14.11 Spring 2006: Putting Social Science to the Test. Lecture Note #3: Incentives and Economic Behavior

David Autor

March 7, 2006

1 Introduction

Homo Economicus is the mythical self-interested, fully rational utility maximizer who inhabits all standard economic models. At the core of this model is the notion that such actors respond optimally to incentives. This assumption is not uncontroversial outside and (although you may not be aware of this) inside of economics. In this lecture, we'll look at some experimental evidence on how incentives shape behavior in market and non-market settings.

We'll begin with one fascinating example. The paper by Gneezy and Rustichini provides experimental evidence on how the existence of explicit incentives may alter social norms.

We'll next move into a canonical topic in the economics of incentives: how do incentives affect behavior in an environment where 'agents' have multiple methods to produce 'results' but only some of these methods are genuinely productive. This is the so-called 'multi-tasking' problem, and it is seen in the Glewwe, Ilias and Kremer paper.

Next, we'll consider the paper by Fehr and Rockenbach, which studies how the availability of explicit punishments affects cooperative behavior.

Finally, the Fehr and Gächter paper on public goods experiments considers whether social norms trump economic incentives in a setting where social norms are costly to enforce and have no direct private payoff. (There are also two other papers by Fehr and coauthors on this week's reading list that we will not have time to discuss in class. These articles are not difficult to follow and they are highly recommended for inspirational reading.)

This lecture also provides an opportunity to consider a core topic in experimental design: statistical power. Statistical power is the ability of an experiment to reject the null hypothesis when the null is false. There are many experiments that 'fail' simply because they do not have the power to provide a definitive statistical answer to the hypothesis they set out to test. For example, if I perform an educational intervention that, on average, reduces the odds of school dropout by 7 percentage points with a 95 percent confidence interval of plus or minus 8 percentage points, then I haven't learned anything useful. I would conclude that my experiment might have had a large effect (7 percentage points is probably large relative to baseline) but I cannot reject the possibility that it had zero effect. Assessing the statistical power of an experimental design prior to conducting the experiment is a critical step to avoiding these heartbreaking disappointments. You must understand this topic to develop your own research designs (your design proposals will need to include power calculations).

In addition to the materials discussed in class on experimental design, we highly recommend the free *Optimal Design* software—and excellent companion manual—available from the University of Michigan at *http://sitemaker.umich.edu/group-based/optimal_ design_ software*. This tool does the mechanics of power calculations for you, though this does not substitute for a firm understanding of the underlying principles of experimental design. Most of the material on power calculations in randomized evaluations will be made available as a set of PowerPoint slides attached to this lecture note.

2 Incentives and Social Norms: Gneezy and Rustichini

Any standard economic model will imply that attaching a fine to an activity will reduce the amount of that activity, all else equal; this is the principle of deterrence. The paper by Gneezy and Rustichini provides an enlightening example where that principle utterly fails.

A group of private Israeli day-care centers were troubled by the frequency at which parents arrived after the 4pm closing time to pick up their children (despite the contracted pickup time of 4pm). Although day-care staff frowned on this behavior, there was no specific sanction for tardiness. The experimental manipulation was, at 6 of 10 day-care centers, to fine parents for being late. The procedure was:

- In the first 4 weeks, collect baseline data—no manipulation
- In weeks 5 through 16, a fine was announced and parents were charged approximately \$3.50 for a delay of 10 minutes or more. (See text of fine in slides.)
- In weeks 17 through 20, the fine was removed without explanation.

The main results are visible in Figure 1:

- In the first five weeks in which the fine was in place, late arrivals approximately doubled at the treatment centers—that is, the fine induced more late arrivals.
- In week seven forward, late arrivals equilibrate at slightly less than twice the initial level.

- After week 17, when the fine was removed, late arrivals *did not* return to normal.

Thus, to summarize: tardiness increased with the fine but did not fall after the fine was removed. These results are not consistent with any standard economic model. Even if 'demand' for tardiness were upward sloping so that a price increased raised tardiness, tardiness should have then fallen when the fine was eliminated.

