Social Beliefs and Social Norms:

II - Incentives, Social Norms and Social Learning

Roland Bénabou

Princeton University

Based in large part on joint work with Jean Tirole

Behavioral Economics Summer School - Louvain - May 2017

Background papers

1. Incentives and Norms (unidim. heterogeneity / signaling)
   - Bénabou, Roland and Jean Tirole "Laws and Norms," NBER. (2011)
2. Social Norms and Social Learning (multidim. signaling)
   - Bénabou, Roland and Jean Tirole "Incentives and Prosocial Behavior," *American Economic Review,* 96(5), 1652-1678
   - S. Nageeb Ali and Roland Bénabou "Image Versus Information: Changing Societal Norms and Optimal Privacy" NBER (2016)

INTRODUCTION

- People's behavior is shaped by their preferences, by explicit incentives (e.g., the law, contracts) and by social norms and informal enforcement (reputation, honor / stigma, etc.)
- These different channels aspects usually studied separately
  - Economists emphasize incentives, norms studied separately
  - Psychologists, sociologists, often skeptical of incentives.
    Fear "crowding out," emphasize persuasion, "norms-based interventions"
- Law scholars somewhere in-between: law is a set of incentives, but also reflects, conveys and adapts to the values of society
- Laws, norms interact, shape each other: need to model together
  - When do incentives undermine or strengthen social norms?
  - Optimal setting of incentives

Example of incentive puzzles: voting

- Panagopoulos (2009): Paying people to vote, from \$2 to \$25, had no significant effect on their turnout.
- Gerber et al. (2008): informing people of who among their neighbors votes, and vice-versa, had large significant effect (30% $\rightarrow$ 38%)
- Funk (2007): removing "mandatory voting" laws (in Swiss cantons) had no effect on turnout where law involved no fine, but negative where a fine of "symbolic" amount ($\approx$ 1 Euro) was involved.

OUTLINE

1. Model combining formal + social incentives
2. The calculus of honor and stigma $\Rightarrow$ social multiplier $\gtrless$ 1
   - Empirical Evidence
3. Optimal incentives with social norms $\Rightarrow$ modified Pigou-Ramsey
4. Persuasion and norms-based interventions $\Rightarrow$ credibility
5. The expressive content of law $\Rightarrow$ informational multiplier
   - Empirical Evidence
6. Models with Multidimensional Heterogeneity / Social Learning

I. BASIC MODEL

Actions

- Agents (one or many) choose action a at cost $C(a)$:
  effort, time, resources.
  - Private-goods context: effort in the firm, non-opportunism...
  - Public-goods context: volunteering, voting, giving blood, helping, contributing to a good cause, not polluting...
- Incentive: receive y per unit of a, from some principal
  - Private-goods context: wage for effort, performance-contingent bonus, penalty for failure, etc.
  - Public-goods context: subsidy, tax, fine, prison
- Action also observed by others: coworkers, friends, rest of society $\Rightarrow$ reputational concerns

Preferences

$$U = (v + y)a - C(a) + \mu E(v|a, y) + e\bar{a}$$

- $v_y \equiv 1$, for now: valuation for money or other "'extrinsic" incentives
- $v_a \equiv v$: "intrinsic motivation" $\sim G(v)$, density $g(v) > 0$.
  - Private-goods context: liking and motivation for the task (e.g., research), work ethic, perfectionism, company spirit, etc.
  - Public-goods context: degree of altruism / prosocial orientation

    ▷ Can be pure or impure, warm glow
- Externality: derives benefit $e\bar{a}$ from aggregate supply $\bar{a}$
- μ : instrumental or hedonic value from being seen as having high v
  - Private-goods context: career concerns $\rightsquigarrow$ valuable to be seen as motivated for the activity in question; as perfectionist, honest, etc.
  - Public-goods context: desirable to be perceived as generous, public minded, reciprocal, good citizen, etc.

Social planner and other principals

- Benevolent planner: given shadow cost of funds λ, maximizes

$$W(y) = \bar{U}(y) - (1+\lambda)\, y\, \bar{a}(y)$$

  - $\bar{U}(y)$: agents' aggregate welfare, in equilibrium under policy y

- More generally: weight $0 \leq \alpha \leq 1$ on agents' $\bar{U}$, private benefit B

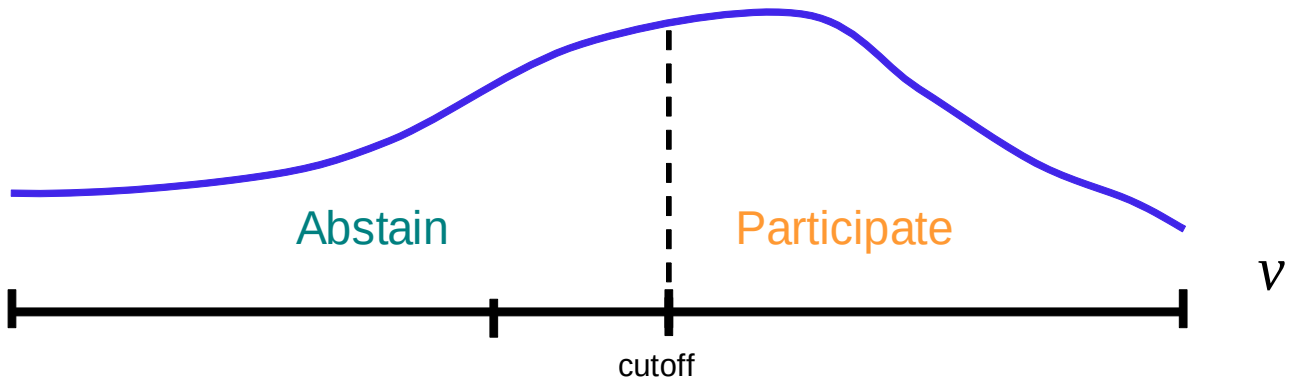
$$W(y) = \alpha \bar{U}(y) + [B - (1+\lambda)y]\, \bar{a}(y)$$

  - NGO, government agency, etc.
  - Purely self-interested, e.g. firm maximizing profits: $\alpha = 0$
  - Can all be reduced to planner's case

- Other policy tools:
  - Sending messages, disclosing information, e.g. about $G(v)$, $\bar{a}$
  - Publicity: making actions more visible: $\mu \uparrow$ (not here)

II. HONOR, STIGMA AND SOCIAL NORMS

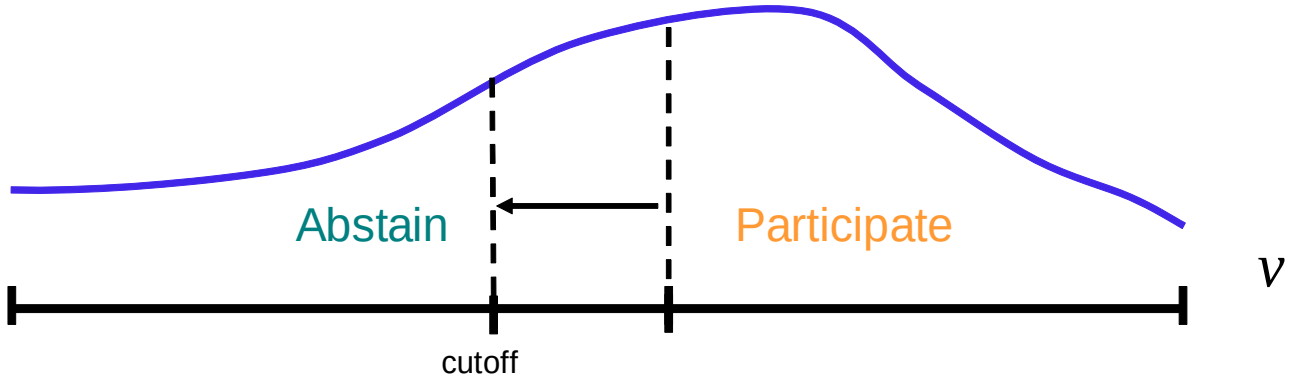
- Source, strength of social norms, impact of incentives?
- Simplest: $a = 0, 1$: work / shirk, contribute / free ride
- Individual participates $(a = 1)$ iff motivation v above cutoff v^*



- Honor: average motivation above cutoff: $\mathcal{M}^+(v^*) = E[\tilde{v} \mid \tilde{v} > v^*]$
- Stigma: average motivation below cutoff: $\mathcal{M}^-(v^*) = E[\tilde{v} \mid \tilde{v} < v^*]$
- Cutoff v^* = point of indifference (when interior):

$$v^* + y + \mu\left[\mathcal{M}^+(v^*) - \mathcal{M}^-(v^*)\right] = c$$

- When more people participate, honor declines, stigma worsens



- Net reputational incentive

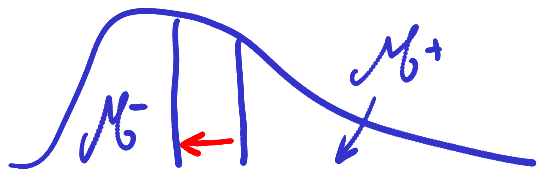
$$\Delta(v^*) \equiv \mathcal{M}^+(v^*) - \mathcal{M}^-(v^*) = \text{Honor - Stigma}$$

may $\searrow$ or $\nearrow$, depending on whether $\mathcal{M}^+$ or $\mathcal{M}^-$ responds more.

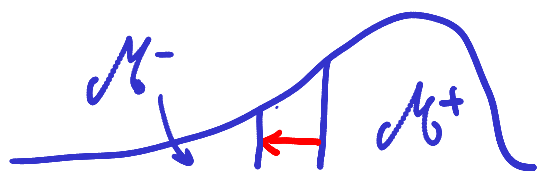
- Key difference between behaviors in which quest for honor versus avoidance of stigma is (endogenously) the main driver of behavior.
- Individuals' actions are
  - Strategic substitutes in first case: $\Delta' > 0 \Rightarrow$ social multiplier < 1
  - Strategic complements in the second: $\Delta' < 0 \Rightarrow$ social multiplier > 1

Role of the distribution of individual preferences

- Expect honor considerations to dominate when there are only a few heroic or saintly types, whom the mass of more ordinary individuals would like to be identified with



- Expect stigma considerations to dominate when the population includes only a few "bad apples" with very low intrinsic values, which most agents will be eager to differentiate themselves from



Jewitt's lemma

Lemma

The shape of $\Delta(v) = \mathcal{M}^{+}(v) - \mathcal{M}^{-}(v)$ *mirrors that of density* $g(v)$:

1. *If* g *is everywhere decreasing (increasing), then* Δ *is everywhere increasing (decreasing)*
2. *If* g *has a unique interior maximum, then* Δ *has a unique interior minimum (but do not coincide)*

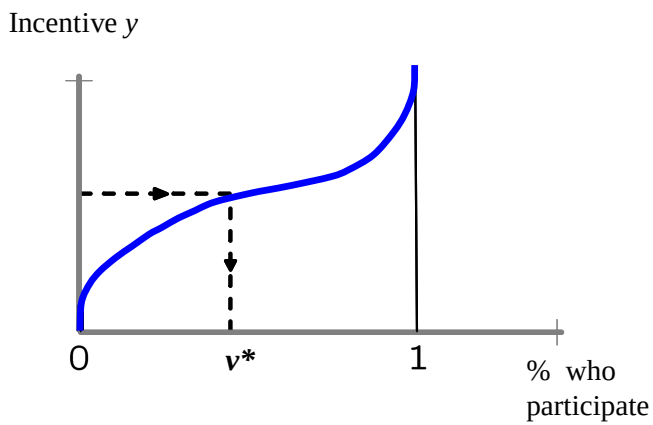
- Will assume strictly unimodal $g(v)$. Covers both SS, SC
- Equilibrium is unique iff $1 + \mu\Delta'(v) > 0$, $\forall v$
- Social multiplier:

$$-\frac{\partial v^*}{\partial y} = \frac{1}{1 + \mu\Delta'(v^*)}$$

The interaction of incentives and norms: summary

When honor motive is dominant:

- Individuals' decisions are substitutes
- Incentives → partial crowding out
  (still work, but weakened)

