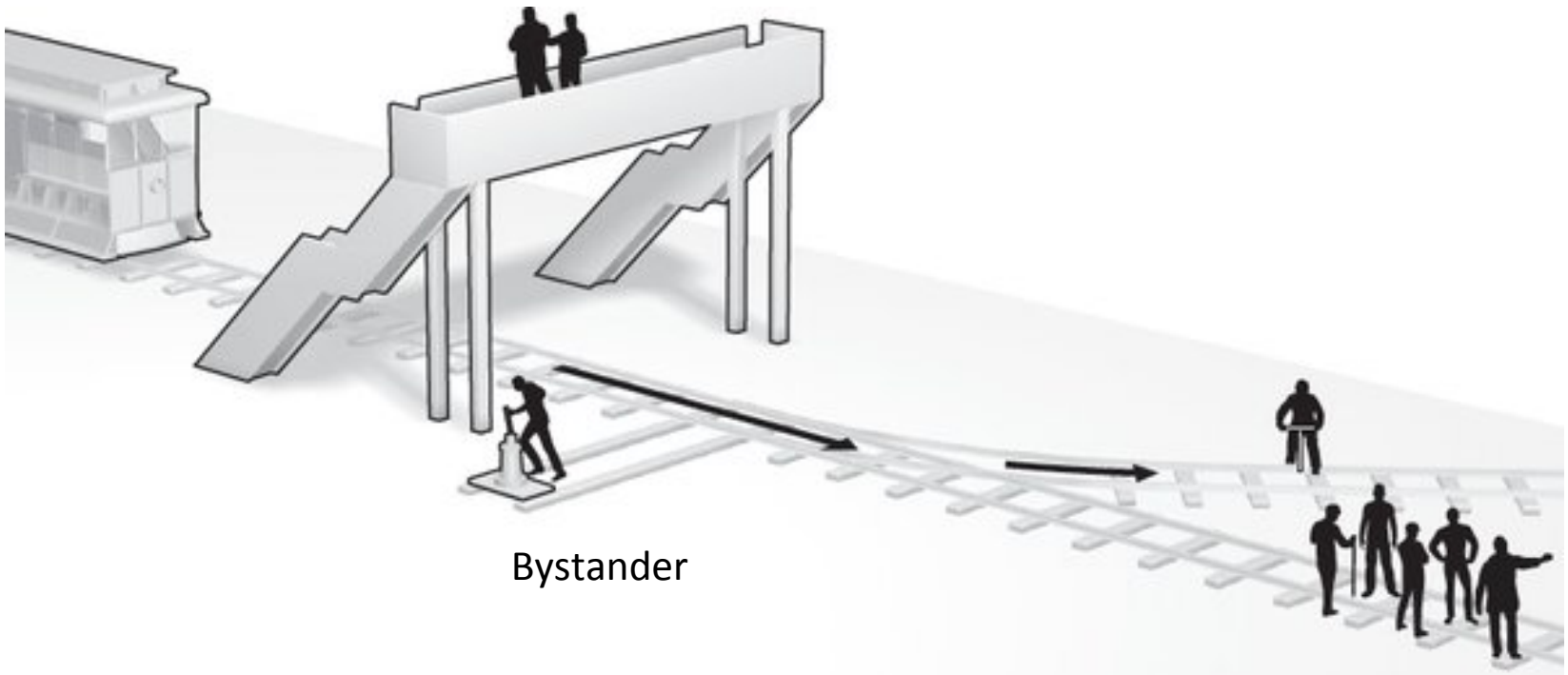
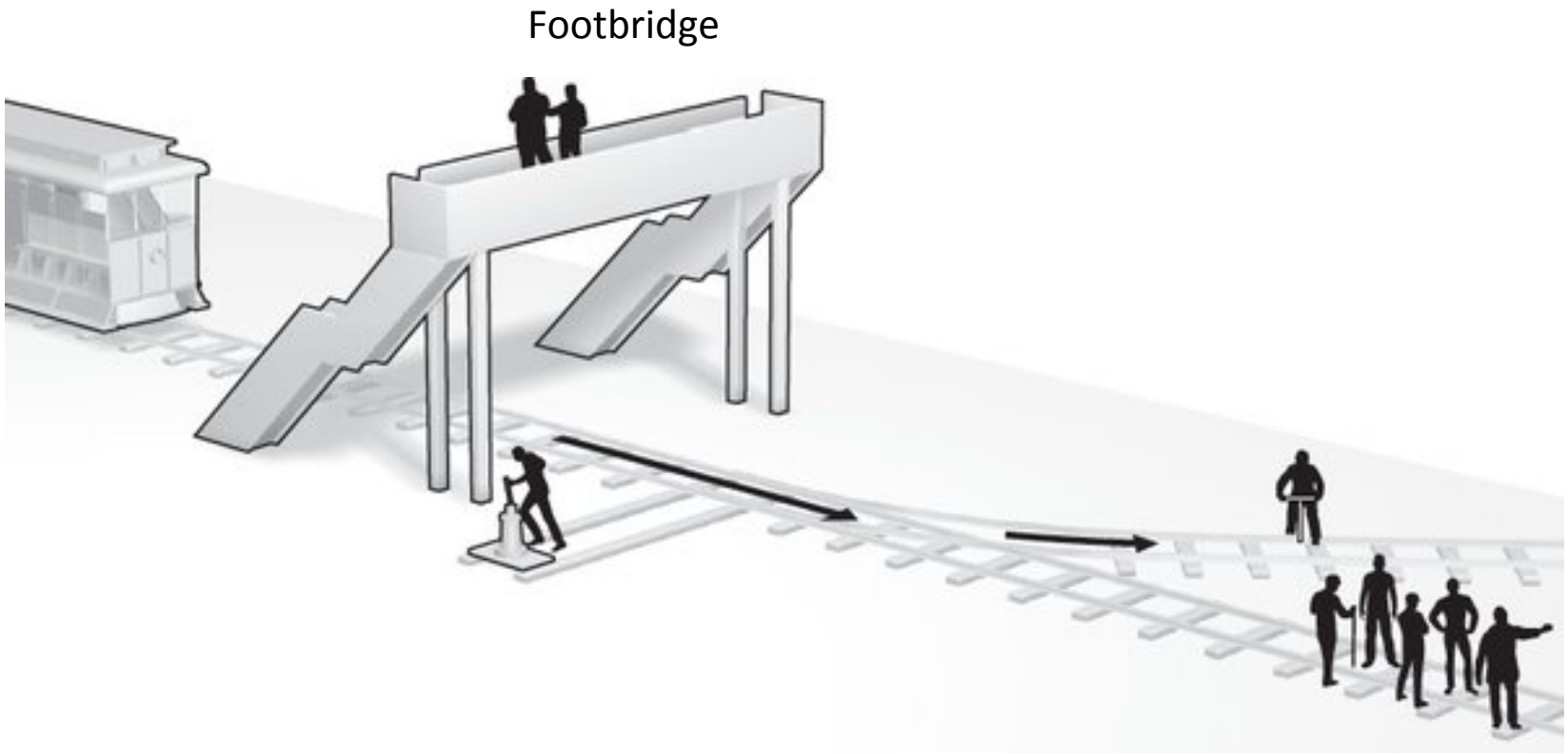
Theory, Evidence, and Relevance of Deontological Motivations

Daniel L. Chen

Moral Trolley Problem



Moral Trolley Problem



Consequentialism vs Deontology

- Social Preferences or Sacred Values?
  - ◂ Social Preferences or Sacred Values? Theory and Evidence of Deontological Motivations
- Experimental Methods
  - ◂ A Theory of Experiments: Invariance of Equilibrium to the Strategy Method of Elicitation
- oTree
  - ◂ oTree: An Open Source Platform for Online, Lab, and Field Experiments
- Political Attitudes
  - ◂ Who Cares? Measuring Attitude Strength in a Polarized Environment
- Multiculturalism and Integration
  - ◂ Non-Confrontational Extremists

Outline

Behavioral questions:

- Are there deontological motivations?
- How would deontological motivations be modeled?
- What implications do deontological motivations have for economics?
- What puzzles do they explain?
- What are the welfare implications?

Generations of Social/Moral Preferences

Domain of preferences:

- Agent's material consequences (homo oeconomicus)
- Agent's and others' material consequences (e.g. Fehr-Schmidt inequity aversion 1999, pure altruism, self-sacrifice, Rabin fairness 1994)
- Agent's and others' material consequences, and social audience (e.g. Andreoni impure altruism 1989, Sugden reciprocity 1984, McCabe et al signalling intentions 2003 or Benabou-Tirole type 2006, Battigalli et al guilt aversion psychological games 2007)
- Agent's and others' material consequences, social audience, and purely internal consequences (e.g., duty/deontological motivations, Smith's impartial spectator 1761)

Typology (Sobel 2005)

- Monetary payoffs of others to enter a decision-maker's utility.
- Chicago School (preferences over general commodities transformed into consumption goods)
- Identity models (utility function over actions and an identity that incorporates the prescriptions that indicate the identity-appropriate behavior Akerlof-Kranton 2000)
- However
  - Agents choose between quantities (in Chicago models)
    - but do not have preferences over *choices* separate from preferences over *quantities*
  - Agents choose acts (in identity models)
    - but do not have preferences over *acts* separate from preferences over *consequences of acts*.

Research Problem

Hypothetical vs. Categorical Imperative

economic models have thus far focused on the *hypothetical imperative*–preferences over acts because of their consequences–rather than the *categorical imperative*–preferences over acts regardless of their consequences

Research question

"Do people have deontological (duty-based) motivations?"

Revealed preference approach as opposed to surveys, vignettes, priming.

Problem:

Consequentialist and deontological motivations are hard to distinguish in normal circumstances. We identify non-consequentialism by varying the probability of a decision being consequential.

Perceived legitimacy motivates obedience to rules irrespective of likelihood of sanction (Tyler 1997)

Kant

Moral problem:

Your friend is hiding in your house from a murderer. The murderer arrives and asks you whether your friend is hiding in your house. Assuming you cannot stay silent, should you lie or tell the truth?

Kant (1793) says you must not lie. ◂ Kant considers uncertainty

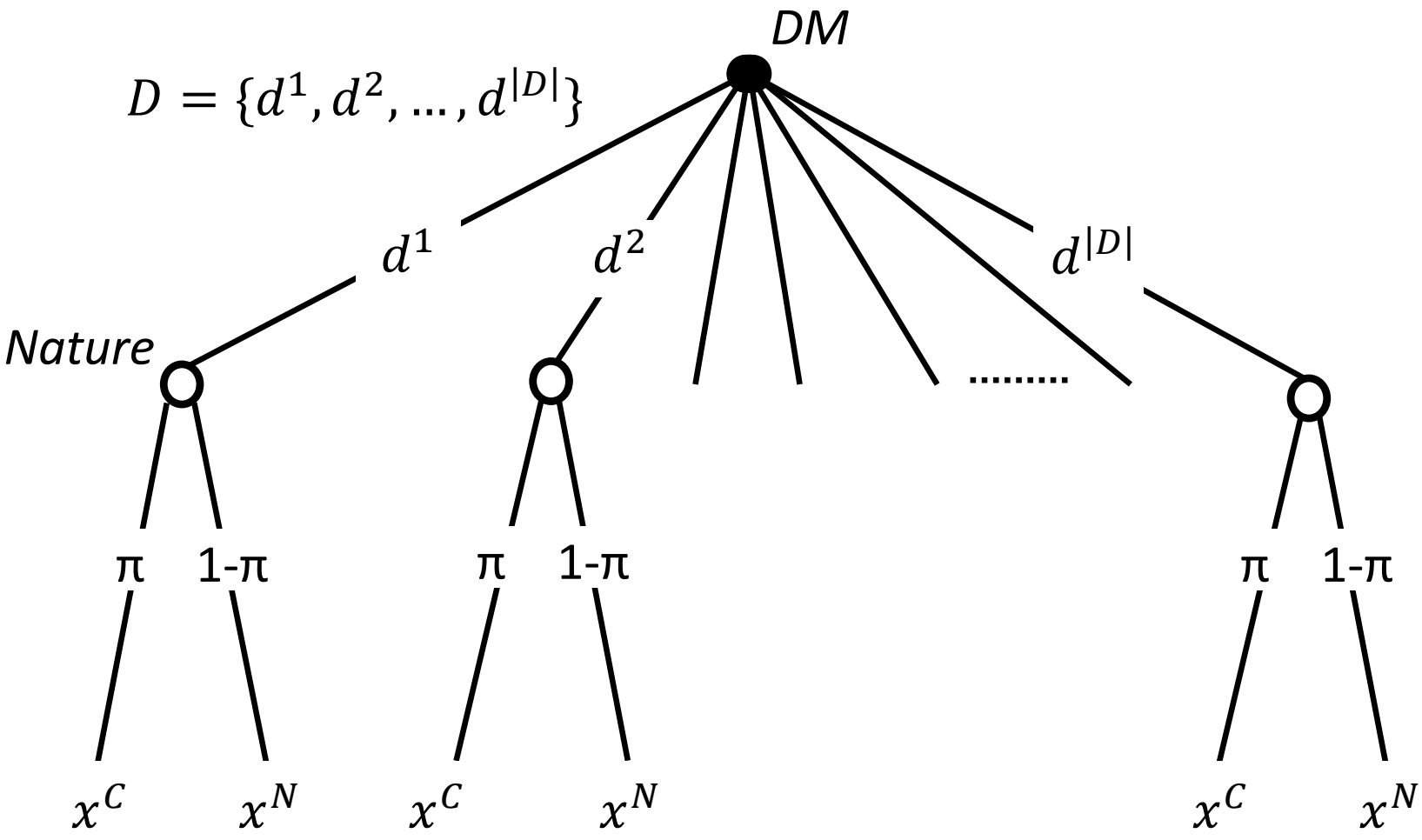
Note:

We are talking here about Kant as a non-consequentialist. The categorical imperative is about what the duties are.

Our model unlinks decisions from consequences

- Thought experiment
  - Separate decision d from the consequences x
- Anything that is a function of x is an outcome (reputation, inferences by other player about $DM's$ intention, etc.)
  - x is a function of the state of nature and decision d
  - In state C, d becomes common knowledge
  - In state N, d remains unknown to anyone except DM
- Consequentialism: Preferences are over lotteries
- Deontological motivations: d matters per se, even in state N

General Idea to detect Kantian duty



$x^C = x^C(d), x^N = constant$

Consequentialist is Invariant to π

Expected Utility (EU) intuition

$$E[u(x_1, x_2)] = \pi u(x_1^C, x_2^C) + (1-\pi)u(x_1^N, x_2^N)$$

- *DM* maximizes the objective function given π.
- One choice variable d.
- Indirect objective function: $V(d) = \pi u(\omega - d, d) + (1-\pi)u(\omega - \kappa, \kappa)$
- $V(d)$ is proportional to $u(\omega - d, d)$
- $\Rightarrow \frac{\partial d^*}{\partial \pi} = 0$

Consequentialist-deontological preferences

Consider preferences over x_1, d:

- $u(x_1, d) = f(x_1) + b(d)$; f, b strictly concave

Objective function

- $E[u(x_1, d)] = \pi(f(x_1^C) + b(d)) + (1-\pi)(f(x_1^N) + b(d))$

Indirect objective function

- $V(d) = \pi f(\omega - d) + (1-\pi) f(\omega - \kappa) + b(d)$

FOC

- $\frac{\partial V(d)}{\partial d} = -\pi f_1(\omega - d) + b_1(d) = 0$

Second derivative

- $\frac{\partial^2 V(d)}{\partial d^2} = \pi f_{11}(\omega - d) + b_{11}(d) < 0$

Using the FOC, by the implicit function theorem

- $\frac{\partial d^*}{\partial \pi} = \frac{f_1(\omega - d^*)}{\pi f_{11}(\omega - d^*) + b_{11}(d^*)} < 0$

Direction of change gives insight into location of the optimand for duty.

Economic methods rely on invariance

- Random lottery incentive (subjects make many choices, but only one of them is implemented)
- Strategy method (subjects make many choices, but only a fraction of decisions count for pay)
- BDM willingness-to-pay elicitation (subjects report a price that is implemented if it is higher than a randomly generated price)
- Contingent valuation (valuation of an environmental good in a hypothetical scenario)
- Market design data (subjects report preferences over, e.g. school, choices whose likelihood of being consequential varies)
- Vickrey auction (bidders submit written bids that are consequential if it is the highest bid)
- Survey data (subjects possibly misreport preferences if their decisions are not consequential)

Experimental evidence may be too prosocial. Not accounting for non-consequentialist motives can bias treatment effects. All economic methods to elicit choices assume that the choice does not enter the utility function separate from its consequences.