The authors propose two potential explanations:

1. **Incomplete contracts.** The formal day-care contract did not specify penalties for coming late, but it was implicitly understood that this was an imposition on the staff and should only occur in emergencies. Because the consequences for violating the contract were not specified, parents did not know how much they stood to be punished for tardiness (i.e., would there children be asked to leave the day-care?). The fine revealed to parents these consequences were very mild. Once the consequences were known, many parents decided they were happy to bear them.

2. **Social norms.** Prior to the fine, parents implicitly understood that arriving late was 'taking advantage' of teachers' generosity, and so refrained from doing so in compliance with a social norm. The fine turned this informal social interaction into a market transaction. The fine conveyed to the parents that they could 'buy' the right to be late, arguably removing the onus from the action. Why did behavior not return to 'normal' when the fine was removed? Perhaps the perception of the social norm was altered by the experience of the fine. Parents continued to think of 'lateness' as an additional service rather than a violation of terms. Quoting the paper:

   > The conclusion seems to be that the absence of a price is not just the limit of very low prices. Mentioning a payment is enough to change the perception of the contract: from a service, which is due from them as subjects in the experiment, to a market exchange. In a similar manner, in the day-care study a fine is enough to change the perception of the obligation to arrive on time.

This paper suggests that *explicit* incentives exist against a backdrop of *implicit* rules and incentives (typically called social norms). By introducing explicit incentives into this milieu, one may inadvertently ‘crowd-out’ or undermine implicit incentives. We’ll see more findings of this flavor in the papers by Fehr. This body of important work builds a case that much of day-to-day human behavior deviates widely from the *Homo Economicus* model.

3 Incentives and Multi-Tasking: Glewwe, Ilias an Kremer (2003)

In Kenya, teacher incentives are extremely weak. Teachers are highly unionized and cannot readily be fired. They earn high salaries (five times GDP) but there is little opportunity for performance to affect promotion or salary. Perhaps partly as a result, teacher absenteeism averages **20 percent** (i.e., 1 missed day per week). Even when teachers are present, it is not clear how much time they spend on classroom activities versus alternative uses of time within the school, e.g., socializing and other non-teaching activities. (An earlier study found that although teachers are absent 20 percent of days from school, they are absent 45 percent of days from classrooms.) This problem is not unique to Kenya. High teacher absence rates are a significant problem in many developing country school systems.

To the extent that incentives exist in Kenyan primary schools, they are based on the results of the national primary school leaving exam (KCPE). The results of this exam are front-page news in Kenya. Every school’s performance is publicly available. KCPE exam scores determine which students will enter which secondary schools (if any). Teachers *do care* about the results of these tests and they devote significant attention to preparing students for the exams. In addition to regular classroom teaching, this may include holding ‘test preps’ outside of class hours as well as devoting class time specifically to test preparation.

3.1 Experimental design

The experiment in this paper takes place in Western Kenya. In 1997 - 1999, fifty randomly selected schools were provided with fairly powerful incentives to improve scores. These schools could win prizes—paid directly to teachers—based on each school’s average performance on exams in grades 4 through 8. All teachers who taught these grades were eligible for a prize. Some details:

- Two categories of prizes: Top-scoring schools and most-improved schools.

- In each category, 3 first prizes, 3 second prizes, 3 third prizes, 3 fourth prizes. Hence, 24 of 50 schools could win a prize.

- Prizes ranged from 21 to 43 percent of typical teacher monthly salaries. So, to fix ideas, for a U.S. teacher earning \$60K per year, this would be a ~\$2,000 prize.

- Prizes based on equally weighted-average of 4th through 8th grade scores; every teacher won simultaneously in school.

- Typical school had 200 kids, 12 teachers. This number of teachers per school provides some opportunity for 'free riding,' but perhaps groups are small enough to provide scope for social pressures from other teachers to limit shirking.

- In order to discourage teachers from encouraging weaker students to repeat, drop out, or not take the exam, students who did not take the exam were assigned very low scores (lower than would be achieved by simply guessing at every question). This design averts an otherwise likely form of gaming.