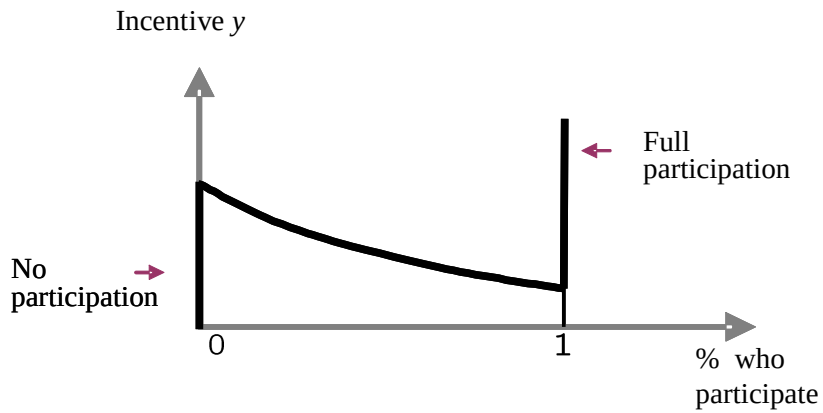


This occurs when:

- Most people are "mediocre", only rare "saintly" types with v well above most others (heroism, organ donation)
- Action is very costly
- There are possible "excuses" for not contributing, and / or one can do it without being noticed (⇒ weak stigma)

When stigma motive is dominant:

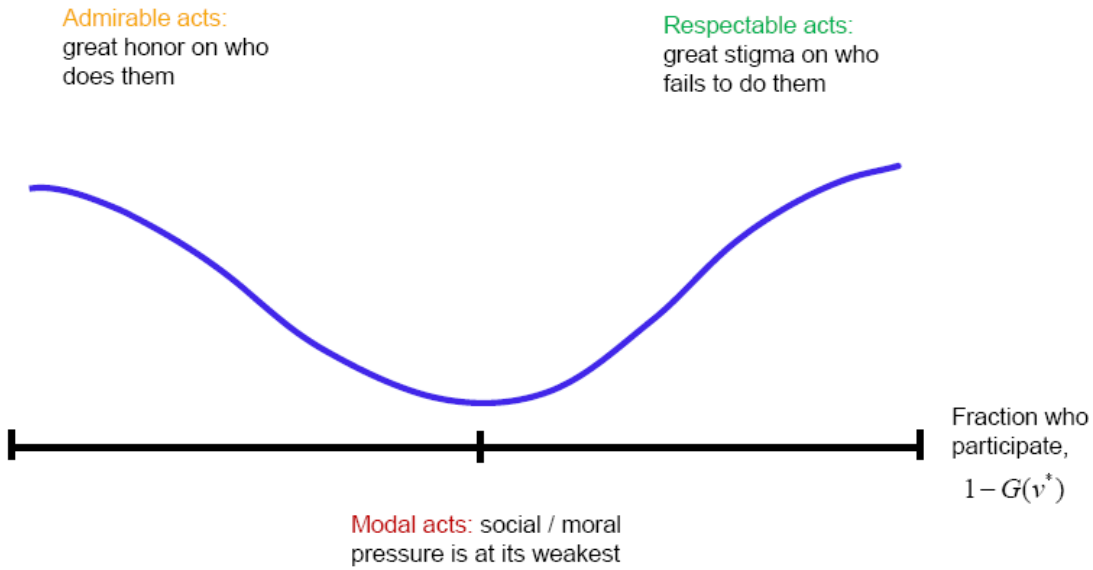
- Individual's decisions are complements
- Multiple norms may coexist
- Small incentives can have large effects: shift norms, crowding in



This occurs when

- Most people are "OK", only a few "rotten apples" with v well below most others (crime, child neglect)
- Action is relatively cheap
- There are possible non-glorious reasons for contributing (e.g., fear of the law), and/or it may go unnoticed (⇒ weak honor)

Classifying behaviors



- Focus now on unique equilibrum. Good behavior $(a = 1)$ is:
  - Respectable if "all but the worst types do it": v^* in the lower tail, so $\Delta'(v^*) < 0$. Not beating your spouse and children

    Such actions are complements (conformity), social multiplier > 1.
  - Admirable if "only the best do it": v^* in the lower tail, so $\Delta'(v^*) > 0$. Donating a kidney to a stranger

    Such actions are substitutes (distinction), social multiplier < 1.
  - Modal if both behaviors are prevalent: v^* in middle range

Implications

1. Material incentives (prizes, law) not very effective to spur "admirable", honor- driven behaviors: y weakens social esteem Δ when v^* is high. Heroism in combat, saving a life...

2. Incentives much more effective to strengthen "respectable", stigma-driven ones: y strengthens social pressure Δ when v^* is low. Corruption, cooperation, being green, political correctness...

3. Small changes in incentives can have large effects, shift social norms, when cost is fairly low and actions observable

4. If stigma / complementarity is strong enough and actions sufficiently visible, there can be multiple, self-sustaining norms

Shifts in prevailing societal values

- Changes in / aggregate uncertainty about preferences of society: v distributed according to

$$G_\theta(v) \equiv G(v-\theta),$$

  i.e. G shifted right by θ. Known or uncertain

- Density $g_\theta(v) = g(v-\theta)$, hazard rate $h_\theta = h(v-\theta)$, mean $\bar{v}+\theta$

- Given θ, reputational return is

$$\Delta_\theta(v) = \Delta(v-\theta)$$

- Known θ : results unchanged, with $g \rightsquigarrow g_\theta, \Delta \rightsquigarrow \Delta_\theta$, $a(y) \rightsquigarrow a_\theta(y)$...

Shifts in societal values

- Participation cutoff $v_\theta^*(y)$ given by

$$v_\theta^*(y) - c + y + \mu\Delta(v_\theta^*(y) - \theta) = 0$$

- Distributional shifts: $v_\theta^*(y) - \theta = v_0^*(y + \theta)$

Proposition

A known shift in θ has same effect on social pressure $\Delta(v_\theta^*(y) - \theta)$ and aggregate behavior $\bar{a}(y) = 1 - G(v_\theta^*(y) - \theta)$ as an increase in y (or a decrease in c) of the same magnitude.

- Societal preference shifts alter norms, act like incentives

- Suggests that perceptions of / messages about θ may be another channel of influence...

New Testable Implications

- When a socially approved behavior is sufficiently prevalent, stigma-avoidance rather than honor-seeking will be the dominant attributionnal concern $\Rightarrow$ formal incentives will have powerful effects on compliance (crowding-in) .
- When a socially approved behavior is sufficiently rare, honor-seeking s rather than stigma-avoidance will be the dominant attributional concern $\Rightarrow$ formal incentives will have weak effects on compliance (partial crowding-out)
- More generally: the more prevalent a socially approved behavior, the larger the effect of formal incentives
  - Cross effect: $\partial a_i / \partial y$ increasing in $\bar{a}$
- Prevalence of good or bad behavior is, of course, endogenous. But know what exogenous / experimentally manipulable factors shift it, e.g., visibility μ, cost c. For instance:
  - The more costly (to most individuals) is a socially approved behavior, the weaker the effects of formal incentives on compliance.

Ethnicity in Children and Mixed Marriages: Theory and Evidence from China (Jia & Persson 2014)

Broad research question:

- How do institutions and policy interventions shape ethnic identification?
  - Existing research suggests identification exhibits both social and individual motives, and both persistence and change
  - Persistent norms: social roots (e.g., Bisin-Verdier 2000
- Material incentives for change: economic roots
  - Bates 1974, Botticini-Eckstein 2007
- But individual and social motives likely interact. Do social norms crowd in or crowd out stronger material incentives?
  - Persson-Jia: very original use and test of Benabou-Tirole 2011 model

Why China?

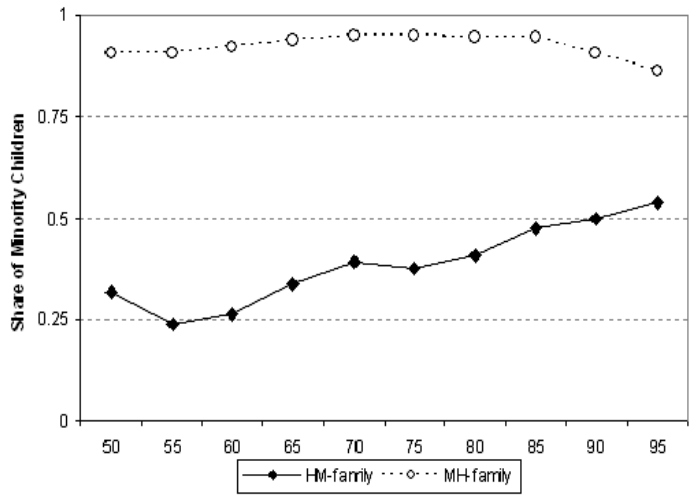
Interesting testing ground for ethnic policies and family choices.

- in 2010: Han ($\sim$1.2 billion) + 55 minorities ($\sim$ 105 million)
- great regional dispersion: minority share from 0.3% (Jiangxi) to 94% (Tibet)
- affirmative-action style interventions by national and provincial governments
- mixed ethnic couples free to choose whichever ethnicity for their children

Two facts on minority children in mixed marriages

Sources: 1982, 1990, 2000 censuses and 2005 mini-census

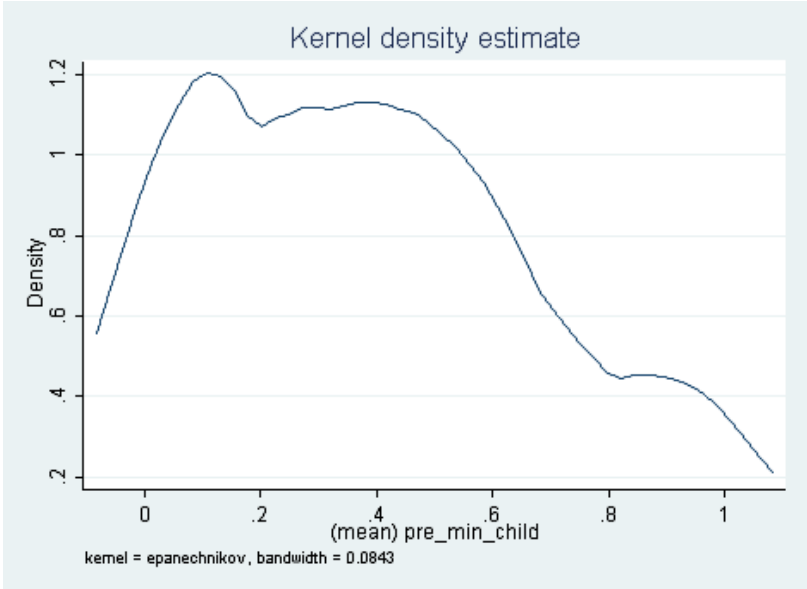
- repeated cross-sectional data for successive cohorts
- can identify location at prefecture (perhaps lower) level
- two types of mixed couples: Han man-Minority woman (HM), Minority man-Han woman (MH)

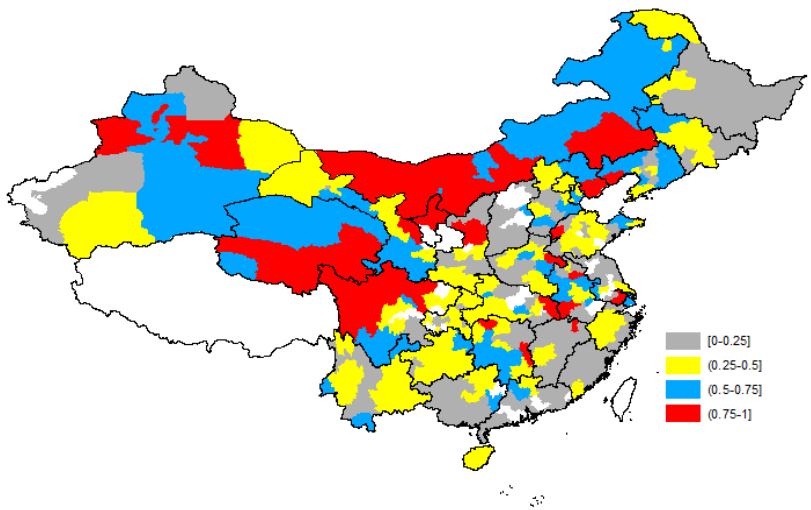


F1: Probability to choose minority identity much higher in MH couples than in HM couples

F2: Probability of minority children clearly increasing in HM couples

Variation in social norms is wide





2. Tests of These Predictions

Data sources

- 1% samples of 1982 and 1990 censuses
- 0.095% sample of the 2000 census
- 1% sample of the 2005 population survey (mini-census)

Information on demographics and socioeconomic status for about 25 million people

- outcomes (minority child or not): individual level
- incentives (b and $e(J)$): region/group/individual level

Test **C1**: Measurement

Material benefits (b) of what type?

