14.770-Fall 2017
Recitation 10 Notes

Arda Gitmez

December 8, 2017

Today: "Why incentives can backfire"

- Holmstrom and Milgrom's multitasking model (1991, JLEO)
- Two papers by Benabou and Tirole (2003, REStud and 2006, AER)

But first, some motivation: recall that in class, we argued that high pay for bureaucrats may be detrimental because it may attract the *wrong* pool of applicants. That's the *screening* story – today, I'll talk about an alternative theory based on *incentives*. (For more econ minded folks, this is getting closer to adverse selection vs moral hazard dichotomy.)

I'll talk about two strand of models:

- The famous "multitasking" model by Holmstrom and Milgrom. Basically, a higher monetary incentive can "shift" agent's efforts towards "wrong" activities. Consider teachers spending all their time on teaching kids how to excel at tests, and no social skills.
- The "intrinsic vs extrinsic motivation" models by Benabou and Tirole. Here, providing monetary incentivizes for an action may indeed reduce the supply of the action, because agents are concerned with looking like they do the action *just because they want to do it*, not because *they're paid to do it*. Many examples of this – see below.

The Multitasking Model

The multitasking model is a very famous one among theorists – it's also such a simple insight that sometimes it's difficult to appreciate how powerful it is. Below is a very very simplified version of the multitasking model:

- Two individuals: a principal and an agent. (Consider government and the teacher.)
- Two tasks: 1 and 2. (Consider teaching how to take tests and teaching social skills.)
- Agent chooses how much effort to spend on each task $(a_1, a_2) \in \mathbb{R}^2$
- Principal observes a single-dimensional measure of total effort $s = \mu_1 a_1 + \mu_2 a_2 + \epsilon$, $\mu_1, \mu_2 > 0$. (Consider a composite exam score – can also add noise.)
- Principal makes a transfer based on the signal: $t(s) = \alpha s + \beta$. (Performance-based pay.)
  - A higher α corresponds to "high-powered incentives".
  - We'll look for the optimal α chosen by the principal, i.e. the wage contract.
  - The linearity of contract is obviously restrictive, but there are justifications for it beyond the scope of this lecture.

Principal's payoff:

$$B_1 a_1 + B_2 a_2 - t(s)$$

where $B_1, B_2 > 0$ are the "benefits". (Consider the relative social benefits from having children who know how to take tests and children with high social skills.)

Agent's payoff:

$$t(s) - \frac{a_1^2}{2} - \frac{a_2^2}{2}$$

Where the $\frac{a_i^2}{2}$ term is the cost of action. Note that, due to the additive cost function, the two effort levels are *substitutes*.

As usual in the moral hazard literature, we'll do the analysis in two steps:

1. Given (α, β), what's the action that agents take? Clearly, looking at agent's payoff function and taking FOCs:
$$a_i(\alpha) = \mu_i \alpha \quad \text{for } i = 1, 2.$$
Since β is just a lump-sum transfer, it doesn't enter into this. Heuristically, the agent spends more effort on an action if it's more visible. (Plausibly μ_i is higher for testing skills than it is for social skills.)
2. Given $a_i(\alpha)$, what's the best α? The principal solves:
$$\max_{\alpha} B_1 a_1 + B_2 a_2 - (\mu_1 a_1 + \mu_2 a_2)$$
subject to
$$a_i = \mu_i \alpha \quad \text{for } i = 1, 2.$$

Taking FOCs and rearranging yields:

$$\alpha^* = \frac{B_1 \mu_1 + B_2 \mu_2}{\mu_1^2 + \mu_2^2} = \frac{||B||}{||\mu||} \cos\theta$$

α^* is high when (B_1, B_2) and (μ_1, μ_2) vectors are *congruent*. Heuristically, high-powered incentives are good when what principal cares about is also what principal measures. However, this is a very demanding assumption: the composite exam score does not necessarily reflect the socially optimal combination of test-taking skills and social skills! In that case, high-powered incentive would backfire: teachers would substitute away from teaching social skills and spend too much time on teaching test-taking skills!

Incentives and Prosocial Behavior

This is a long paper – I'll only cover the gist of it here. The basic idea is simple: agents have concerns for social reputation or self-respect, and consequently they want to *signal* that they are altruistic type. They can:

- Signal their type by working hard even if pay is low.
- Nevertheless, the signaling value of such an action decreases when individuals are paid to work hard, so we may end up with lower effort.