Two Field Analogs

- d = Bone marrow donotion sign-up
  - π varies by need & genetic match (Bergstrom et al. 2009)
  - Decision to sign-up to be a bone marrow donor is 3.5 times more likely, as probability your decision is implemented falls (from .3 to .1%)

- d = Not abort a fetus with Down Syndrome
  - π varies by survey time & genetics (prospective vs. likely vs. 100%) (Choi et al. 2012)
  - 60% less likely to abort when decision to abort is hypothetical
- d not irrevocable, not anonymous, π not exogenous

Invariance Theorem

Theorem

If there exist $x, x', x'' \in X$ *and* $\pi \in (0;1]$ *such that* $\pi x + (1-\pi)x''$ $\succcurlyeq \pi x' + (1-\pi)x''$, *and if DM satisfies the assumptions Preference Relation, FOSD (and Strict FOSD), then for all* $\pi' \in (0;1]$ *:*
$\pi' x + (1-\pi')x'' \succcurlyeq \pi' x' + (1-\pi')x''$

1. ◂ Proof Graphically
2. ◂ Proof
3. ◂ Theorem is not Savage's Sure Thing Principle
4. ◂ FOSD is weaker than Independence
5. ◂ Strict FOSD does not imply FOSD

Deontologicalism

"deontological moralities, unlike most views of consequentialism, leave space for the supererogatory. A deontologist can do more that is morally praiseworthy than morality demands. A consequentialist cannot. For the consequentialist, if one's act is not morally demanded, it is morally wrong and forbidden. For the deontologist, there are acts that are neither morally wrong nor demanded." (Stanford Encyclopedia of Philosophy)

This can be formalized as a lexicographic preference, with deontological before consequentialist motivations.

Definition: Deontological Preference

A preference is called *deontological* if there exist u, f such that $u = u(d)$, and $f = f(x)$, and f.a. $(x,d),(x',d')$: $(x,d) \succsim (x',d')$ if and only if $u(d) > u(d')$ or $[u(d) = u(d')$ and $f(x) \geqq f(x')]$.

For purely deontological preferences the optimal decision is constant in π

Can d^* vary in π?

- Under both consequentialist and deontological preferences, d^* invariant to probability.

Definition: Consequentialist-Deontological Preference

A preference is called *consequentialist-deontological* if there exists a utility representation u such that $u = u(x, d)$.

Confounds

Ex-ante fairness

What if people value some kind of ex-ante fairness?

Utility function for ex-ante fairness:

$U = f(E[u(x_1)], E[\widetilde{u}(x_2)])$

- $u, \tilde{u}$ strictly increasing, concave
- $E[u(x_1)], E[\widetilde{u}(x_2)]$ are normal goods

Fact:

If the *DM* maximizes ex-ante fairness then the sign of $\frac{\partial d^*}{\partial \pi}$ is the same as that of $\kappa - d^*$.

- ◂ Cognition Costs
- ◂ Revealed preference and psychological method to distinguish between internal consequences

Evidence

- Lab Experiment
  - ◂ Subjects became 50% more charitable when the decision was hypothetical.
- Online Experiment
  - ◂ Subjects became 33% more charitable as the decision became hypothetical.
- Structural Estimation
  - ◂ Using variation in our data generated by the experiment

Congratulations, you did well!

Therefore you have earned CHF20.00 to be divided between the charity *Doctors Without Borders* and you.

Look at the paper in the envelope on your desk. To indicate your donation decision mark the appropriate box, then seal the envelope.

After everyone has made their decision, we will spin the wheel of fortune. Depending on where it stops your decision will be implemented or not:

If the wheel stops on 1, 10 or 13:
Your envelope will be brought into the experimenter's room and **opened** there. The **donation** will be made **according to your instructions.**

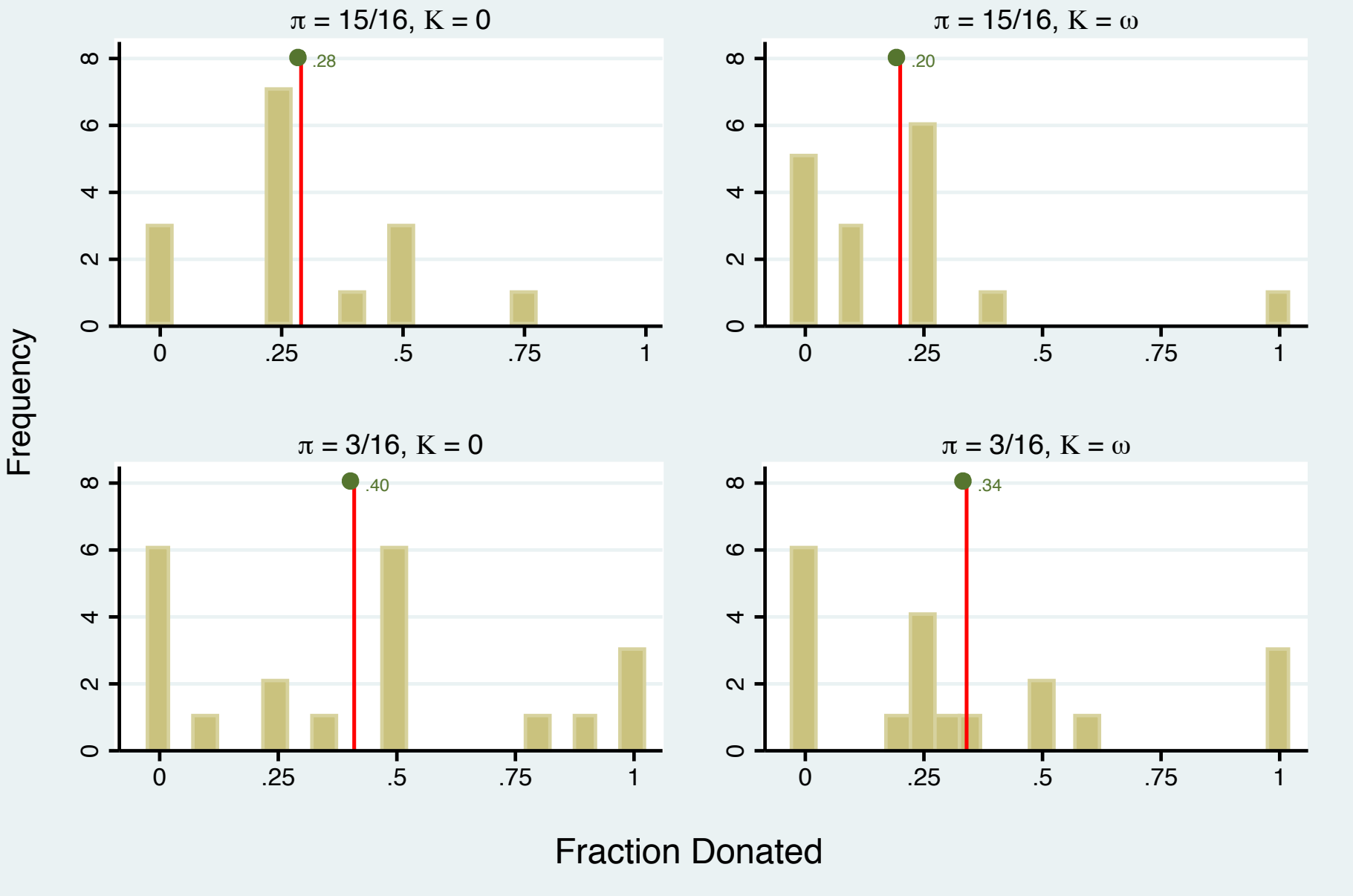
If the wheel stops on 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16:
Your envelope will be **shredded. No one will ever know how much you wanted to donate.** Instead, by default, the charity gets CHF0.00 and you get CHF20.00.

Please make your choice now, mark the appropriate box on the paper, place the paper in the envelope, and seal the envelope.

After you have sealed the envelope, please press OK.

OK

Donation (Shredding)

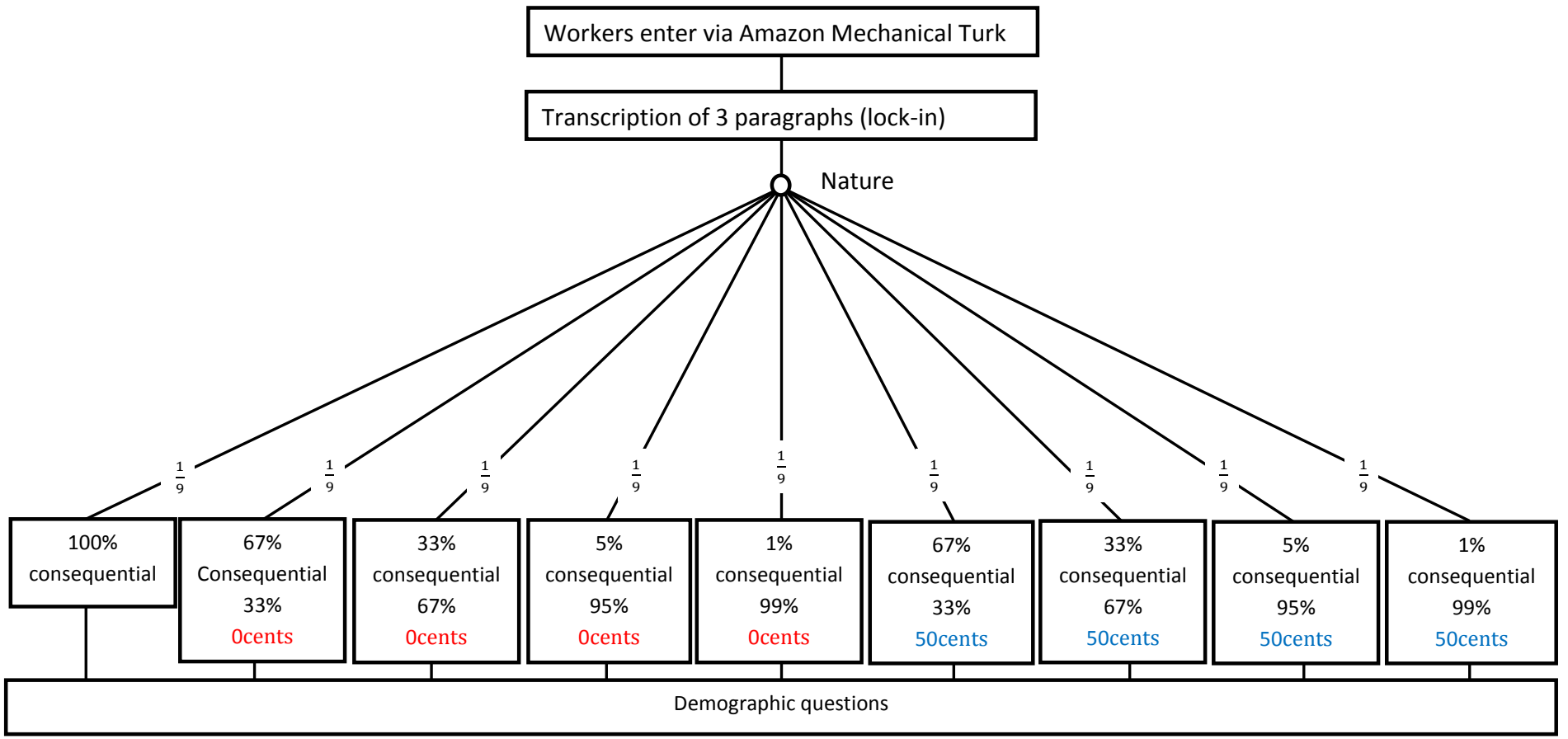


Recall: ex-ante consequentialism => sign $\frac{\partial d^*}{\partial \pi}$ = sign $(\kappa - d^*)$; N = 71 (of 264 invited)

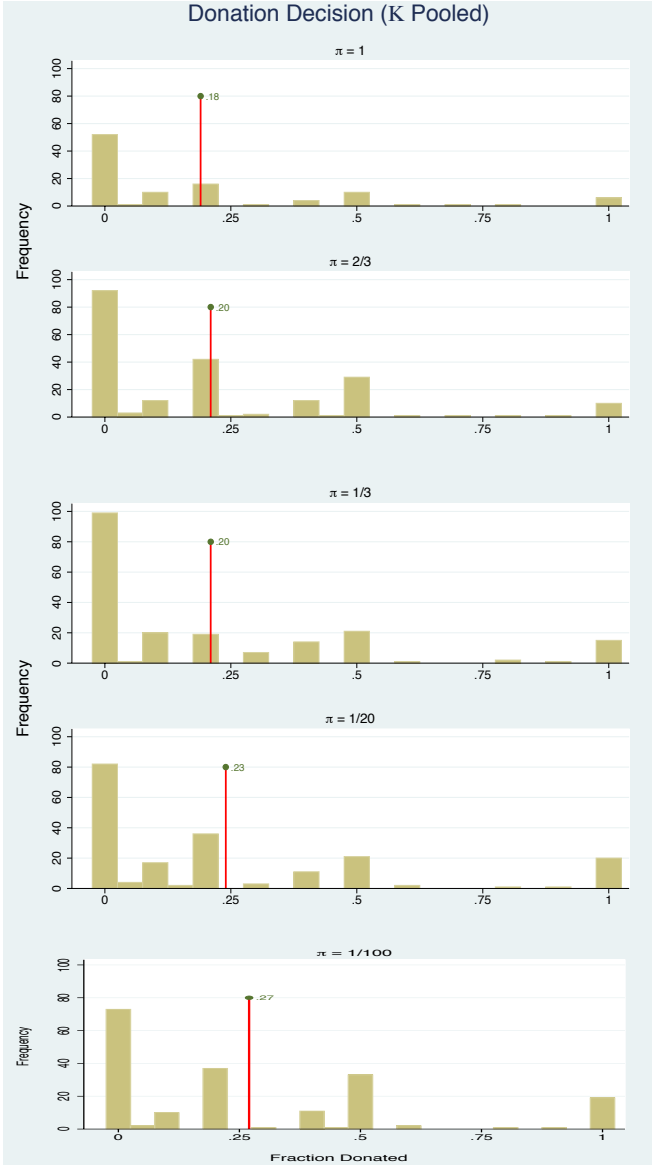
Additional Checks

- ◂ Not targeting expected income of recipient
- ◂ Not targeting expected giving
- ◂ Regression and non-parametric

Donation (MTurk)



Donation (MTurk)



Red: Mean

N = 900

Homo oeconomicus

Example

Homo oeconomicus and Duty Bliss point

$u(x_{1,}, d) = \lambda(x_{DM}) + (1-\lambda)(-(\delta - d)^2) = \lambda(\omega - d) + (1-\lambda)(-(\delta - d)^2)$

FOC: $0 = \pi\lambda(-1) + 2(1-\lambda)(\delta - d)$

Linear: $d^* = \frac{-\lambda}{2(1-\lambda)}\pi + \delta$

Estimate of -0.073 implies that $\lambda = 0.13$.

Small weight on homo oeconomicus is intuitive since many donate more than the bliss point of 25%

Fehr-Schmidt

Example

Fehr-Schmidt preferences and deontological bliss point

$u(x_{1,}, x_2, d) =$
$\lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (1-\lambda)(-(\delta - d)^2).$

Linear: $d^* = \frac{\lambda(-2\alpha-1)}{2(1-\lambda)}\pi + \frac{\lambda(\alpha+\beta)}{(1-\lambda)}\pi 1[\frac{\omega}{2} > d] + \delta$

Fehr-Schmidt

Table 8: Trading Off Consequentialist-Deontological Motivations (AMT Experiment)

	OLS	IV	IV
	(1)	(2)	(3)
	Decision (d)		
Mean dep. var.	0.23		
% Consequential (π)	-0.239***	-0.363***	-0.368***
	(0.0249)	(0.0548)	(0.139)
π * $1(d \geq w/2)$	0.870***	1.516***	1.542**
	(0.0412)	(0.250)	(0.714)
Constant (Duty Bliss Point)	0.251***	0.249***	0.249***
	(0.0116)	(0.0131)	(0.0134)
IV	N	π, Indian	π, Age $\leq$ 25
Observations	902	902	902
R-squared	0.336	0.155	0.140

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Bliss point is still 25%

The interpretation of the reduced form coefficients is sensible, but our coefficients also have a structural interpretation.

Fehr-Schmidt

Example

Fehr-Schmidt preferences and deontological bliss point

$u(x_{1,}, x_2, d) =$
$\lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (1-\lambda)(-(\delta - d)^2)$.