- bundle of policies: family planning, entrance to college, employment
- (i) **timing: pre- and post-1980**
- (ii) one-child policy: rollout or revealed fertility
- (iii) heterogeneous benefits: Zhuang vs. other minorities

Social norms ($\frac{d\Delta(\varepsilon_H^*)}{d\varepsilon^*}$) in which peer group?

- need to avoid the reflection problem (Manski, 1993)
- (i) 1970s cohort in same prefecture and ethnic group
- (ii) **previous cohort in same prefecture and ethnic group**
- (ii) same residency and previous cohort in same prefecture and ethnic group

Test **C1**: Results in Table 2A

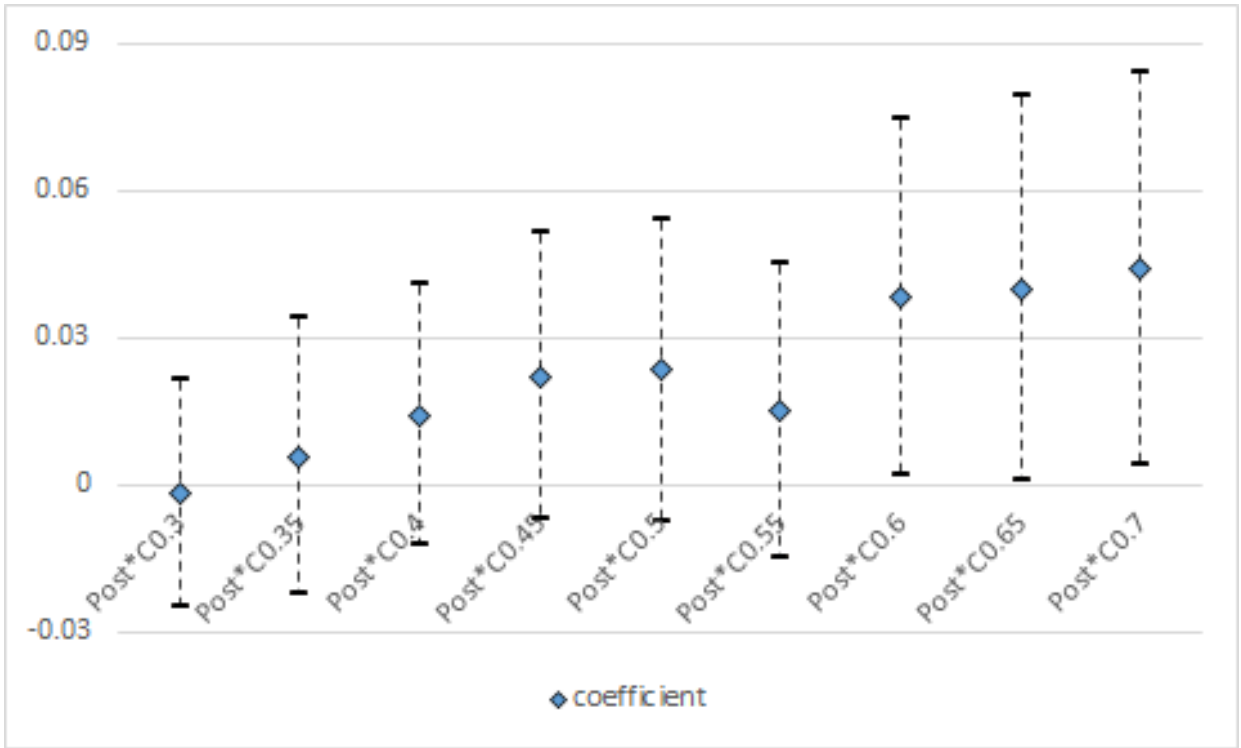
Higher social multiplier with fewer minority kids?

$$MinChild_{h,p,t} = \beta_b \mathrm{I}(\leq 0.\mathrm{X})_{p,t-1} \times Post1980_t + pref_p + birth_t + prov \times t + \varepsilon_{h,p,t}$$

	(1)	(3)	(4)	(5)
$\mathrm{I}(\leq 0.55) \times Post1980$		0.015		
		(0.015)		
$\mathrm{I}(\leq 0.60) \times Post1980$			0.038**	
			(0.018)	
$\mathrm{I}(\leq 0.65) \times Post1980$				0.040**
				(0.020)
$Post1980$	0.081***			
	(0.010)			
Prefecture FE	Y	Y	Y	Y
Birth Cohort FE		Y	Y	Y
Province Trends		Y	Y	Y
# of clusters	346	346	346	346
# observations	97399	97399	97399	97399

Test **C1**: Results (continued) in Figure 6

$$MinChild_{h,p,t} = \beta_b \mathrm{I}(\leq 0.\mathrm{X})_{p,t-1} \times Post1980_t + pref_p + birth_t + prov \times t + \varepsilon_{h,p,t}$$



Test **C1'**: Results in Table 2B

	(2)	(3)
I(0-0.25)$\times$*Post*1980	0.061***	0.046*
	(0.020)	(0.026)
I(0.25-0.5)$\times$*Post*1980	0.094***	0.050*
	(0.031)	(0.029)
I(0.5-0.75)$\times$*Post*1980	0.084***	0.036
	(0.030)	(0.035)
Prefecture FE	Y	Y
Birth Cohort FE	Y	Y
Province Trends		Y
# of clusters	346	346
# observations	97399	97399

Test **C2**: Results in Table 3

Smaller effect of smaller benefit?

$$MinChild_{i,p,t} = \beta_z Post1980_t \times ZhuangWife_i + \gamma ZhuangWife_i + \beta_b Post1980_t + pref_p + birth_t + prov \times t + \varepsilon_{i,p,t}$$

	(1)	(2)	(3)	(4)	(5)
Zhuang Wife × Post	-0.060***	-0.054***	-0.026**	-0.023*	-0.044**
	(0.014)	(0.014)	(0.011)	(0.012)	(0.011)
Zhuang Wife	-0.133***	-0.134***	-0.144***	-0.157***	-0.138***
	(0.039)	(0.039)	(0.034)	(0.035)	(0.035)
Post	0.092***				0.022***
	(0.012)				0.008
Prefecture FE	Y	Y	Y	Y	Y
Birth Cohort FE		Y	Y	Y	Y
Province Trends			Y	Y	Y
# of clusters	346	346	346	339	339
# observations	97399	97399	97399	95753	95753

migration minimized in (4), one-child policy rollout not *Post*1980 in (5).

Test **C3**: Measurement

Intrinsic costs (e) of what type?

- son versus daughter
- wife from religious minority

$$\begin{aligned} MinChild_{i,p,t} &= \beta_s Post1980_t \times Son_i + \delta Son_i \\ &\quad + pref_p + birth_t + prov \times t + \varepsilon_{i,p,t} \end{aligned}$$

Test **C3**: Results, Table 4

Smaller effect of material benefits at higher interinsic costs?

	(2)	(3)	(6)	(7)
Son × *Post*1980	-0.016**	-0.007		
	(0.006)	(0.006)		
Religious Wife×*Post*1980			-0.037**	-0.009
			(0.016)	(0.012)
Son	0.000	-0.009**		
	(0.004)	(0.004)		
Religious Wife			0.111***	0.093***
			(0.017)	(0.017)
Prefecture FE	Y	Y	Y	Y
Birth Cohort FE	Y	Y	Y	Y
Province Trends		Y		Y
# of clusters	346	346	346	346
# observations	97399	97399	95578	95578

III. WELFARE AND OPTIMAL INCENTIVES

- Net social value of an individual contribution, e.g., buying a Prius?
- Agent gets
  - Cost to individual: $-c$
  - Intrinsic value v: how much he values the improvement in public good (air quality) that his action brings about + pure "joy or giving"
  - Extrinsic reward: y. Subsidy, tax rebate, penalty avoided, etc.
  - Improved (self) image: $\mu \times$ (Honor $-$ Stigma)
- Others get
  - Benefit e created by unit increment to the public good, $\bar{a}$
  - Incentive payments: $-y(1+\lambda)$, from taxes or private sources
  - Loss of self image: stigma of non-contributors rises, honor of contributors falls (SUV owners, but also Prius owners)
- Pursuit of esteem is a zero-sum game: average reputation in society remains fixed, since distribution of types is fixed.
- Esteem, or even self-esteem is, by its very nature, a positional good

Welfare calculus

- Agents' behavior always characterized by a cutoff v^*
- Average utility

$$\begin{aligned}\bar{U}(v^*; y) &= \int_{v^*}^{+\infty} \left(e + v - c + y + \mu E\left[\tilde{v} \mid \tilde{v} \geq v^*\right]\right) \, g_\theta(v) \, dv \\ &+ \int_{-\infty}^{v^*} \mu E\left[\tilde{v} \mid \tilde{v} \leq v^*\right] g_\theta(v) dv \\ &= \int_{v^*}^{+\infty} \left[e + v - c + y\right] g_\theta(v) dv + \mu \bar{v}_\theta\end{aligned}$$

- Shows (linear) reputation as zero-sum game, positional good
- Principal maximizing social welfare

$$W = \bar{U} - (1 + \lambda) \, y \, \bar{a}(y) = \int_{v^*}^{+\infty} \left(e + v - c - \lambda y\right) g_\theta(v) dv + \mu \bar{v},$$

but extends to non-benevolent principals

Optimal incentives with known societal preferences

- Symmetric information about θ : $y \longrightarrow$ cutoff $v^* = v_\theta^*(y)$
- Planner sets y to maximize

$$W_\theta^{FI}(y) = \int_{v_\theta^*(y)}^{+\infty} (e + v - c - \lambda y)\, g_\theta(v)\, dv + \mu \bar{v}$$

- Optimality condition

$$\underbrace{\frac{e + v_\theta^*(y) - c - \lambda y}{1 + \mu \Delta_\theta'(v_\theta^*(y))}}_{\text{social multiplier}} \times \; g_\theta(v_\theta^*(y)) = \lambda \left[1 - G_\theta(v_\theta^*(y))\right]$$