3.2 What does theory predict?

The experiment creates strong incentives for raising test scores. However, it is not necessarily test scores that we most want to maximize. Other outcomes we might care about include: knowledge acquisition, general non-test learning, knowledge retention (things that show up on the exam may not be reflected outside of the exam.) We might also care about actual teacher inputs, such as attendance.

How will teachers respond to these incentives? Suppose that test scores are a function of student learning and teacher 'signaling' effort, which may include activities that only affect test scores but do not affect learning:

$$T_t = L\left(e_t, e_{t-1}, e_{t-2}...\right) + \gamma\left(s_t\right) + \varepsilon_t,$$

where T is the test scores of students in period t, L denotes student learning and, e_t denotes teacher effort on *long-run* learning in period t, s_t is signaling effort, and ε_t is a random shock.

Assume that teachers' utility is given by

$$U = M - C\left(e, s\right),$$

where M is teacher pay and $C\left(\cdot\right)$ is a utility cost that depends on teaching and signaling effort. It is important to note that e and s may be either substitutes or complements. They will be substitutes if they are mutually exclusive activities and teachers face a time constraint. On the other hand, if teachers face a fixed cost of attending school at all, then these activities may be complements (i.e., if the teacher shows up, it is easy to produce both).

Suppose that teacher pay is given by

$$M_t = \alpha + \beta T_t.$$

If $\beta = 0$, so that teacher pay is independent of performance, teachers will choose e and s so that the marginal product of each is zero. This doesn't necessarily mean that teaching effort will be zero. Teachers may enjoy teaching for its own sake, may care about students, or may want to uphold social norms; thus, exerting no effort may have *negative* utility cost. However, teachers will not optimally exert effort beyond the point where its marginal cost in utility terms is greater than zero.

Our experimental manipulation involves raising $\beta \gg 0$, so now teachers face an incentive to exert positive (i.e., costly) effort. Specifically, a utility maximizing teacher will choose:

$$\max_{e,s} \; U\left(M, e, s\right) = \alpha + \beta\left[L\left(e_t, e_{t-1}, e_{t-2}...\right) + \gamma\left(s_t\right)\right] - C\left(e, s\right),$$

with first order conditions:

$$\frac{\partial L}{\partial e}\beta = \frac{\partial C\left(e, s\right)}{\partial e} \text{ and } \frac{\partial \gamma}{\partial s}\beta = \frac{\partial C\left(e, s\right)}{\partial s}.$$

Notice that if e and s are complements in the cost function ($C_{es} < 0$) or if utility is additively separable ($C_{es} = 0$), then both types of effort will rise. If these activities are substitutes, one could rise at the expense of the other. For example, teachers might choose to stop *all* productive learning to instead focus on test-score signaling. Thus, although effort spent on either learning *or* signaling must rise, it is not necessarily the case that both must rise and it is plausible that one or the other will actually fall.

This model points out the possible pitfalls of 'paying for A when you are hoping to get B.' That is, if we are hoping to maximize learning but we are paying for test scores, we might only get test scores without learning (or not—that's only a possibility). This point is made formally in a famous paper by Holmstrom and Milgrom in 1991.

How can the authors of this study distinguish this possibility?

1. One is to study the incidence of absenteeism and test 'preps.' It's possible that reduced absenteeism reflects higher effort on learning whereas more preps reflect greater effort on signalling.

2. One can also look at the persistence of any test score gains. Gains that are not retained are likely to reflect primarily signaling.

3. One could also check if gains are concentrated in subjects in which memorization is most important.

Of course, parents may be happy with test score gains, whatever their provenance. However, if gains are purely due to gaming of the tests—not human capital—these gains will not increase student knowledge or productivity. In the long, the society does not benefit, and the student may not benefit either.

3.3 What actually happened?

The initial news is good:

1. Exam participation rises (see Table 8)

2. Test scores rise (see Table 9)

But there is much to suggest that these gains are not what they seem.