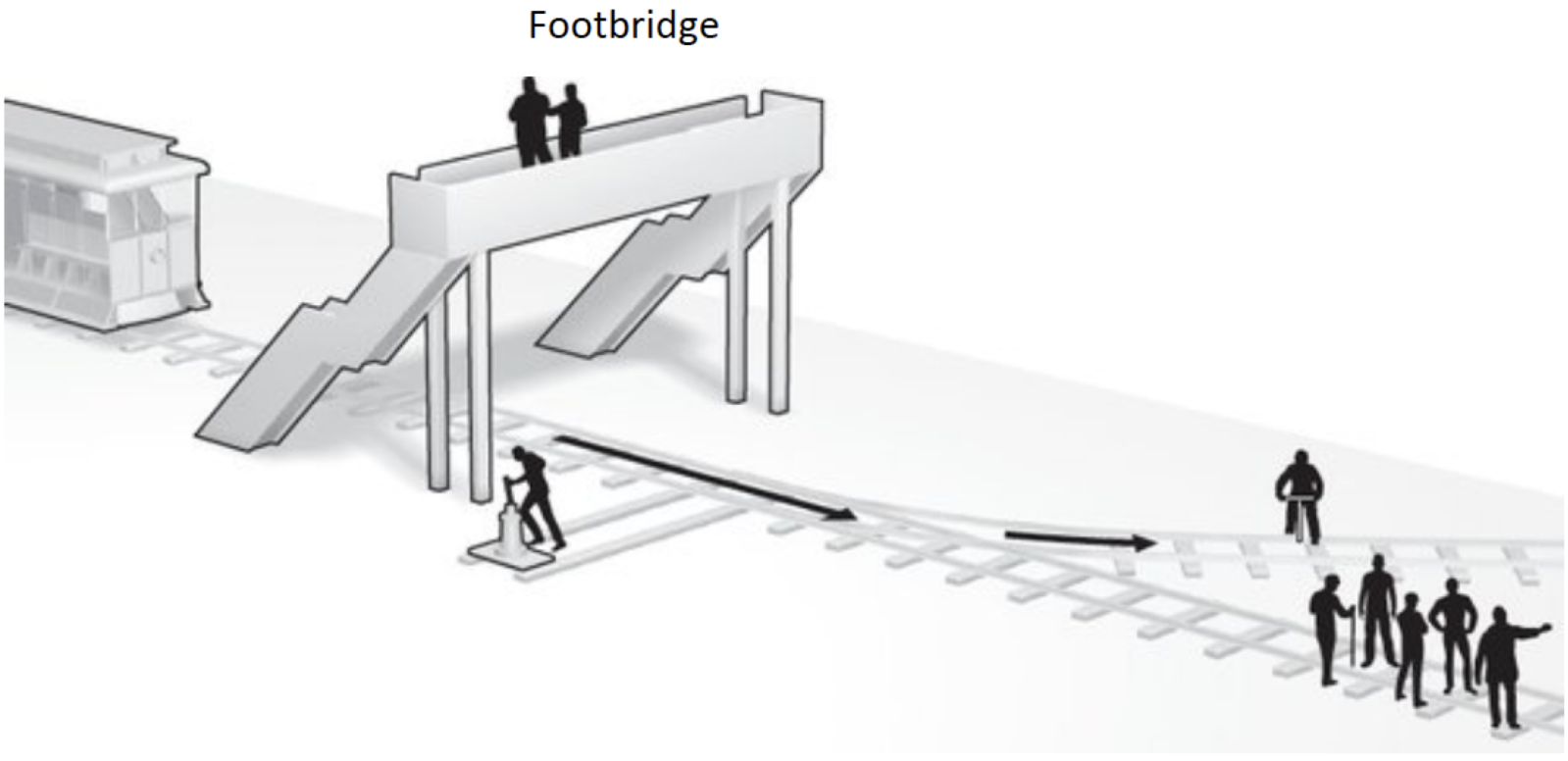
$\frac{\lambda(2\beta-1)}{2(1-\lambda)} = -0.36$ (I) and $\frac{\lambda(-2\alpha-1)}{2(1-\lambda)} = 1.16$ (II). 2 equations and 3 unknowns.
For a range of plausible values of $\alpha > \beta > 0$, the variance with implementation probability in our data is inconsistent with the joint hypothesis of consequentialist motivations being Fehr-Schmidt and a non-zero weight on consequentialist motivations.

Intermediate Signpost

- Theory
  - Formal interpretation of major moral philosophies
    - Approximation of lexicographic = concave cost of deviating from duty
      ("what the hell" experiments -> "cave in" on duties when pressure is high)
- Method
  - Revealed preferences method to detect deontological motivations
    - Gambler's fallacy, negatively autocorrelate trying to be fair
  - The direction of the decision changes gives insight into the location of the optimand for one's "greatest" duty
- Evidence
  - Suggests both consequentialist and deontological motivations

Consequentialism vs. Deontology

Moral Trolley Problem



Screen for deontological motivations in business leaders, politicians, or judges (Besley 2005)

Criticized in optimal policy design as necessarily harming some individuals (Kaplow et al 2006)

Intermediate Signpost

- Relevance for economics
  - Contingent valuation; psychological vignettes $\precsim$ consequential decisions
  - Negative framing
    - Methods like the random lottery method for moral decisions may reveal decisions that are 'too' moral (and treatment effects may be larger)
  - Positive framing
    - Strategy method incidentally turns out to be a method to investigate non-outcome based preferences
  - Relevance for policy
    - *mens rea* (intention) vs. *actus reus* (outcome) when it matters
    - moral rights in copyright (litigate for non-consequential reasons)
  - Welfare economics with 'sacred values'
    - Multiculturalism, conservative and liberal societies, integration
    - Concave costs silence undercurrents of dissent (despite free speech)

Kant on lying

- Kant does discuss uncertainty
- Kant says one must not lie (one may remain silent)

"It is indeed possible that after you have honestly answered Yes to the murderer's question as to whether the intended victim is in the house, the latter went out unobserved and thus eluded the murderer, so that the deed would not have come about. However, if you told a lie and said that the intended victim was not in the house, and he has actually (though unbeknownst to you) gone out, with the result that by so doing he has been met by the murderer and thus the deed has been perpetrated, then in this case you may be justly accused as having caused his death. For if you had told the truth as best you knew it, then the murderer might perhaps have been caught by neighbors who came running while he was searching the house for his intended victim, and thus the deed might have been prevented. [...] ***To be truthful (honest) in all declarations is, therefore, a sacred and unconditionally commanding law of reason that admits of no expediency whatsoever."***

- Kant (1799) : "On a supposed right to lie because of philanthropic concerns". ◂ Kant

Invariance Theorem

Theorem

If there exist $x, x', x'' \in X$ *and* $\pi \in (0;1]$ *such that* $\pi x + (1-\pi)x'' \succcurlyeq \pi x' + (1-\pi)x''$, *and if DM satisfies the assumptions Preference Relation, FOSD (and Strict FOSD), then for all* $\pi' \in (0;1]$ *:*
$\pi' x + (1-\pi')x'' \succcurlyeq \pi' x' + (1-\pi')x''$

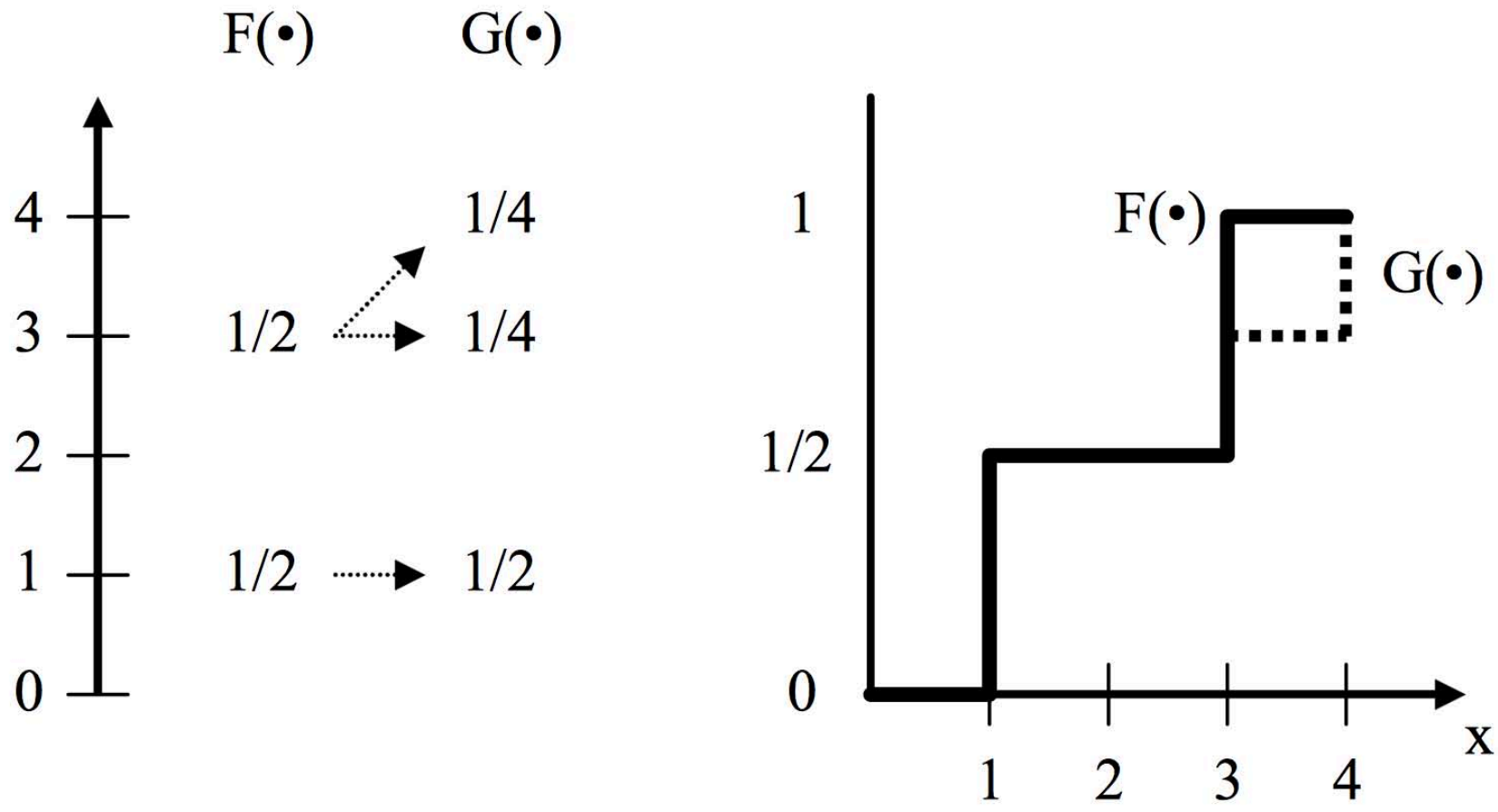
Proof.

(i) $x \succsim x'$: Suppose not, then $x' \succ x$, and therefore $\pi x' + (1-\pi)x''$ strictly FOSD $\pi x + (1-\pi)x''$. Then by Strict FOSD,
$\pi x' + (1-\pi)x'' \succ \pi x + (1-\pi)x''$, a contradiction.
(ii) Since $x \succsim x'$, $\pi' x + (1-\pi')x''$ first-order stochastically dominates $\pi' x' + (1-\pi')x''$. Thus by FOSD $\pi' x + (1-\pi')x'' \succcurlyeq \pi' x' + (1-\pi')x''$.

◂ Invariance Theorem

■

First-Order Stochastic Dominance



G first-order stochastically dominates F with respect to $\succsim$ if for all x':
$\sum_{x:x' \succsim x} G(x) \leq \sum_{x:x' \succsim x} F(x)$

◂ Invariance Theorem

First-Order Stochastic Dominance

Definition

(FOSD) p first-order stochastically dominates q with respect to $\succsim$ if for all x': $\sum_{x:x' \succsim x} p(x) \leq \sum_{x:x' \succsim x} q(x)$.

Assumption: FOSD If p FOSD q with respect to $\succsim$, then $p \succsim q$.

Note: Social preferences not monotone, so we define FOSD with respect to ordering induced by preferences.

Strict FOSD

Definition

(Strict FOSD) p strictly FOSD q with respect to $\succsim$ if p FOSD q with respect to $\succsim$, and
there exists x': $\sum_{x:x' \succsim x} p(x) < \sum_{x:x' \succsim x} q(x)$.

Assumption: Strict FOSD If p strictly FOSD q with respect to $\succsim$, then $p \succ q$.

Notes:

- p strictly FOSD q, then p FOSD q.
- If p FOSD q but not strictly so, and q FOSD p but not strictly so, this does not imply $p = q$.
- Assumption Strict FOSD satisfied does NOT imply the assumption FOSD satisfied.

Strict and Weak FOSD

Example

Preferences that satisfy strict FOSD but violate weak FOSD.

- An indivisible treat that mom can allocate to one of her two children x or y. Mom would like to be exactly fair, thus her most preferred lottery is $(x;\frac{1}{2},y;\frac{1}{2})$, she is indifferent between all other lotteries. For all $\pi,\pi' \in [0;1] \setminus \frac{1}{2}$: $(x;\pi,y;1-\pi) \sim (x;\pi',y;1-\pi')$ and $(x;\frac{1}{2},y;\frac{1}{2}) \succ (x;\pi,y;1-\pi)$. These preferences are complete and transitive. Strict FOSD is trivially satisfied since there is no lottery that strictly first-order stochastically dominates another lottery.
- However, axiom WFOSD is violated: $(x;\frac{2}{3},y;\frac{1}{3})$ weakly first order-stochastically dominates $(x;\frac{1}{2},y;\frac{1}{2})$, but $(x;\frac{1}{2},y;\frac{1}{2}) \succ (x;\frac{2}{3},y;\frac{1}{3})$.
- When does strict FOSD imply weak FOSD?

Continuity and FOSD

Definition

$\succsim$ is **continuous** if for all $p, q, r \in P$ the sets $\{\alpha \in [0,1] : \alpha p + (1-\alpha)q \succsim r\}$ and $\{\alpha \in [0,1] : r \succsim \alpha p + (1-\alpha)q\}$ are closed in [0,1].

Note that: $\{\alpha \varepsilon [0,1] : x \succsim \alpha x + (1-\alpha)y\} = [0; \frac{1}{2}) \cup (\frac{1}{2}, 1]$. But continuity is not enough.

Example

Again mom would like to be fair, but now between two unfair lotteries she prefers the one that is more fair. For all $\pi, \pi' \in [0;1]$:
$\pi \cdot (1-\pi) \geq \pi' \cdot (1-\pi')$ if and only if $(x; \pi, y; 1-\pi) \succsim (x; \pi', y; 1-\pi')$.
Strict FOSD and continuity are satisfied. But WFOSD is violated:
$(x; \frac{2}{3}, y; \frac{1}{3})$ weakly first order-stochastically dominates $(x; \frac{1}{2}, y; \frac{1}{2})$, but $(x; \frac{1}{2}, y; \frac{1}{2}) \succ (x; \frac{2}{3}, y; \frac{1}{3})$.

Rich Domain, Continuity and FOSD

However, if there are also two outcomes $x, y \in X$ such that $x \succ y$, then strict FOSD implies weak FOSD.

Proof.

Suppose p weakly first-order stochastically dominates q. We need to show that $p \succsim q$.
Suppose not, that is $q \succ p$. Since X is finite there exists an $\overline{x}$, $\underline{x}$ such that for all x: $\overline{x} \succsim x$, and an $x \succsim \underline{x}$. By RICH $\overline{x} \succ \underline{x}$.
At least one of the following three cases is satisfied: (i) $\overline{x} \succ q$, (ii) $p \succ \underline{x}$ or (iii) $q \succsim \overline{x} \succ \underline{x} \succsim p$.
(i) Since p weakly first-order stochastically dominates q, and $\overline{x} \succ q$, for any $\alpha > 0$ the lottery $\alpha\overline{x} + (1-\alpha)p$ strictly first-order stochastically dominates q. But then $\{\alpha : \alpha\overline{x} + (1-\alpha)p \succsim q\} = (0,1]$, a violation of continuity. (ii) and (iii) similarly.

■

◂ Invariance Theorem

FOSD is not Sure-Thing Principle

Savage's Sure-Thing Principle is not FOSD. If invariant, then probability does not matter.

7 The sure-thing principle

A businessman contemplates buying a certain piece of property. He considers the outcome of the next presidential election relevant to the attractiveness of the purchase. So, to clarify the matter for himself, he asks whether he would buy if he knew that the Republican candidate were going to win, and decides that he would do so. Similarly, he considers whether he would buy if he knew that the Democratic candidate were going to win, and again finds that he would do so. Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event obtains, or will obtain, as we would ordinarily say. It is all too seldom that a decision can be arrived at on the basis of the principle used by this businessman, but, except possibly for the assumption of simple ordering, I know of no other extralogical principle governing decisions that finds such ready acceptance.

Machina and Schmeidler (1992) formulation of Sure-Thing Principle is about $\frac{\partial d^*}{\partial \kappa} = 0$.

◂ Invariance Theorem

FOSD is not Independence

(IND) $\succsim$ satisfies independence if for all lotteries p, q, r in P : $p \succsim q \Leftrightarrow \alpha p + (1-\alpha) r \succsim \alpha q + (1-\alpha) r$.

Example

Cumulative Prospect Theory (Rank Dependent Expected Utility) satisfies FOSD and allows for Allais Paradox, but not Independence.

If the cardinality of the outcome space is 2, that then independence is as weak an axiom as first-order stochastic dominance.