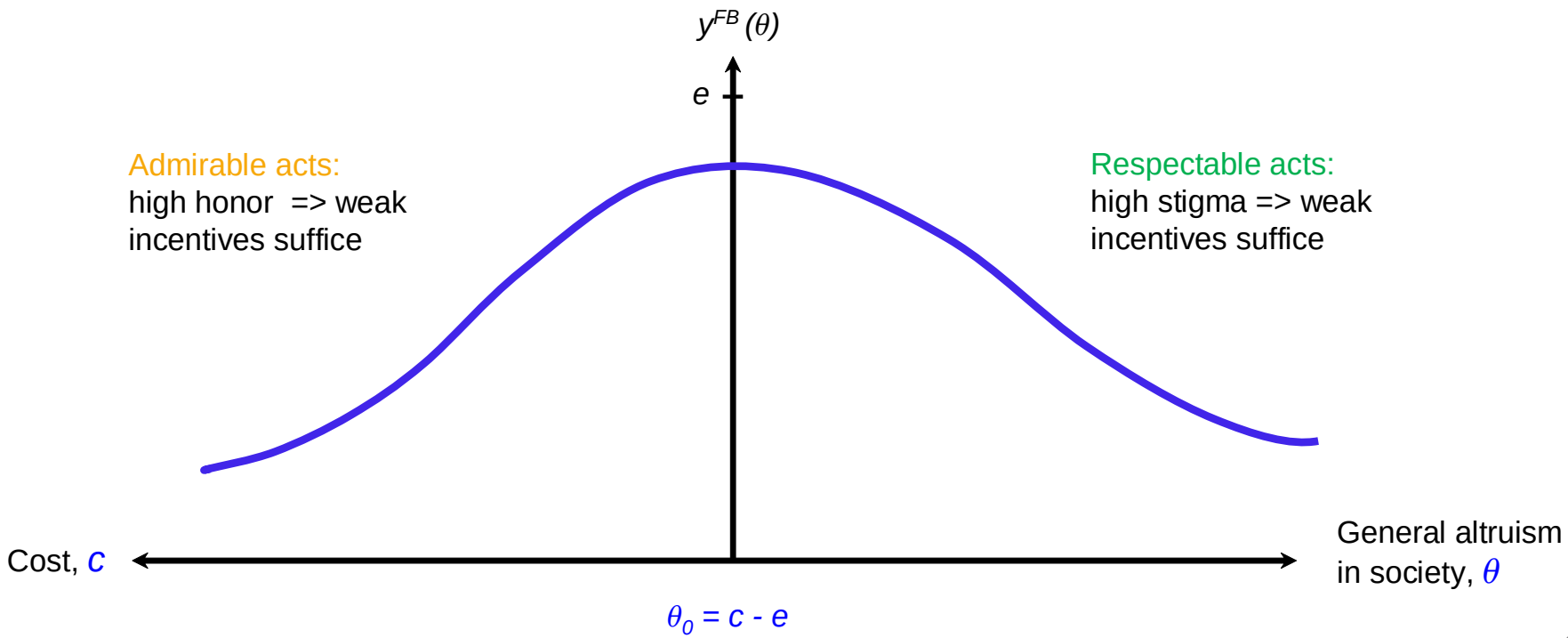
- Ramsey-like taxation
  - LHS = Net social marginal benefit of raising y by \$1, inducing $da_\theta = (-\partial v^* / \partial y) \times g_\theta$ new agents to participate
  - RHS = deadweight loss from paying \$1 more to inframarginal contributors

Proposition (modified Pigou)

The first-best subsidy $y^{FB}(\theta)$ *under symmetric information and no tax distortion* $(\lambda = 0)$ *is*

$$y^{FB}(\theta) = \underbrace{e}_{\text{externality}} \underbrace{- \mu\Delta(c - e - \theta)}_{\text{reputation tax}}$$

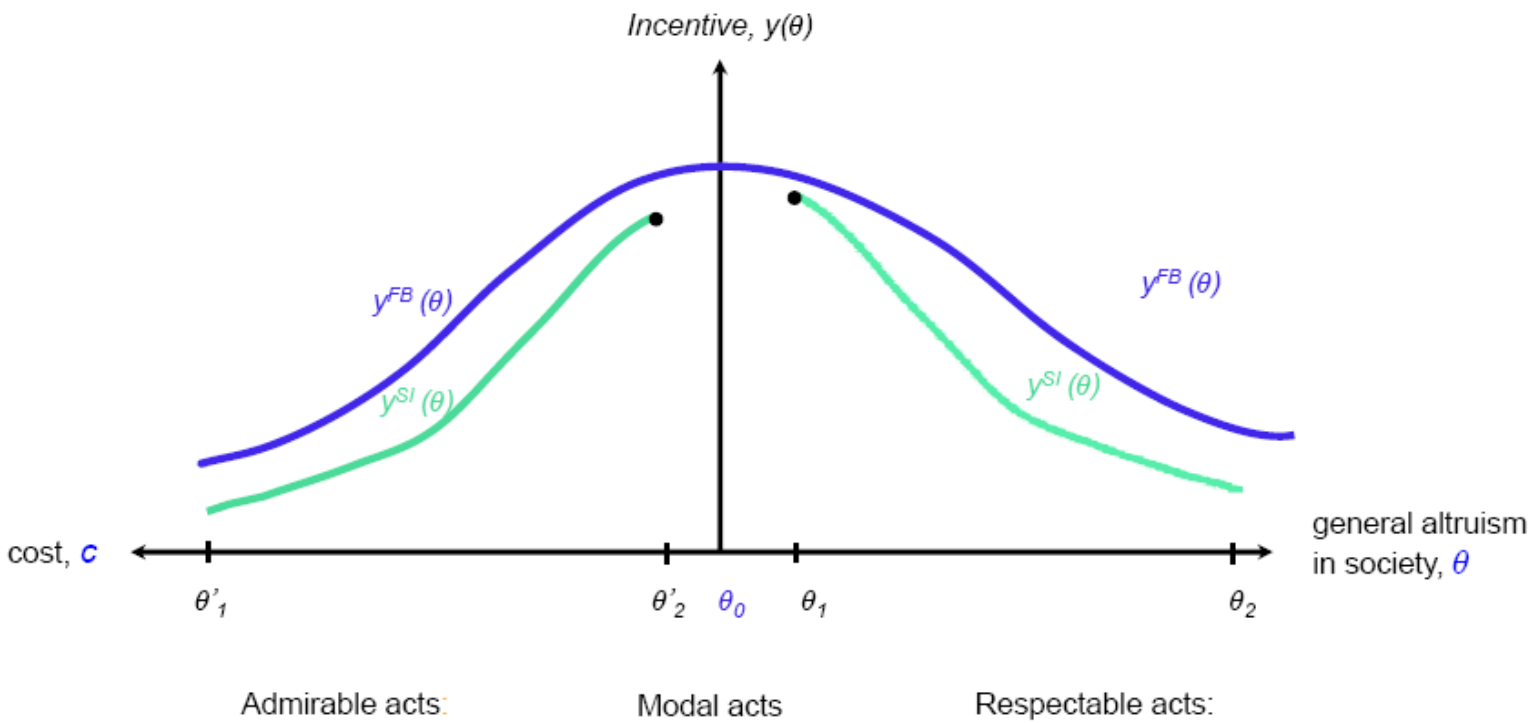
It is unimodal with respect to θ *and* c*, and maximized at* $\theta_0 \equiv c - e$.



Proposition (second best - cost of public funds)

Let (θ_1, θ_2) be any interval not containing θ_0. For $\lambda > 0$ low enough,

1. *The symmetric-information policy $y^{FI}(\theta)$ is uniquely defined on (θ_1, θ_2), with $0 < y^{FI}(\theta) < y^{FB}(\theta)$*
2. *The incentive $y^{FI}(\theta)$ strictly increasing in θ when $\theta_2 < \theta_0$ and strictly decreasing when $\theta_0 < \theta_1$.*



V. The expressive function of law

- Large (informal) literature arguing that laws have a dual role:
  - Not just a menu with "prices" for good or bad behaviors
  - Also express society's values: what it approves of or chooses to punish, how it chooses to punish; this expressive function is important
- Expressive considerations used to argue for both
  - tougher laws (even inefficiently so), e.g. prison vs. fines or community service.
  - gentler hand, e.g. limiting severity of sanctions: corporal punishments, torture, shaming, death penalty
- Other examples
  - Prohibition / legalization of "soft" drugs, or flag burning
  - Gay marriage vs. equivalent civil union. Earlier: Georgia's anti-sodomy law, unenforced but remained on the books; antimiscegenation laws
  - No price / market for organs, adoption, etc.

Modeling expressive law

- Social planner knows / has information on aggregate preference of society or "community standards" θ, hence $G_\theta(v)$
  - May have observed behavior of a representative sample; polls
  - Law, incentives, will then inevitably convey message about it

- Individuals in society only know that

  (i) $\theta \in (\theta_1, \theta_2)$ to the left of peak $\theta_0 \equiv c - e$.
  Alternatively, that $\theta \in (\theta_1, \theta_2)$ to the right θ_0.

  Thus, agents have broad sense of whether some behavior is rare and admirable or common and merely respectable

  (ii) Planner sets incentive $y^{AI}(\theta)$ to maximize social welfare

Equilibrium

- Look for separating equilibrium where $y^{AI}(\theta) \nearrow$ on (θ_1, θ_2) if lies to the left of θ_0, $\searrow$ if lies to the right
- Agents invert the policy, infer θ as solution $\hat{\theta}(y)$ to $y^{AI}(\hat{\theta}(y)) \equiv y$.
- Resulting cutoff for participation: $v^*_{\hat{\theta}(y)}(y) \Rightarrow$ planner maximizes

$$W^{AI}_{\theta}(y) = \int_{v^*_{\hat{\theta}(y)}(y)}^{+\infty} (e + v - c - \lambda y)\, g_{\theta}(v) dv + \mu(\bar{v} + \theta)$$

- FOC + Eqbm:

$$\underbrace{\frac{e - c - \lambda y + v^*_{\theta}(y)}{1 + \mu \Delta'_{\theta}(v^*_{\theta}(y))}}_{\text{social multiplier}} \times \underbrace{[1 - \mu\, \Delta'_{\theta}(v^*_{\theta}(y))\ \hat{\theta}'(y)]}_{\text{informational multiplier}} = \frac{\lambda}{h_{\theta}(v^*_{\theta}(y))}$$

- FOC = implicit DE in $\hat{\theta}(y)$, or its inverse, $y(\theta)$

$$\frac{e - c - \lambda y(\theta) + v_\theta^*(y(\theta))}{1 + \mu \Delta'_\theta(v_\theta^*(y(\theta)))} \times \left[1 - \frac{\mu \, \Delta'_\theta(v_\theta^*((\theta)))}{y'(\theta)}\right] = \frac{\lambda}{h_\theta(v_\theta^*(y(\theta)))}$$

- This is the "expressive content of the law" $\rightsquigarrow$ new multiplier

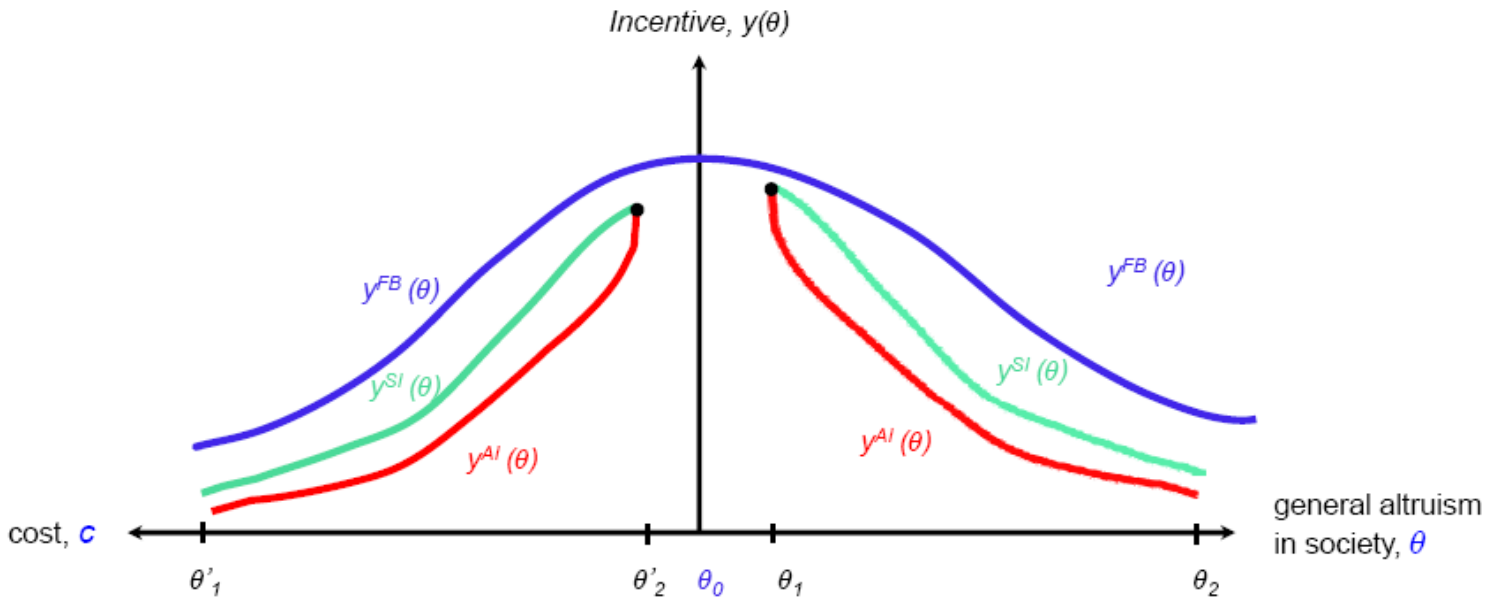
- Reflects planner's taking into account that agents will make inferences from chosen policy, about:
  - Where societal values lie: $\hat{\theta}'(y) = 1/y'(\theta)$
  - Social norms / sanctions will face as a result: $\mu \, \Delta'_{\hat{\theta}}(v_{\hat{\theta}}^*(y(\theta)))$

- For $\lambda = 0$, the first-best solution, $y^{FB}(\theta) = e - \mu\Delta(c - e - \theta)$, is the unique separating equilibrium
  - Intuitive: no need for expressiveness

Proposition (law expressing societal standards)

Whether the prosocial action is of a respectable or admirable nature ($\theta_0 < \theta_1$ *or* $\theta_2 < \theta_0$)*, for all* $\lambda > 0$ *low enough:*

1. *Principal always sets lower-powered incentives under asymmetric information:* $y^{AI}(\theta) < y^{FI}(\theta)$ *for all* $\theta \in (\theta_1, \theta_2)$.
2. *Participation / compliance is lower than under full information.*



Intuition

- Why is $y^{AI} < y^{FI}$, whether a high y signals a high or a low θ?