That is, *extrinsic incentives may crowd out intrinsic motivation.* Some examples of such crowding out:

- Gneezy and Rustichini (2000) found that schoolchildren collected less money when given performance incentives.
- Titmuss (1970) argued that paying blood donors could reduces the supply
- Gneezy and Rustichini (2000) "A Fine is a Price" found that fining parents for picking up their children late from day-care centres resulted in more late arrivals.

Here's a basic model which may explain such phenomenon. There is only one action a. Each individual selects a participation level a from choice set $A \subset \mathbb{R}$. Participation is not free: there is an effort cost $C(a)$. Assume that there is a material reward ya for choosing a.

Let ν_a and ν_y denote agent's intrinsic valuations for contributing to the social good and for money (greed), so contribution gives a benefit of

$$(\nu_a + \nu_y y) a - C(a)$$

Individuals' types $\mathbf{v} = (\nu_a, \nu_y) \in \mathbb{R}^2$ are drawn independently from $f(\mathbf{v})$ with mean $(\bar{\nu}_a, \bar{\nu}_y)$. As usual, types are private information.

In this model, decisions carry reputational costs and benefits, i.e. there is a value of *reputation*. Such value may be instrumental (e.g. marriage market) or affective (shame as hedonic good) – we'll remain agnostic about it. Assume the following functional form for reputational benefits:

$$r(a, y) = \mu_a \mathbb{E}(\nu_a \mid a, y) - \mu_y \mathbb{E}(\nu_y \mid a, y), \ \mu_a \geq 0, \mu_y \geq 0$$

so that the decision problem of agent is:

$$\max_{a \in A} \{(\nu_a + \nu_y y) a - C(a) + \mu_a \mathbb{E}(\nu_a \mid a, y) - \mu_y \mathbb{E}(\nu_y \mid a, y)\}$$

The FOC gives us

$$C'(a) = \nu_a + \nu_y y + \mu_a \frac{\partial E(\nu_a \mid a, y)}{\partial a} - \mu_y \frac{\partial E(\nu_y \mid a, y)}{\partial a}$$

Just through eyeballing, you can see that optimal a reveals three motivations: *intrinsic, extrinsic* and *reputational*.

The reputation is hardwired into this equation because individuals try to infer ν_a and ν_y from the actions taken. Here's a critical observation: A higher incentive rate y reduces informativeness of actions about ν_a, but increases it about ν_y. In other words, stronger incentives may reduce effort because high effort in this case makes you look greedy.

Here's a graph which captures it all:

Copyright Roland Bénabou, Jean Tirole, and the American Economic Association reproduced with permission of the American Economic Review.

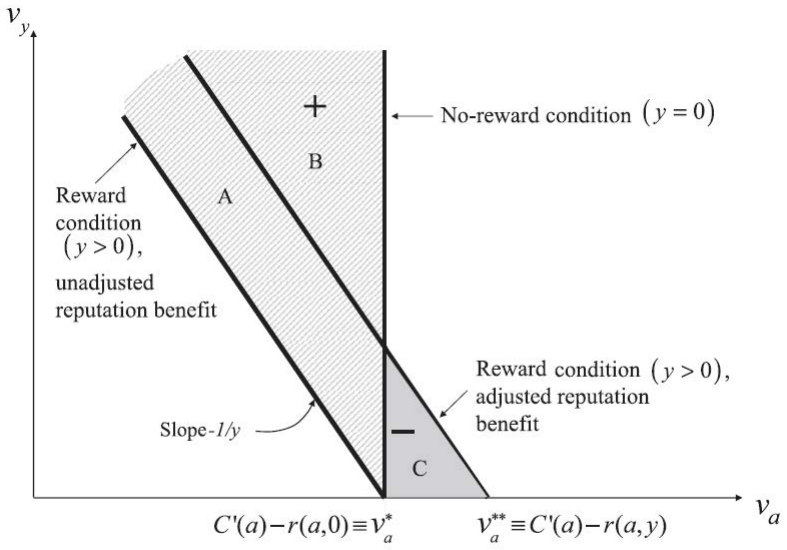


FIGURE 1. THE EFFECTS OF REWARDS ON THE POOL OF PARTICIPANTS

There are two reputational effects in equilibrium in this graph:

- New high-contributors have lower ν_a's than old ones – they drag down the group's reputation for prosocial orientation.
- New high-contributors are "greedy" types whereas those who still contribute below a after the reward is introduced reveal that they case less about money than average.