◂ Invariance Theorem

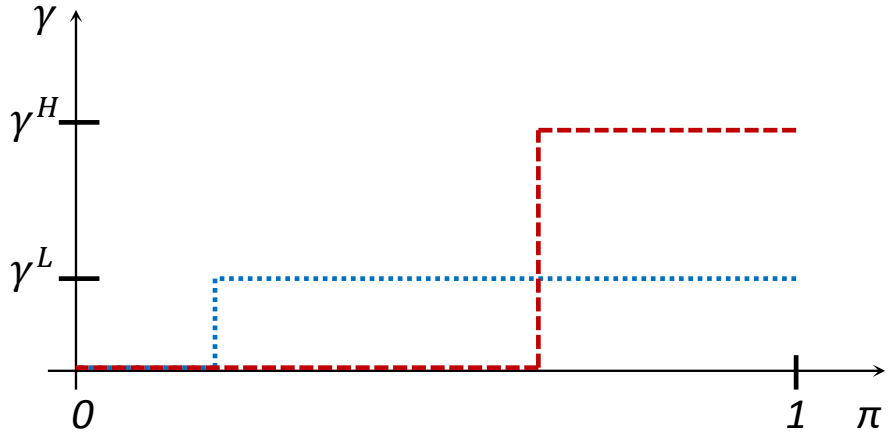
Non-Additive Utility

- It is possible to generate examples where $\frac{\partial d^*}{\partial \pi} > 0$
- Let $u = u(x_1, d)$, $u_1, u_2 > 0$ and $u_{11}, u_{22} < 0$ (risk-aversion).
- For sufficiently negative $u_{12}(\omega - d, d)$ we can get $\frac{\partial d^*}{\partial \pi} > 0$.
- SOC: need u_2 sufficiently positive and sufficiently negative u_{22}.
- But these are not interpretable under uncertainty. u_{12} can easily change sign if, for example, you take the log or square or other strictly monotone transformation of the utility function.

◂ Consequentialist-deontological preferences

Cognition Costs

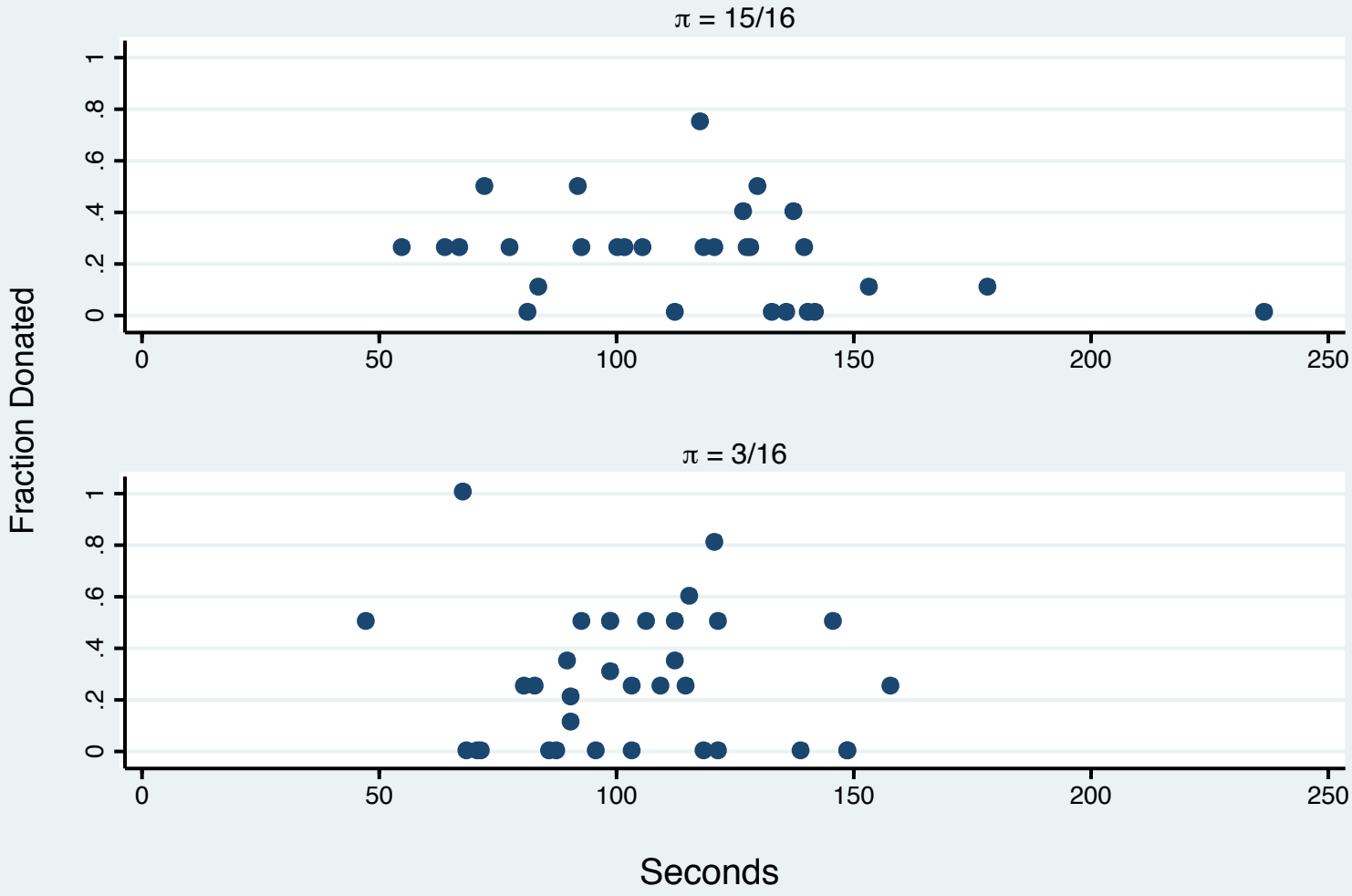
The *DM* can compute the optimal decision, but to do so, she incurs a cognition cost $\gamma \geq 0$, or otherwise she can make a heuristic (fixed) choice $\bar{d}$ for which (normalized) costs are 0.



Note:

Cognition cost model predicts that time spent on the survey also changes as d changes with π.

Time Spent

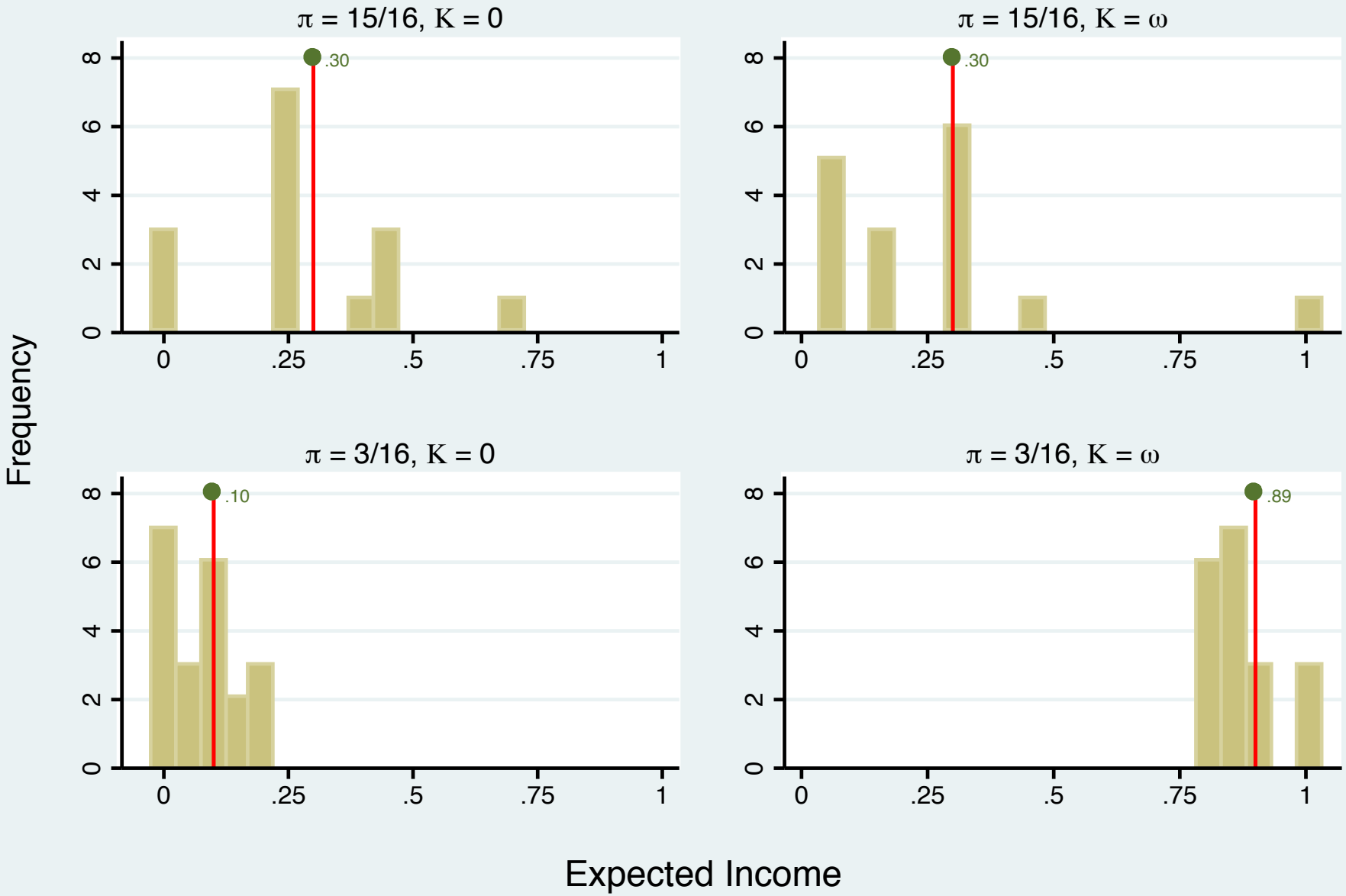


◂ Confounds

Revealed preference method cannot distinguish different internal "consequences"

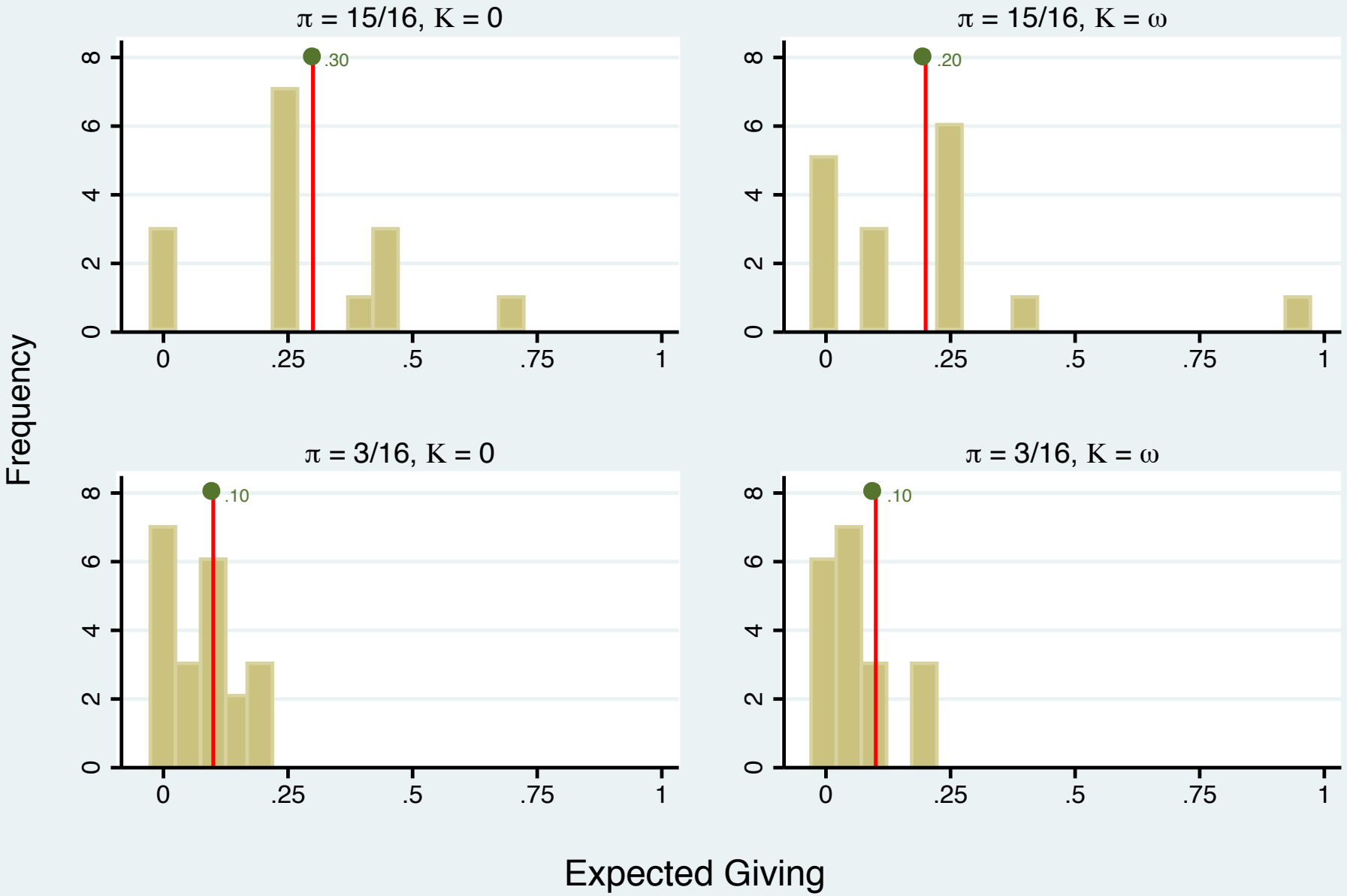
- One response: it may be a semantic difference
  - Even purely deontological preferences likely have some neurobiological "consequence"
  - What drives deontological motivations, is it conscience or guilt or ?
- Second response: Is self-image long-term?
  - Self-signaling (Benabou and Tirole 2006)
  - Perhaps distinguish through forgetting or cognitive load
  - or faster speeds (Spontaneous Giving and Calculated Greed; Rand et al 2012)
- Third response: Shredding removes the experimenter
  - Foreigner presence increased generosity by 20% (Cilliers, Dube, Siddiq 2015)
- ◂ Confounds

Expected Income E(x_2) (by κ)



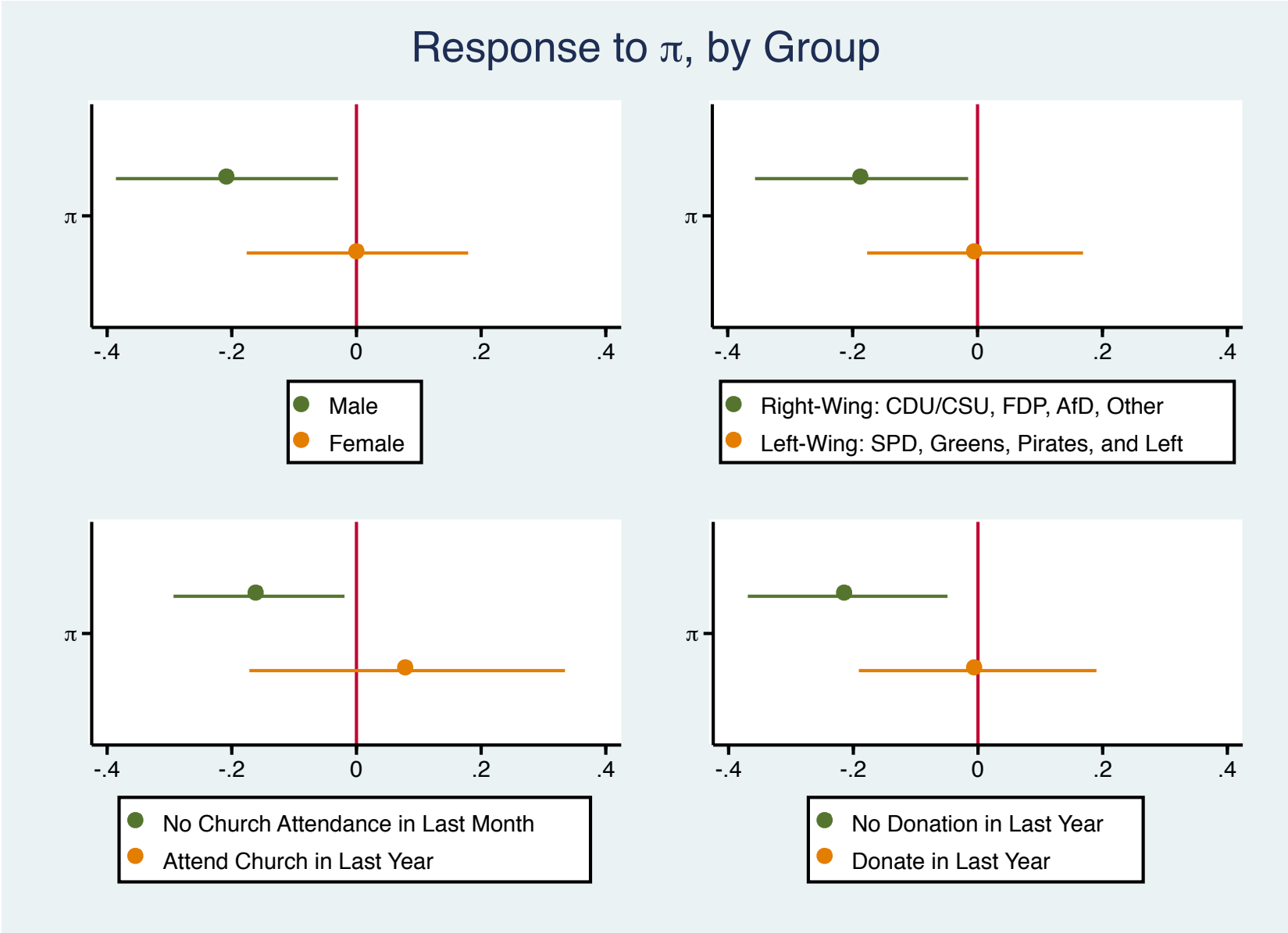
Recall: ex-ante fairness implies sign $\frac{\partial d^*}{\partial \pi} = \text{sign}\,(\kappa - d^*)$ ◂ Additional Checks