- Respectable activities / SC: lower y conveys the message: "everyone does it, except the most disreputable people who suffer great stigma This is why we need not provide strong extra incentives"

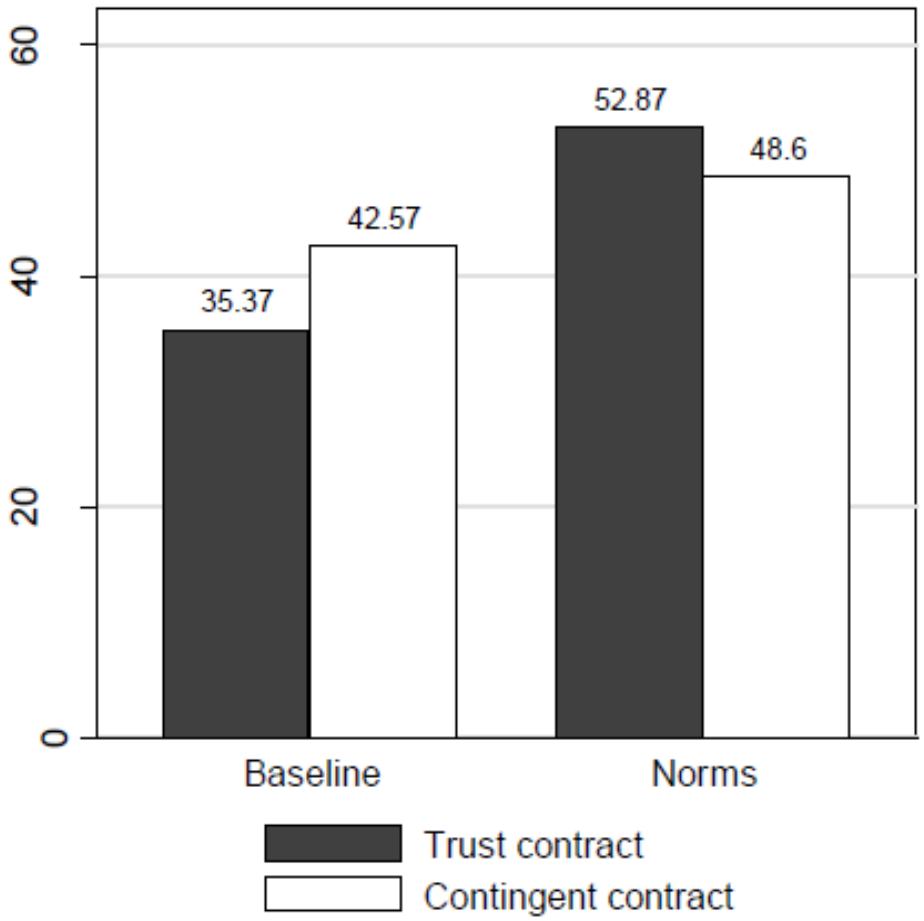
- Admirable activities / SS: lower y conveys the message "the glory suffices: contributors are rare beings, who reap such honor and social esteem that no additional incentives are necessary"

- While "gentler", expressive law is more responsive to changes in societal values than "standard" law. On both sides of the peak,
  - Level: $y^{AI} < y^{FI}$ everywhere
  - Sensitivity: average slope over (θ_1, θ_2) is steeper for y^{AI} than for y^{FI} (especially at the origin)

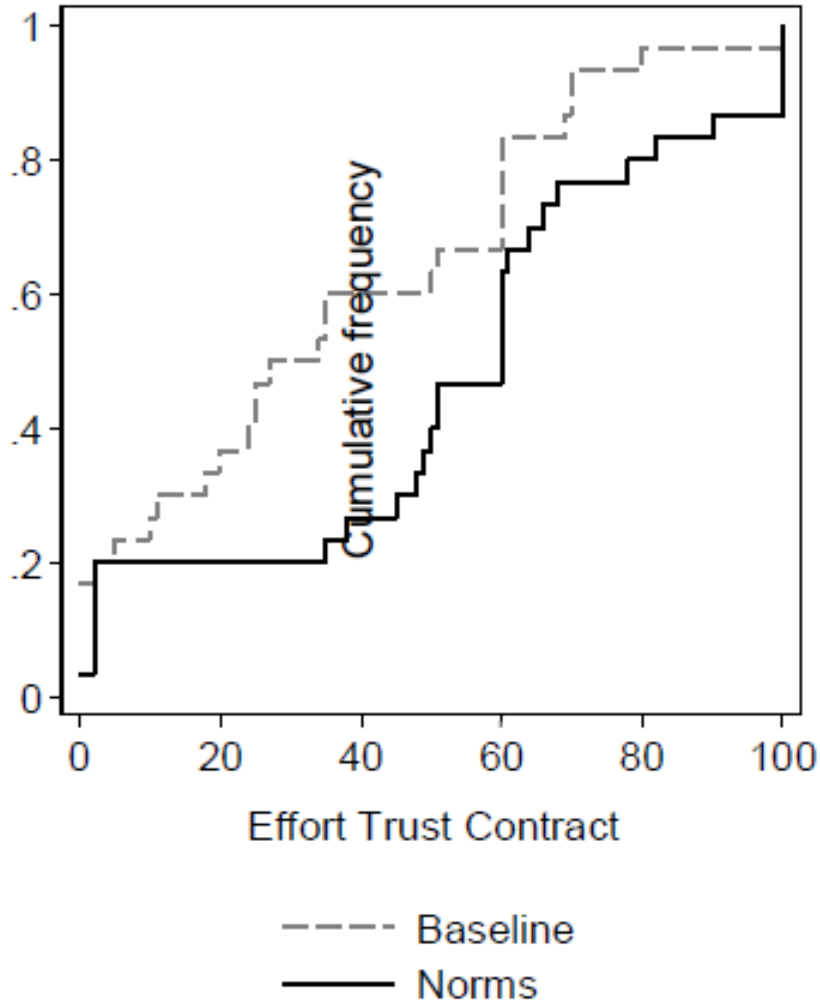
A. Danilov and D. Sliwka (2013) "Can Contracts Signal Social Norms?"

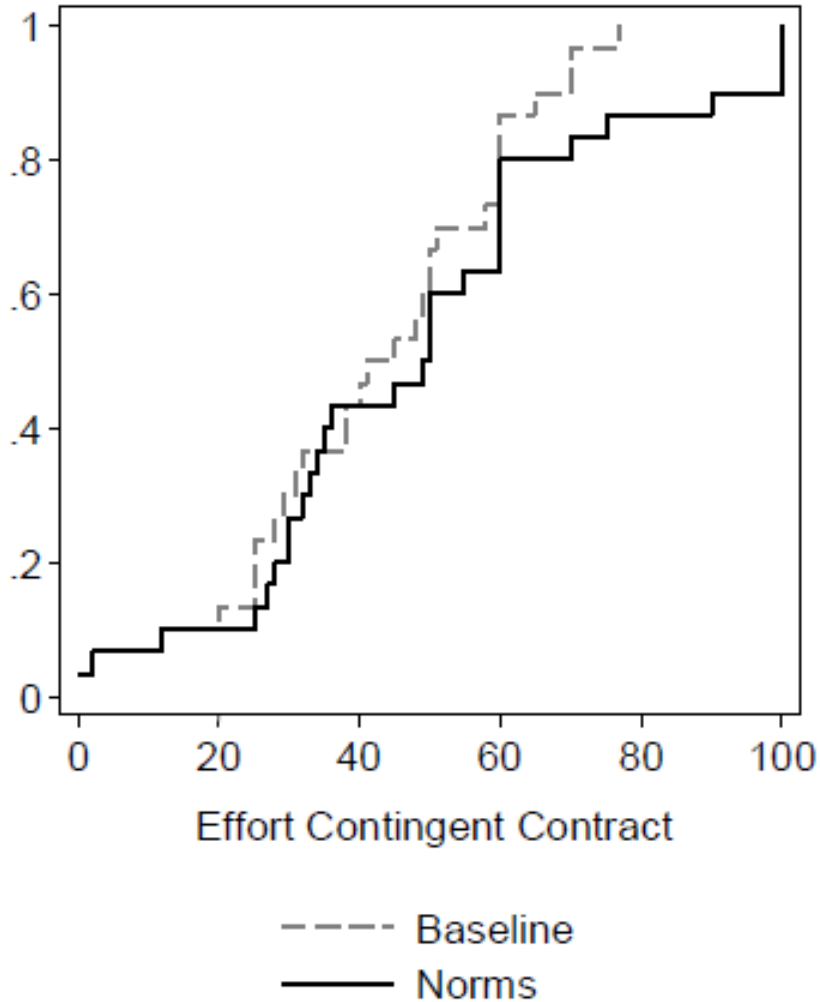
- Agent chooses $a \in [0, 100]$, at cost $C(a) = a^2/2$.
- Principal earns 12 Euros with probability a, nothing otherwise
- Principal chooses between:
  - "Trust contract": unconditional wage of 5 Euros
  - "Contingent" or incentive contract" agent gets bonus $b = 5$ Euros iff Principal receives 12 Euros
- Agent's efforts elicited for both contracts, using the strategy method
- Two informational conditions, payoffs unchanged:
  - "Baseline": as described above
  - "Norms": before choosing contract, Principal sees decisions taken by 10 agents from previous baseline condition.

    Agent knows Principal selecting his contract has seen such information.

