.

Hypotheses

Hypothesis 1: When a minor repair is needed, more sellers make pro-social choices in the REWARD condition and the SOCIAL conditions compared to the PRIVATE condition, and the effect is larger in SOCIAL STRONG than in SOCIAL WEAK.

Hypotheses

Hypothesis 1: When a minor repair is needed, more sellers make pro-social choices in the REWARD condition and the SOCIAL conditions compared to the PRIVATE condition, and the effect is larger in SOCIAL STRONG than in SOCIAL WEAK.

Hypothesis 2: When a major repair is needed, more sellers make Pareto-damaging choices in the REWARD condition and the SOCIAL conditions compared to the PRIVATE condition, and the effect is larger in SOCIAL STRONG than in SOCIAL WEAK.

Hypotheses

Hypothesis 1: When a minor repair is needed, more sellers make pro-social choices in the REWARD condition and the SOCIAL conditions compared to the PRIVATE condition, and the effect is larger in SOCIAL STRONG than in SOCIAL WEAK.

Hypothesis 2: When a major repair is needed, more sellers make Pareto-damaging choices in the REWARD condition and the SOCIAL conditions compared to the PRIVATE condition, and the effect is larger in SOCIAL STRONG than in SOCIAL WEAK.

Results: Seller's choices are informative about the seller's honesty

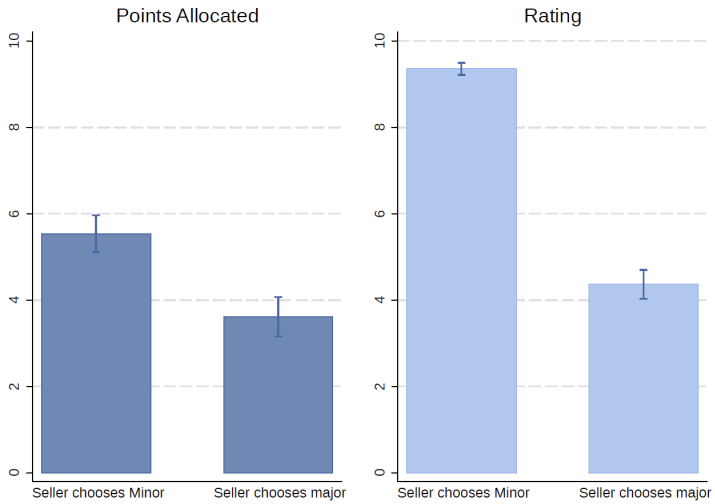
Table 3: Distribution of sellers' choices, by treatment and condition

Treatment	Seller's implemented choice	
	Minor repair	Major repair
Percentage of times minor needed		
SOCIAL WEAK	96	70
SOCIAL STRONG	97	60
REWARD	92	55

- The probability of the seller's honest behavior conditional on her choice of minor repair is above 90% in all treatments; it is below 50% in all treatments when she recommends a major repair.

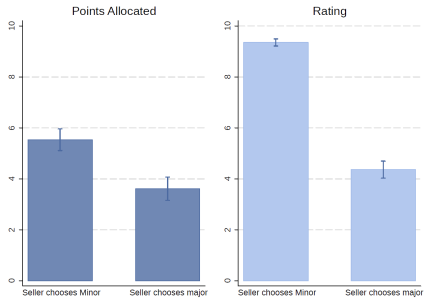
Results: Buyers appreciate minor repairs

Figure 4: Average number of points allocated by the buyers to the seller in the REWARD treatment and average rating of sellers in the SOCIAL STRONG treatment



Results: Buyers appreciate minor repairs

Figure 4: Average number of points allocated by the buyers to the seller in the REWARD treatment and average rating of sellers in the SOCIAL STRONG treatment



- In REWARD treatment, buyers allocate on average 1.92 fewer points to the seller after getting a major repair compared to a minor repair.
- In SOCIAL STRONG treatment, buyers rate sellers on average by 4.99 points lower after getting a major rather than a minor repair.
- Both differences are highly significant ($p < 0.001$, paired t-test).

Results: Strength of image concerns matters

Table 4: Fraction of pro-social and Pareto-damaging choices, by treatment.

Treatment	Pro-social choices	Pareto-damaging choices
PRIVATE	0.37	0.05
	(0.05)	(0.02)
SOCIAL WEAK	0.49	0.04
	(0.04)	(0.02)
SOCIAL STRONG	0.70	0.10
	(0.05)	(0.04)
REWARD	0.83	0.34
	(0.04)	(0.06)

Notes: s.e. in parentheses

- The frequency of pro-social choices increases with the strength of image concerns (all differences highly significant).
- The frequency of Pareto-damaging choices significantly increases only in the REWARD condition.

Results: Strength of image concerns matters

Table 5: Determinants of the sellers' probability to choose a minor repair.