Expected Giving (πd^*) (by κ)



Recall: ex-ante => sign $\frac{\partial d^*}{\partial \pi}$ =sign $(\kappa - d^*)$ ◂ Additional Checks

Heterogenous Treatment



N = 173 (of 975 invited) ◂ Additional Checks

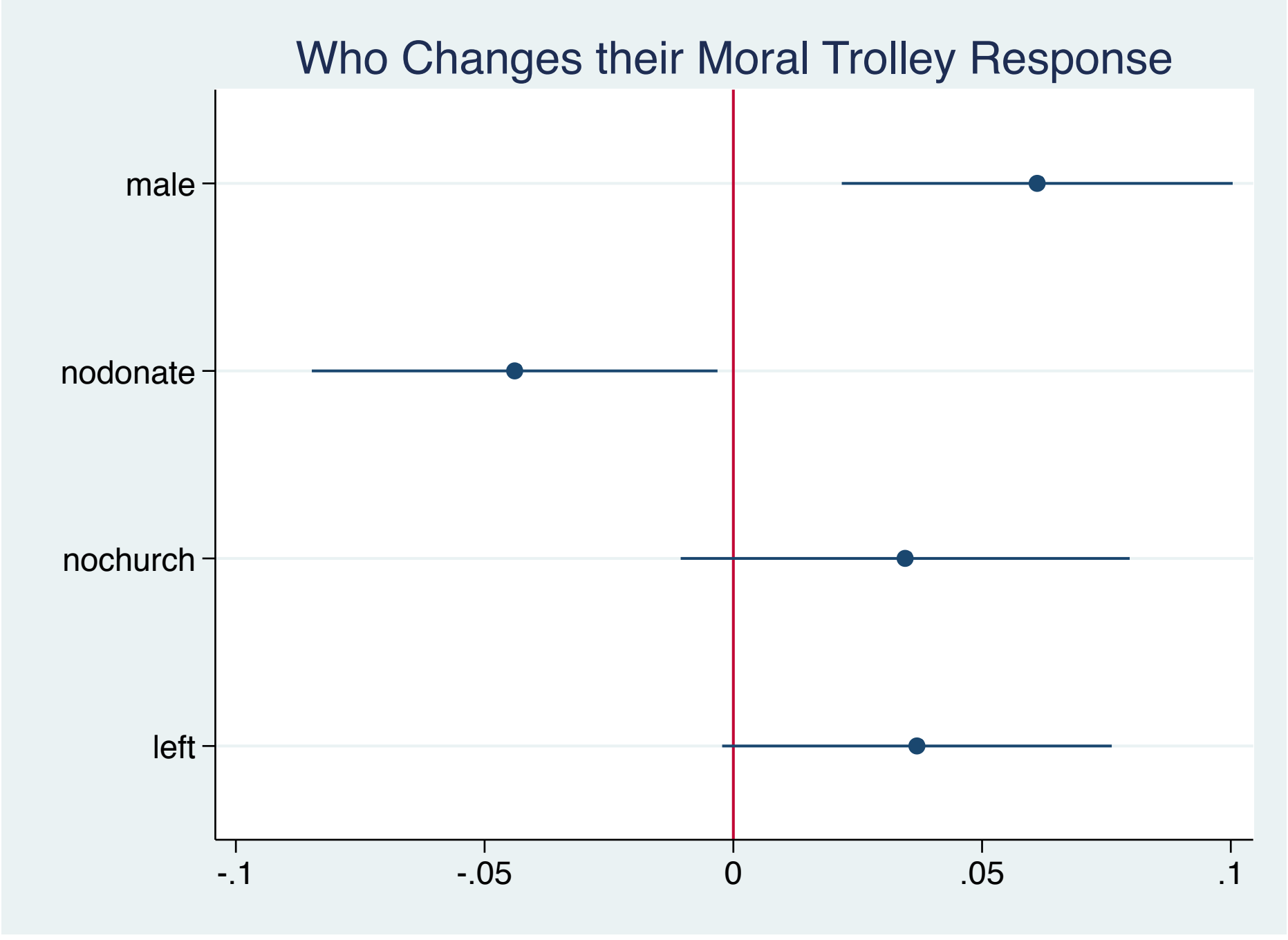
Moral Trolley

You are a solider and you and a few other soldiers are taken prisoner by the enemy. After a year in captivity, your group has tried to escape, but was caught. The enemy overseer decided to hang you and your group. At the gallows he loosens the noose around your neck and proclaims that if you pull the chair under one of the other soldiers in your group, you and the other **19 soldiers** in your group will be released. If you do not, everyone will be hung. The warden says this seriously and will keep his promise. Would you remove the chair in this situation?

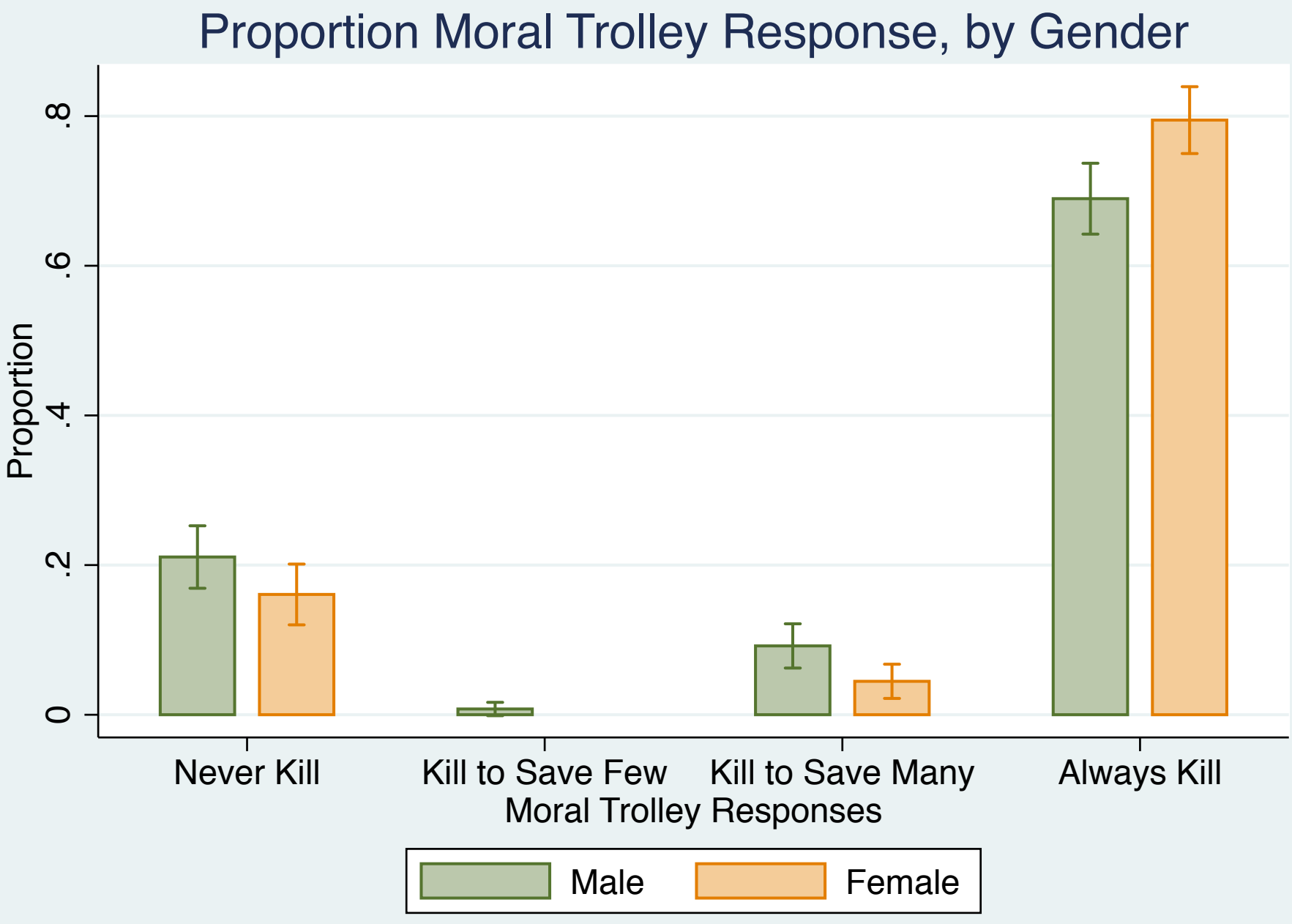
Moral Trolley

You are a solider and you and a few other soldiers are taken prisoner by the enemy. After a year in captivity, your group has tried to escape, but was caught. The enemy overseer decided to hang you and your group. At the gallows he loosens the noose around your neck and proclaims that if you pull the chair under one of the other soldiers in your group, you and the other **5 soldiers** in your group will be released. If you do not, everyone will be hung. The warden says this seriously and will keep his promise. Would you remove the chair in this situation?

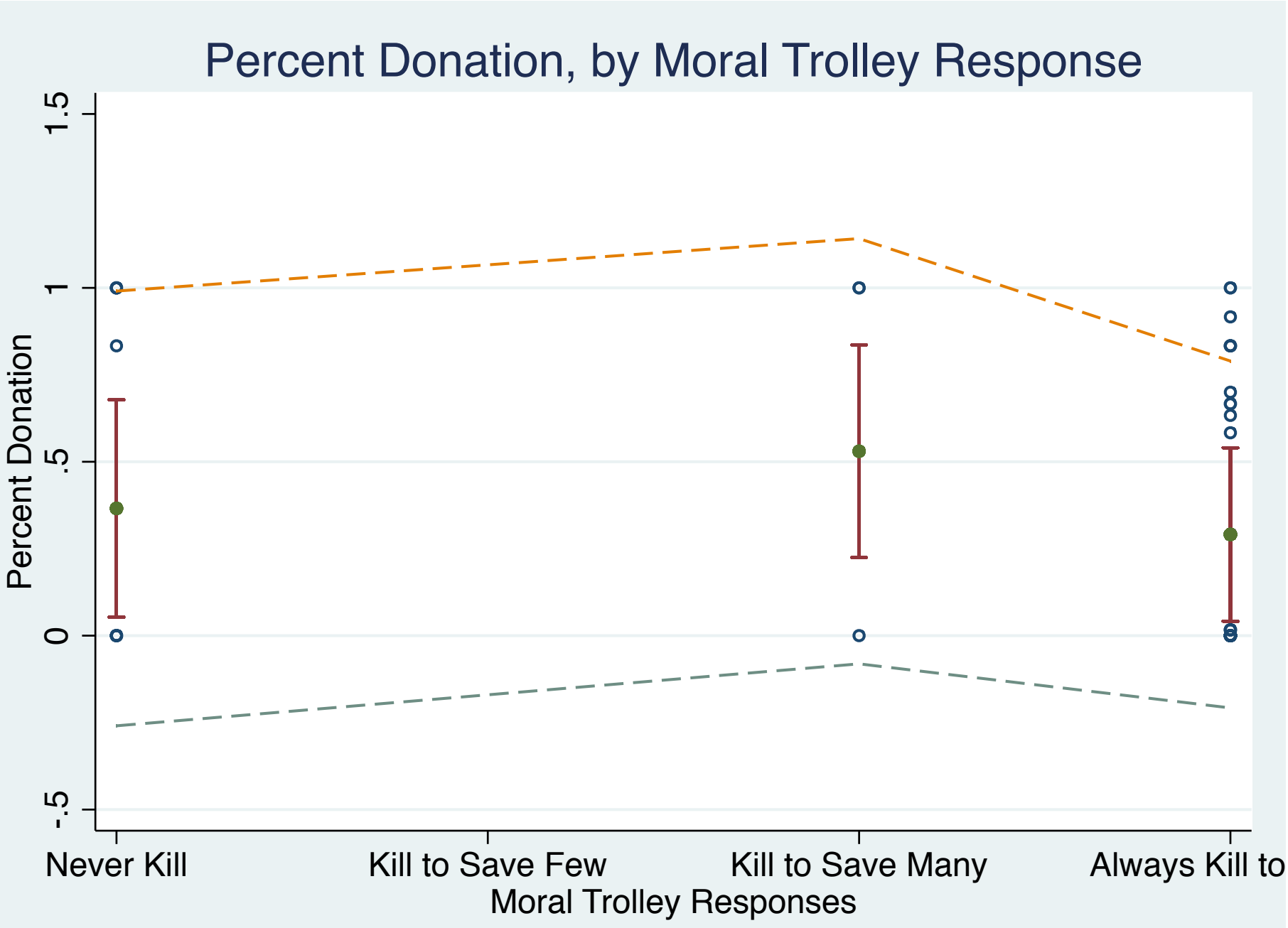
Moral Trolley



Moral Trolley



Moral Trolley



Econometric Specifications

- Basic regression: $d_i = \beta_0 + \beta_1 Treatment_i + \beta_2 X_i + \varepsilon_i$
- Wilcoxon test
- Sensitivity d_i^* to π as predicted from demographics
  - $d^*{}_i = \beta_0 \pi_i + \beta_1 \mathbf{X}_i \pi_i + \alpha \mathbf{X}_i + \varepsilon_i$
- Structural estimation (assume specific mixed preference and get estimates on relative importance)

Pooled Results

	Ordinary Least Squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected Income E(x_2)		Expected Giving (πd^*)	
Mean dep. var.	0.30		0.39		0.12	
% Consequential (π)	-0.176*	-0.159*	-0.259**	-0.278***	0.212***	0.219***
	(0.0978)	(0.0855)	(0.108)	(0.0802)	(0.0484)	(0.0452)
K Fixed Effects	N	Y	N	Y	N	Y
Observations	71	71	71	71	71	71
R-squared	0.045	0.292	0.077	0.506	0.218	0.339

Notes: Standard errors in parentheses. Raw data shown in Figures 4 and 5. * p < 0.10, ** p < 0.05, *** p < 0.01

◂ Conclusion

Pooled Results

	Ordinary Least Squares					
	(1)	(2)	(3)	(4)	(5)	(6)
	d^*		Expected Income E(x_2)		Expected Giving (πd^*)	
Mean dep. var.	0.23		0.34		0.07	
% Consequential (π)	-0.0725**	-0.0684*	-0.224***	-0.219***	0.194***	0.213***
	(0.0288)	(0.0390)	(0.0334)	(0.0299)	(0.0132)	(0.0181)
K Fixed Effects	N	Y	N	Y	N	Y
Controls	N	Y	N	Y	N	Y
Observations	902	900	902	900	902	900
R-squared	0.007	0.059	0.048	0.604	0.194	0.214

Notes: Standard errors in parentheses. Raw data shown in Figure 10. Controls include indicator variables for gender, American, Indian, Christian, Atheist, aged 25 or younger, and aged 26-35 as well as continuous measures for religious attendance and accuracy in the lock-in data entry task. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

◂ Conclusion

Disaggregated Results

	Ordinary Least Squares							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Decision (d)		Decision (d)		Decision (d)		Decision (d)	
	K = Unknown		K = 10¢		K = 0¢		K = 50¢	
Mean dep. var.	0.26		0.22		0.20		0.22	
% Consequential (π)	-0.0778	-0.0654	-0.0525	-0.0321	-0.0711	-0.0708	-0.0644	-0.0675
	(0.0523)	(0.0523)	(0.0526)	(0.0536)	(0.0464)	(0.0466)	(0.0462)	(0.0456)
Male		-0.0909**		-0.0474		0.0108		0.0178
		(0.0399)		(0.0430)		(0.0395)		(0.0362)
American		0.0241		-0.0539		0.0838		0.117*
		(0.0524)		(0.0539)		(0.0664)		(0.0598)
Indian		-0.0672		-0.0785		-0.0673		-0.0626
		(0.0566)		(0.0560)		(0.0630)		(0.0590)
Christian		-0.0295		0.0584		-0.0215		-0.000293
		(0.0483)		(0.0503)		(0.0494)		(0.0479)
Atheist		-0.0188		0.00480		0.0113		-0.0927
		(0.0644)		(0.0649)		(0.0802)		(0.0725)
Religious Services Attendance		-0.00614		0.000508		0.00367		-0.00546
		(0.0145)		(0.0156)		(0.0137)		(0.0137)
Ages 25 or Under		-0.0207		-0.122**		-0.0109		-0.113**
		(0.0518)		(0.0570)		(0.0493)		(0.0474)
Ages 26-35		0.00271		-0.110*		-0.00105		-0.111**
		(0.0548)		(0.0593)		(0.0493)		(0.0480)
Own Errors		-0.000192		-0.000186		0.000220		-0.000148
		(0.000193)		(0.000163)		(0.000194)		(0.000143)
Observations	260	260	218	218	256	255	271	270
R-squared	0.009	0.069	0.005	0.081	0.009	0.052	0.007	0.097

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Non-Parametric Tests

	Wilcoxon-Mann-Whitney 2-sided test (p-values)		
	(1)	(2)	(3)
Thresholds	K Unknown or 10¢	$K = 0$¢ or 50¢	K Pooled
$\pi = 1$ vs. $\pi \leq 0.67$	0.91	0.05	0.11
$\pi \geq 0.67$ vs. $\pi \leq 0.33$	0.07	1.00	0.20
$\pi \geq 0.33$ vs. $\pi \leq 0.05$	0.05	0.10	0.01
$\pi \geq 0.05$ vs. $\pi = 0.01$	0.15	0.02	0.01

	π Pooled
$K \geq 10$¢ vs. $K = 0$¢	0.40
$K = 50$¢ vs. $K \leq 10$¢	0.11

Non-Parametric Tests

	Non-parametric test for equality of medians, 2-sided test (p-values)
Thresholds	Pooled
$\pi = 3/16$ vs. $\pi = 15/16$	0.04
K = 0 vs. K = Max	0.01

Who responds to π?