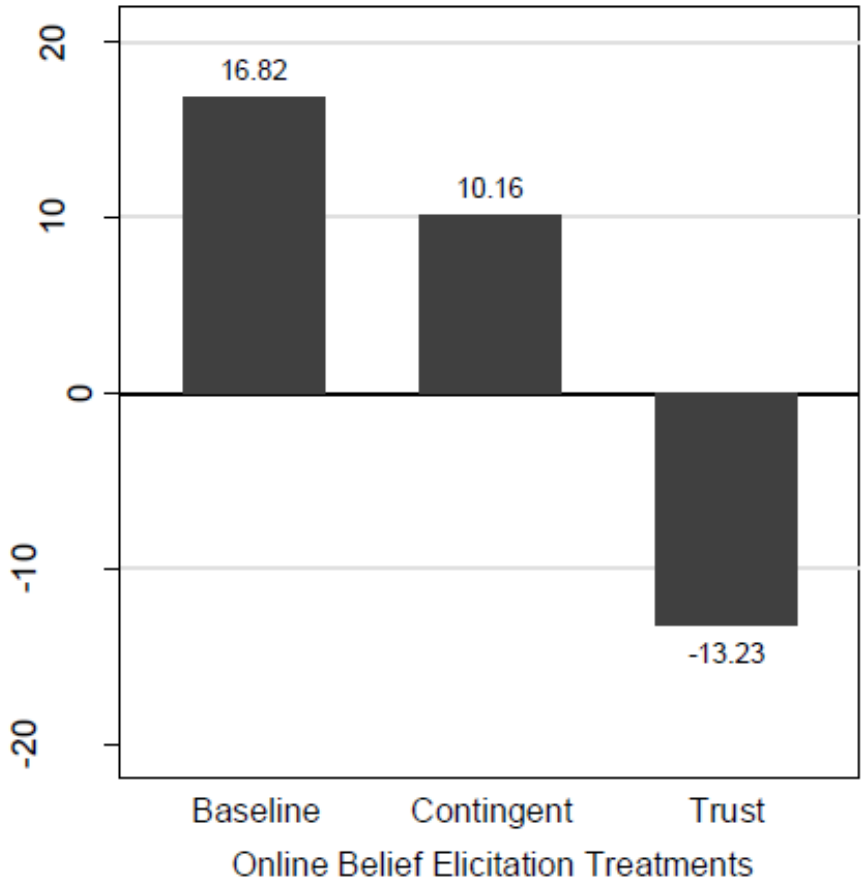


Average Effort for the Trust and Contingent Contracts

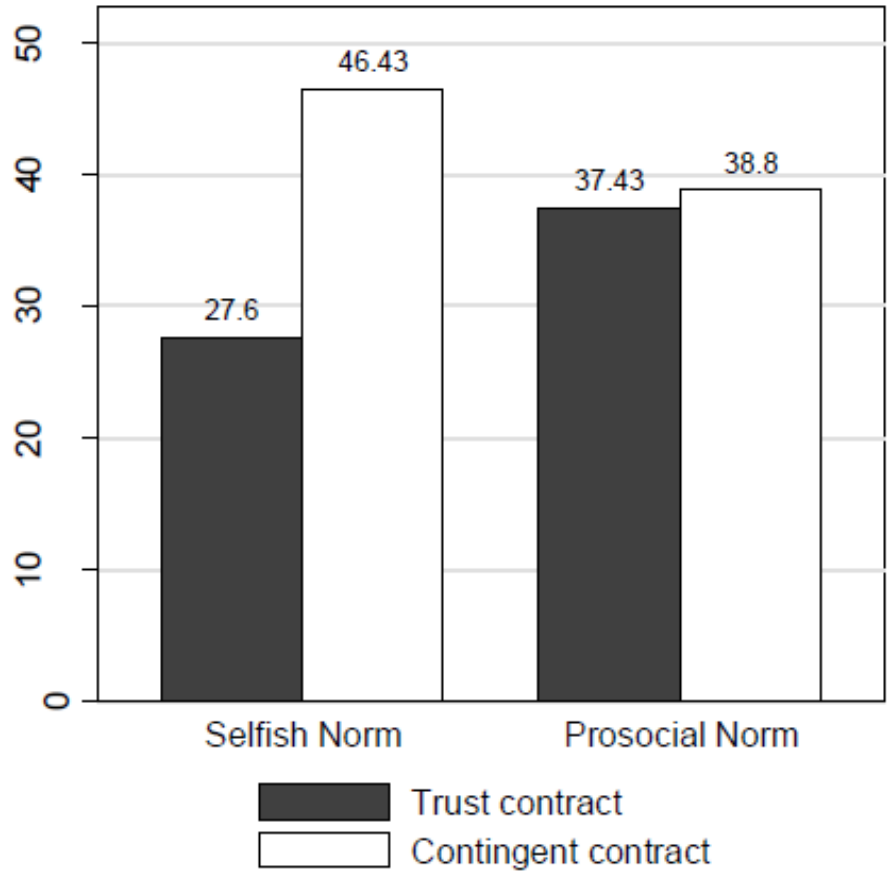

Cumulative frequency
0
.2
.4
.6
.8
1
0
20
40
60
80
100
Effort Trust Contract
Baseline
Norms


0
.2
.4
.6
.8
1
0
20
40
60
80
100
Effort Contingent Contract
Baseline
Norms

Elicited beliefs and actions

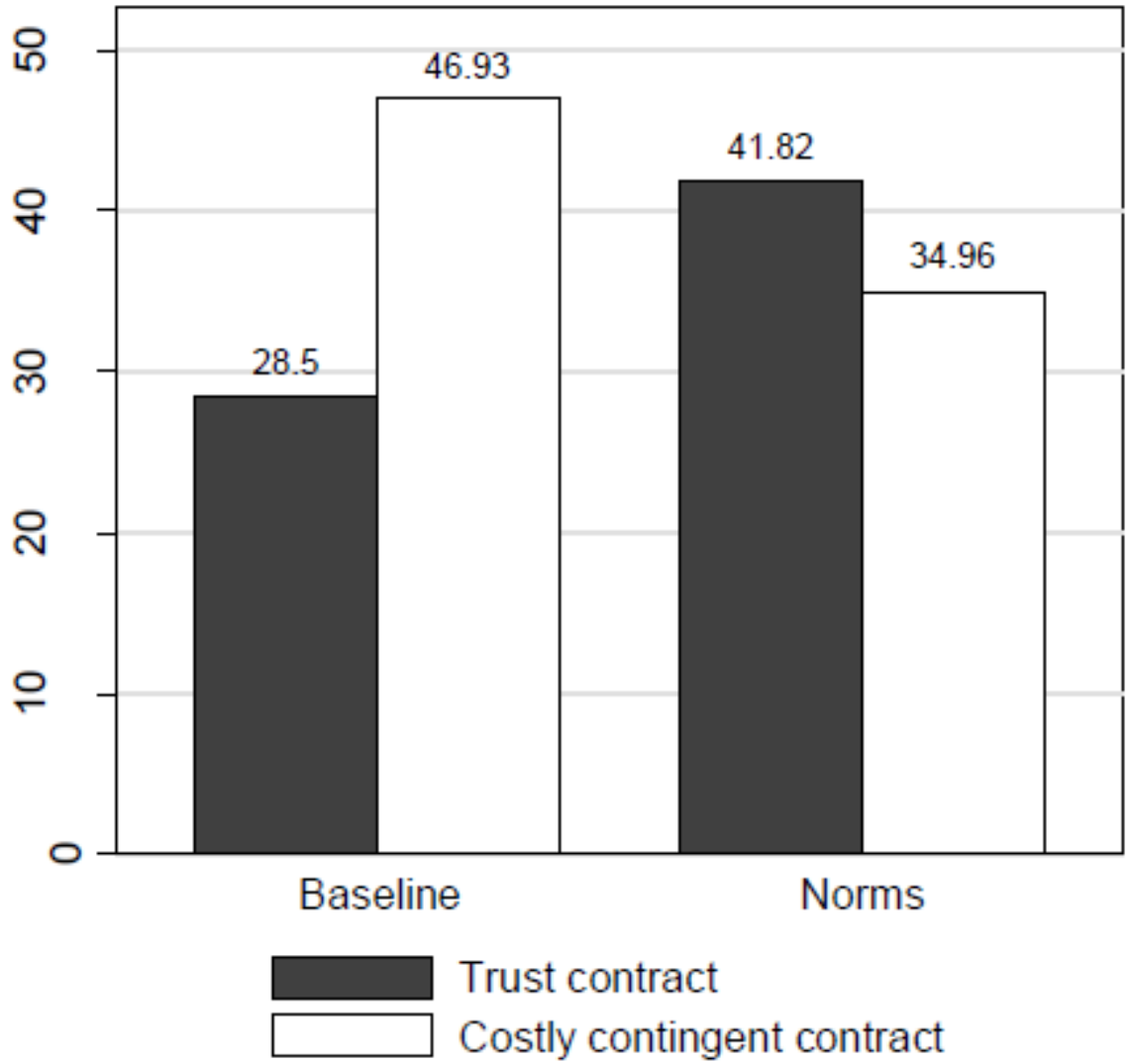


Average Difference in Estimated Efforts



Average Effort, "Induced Norms" Treatment

Varying the strength of the Principal's signal



Average Efforts, When Contingent Contract is Costly

Spillovers across spheres of behavior

- Two activities, a and b, both 0 - 1 decisions,
- Informal interactions: individual's $a-$behavior is observed by other private citizens, but not by principal / gvt.
  - Cooperating, helping, public goods contributions, not rent-seeking
  - Informational costs, activity done privately, observable not verifiable

$$y_a = 0, \quad \mu_a = \mu > 0$$

- Formal interactions: individual's $b-$behavior is observed by principal / gvt., but not by other private citizens
  - Transactions involving principal: paying / evading taxes, bureaucrats' honesty or corruption; employee productivity
  - Or, other agents less able than principal to sort through excuses

$$y_b = y > 0, \quad \mu_b = 0$$

- For simplicity, a person has same $v_a = v_b = v$ in both activities: general degree of prosociality (just need correlated G's)

- Two cutoffs:
  - $v_b^*(y) = c - y$ and $v_a^*(y) - c + \Delta_{\hat{\theta}(y)}(v_a^*(y)) = 0$
  - v_a^* depends on y only through inferences on θ

- Gvt. or other principal maximizes

$$W_\theta^{AI}(y) = \int_{v_b^*(y)}^{+\infty} (e_b + v - c_b - \lambda y)\, g_\theta(v)dv + \int_{v_a^*(y)}^{+\infty} (e_a + v - c_a)\, g_\theta(v)dv + \mu(\bar{v} + \theta),$$

$$\frac{\partial W_\theta^{AI}(y)}{\partial y} = (e_b + v_b^*(y) - c_b - \lambda y)\, g_\theta(v_b^*(y)) - \lambda\left[1 - G_\theta(v_b^*(y)\right] - (e_a - c_a + v_a^*(y))\, g_\theta(v_a^*(y)) \left(\frac{\partial v_a^*(y)}{\partial y}\right)$$

The expressive spillovers of law

- Social cost of raising incentive rate y for b behavior by \$1 includes:
  - Standard: must pay that extra \$1 to all *who* were complying anyway
  - New: less $\bar{a}$ compliance, as people infer that they face "worse" society, hence weaker social enforcement in other realms of behavior

Proposition (expressive spillovers)

Let the norms-enforced behavior (a) be of a respectable nature $(\Delta' < 0)$:

1. *Principal sets lower-powered incentives for the incentivized action b under asymmetric information:*

$$y^{AI}(\theta) < y^{FI}(\theta) \quad \text{for all } \theta,$$

2. *Participation in b is lower than under full information, participation in a is unchanged*

Why economists are unpopular

- Common resistance to economists' positive and normative messages about power of / need for incentives, markets).
- "Putting a price on everything": expresses bad news about human nature: low altruism v_a ($\sim$ low θ), high greed v_y.

1. Society may just not want to hear bad news about itself.
   - Often does not. Ideology, groupthink, identity...
2. Economists may be focussing on b -type behaviors, where incentives are easily available and social norms weak.
   - Perhaps less attention to / data on a -type behaviors, in which incentives are unavailable and social norms are strong.
   - Espousing, making salient a dim view of human nature, by stating / signaling that strong incentives are effective or needed in a, undermines the social norms in b. Creates need for incentives there, but may be less cost-effective way of achieving compliance

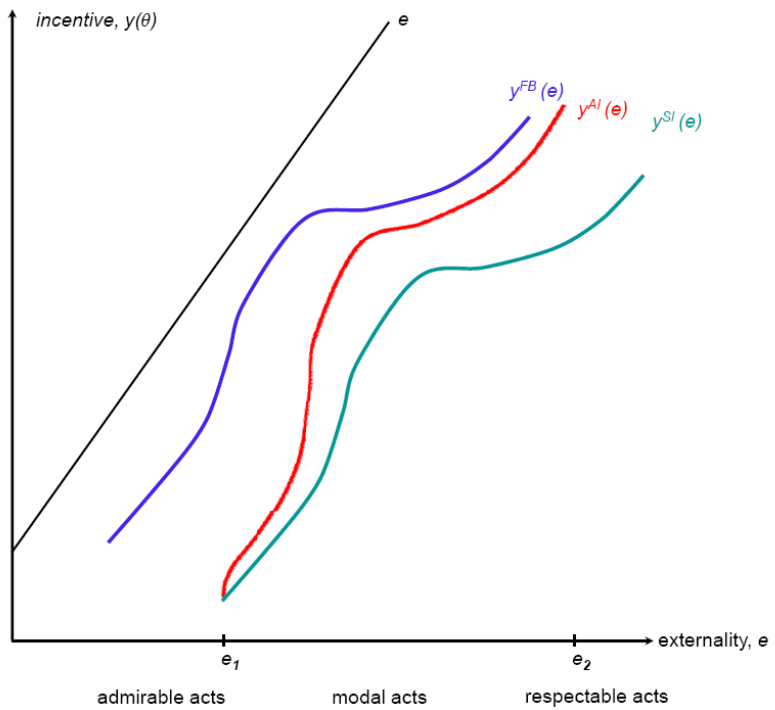
When expressiveness strengthens the law

- (When) can expressive content make law / incentives more strict rather than more lenient, i.e. $y^{AI} > y^{FI}$?
  - "Lock them up and throw away the key. We need to send a message"
- People's intrinsic motivation "should" be linked to how useful their action is for others: making one's contribution to the firm, to public goods that others enjoy, to social welfare. Thus:
- Let intrinsic motivation now be *ve*, with $v \sim G(v)$
- Reputation / self-image still bears on v = degree of social concern
- Principal knows e : how damaging are CO_2 emissions, how much good \$1 can do in poor countries, negative externalities from drunk driving, drugs, how important to firm is quality / customer service...