	(1)	(2)	(3)	(4)
	Pro-social	Pareto-damaging	Pro-social	Pareto-damaging
Sample	All	All	Non-friends	Non-friends
SOCIAL WEAK (a)	0.119**	-0.001	0.069	0.002
	(0.048)	(0.024)	(0.049)	(0.025)
SOCIAL STRONG (b)	0.337***	0.054	0.317***	0.057
	(0.061)	(0.042)	(0.063)	(0.043)
REWARD	0.466***	0.295***	0.459***	0.317***
	(0.061)	(0.063)	(0.063)	(0.068)
Constant	0.368***	0.046**	0.368***	0.046**
	(0.052)	(0.023)	(0.052)	(0.023)
Wald Test (a)=(b) (p-value)	<0.001	0.135	<0.001	0.146
Observations	832	832	739	739
R-squared	0.112	0.127	0.126	0.138

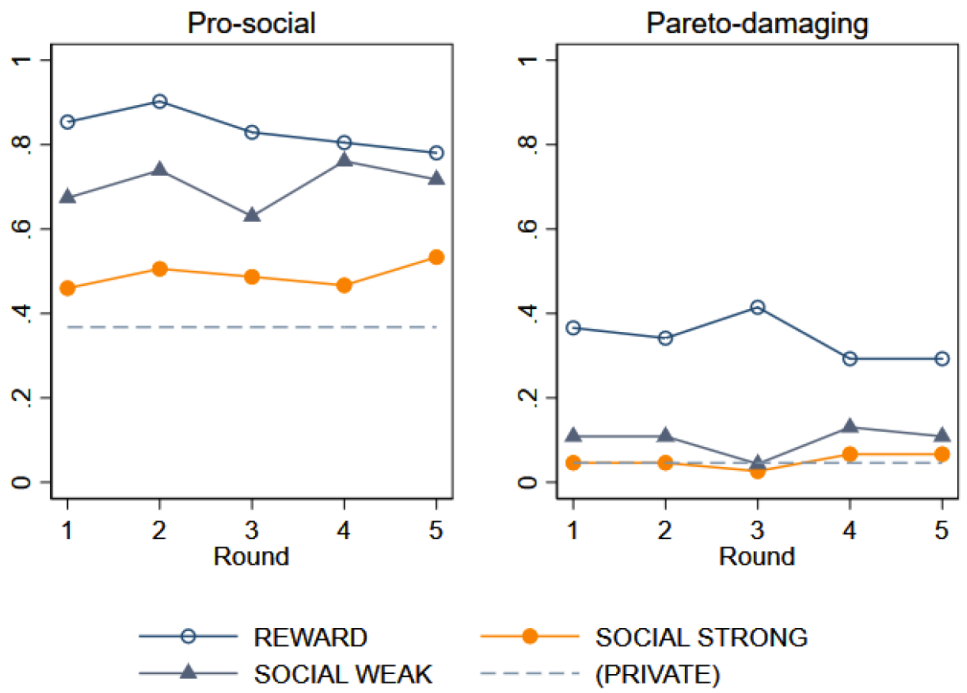
Notes: Standard errors (clustered at the seller and buyer level) in parentheses. ***p<0.01, ** p<0.05, * p<0.1. Excluded treatment is PRIVATE.

Results: Which sellers engage in Pareto-damaging?

- We call a seller *social* if he chooses the kind of repair the buyer needs in PRIVATE treatment. The fraction of social players is 36%.
- We call a seller *selfish* if he chooses a major repair in PRIVATE treatment regardless of the buyer's true needs. The fraction of selfish players is 60%.
- Selfish types engage in Pareto-damaging 38% of the times vs. 25% for the social types, but the difference is not significant.

Results: Little evidence of learning

Figure 5: Evolution of the fraction of pro-social and Pareto-damaging choices over time by treatment



Summary and conclusions

- We construct a buyer-seller model that captures "bad reputation" effects without any explicit dynamics.
- The key assumption is uncertainty about which kind of good/service the buyer needs (credence goods context).
- Heterogeneity in the strength of image concerns plays a key role:
  - only a fraction of agents get engaged in Pareto-damaging quest for image **and**
  - only if the strength of image concerns is large compared to intrinsic motivation for honest behavior.

Summary and conclusions

- Experimental results are in line with the model's predictions:
  - social pressure per se, while it increases pro-social honest behavior, does not generate Pareto-damaging quest for image;
  - once image concerns are reinforced further and get instrumental dimension, a fraction of sellers get involved in Pareto-damaging undertreatment for image-seeking reasons.
- Downward deception that we observe in our REWARD treatment has been elusive in experiments so far.
- Further extensions (e.g., explicit modeling of seller competition, possibly with buyers' search costs) seems a promising venue.