	Ordinary Least Squares									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Decision (d)									
Mean dep. var.	0.23									
% Consequential (π)	-0.100**	-0.0493	-0.124**	-0.0500	-0.0522	-0.0774	-0.0618	-0.0548	-0.0839**	-0.0190
	(0.0494)	(0.0429)	(0.0506)	(0.0436)	(0.0403)	(0.0616)	(0.0467)	(0.0443)	(0.0407)	(0.126)
π * Male	0.0612									0.0490
	(0.0577)									(0.0611)
π * American		-0.0675								0.0370
		(0.0627)								(0.0911)
π * Indian			0.0990*							0.0426
			(0.0574)							(0.0963)
π * Christian				-0.0599						-0.0658
				(0.0632)						(0.0783)
π * Atheist					-0.133					-0.145
					(0.0837)					(0.108)
π * Religious Services Attendance						0.00394				-0.00739
						(0.0210)				(0.0224)
π * Ages 25 or Under							-0.0149			-0.0815
							(0.0576)			(0.0787)
π * Ages 26-35								-0.0386		-0.0878
								(0.0597)		(0.0808)
π * Own Errors									0.000402	0.000319
									(0.000299)	(0.000307)
K Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	900	900	900	900	900	900	900	900	900	900
R-squared	0.061	0.061	0.063	0.060	0.062	0.059	0.059	0.060	0.061	0.068

Notes: Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Homo oeconomicus

Example

Homo oeconomicus and Duty Bliss point

$u(x_{1,}, d) = \lambda(x_{DM}) + (1-\lambda)(-(\delta - d)^2) = \lambda(\omega - d) + (1-\lambda)(-(\delta - d)^2)$

FOC: $0 = \pi\lambda(-1) + 2(1-\lambda)(\delta - d)$

Linear: $d^* = \frac{-\lambda}{2(1-\lambda)}\pi + \delta$

Estimate of -0.073 implies that $\lambda = 0.13$.

Many people donate more than the bliss point of 25%

Fehr-Schmidt

Example

Fehr-Schmidt preferences and deontological bliss point

$u(x_{1,}, x_2, d) =$
$\lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (1-\lambda)(-(\delta - d)^2).$
GMM:
$E\left[\pi\left(1[\frac{\omega}{2} > d]\left[d - \pi\frac{\lambda(2\beta-1)}{2(1-\lambda)} - \delta\right] + 1[\frac{\omega}{2} \leq d]\left[d - \pi\frac{\lambda(-2\alpha-1)}{2(1-\lambda)} - \delta\right]\right)\right] = 0$

Linear: $d^* = \frac{\lambda(-2\alpha-1)}{2(1-\lambda)}\pi + \frac{\lambda(\alpha+\beta)}{(1-\lambda)}\pi 1[\frac{\omega}{2} > d] + \delta$

Fehr-Schmidt

Table 8: Trading Off Consequentialist-Deontological Motivations (AMT Experiment)

	OLS	IV	IV
	(1)	(2)	(3)
	Decision (d)		
Mean dep. var.	0.23		
% Consequential (π)	-0.239***	-0.363***	-0.368***
	(0.0249)	(0.0548)	(0.139)
π * $1(d \geq w/2)$	0.870***	1.516***	1.542**
	(0.0412)	(0.250)	(0.714)
Constant (Duty Bliss Point)	0.251***	0.249***	0.249***
	(0.0116)	(0.0131)	(0.0134)
IV	N	π, Indian	π, Age $\leq$ 25
Observations	902	902	902
R-squared	0.336	0.155	0.140

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Fehr-Schmidt

Example

Fehr-Schmidt preferences and deontological bliss point

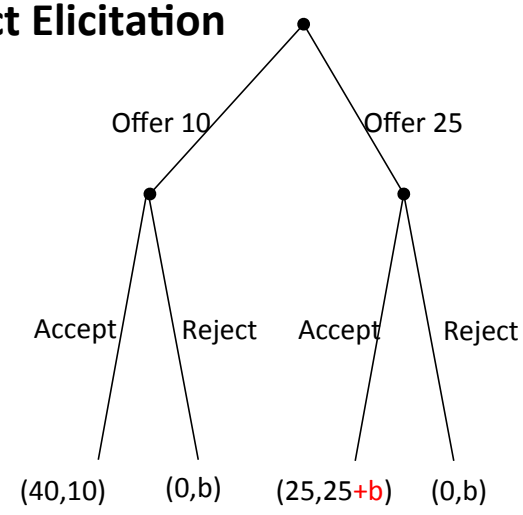
$u(x_{1,}, x_2, d) =$
$\lambda(x_1 - \alpha max\{x_2 - x_1, 0\} - \beta max\{x_1 - x_2, 0\}) + (1-\lambda)(-(\delta - d)^2)$.

$\frac{\lambda(2\beta-1)}{2(1-\lambda)} = -0.36$ (I) and $\frac{\lambda(-2\alpha-1)}{2(1-\lambda)} = 1.16$ (II). 2 equations and 3 unknowns.
Same problem: $\lambda < 0$ if $\alpha > \beta > 0$.
Linear form of Fehr-Schmidt leads to bang-bang solution. Either consequentialist donates nothing, but many people donate more than bliss point, or, consequentialist donates 50-50, which is too much so $\lambda < 0$.

Self-Image

Simplified ultimatum game

Direct Elicitation



Strategy Method

		x≥10 (AA')	x≥25 (RA')	(AR')	(RR')
p	10	(40,10)	(0,b)	(40,10)	(0,b)
1-p	25	(25,25)	(25,25+b)	(0,b)	(0,b)

Can we find an example of a non-consequentialist preference that predicts different outcomes under Direct Elicitation vs. Strategy Method?

Self-image: If I did not commit or in fact accept the unfair offer, I get an additional psychic benefit of 0<b<10.

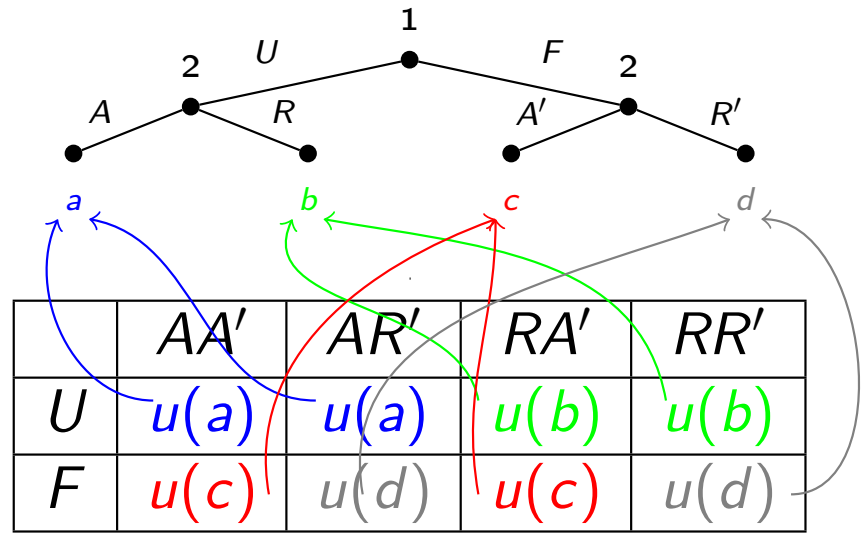
DE: offer 25 -> accept -> utilities (25,**25+b**)
DE: offer 10 -> accept -> utilities (40,10)

SM: strategy accept x≥10 yields: p*10 + (1-p)25 = 25-15p
SM: strategy accept x≥25 yields: p*(0+b) + **(1-p)(25+b)** = 25+b-25p

Act on self-image (x≥25) **iff** p<0.1b (**low prob of bearing consequences of obtaining self-image**)

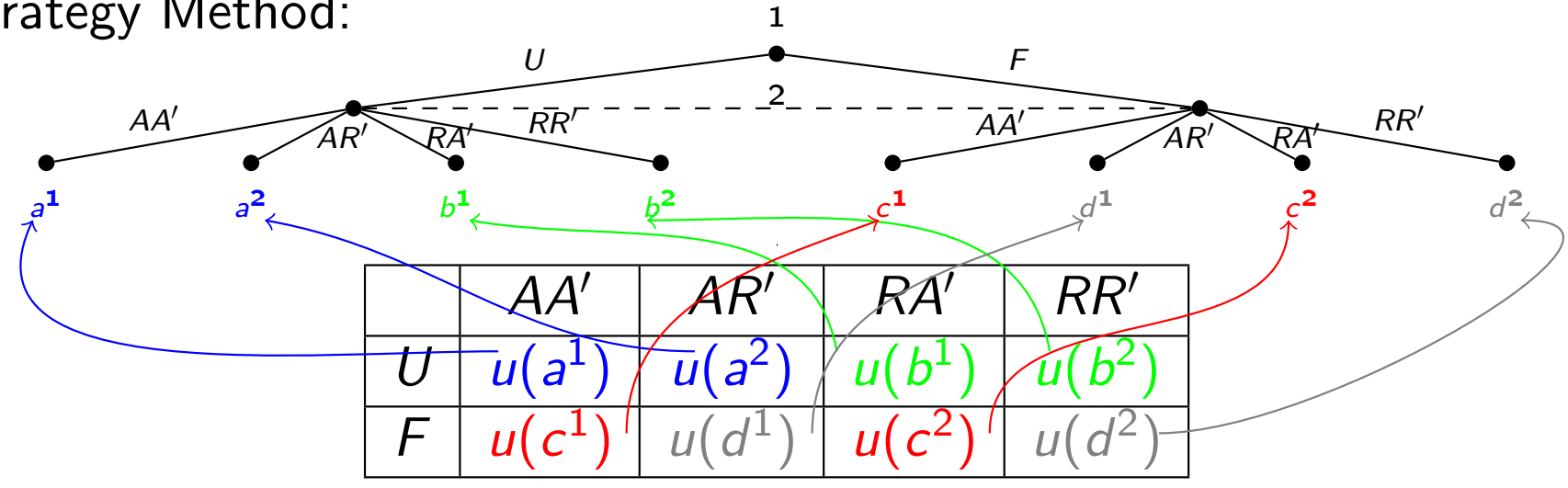
General Idea

Direct Elicitation:



	AA'	AR'	RA'	RR'
U	$u(a)$	$u(a)$	$u(b)$	$u(b)$
F	$u(c)$	$u(d)$	$u(c)$	$u(d)$

Strategy Method:



	AA'	AR'	RA'	RR'
U	$u(a^1)$	$u(a^2)$	$u(b^1)$	$u(b^2)$
F	$u(c^1)$	$u(d^1)$	$u(c^2)$	$u(d^2)$

Research Questions

1. Is it really true that from a game theoretical perspective direct elicitation (DE) and the strategy method (SM) *should* yield the same equilibrium outcome?
   - No: Not in general
2. If not, can we find conditions such that DE and SM should yield the same equilibrium outcome?
   - Yes: But they are on preferences

Theorem

$G_\pi^{DE} = (\Gamma^{DE}, \pi^{DE} : Z^{DE} \to \mathbb{R}^{\mathrm{I}})$ — Strat. iden. $\Leftrightarrow$ — $G_\pi^{SM} = (\Gamma^{SM}, \pi^{SM} : Z^{SM} \to \mathbb{R}^{\mathrm{I}})$

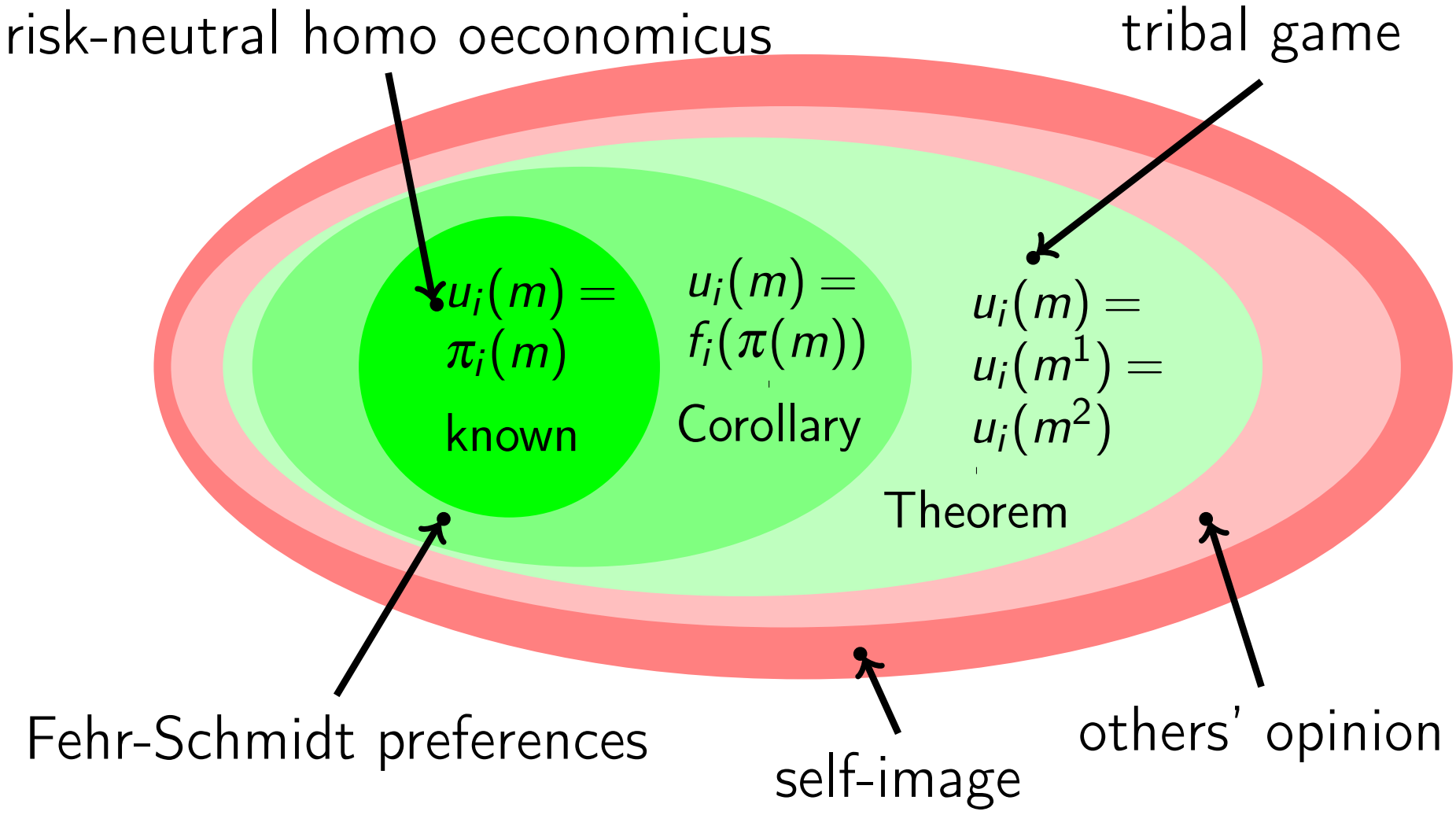
$\uparrow$ equilibrium may change $\downarrow$ — $\uparrow$ equilibrium may change $\downarrow$

$G^{SM} = (\Gamma^{DE}, u^{DE} : Z^{SM} \to \mathbb{R}^{\mathrm{I}})$ — Thm. $\Leftrightarrow$ — $G^{SM} = (\Gamma^{SM}, u^{SM} : Z^{SM} \to \mathbb{R}^{\mathrm{I}})$

- Strategic equivalence if their strategic forms are identical up to a positive affine transformation of each player's Bernoulli **utility**

(Harsanyi et al. 1988)

Venn diagram



Experimental Evidence

See if off-equilibrium payoffs affects behavior depending on whether the strategy and direct response method is used in:

- Lab, online, and meta-analysis of ultimatum game
- (More complex) 3-player prisoner's dilemma with varying salience of off-equilibrium payoffs

Meta-Analysis (Ultimatum Game)