Proposition (law expressing magnitude of externalities)

Let AI bear on e, and intrinsic values be ve. Whether the prosocial action is of a respectable or admirable nature, for all λ low enough:

1. *The principal sets higher-powered incentives under asymmetric information: $y^{AI}(e) > y^{FI}(e)$ for all e.*
2. *Participation / compliance is higher than under full information.*



Lessons So Far...

1. Laws and norms shape each other, and behavior
   - Admirable acts: few people do, SS, incentives $\rightsquigarrow$ partial crowding out
   - Respectable acts: most people do, SC, incentives $\rightsquigarrow$ partial crowding in
2. Optimal incentives with norms - symmetric info:
   - Social or self esteem is a positional good. Prosocial actions inefficiently distorted toward the most visible
   - Pigou - Ramsey adjusted by reputation tax $\Rightarrow$ hill shaped
3. Norms based interventions: communication on $\bar{a}, \theta, e, \mu$. Credibility.
4. Optimal incentives with norms, asymmetric info: expressive law
   - Weakens optimal incentives when informative about society's general "goodness" θ, or "cruelty" κ. Strengthens them when informative about importance of externalities e
   - What is expressed concerning θ by law or incentives bearing on one activity carries over to people's attitudes and behavior in others
5. Resistance to economists' discourse about incentives

VI. THE BROADER MODEL

$$U = (v_a + v_y y)a - C(a) + \mu_a E(v_a|a, y) - \mu_y E(v_y|a, y) + e\bar{a}$$

$$W = \alpha \bar{U}(y) + [B - (1 + \lambda)y]\,\bar{a}(y)$$

1. Incentives and intrinsic motivation: y affects perceived v_a or $C(a)$
   - Private P-A setup: $e = 0$, $\mu_a = \mu_y \equiv 0$, $v_y \equiv 1$, AI on $\bar{v}_a$; $\alpha = 0$
2. Incentives and social norms: y affects $\mu_a E(v_a|a, y)$ via what reveals about people's general behavior / preferences, e.g., $\bar{a}$, $g(v_a)$
   - Public-goods setup with unidimensional type uncertainty: $e > 0$, $\mu_a > 0 = \mu_y$, $v_y = 1$, $v_a = v \sim G(v)$; $\alpha = 1$
3. Incentives "sully the meaning" of good actions: y affects attribution of a to intrinsic motivation v_a vs. greed v_y, or image-seeking, μ.
   - Need multidimensional type uncertainty about $(v_a, v_y; \mu_a; \mu_y)$

"Incentives and Prosocial Behavior" (B-T, AER 2006)

$$U = (v_a + v_y y)a - C(a) + \mu_a E(v_a|a,y) - \mu_y E(v_y|a,y) + e$$

- Actions a now vary over $\mathbb{R}$, cost $C(a) = ka^2/2$. FOC:

$$v_a + v_y y + \underbrace{\mu_a \frac{\partial E\left(v_a|a,y\right)}{\partial a} - \mu_y \frac{\partial E\left(v_y|a,y\right)}{\partial a}}_{\text{reputational return}} = ka$$

- Agents' valuations (v_a, v_y) are distributed in the population as

$$\begin{pmatrix} v_a \\ v_y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \bar{v}_a \\ \bar{v}_y \end{pmatrix}, \begin{bmatrix} \sigma_a^2 & \sigma_{ay} \\ \sigma_{ay} & \sigma_y^2 \end{bmatrix} \right), \quad \bar{v}_a \gtrless 0, \quad \bar{v}_y > 0,$$

- Focus here on case where everyone has same reputational concerns $(\bar{\mu}_a, \bar{\mu}_y)$ $\rightsquigarrow$ study material rewards
  - Paper also analyses case where $\boldsymbol{\mu}$ is also normally distributed across individuals $\rightsquigarrow$ study image rewards

Parsing out motivations

- Common $\boldsymbol{\mu} = \bar{\mu}$ $\Rightarrow$ same reputational motivation for all agents

$$\bar{r}(a, y) \equiv \bar{\mu}_a \frac{\partial E\left(v_a | a, y\right)}{\partial a} - \bar{\mu}_y \frac{\partial E\left(v_y | a, y\right)}{\partial a}$$

- So by FOC $v_a + v_y y + \bar{r}(a, y) = ka \Rightarrow$ agent's choice of a reveals the combination

$$v_a + v_y y = ka - \bar{r}(a, y)$$

- Signal extraction with normal random variables $\Rightarrow$

$$E\left(v_a | a, y\right) = \bar{v}_a + \rho(y) \cdot \left[ka - \bar{v}_a - \bar{v}_y \ y - \bar{r}(a, y)\right]$$

$$E\left(v_y | a, y\right) = \bar{v}_y + \chi(y) \cdot \left[ka - \bar{v}_a - \bar{v}_y \ y - \bar{r}(a, y)\right]$$

$$\rho(y) \equiv \frac{\sigma_a^2 + y\sigma_{ay}}{\sigma_a^2 + 2y\sigma_{ay} + y^2\sigma_y^2} \qquad \textit{and} \quad y\chi(y) \equiv 1 - \rho(y)$$

Proposition

Let all agents have the same image concern $(\bar{\mu}_a, \bar{\mu}_y)$.

1. *There is a unique (linear) equilibrium, in which an agent with preferences* (v_a, v_y) *contributes*

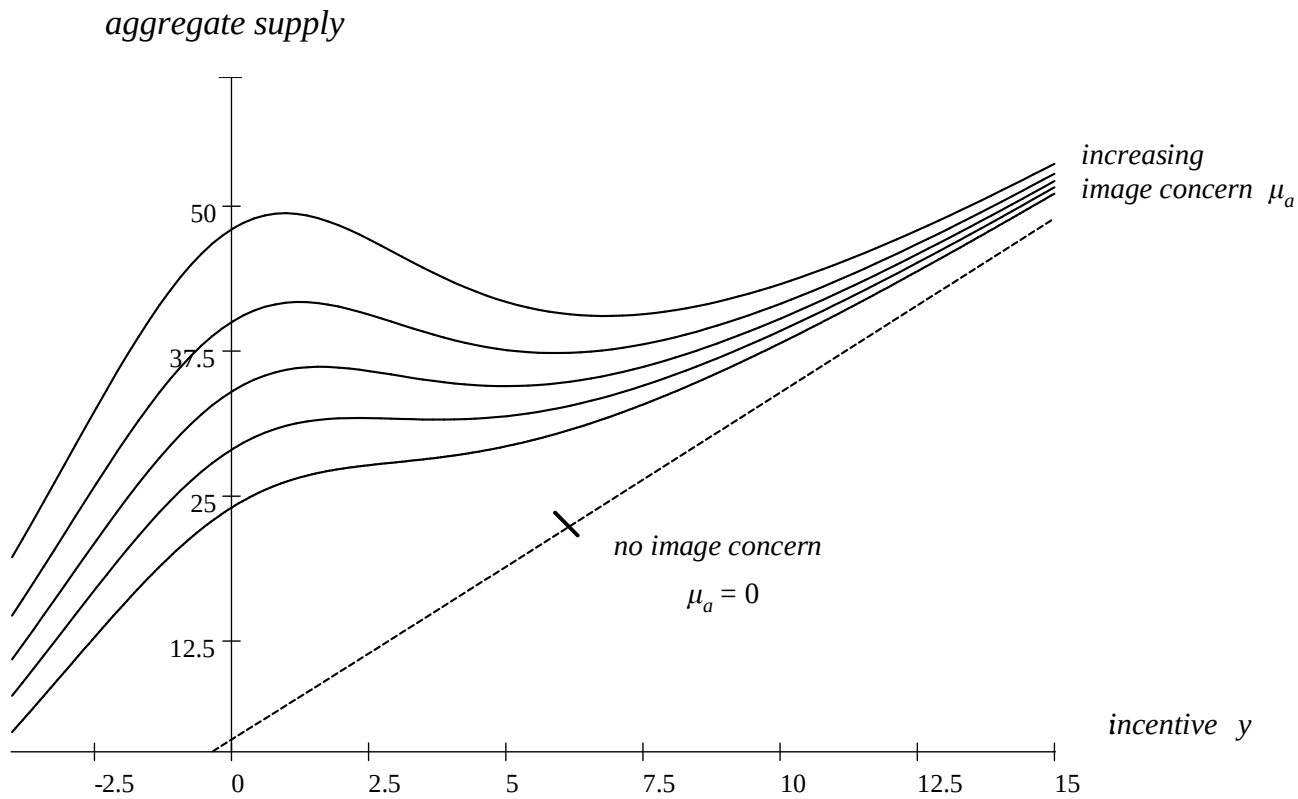
$$a = \frac{v_a + v_y\ y}{k} + \bar{\mu}_a \cdot \rho(y) - \bar{\mu}_y \cdot \chi(y),$$

   with $\rho(y)$ *and* $\chi(y)$ *correlation coefficients defined earlier.*

2. *Marginal reputational return is* $\bar{r}(y) = k\left[\bar{\mu}_a \cdot \rho(y) - \bar{\mu}_y \cdot \chi(y)\right]$.

- Effects of extrinsic incentives on inferences and behaviors:
  - Higher y increases direct payoff from contributing, $v_a + v_y\ y$
  - But also alters signaling value, along both dimensions

- With $\sigma_{ay} = 0$: $\bar{a}(y) = \frac{\bar{v}_a + \bar{v}_y \ y}{k} + \frac{1}{1+y^2\sigma_y^2/\sigma_a^2}\left(\bar{\mu}_a - \bar{\mu}_y \frac{y\sigma_y^2}{\sigma_a^2}\right)$



- Drawn for $\mu_a \nearrow$, with $\bar{\mu}_y = 0$: no stigma on greed / neediness
- When y increases, pro-social behavior is becomes increasingly ascribed to greed rather than altruism

Proposition (overjustification and crowding out)

Let $\sigma_{ay} = 0$. For all $\bar{\mu}_a$ above some threshold μ_a^*, there is a range $[y_1, y_2]$ where incentives are counterproductive: $\bar{a}(y)$ is decreasing on $[y_1, y_2]$, and increasing elsewhere.