The net effect of a financial reward therefore is ambiguous: supply curve of contributions can be locally downward or upward-sloping.

Here's a more formal analysis:

- Assume $C(a) = \frac{ka^2}{2}$.
- Valuations ν_a, ν_y distributed normally with covariance σ_{ay}

In this case, standard results for normal random variables yield

$$E(\nu_a \mid a, y) = \bar{\nu}_a + \rho(y) \cdot (ka - \bar{\nu}_a - \bar{\nu}_y y - r(a, y))$$

$$E(\nu_y \mid a, y) = \bar{\nu}_y + \chi(y) \cdot (ka - \bar{\nu}_a - \bar{\nu}_y y - r(a, y))$$

where

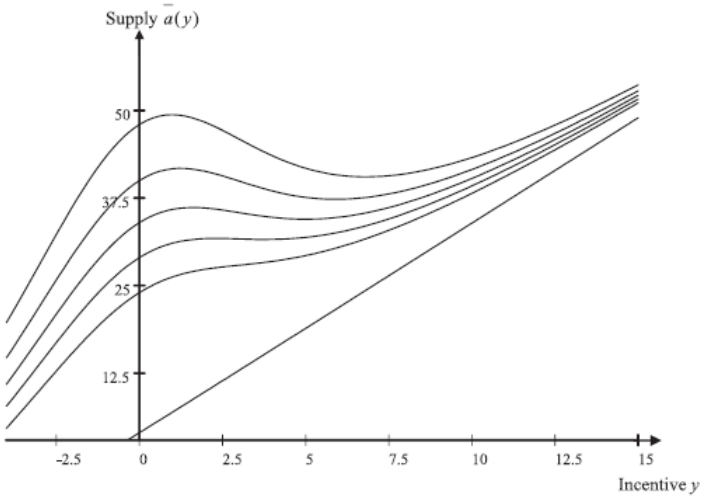
$$\rho(y) = \frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2},\ \chi(y) = \frac{1 - \rho(y)}{y}$$

Note that $\rho(0) = 1$: when there are no financial rewards, actions are fully informative about ν_a.

Proposition 1. *There is a unique (differentiable-reputation) equilibrium, in which an agent with preferences (ν_a, ν_y) contributes at the level*

$$a = \frac{\nu_a + \nu_y y}{k} + \mu_a \rho(y) - \mu_y \chi(y)$$

In this case, higher y increases agents' direct payoff from contributing but reduces associated signaling value along both dimensions. Incentives indeed backfire over some range (provided sufficient reputational concern), see:



A. Varying $\bar{\mu}_a$ (with $\bar{\mu}_y = 0$). The straight line corresponds to $\bar{\mu}_a = 0$ (no reputation concern).

Intrinsic and Extrinsic Motivation

Let me just tell you the story of the model, and you can read about it if you want to.

We have the 14.770 Final Exam in $\sim$ 10 days from now on. You'll choose how much to study for the final; and clearly both you and me want you to do well in the final.

To be honest, I've seen (at least part of) the final exam by now, so I know how difficult it would be. So how would you feel if I suddenly came up, picked one of you and told "I'll pay you $ 100 if you take a 90 or above from the final"?

- On the one hand, this incentivizes you to work harder for final for obvious reasons.
- On the other hand, this statement says something about the *difficulty* of the final and my perception about the person I picked – it plausibly tells you that I know the final exam is difficult, and that person needs some extra incentives to work enough for the final. But this may discourage him from studying as well!

Which effect dominates? Depends, but it's easy to come up with scenarios where offering additional monetary incentives can backfire.

Disclaimer: I'll not pay anything to any of you, regardless of your performance in the final exam.

How do you feel now? (Which is how you finish a semester. BA-DUM-TSS.)

MIT OpenCourseWare
https://ocw.mit.edu

14.770 Introduction to Political Economy

Fall 2017

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.