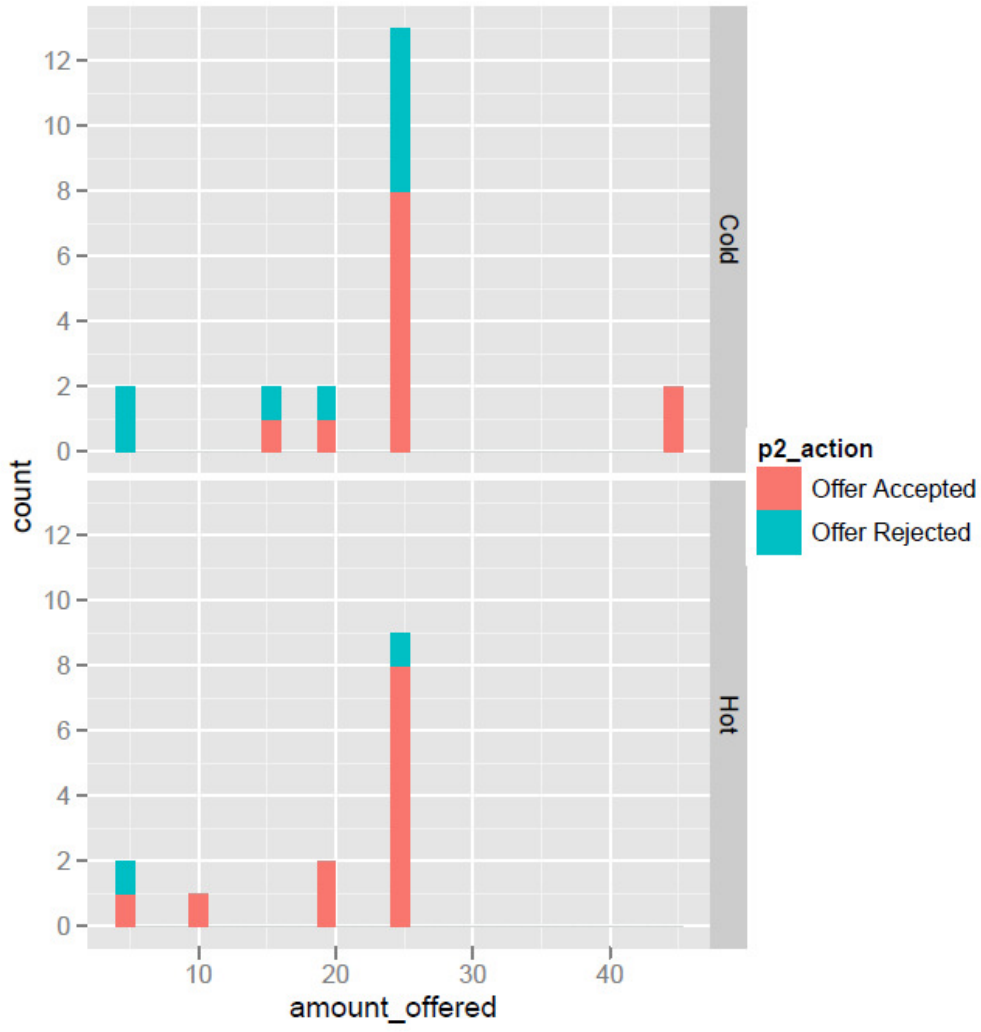
	(1) rejectionrate	(2) rejectionrate	(3) rejectionrate	(4) rejectionrate	(5) rejectionrate	(6) rejectionrate
hot	-0.225***	-0.214***	-1.595**	-1.599**	-0.0732	-0.124*
	(0.000)	(0.001)	(0.009)	(0.010)	(0.660)	(0.021)
offer		-0.373	-3.609*	-3.609*	-0.227	-0.377
		(0.280)	(0.014)	(0.016)	(0.408)	(0.241)
poor		-0.0160				
		(0.697)				
hotoffer			3.438*	3.443*		
			(0.022)	(0.024)		
pooroffer				0.00719		
				(0.939)		
_cons	0.348***	0.497**	1.789**	1.789**	0.303	0.397**
	(0.000)	(0.001)	(0.003)	(0.003)	(0.139)	(0.005)
N	49	49	49	49	49	48

p-values in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

- 20% additional acceptances occur in the hot setting
- (2) add linear time trend and dummy indicator for country economic development
- (3-4) add interaction between offer and hot setting, poor setting
- (5) weighted with citations; (6) weighted by number of observations in experiment
- Among the 16 cold experiments, 9 reported the threshold only. The 9 studies had a higher average threshold, which suggests if the 9 studies had also reported acceptance rates, the difference between hot and cold may have been even larger.
- Offer was not significantly influenced by elicitation method

Ultimatum Game (DE vs. SM)



- Direct elicitation results in more acceptances

Ultimatum Game (DE vs. SM x Emo vs. Neu)

- 418 subjects, roughly 70 per treatment
- Direct elicitation results in more acceptances
- This effect is larger when off-equilibrium payoffs are salient
- No significant differences between strategy and threshold method
- Note: Offers are slightly, but not significantly, lower in direct elicitation setting;

◂ Control for Offer

3-player prisoner's dilemma (Falk, Fehr, Fischbacher)

- Strategy method increased deductions
  - ◂ Deductions differentially affected by emotional prime
  - ◂ Controlling for first stage outcome
- Differences between Direct Elicitation and Strategy Method (and differences-in-differences) were greater:
  - ◂ for deductions of D1
  - ◂ if you are a contributor

Implications

"For example, the notion of subgame perfect equilibrium is lost in the transition from the extensive to the strategic form of the game, since there are no subgames in a game in which players state their strategies simultaneously."
Alvin E. Roth (ch.4, Handbook of Experimental Economics, 1995)

- We have learnt that there is a much more fundamental problem:
  - Strategic equivalence is often lost in the transition from direct elicitation to strategy method elicitation.

Do people have deontological (duty-based) motivations?

Research question

Revealed preference in lab settings. How about in the field?

Research question

How should we model deontological motivations? Any relevance for theory?

Fooled by Randomness

How people often imagine a sequence of coin flips:

0101001011001010100110100

A real sequence of coin flips:

0101011111011000001001101

Fooled by Randomness

Law of Small Numbers (Rabin 2002)

- Expect very small samples/short sequences to resemble the population
- Expect alternation even though streaks often occur by chance

Gambler's Fallacy (Tversky and Kahneman 1974)

- Seeing a 0 or 0s increases the odds of the next draw being a 1 and vice versa (e.g. a "fair" slot machine)

Decision-Making under the Gambler's Fallacy

Large literature explores these misperceptions of randomness

- Many studies focus on **predictions** in lab settings or **betting** behavior in casinos
- Little field research on how misperceptions of randomness can affect agents making sequential **decisions** under uncertainty

Hypothesis:

- Gambler's fallacy $\Longrightarrow$ Negatively autocorrelated decisions, avoidance of streaks

The Decision-Maker's Problem

Suppose an agent makes 0/1 decisions on randomly ordered cases

- If decisions are based on case merits, decision on the previous case should not predict decision on the next case (controlling for base rates)

If the decision-maker suffers from the gambler's fallacy

- After deciding 1 on the previous case, will approach the next case with a *prior belief* that it is likely to be a 0 (and vice versa)
- Also receives a noisy signal about the quality of the current case
- Decisions will be negatively autocorrelated if they depend on a mixture of prior beliefs and the noisy signal
- Similar patterns if agent is rational but judged by behavioral others

Decisions vs. **predictions/betting**: Greater confidence in the noisy signal $\Longrightarrow$ less negative autocorrelation in decisions

Three High-Stakes Real World Settings

1. Refugee court judge decisions to grant or deny asylum
   - Random assignment to judges and FIFO ordering of cases
   - High stakes decisions determining whether refugees are deported

2. Loan officer decisions to grant or deny loan applications
   - Field experiment with random ordering of loan files (Data from Cole, Kanz, and Klapper 2013)
   - Randomly assigned incentive schemes

3. Umpire calls of strike or ball for pitches in baseball games
   - Exact pitch location, speed, etc. to control for pitch quality
   - Know whether the decision was correct

Asylum Judges: Data

High stakes: Applicant reasonably fears imprisonment, torture, or death if forced to return to her home country (Stanford Law Review 2007)

Cases filed within each court are randomly assigned to judges, and judges review the queue of cases following "first-in-first-out"

- Control for time-variation in court-level case quality using recent approval rates of other judges in same court (tends to be slow-moving positive autocorrelation)
- Control for time of day fixed effects

Judges have a high degree of discretion

- No formal or advised quotas (substantial heterogeneity in grant rates across judges in the same court)
- Serve until retirement, fixed wage schedule w/o bonuses

Asylum Judges: Baseline Results

	Grant Asylum Dummy				
	(1)	(2)	(3)	(4)	(5)
Lag grant	-0.00544*	-0.0108***	-0.0155**	-0.0326***	
	(0.00308)	(0.00413)	(0.00631)	(0.00773)	
β_1: Lag grant - grant					-0.0549***
					(0.0148)
β_2: Lag deny - grant					-0.0367**
					(0.0171)
β_3: Lag grant - deny					-0.00804
					(0.0157)
p-value: $\beta_1 = \beta_2 = \beta_3$					0.0507
p-value: $\beta_1 = \beta_2$					0.290
p-value: $\beta_1 = \beta_3$					0.0214
p-value: $\beta_2 = \beta_3$					0.0503
Exclude extreme judges	No	Yes	Yes	Yes	Yes
Same day cases	No	No	Yes	Yes	Yes
Same defensive cases	No	No	No	Yes	Yes
N	150,357	80,733	36,389	23,990	10,652
R^2	0.374	0.207	0.223	0.228	0.269

- Judges are up to 5 percentage points less likely to grant asylum if the previous case(s) were granted
- Up to 17% decline relative to the base rate of asylum grants

Asylum Judges: Heterogeneity

Stronger negative autocorrelation

- Consecutive cases with applicants of the same nationality
- Moderate judges (grant rate, excluding current observation, is between 0.3 and 0.7)

Weaker negative autocorrelation

- More experienced judges (8+ years)

Abbreviated Results

Negative autocorrelation in decisions and avoidance of streaks

- Up to ◂ 5% of decisions are reversed are reversed due to the gambler's fallacy
- Stronger bias for moderate decision-makers, similar or close-in-time cases
- Weaker bias for experienced or educated decision-makers, under strong incentives for accuracy

Unlikely to be driven by potential alternative explanations

- Preference to be equally nice/fair to two opposing teams
- ◂ Sequential contrast effects
- ◂ Quotas and/or learning
- ◂ Not driven solely by concerns of external perceptions

Preference for Randomization

Agents may prefer to alternate being "mean" and "nice" over short time horizons

- Loan officers in the experiment are told that their decisions do not affect actual loan origination

- More generally, the gambler's fallacy and desire to "do right" may be the reason why agents feel more guilty after "1100" than "1010"

Roadmap

Social Preferences or Sacred Values? Theory and Evidence of Dentological Motivations

A Theory of Experiments: Invariance of Equilibrium to the Strategy Method of Elicitation and Implications for Social Preferences

Decision-Making Under the Gambler's Fallacy: Evidence From Asylum Courts, Loan Officers, and Baseball Umpires

Ideological Perfectionism on Judicial Panels

Deontological Motivations

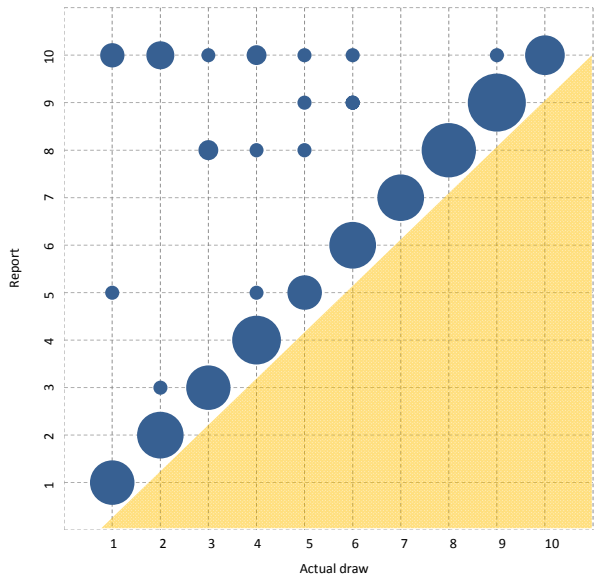
- How do individuals perceive the cost of taking actions they disagree with politically or morally?
- Economics tends to gravitate towards the assumption that costs – be they economic, effort or cognitive – are convex
- One rationale for this assumption is that it makes theoretical models analytically tractable
- Another rationale is that it seems intuitively plausible. However, such intuition has proved fragile following a number of recent experiments
  - even small deviations from convictions are perceived to be very costly, but once a small deviation has been made, further deviations will entail relatively little additional cost
  - implies individuals tend to give up on their morals if they cannot follow them fully, suggesting a concave cost of deviation
- Concave ideological preferences explains a novel and puzzling phenomenon in judicial decisions along with a number of related empirical facts

Lab Experiments

- For instance, individuals with concave moral costs will tend to give up on their morals if they cannot follow them fully
- This pattern of behavior has been popularly labeled the "what-the-hell-effect" (Ariely 2012; Baumeister et al. 1996)
  - The decision whether to lie is often insensitive to the outcome of lying once it it preferred over the outcome of being truthful (Hurkens et al. 2009; Gneezy et al. 2013) ◂ (Abeler, Nosenzo, Raymond)
  - Once individuals are induced to cheat, they succumb to full-blown cheating (Gino et al. 2010)
  - Once induced to kill mice, indifferent to the number of mice killed ◂ (Falk and Szech 2013)
  - In politics it may be more sensible to assume concave preferences. A voter on the far right would be more or less indifferent between two candidates on the left (both are equally bad), but would care greatly about which of the right wing candidates wins (Osbourne 1995)
- What are the implications of concave preferences for important real world decision situations

Lab Experiments

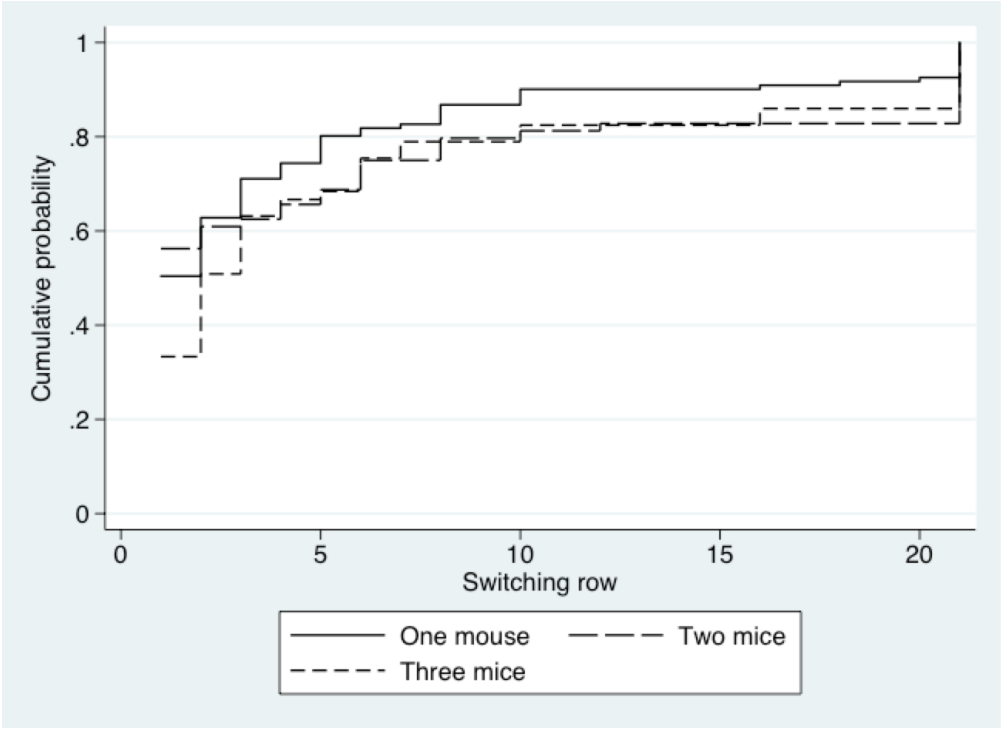
Abeler/Nosenzo/Raymond - experiment



- **Reports do depend on actual draws**.
- 83% of subjects report their true draw.
- Strong positive correlation between draws and reports (Spearman's rho = 0.657, p = 0.000).
- **None underreports their draws**.