- Focussed here on the crowding-out case, as has received more attention, more paradoxical.
- But, should not be overemphasized, e.g. can also get crowding-in, when $\sigma_{ay} < 0$
- Testable implications:
  - People contribute more when observed by others: $\partial \bar{a} / \partial \mu > 0$, but
  - This should attenuate when they are (known to be) rewarded for doing it: $\partial^2 \bar{a} / \partial y \partial \mu < 0$
  - Equivalently, effectiveness of incentives y smaller, or even reversed when both contribution and reward are observed

"Click for Charity" (Ariely, Bracha, Meier, AER 2007)

- Task: sequentially pressing keys X and Z on the keyboard for up to 5 minutes.
- For every $X - Z$ pair, pay money in participant's name to an assigned charity: 1 cent for each of first 200 pairs, 0.5 cents for each of next 200 pairs, 0.25 cents for each of next 200 pairs,... 0.01 cents for each above 1,200.
- Design: $2 \times 2 \times 2$:
  - "Good" or "Bad" Charity: American Red Cross, National Rifle Association
  - Incentives: either no payment to self, or same schedule as for charity,. Implemented with random draw
  - Private vs. public condition: anonymous, vs. at the end, must tell other participants which charity was assigned to, $ earned for it and for oneself
- 161 subjects

Figure 1: Effect of Private Incentive for "Good" Charity

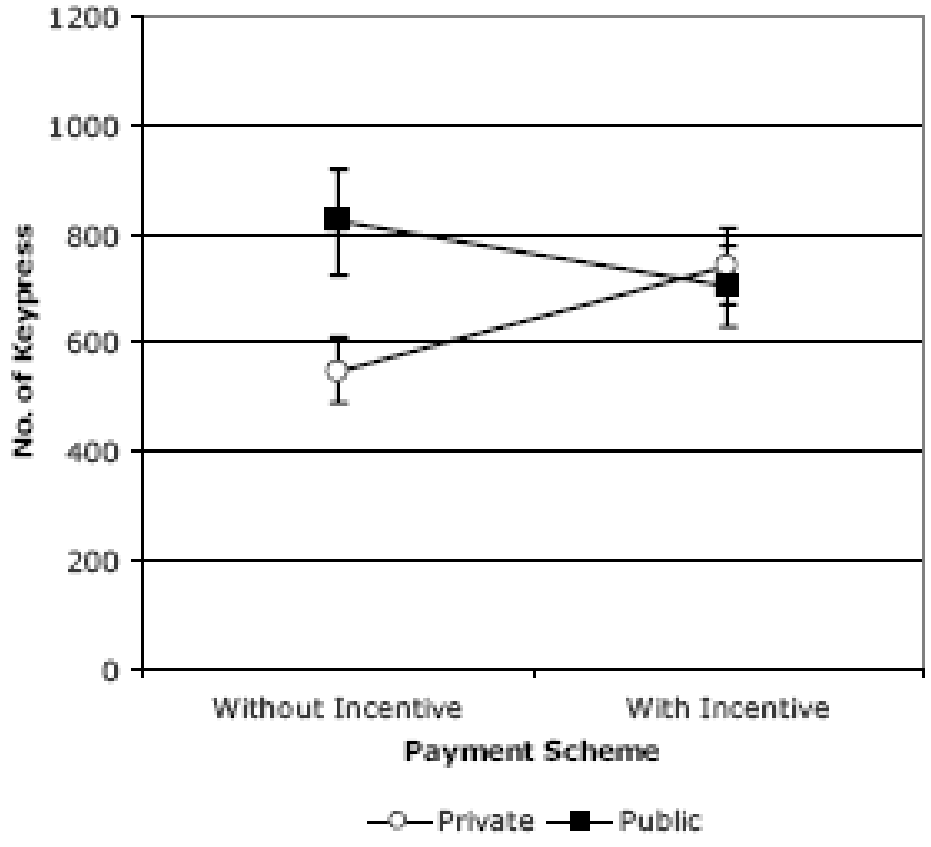
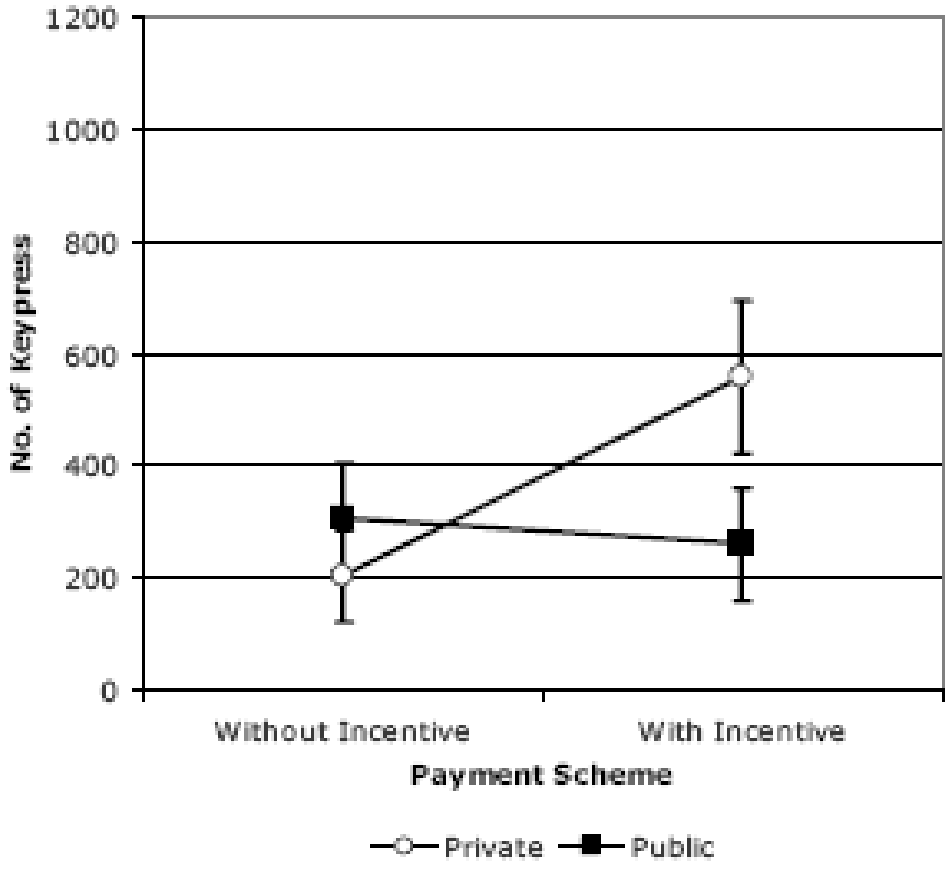


Figure 2: Effect of Private Incentive for "Bad" Charity



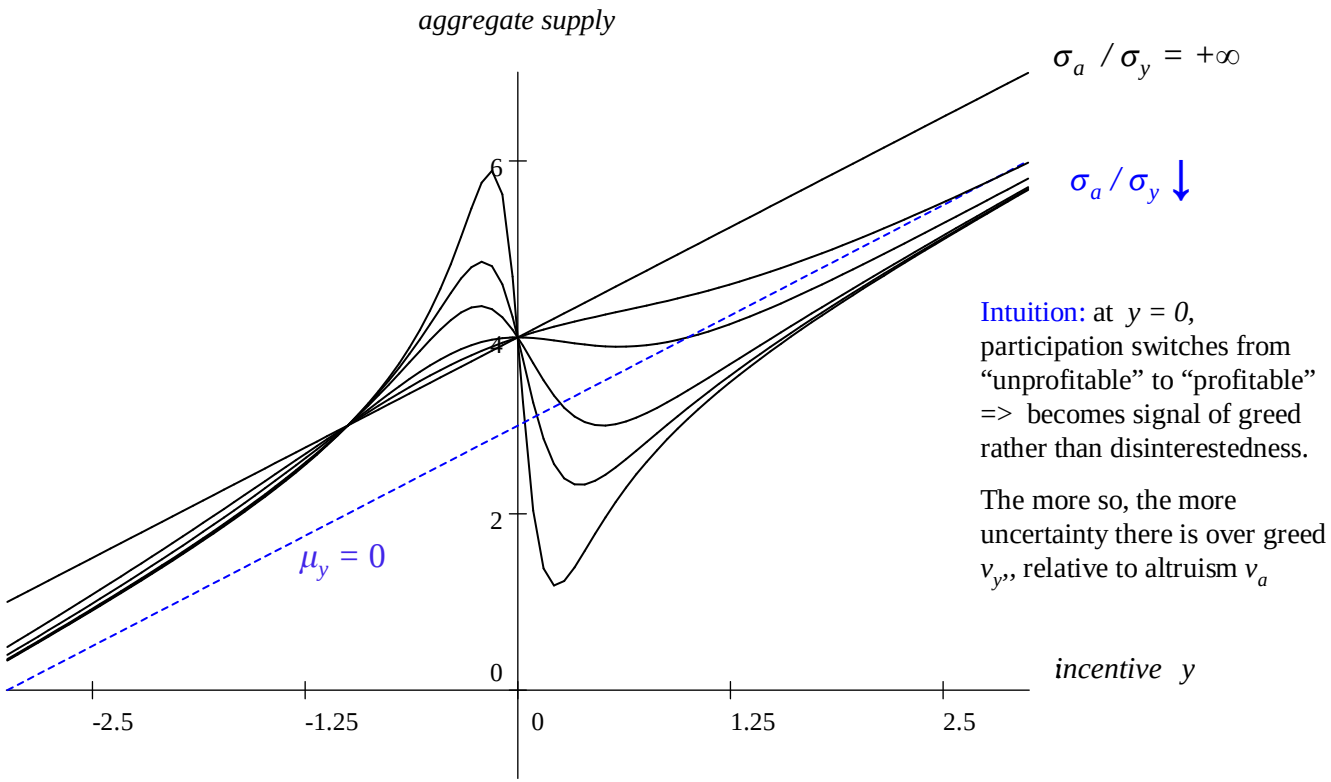
The case of "small rewards"

- Some studies find crowding out ($\bar{a}(y) \searrow$) to occur mostly at low $ amounts. Then, why relevant?
- Sometimes suggested that the main effect is a discontinuity at zero in subjects' response to incentives. Appeal to framing.

  (e.g., Gneezy-Rustichini 2000b, Bowles-Reyes 2009)
- Is there something qualitatively different between "unrewarded" and "rewarded" activities that could cause rational agents to behave in this way?
- Show that there is. But also that relevant notion of "small" rewards likely to be quite different in real-world .vs. lab.

‘

- With $\sigma_{ay} = 0$, $\bar{a}'(0) = \frac{\bar{v}_y}{k} - \bar{\mu}_y \left(\frac{\sigma_y}{\sigma_a}\right)^2$



- Illustrate with $\bar{\mu}_y > 0 = \mu_a$: no concern to appear prosocial, just not greedy
- In situations with much more uncertainty (more to learn) about individuals' desire for money than about their motivation for task at hand, even minimal concern about appearing greedy (small $\bar{\mu}_y > 0$) is sufficient to cause sharply negative response to small incentives $\rightsquigarrow$ downward discontinuity in supply

Small rewards and signal-reversal

Proposition (signal-reversal)

1. *Small incentives are counterproductive,* $\bar{a}'(0) < 0$, *whenever*

$$\frac{\bar{v}_y}{k} < \bar{\mu}_a \left(\frac{\sigma_{ay}}{\sigma_a^2} \right) - \bar{\mu}_y \left(\frac{\sigma_y^2 - 2\sigma_{ay}^2/\sigma_a^2}{\sigma_a^2} \right)$$

2. *Let* v_a *and* v_y *be uncorrelated, or not too correlated. As* $\sigma_a/\sigma_y \to 0$, *the supply function's slope at* $y = 0$ *tends to* $-\infty$.

3. *Let participation entails unit opportunity cost with monetary value* $\tilde{y}$. *Then* $\bar{a}'(\tilde{y}) < 0$ *and* $\bar{a}'(\tilde{y}) \to -\infty$ *under conditions (1) and (2).*

- Signal-reversal effect due to $\mu_y > 0$ creates, around zero net reward, additional source of crowding out on top of signal-jamming ($\rho(y) \downarrow$), which operates at all y's for acts with $\mu_a > 0$

Remarks

- Result on adverse effects of small incentives (when $\mu_y > 0$) applies whether or not the task is prosocial($\bar{\mu}_a \geq 0$)
  - Explains why adverse effects of small rewards found for both private, tasks and for public-goods provision (raising money for charity)
- Shows that relevant "tipping point" is not really zero –except in lab, where subjects have no alternative uses of time. It is instead agents' opportunity cost of time or effort, can be significant + more relevant
  - Suggests future work should involve situations where opportunity costs are (known to be) non-trivial and vary across subjects
- Both results (signal-jamming and signal-reversal) $\Rightarrow$
  - In field experiments, key question to ask = whether beneficiaries and observers of some activity (especially, prosocial) know or not that the person performing it is being incentivized

"When I was making money, I made the most money, and now that I'm spiritual I'm the most spiritual."