1. Teacher attendance did not increase, which leads one to wonder how much new learning could have gone on. (see Table 3)

2. Homework assignments did not increase (see Table 4).

3. Dropouts and grade repetition did not improve (see Table 7)—which actually puts a somewhat negative spin on the increase in test participation; teachers may have encouraged test participation without encouraging students to pursue education.

4. There are no changes in pedagogy except for more test prep sessions (see Tables 5 and 6)

5. Perhaps most damningly, the exam gains these students make in the two years of the program completely evaporate in year 3 when the program ends (again see Table 9). This suggest that whatever mechanisms teachers used to raise scores, this method did not produce durable gains in learning.

3.4 Interpretation

This paper shows the bright and dark sides of the efficacy of incentives to produce results. The intervention did raise test scores. However, it's generally extremely difficult for a principal to design an evaluation metric that exactly captures the behavior she wishes to 'incentivize' among agents. Normally, there is some scope for agents to game the measures so that she is rewarded for outcomes without necessarily producing desired outputs. This paper provides a rather extreme example of this type of distortion: teachers did raise test scores but they did not appear to improve learning, pedagogy or even attendance. Thus, the 'incentive' works in the narrowest sense and fails in any meaningful sense. The program paid for A and obtained A, but the program designers may have been wishing for B instead.

4 The Effect of Sanctions on Altruism: Fehr and Rockenbach (2003)

The setting of this experiment is a classic anonymous 'trust game' treatment. Both an 'investor' and a 'trustee' are given 10 monetary units, which we'll call dollars. The 'investor' can keep the $10 or give any share to the anonymous 'trustee.' Whatever amount the investor gives to the trustee is tripled by the experimenter. Hence, if the investor gives $10, the trustee receives $30. Along with his transfer, the investor must also specify the 'desired back-transfer,' which is the money he'd like in return. This desired level can be anywhere between $0 and three times the amount invested. Finally, the trustee decides whether and how much money to return to the investor.

The key features of this game are:

1. Both parties can be made strictly better off by the investor giving any positive amount to the trustee and the trustee returning more than one-third of that amount.

2. The trustee has an incentive to keep any amount given and return none. A purely income-maximizing trustee will always do this.

3. Anticipating this fact, an investor may choose not to give money.

4. In the unique subgame perfect Nash equilibrium of this game, the investor gives no money.

Thus, the game presents a sequential prisoners' dilemma in which both parties are better off if both cooperate than if neither cooperates, but if one cooperates, the other is *even better off* not to cooperate.

Much experimentation has found that many subjects do *not* play the unique equilibrium of the game. Typically, about 40 percent of subjects engage in 'trust' behavior, despite the fact that there is no opportunity for sanction or for withholding of future cooperation in future interactions (and remember that the identities of participants are anonymous).

The innovation of this study is to consider how behavior changes when explicit sanctions ('incentives') are added to the environment. The 'incentive' condition is identical to the one above except that investors have an option to impose a fine of $4 on the trustee if less than the desired amount is sent back. The investor does not have to impose a fine, however. He can stipulate both the amount of the back-transfer desired *and* the fine that will be imposed (which may include $0) if less than that amount is provided. Hence, the trustee knows the investor's expectation and the sanction that will follow for not meeting this expectation.

To avoid evocative language the experimental instructions did not include value-laden terms like fine or punishment. Instead, the fine was described, more neutrally, as a deduction from the trustee's payoff.

4.1 Results

Figure 1 and Table 1 tell almost the whole story.

- In both treatment conditions, the amount of the back-transfer is rising in the investor's initial transfer. This shows that on average (though not necessarily in the modal case), there is spontaneous cooperation. [The authors call this altruism, but I'm not so clear that this is the right term.]

- Under the incentive condition where a *fine is imposed*, transfers are on average *lower* than in the non-incentive condition.

- Under the incentive condition where a *fine is not imposed*, transfers are on average *higher* than in the non-incentive condition.

A glance at Table 1 shows that investors who chose *not* to use incentives actually made larger investments but asked for a slightly smaller percentage in return than did investors who invested with a fine. However, the back-transfer as a percentage of transfer was highest in the case when investors chose *not* to use incentives.

Thus, is appears that explicit incentives may *undermine* voluntary cooperation in this setting. Why? One further clue about what is going on is found in Figure 2. Here, the authors categorize the desired back-transfers by whether or not they would cause the investor to earn more ('high request') or less ('low request') than the trustee. In two of three conditions (no incentive possible; fine imposed), trustees appear to not react positively to 'high' requests—that is, their percentage returned of the transfer falls relative to the low request (though perhaps not their absolute return; this info is not given). In the third case, however, where a fine is not imposed, trustees increase their percentage returned of the transfer—that is, they appear to comply (at least somewhat) with the request. This suggest that they may be reciprocating a 'fair' request with a 'fair' response.

4.2 Discussion

This paper exhibits another subtle feature of incentives. In the Gneezy and Rustichini paper, market prices appeared to displace social norms. In this study, explicit incentives appear to undermine cooperation. The explanation that the authors favor hinges on perceptions of fairness. Many people appear to have an intrinsic sense of fairness, altruism or desire to cooperate that causes them to respond to generous acts generously—and to initiate spontaneous generosity—even where there is no direct payoff to generosity. Explicit incentives may undermine these responses if participants view the incentives as an indication of distrust. By contrast, declining to use explicit incentives when they are available may be seen as a signal of extra-trust (or special 'fairness'). This may cause cooperative people to reciprocate in kind.

These ideas are informal and are not fleshed out in the article. If interested, you should consult the large sociological and influential economics literature (that starts with Nobel Laureate George Akerlof) on the cultural phenomenon of 'gift exchange.' Gift exchange is a setting where parties provide valuable up-front 'gifts' to recipients who have no explicit incentive to reciprocate—and yet recipients often do so. Akerlof argues that gift exchange is an important component of many market interactions, most especially in labor markets (where, he theorizes, that employers typically pay workers above their reservation (outside) wages and workers reciprocate by working harder than absolutely necessary to avoid being fired). The Fehr and Rockenbach results are consistent with gift exchange.

5 Another Side of Sanctions: Fehr and Gächter (2000)

The prior papers might lead you to believe that economic incentives always have perverse effects. This is very far from the truth. While it's hard to make a rigorous statement of this fact, it's probably the case that incentives work as intended in most market settings—and this is why markets appear to work so well. However, the 2000 paper by Fehr and Gächter demonstrate an interesting case where sanctions work well even where they arguably should *not*.

The setting is a public goods experiment. This is stratified 2 x 2 random assignment design with two pairing mechanisms and two experimental manipulations:

- Pairing: In the 'stranger' treatment, groups of size 24 are randomly reallocated into 6 subgroups of 4 in each period for 10 successive periods. In the 'partner' treatment, they are randomly allocated into groups of 4 that are then stable for 10 periods.

- Manipulations: In one set of treatments, participants had the opportunity to 'punish' one another; in another treatment, they did not.

The setting is a 'public goods' game. Participants are given an endowment of y tokens which they can either keep for themselves or invest some subset (including all) into a project g. These investment decisions are made simultaneously, so there is no opportunity for coordination. Participants are anonymous. Each investment in g benefits all group members in net by more than the cost to the investor, but the net benefit to the investor of his own contribution to g is negative.

Specifically, the payoff for individual i is:

$$\pi_i = y_i + a \cdot \sum_{j=1}^{n} g_j,$$

where $0 < a < 1 < n \cdot a$. Thus, total payoffs are rising in g since $n \cdot a > 1$, but individual payoffs are declining in g_i since $\partial \pi_i / \partial g_i = -1 + a < 0$. It is again easy to see that the unique subgame perfect Nash equilibrium of this game is that no player contributes to the public good.

The punishment condition adds a twist to this environment. After the first-stage decisions are announced (how much each subject has contributed to g), subjects are offered an opportunity to (simultaneously) punish one-another by assigning punishment points. Each punishment point assigned to another player reduces his payoff (π) by 10 percent. The cost to the punisher of assigning points is an increasing, convex function of the number of points selected (see Table 2).

Notice that the standard logic of backward induction says that no one will punish in this game. Clearly, in the last period, there is no incentive for one player to punish another since there can be no further cooperation in earlier rounds. Recognizing this, there will be no cooperation in the final round. However, if there is going to be no cooperation in the final round, then the question is whether there will be cooperation in the second-to-final round. Clearly not, because there will be no further opportunities for cooperation at that point (since we no there is no cooperation in the final round). Repeating this logic 10 times, we arrive at the prediction that there is no cooperation in any round. This prediction looks more plausible in the 'stranger' treatment (where all interactions are one-shot) than in the 'partner' treatment (where there are 10 rounds).

5.1 What happened?

Main results:

- Existence of punishment opportunities causes a large rise in the average group contribution level in both the stranger and partner treatments.
- Notably group contributions are considerably higher in the partner condition.
- But both partner and stranger conditions tend to converge towards full 'free riding' (no public goods) over the course of 10 periods. This suggests that the basic economic model of selfish maximizing behavior holds if there are no punishments.
- A considerable amount of punishment actually occurs in this game, despite the fact that punishments should rarely or never occur according to the standard theory—esp. in the stranger treatment. As shown in Figure 5, subjects who deviate from the group average are likely to get punished. And the larger the deviation, the larger the expected punishment.

What are the consequences of punishment for 'efficiency'—that is, the total payoff of the group? Here's what the authors say:

- In the Partner treatment, in particular, contributions are lower in the early periods of the punishment condition than during the later periods, and this caused much more punishment activities in the early periods.
- Contributions gradually decline over time in the no-punishment condition.
- Taken together, the results suggest that the presence of punishment opportunities eventually leads to pecuniary efficiency gains. To achieve these gains, however, it is necessary to establish the full credibility of the punishment threat by actual punishments.

5.2 Interpretation

> "This paper provides evidence that spontaneous and uncoordinated punishment activities give rise to heavy punishment of free-riders. In the Stranger-treatment this punishment occurs, although it is costly and provides no future private benefits for the punishers. The more an individual negatively deviates from the contributions of the other group members, the heavier the punishment."

Thus, cooperation appears to be spontaneous and here it is *complemented* rather than substituted (as in Fehr and Rockenbach) by the existence of explicit incentives. Two parting thoughts:

1. Why do punishments increase cooperation here and reduce them in the previous case? It's possible to interpret both studies through a lens of public spiritedness vs. selfishness. In the current study, subjects are punished for failure to contribute to a *public good.* Thus, agents who engage in punishment are not acting selfishly—in fact, they are contributing a public good by punishing since their payoffs are reduced. Administering punishments here can be thought of as enhancing cooperation. By contrast, in the prior study, punishments were arguably only used by agents to achieve selfish goals, i.e., to increase private earnings.

2. There is also an irony here. Incentives do work to generate public goods contributions in this setting, which is *sort of* what the theory would predict. More precisely, the theory predicts that if punishments are administered (or credible), then public goods contributions will rise. But the theory also predicts that punishments *will not* be used.

6 Wrap-up

Thus ends our brief tour of incentives. What should you conclude from these papers? A theme that much recent economic analysis emphasizes (and something I will also discuss in later lectures) is that there are already many constraints on human behavior operative that do not look like individual maximization. That is, many agents appear to be complying with norms of fairness and cooperation in the absence of explicit incentives (or perhaps these norms provide strong incentives but we don't know how to adequately account for them). The insertion of economic incentives in such an environment may work exactly as neoclassical theory predicts or it may instead generate perverse consequences (or, as in the public goods experiment, unexpected *positive* consequences) by interacting strongly with the implicit constraints and incentives already in place. Understanding these implicit incentives and constraints may help to interpret much human behavior that economic models do not seem to accommodate.

On the other hand, there is a counter-argument to the view that 'fairness' and 'social preferences' are important to economic behavior. We will see this counter-argument in Lecture V. The main critique of the work above questions the external validity of these laboratory experiments on fairness and cooperation.