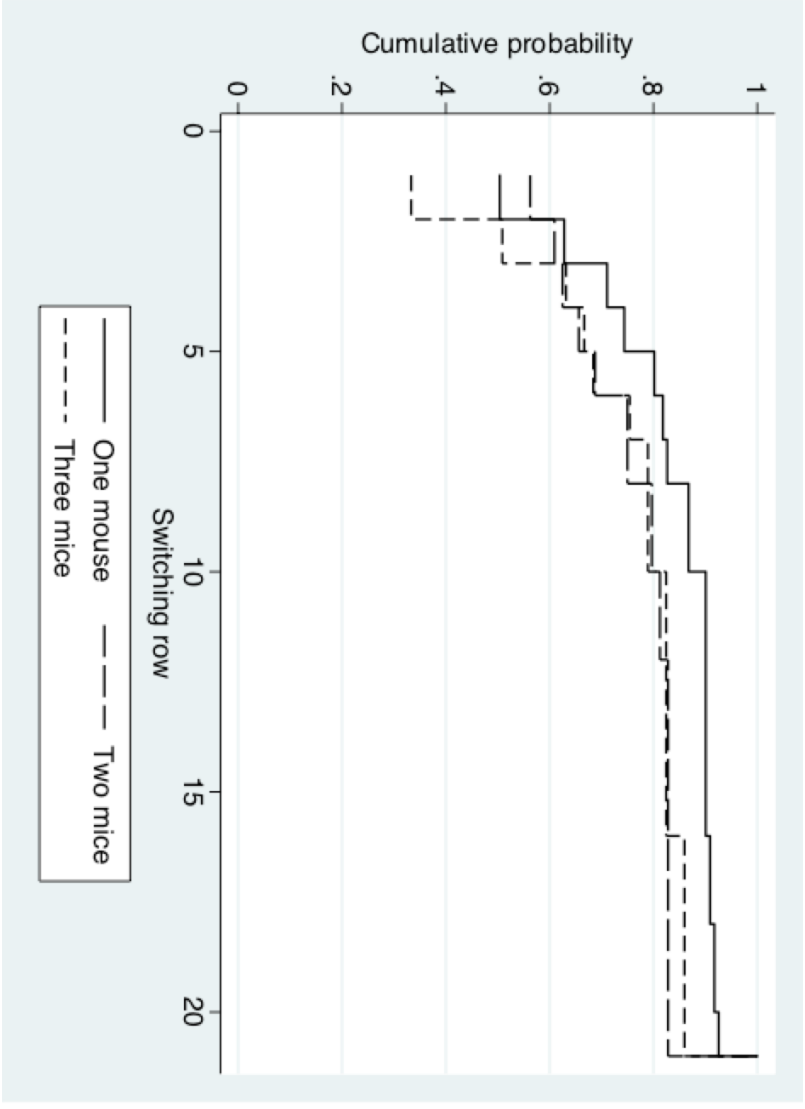
- Subjects did a computer task and at the end of the task they randomly drew their reward from a uniform distribution between 1 and 10 (they knew the distribution).
- The number they sampled appeared on their screen and they were asked to copy it on a piece of paper and hand it out to the experimenter on their way out of the lab in order to get paid. ◂ Lab

Lab Experiments

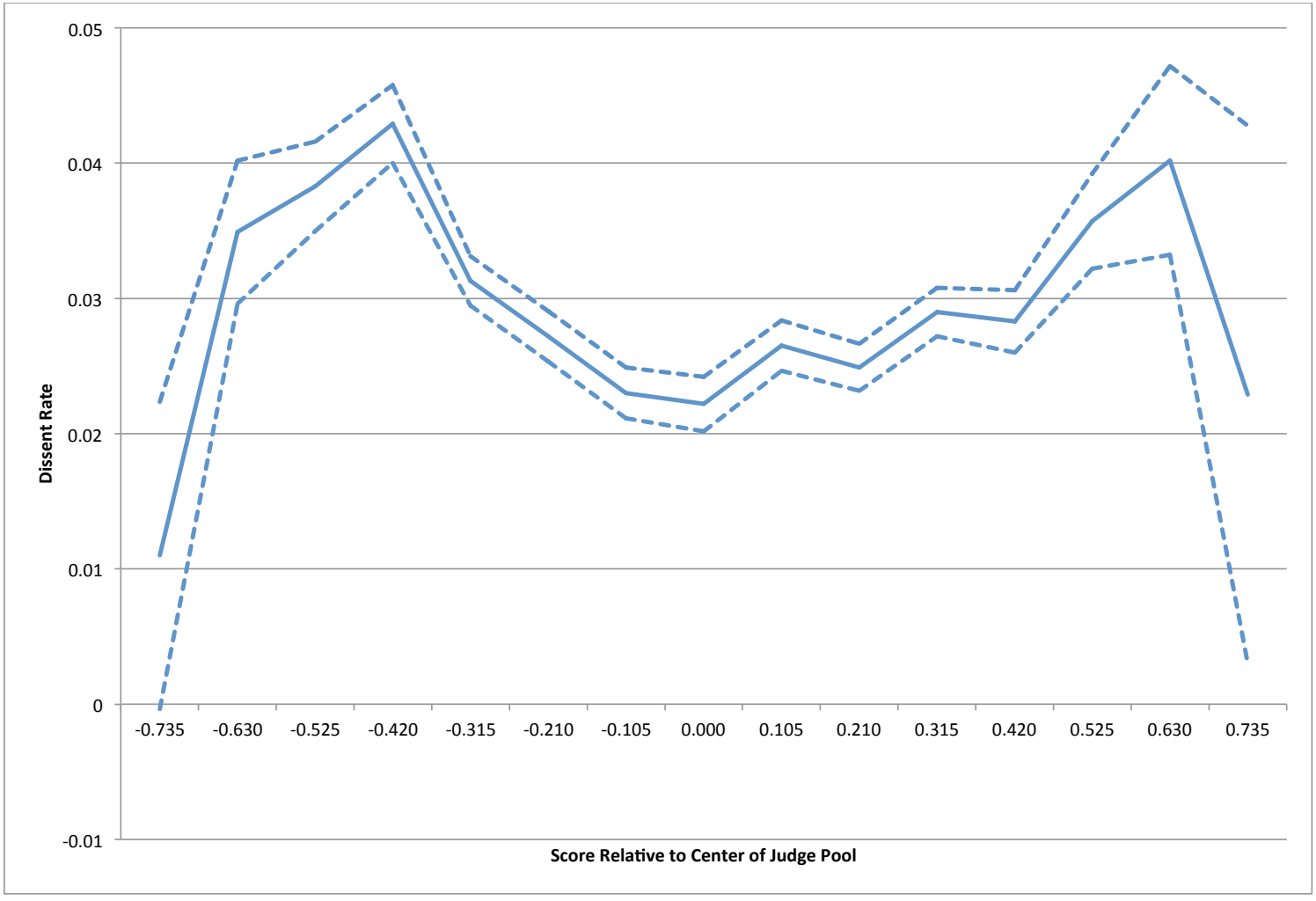


- Subjects were shown a list of binary alternatives, labeled Option A and Option B
- Option A implied that the mouse would be saved and that the subject would receive no money. Option B implied the killing of the mouse and a monetary amount. Monetary amounts associated with killing increased from row to row, starting from 5 up to 100 euros, in steps of 5 euros.
- Earlier a subject switches, the less he or she values the life of his or her mouse relative to earning money

Lab Experiments



The "Spider" Result



Intuition

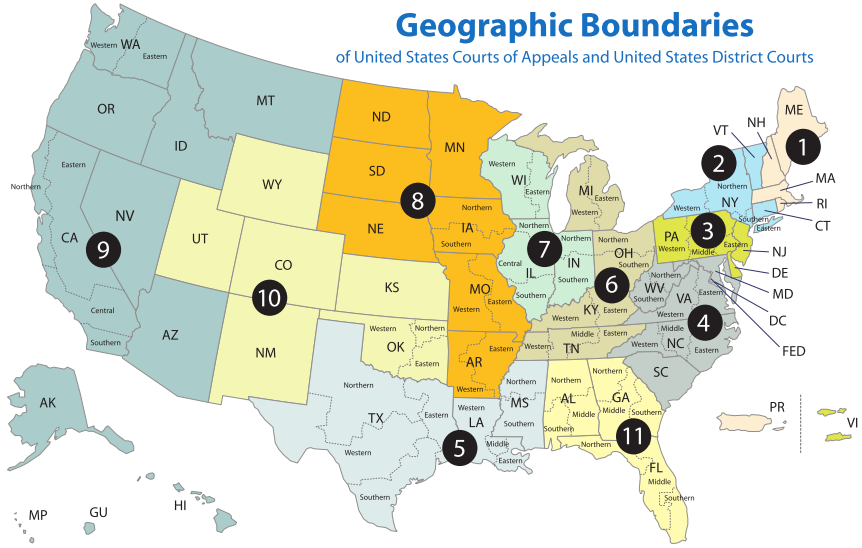
- Judges feel bad when signing "unfavorable" verdicts
- They are ideological perfectionists: signing even one unfavorable verdict comes at high cost, signing many is marginally less costly
- Not signing (i.e., dissenting on a 3-judge panel) implies a collegial pressure

Intuition

- For extreme judges, the marginal cost of signing unfavorable verdicts falls while the marginal benefit of signing unfavorable verdicts stays high, so you just sign all of others' verdicts.
- For moderate and centrist judges, the marginal cost of signing unfavorable verdicts remains high while the marginal benefit is low, so the # of dissents is determined by the natural normal distribution of judge scores.

Circuit Courts

- 12 Circuit courts decide on appeals from lower courts
  - Three judges are randomly picked to a case
  - Set precedent for future cases
  - Between 8 to 40 judges in each circuit, politically appointed by president, for life
- The opinion (interpretation of the law) is what sets precedent and is a continuous variable
- Judges can "**dissent**" by not signing opinion and then write motivation why



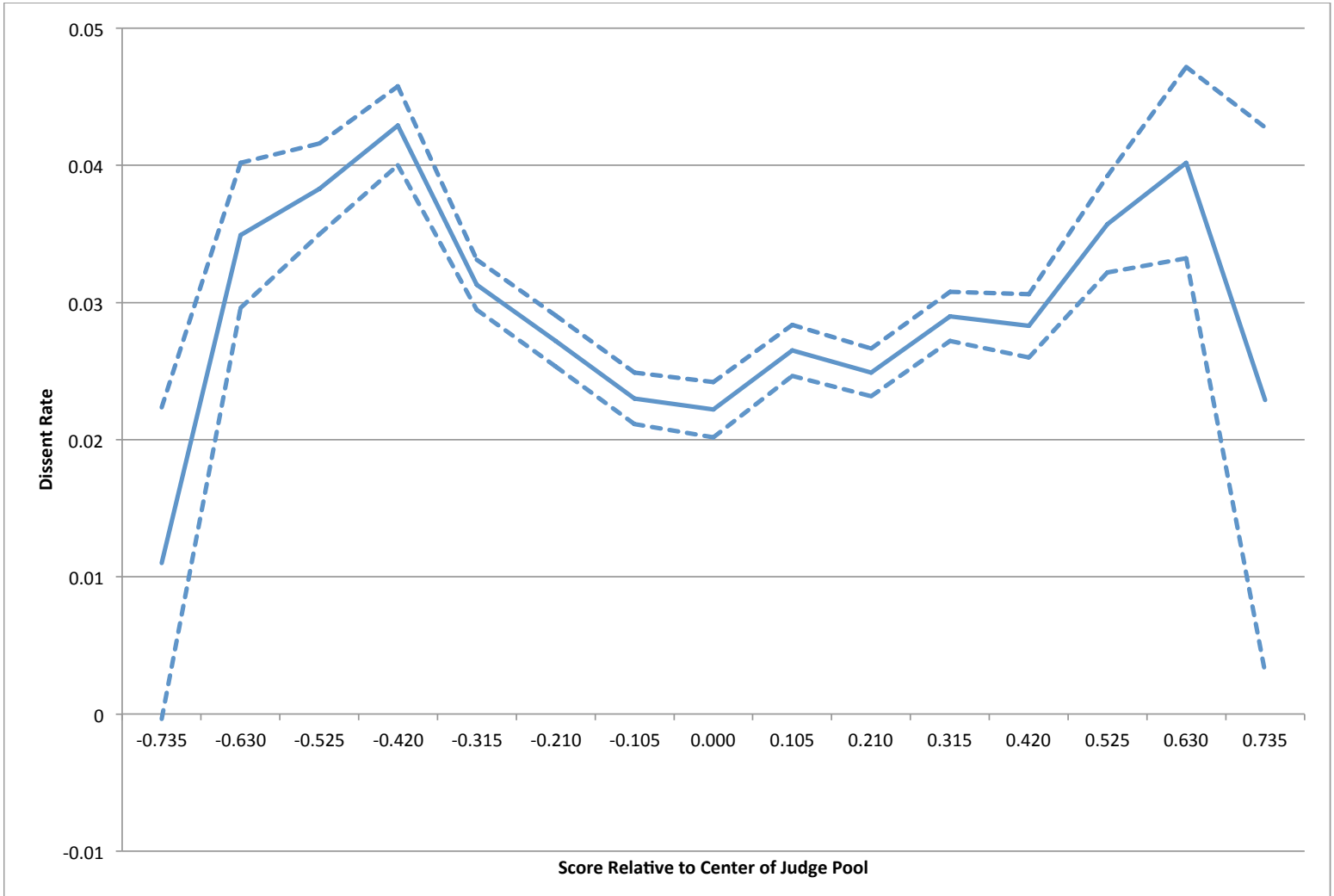
Ideology

- What role does ideology play in determining whether a judge dissents?

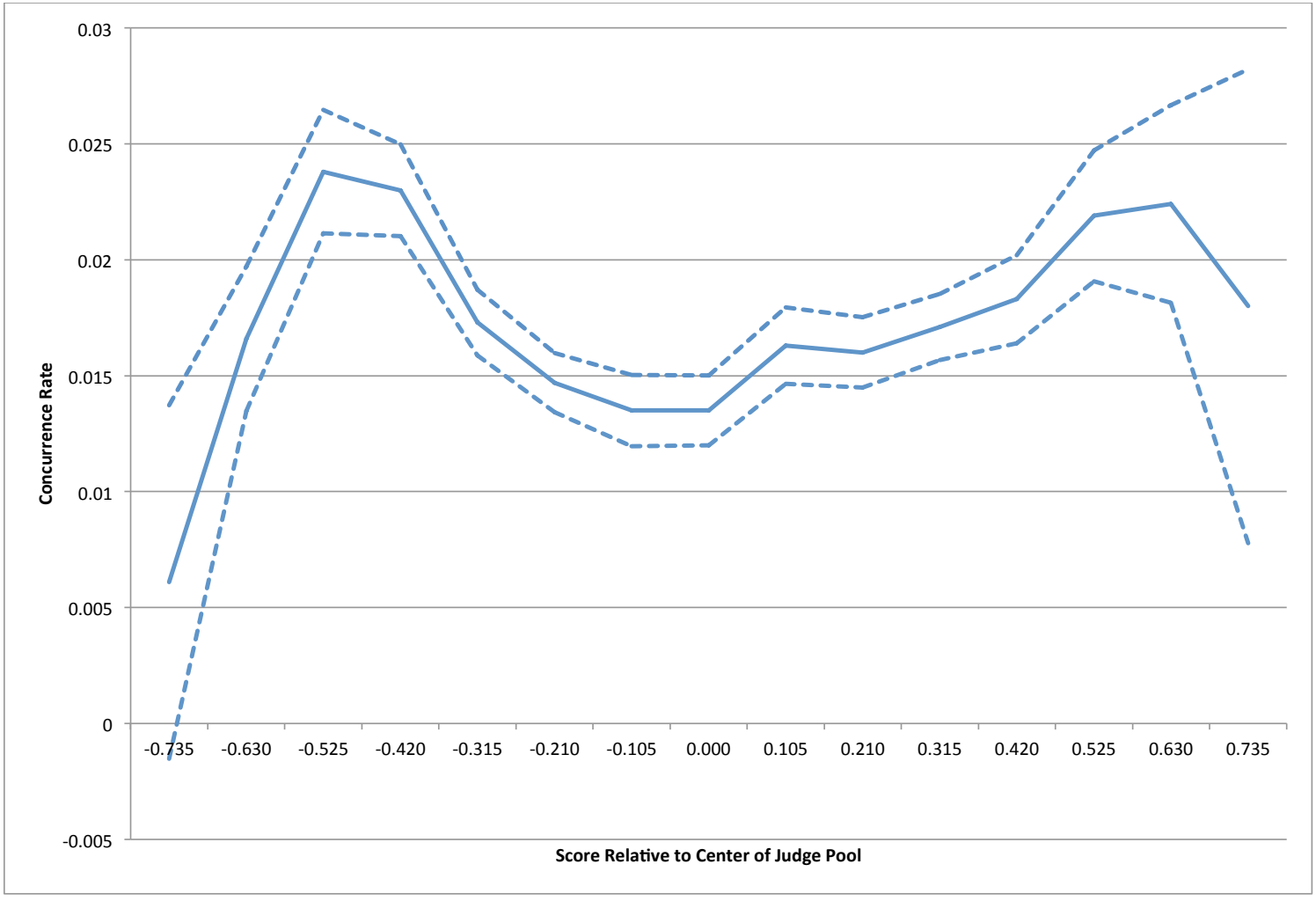
- Use proxy for ideology by weighing voting behavior of appointing president and voting behavior of home state senators (Judicial Common Space)
- Yearly data 1950-2007 (Openjurist), 5% sample (1925-2002) (Songer-Auburn)
- Proxy goes from -1 (leftist/liberal/democrat) to +1 (rightist/conservative/republican)

The "Spider" Result for Dissents

- Dissent is as a non-monotonic function of ideological extremeness:
  - Centrists dissent seldom
  - Moderates dissent often
  - Extremists dissent seldom



The "Spider" Result for Concurrences



How Can the Spider Be Explained?

- Note: The result is driven by ideological distance between judges, not by the ideology per se.
- Example: A very liberal judge (-1) will dissent seldom when in circuit of very conservative judges (+1), but dissent often in circuit of moderate liberals (-0.5).
- This is about interaction between judges with different ideologies.

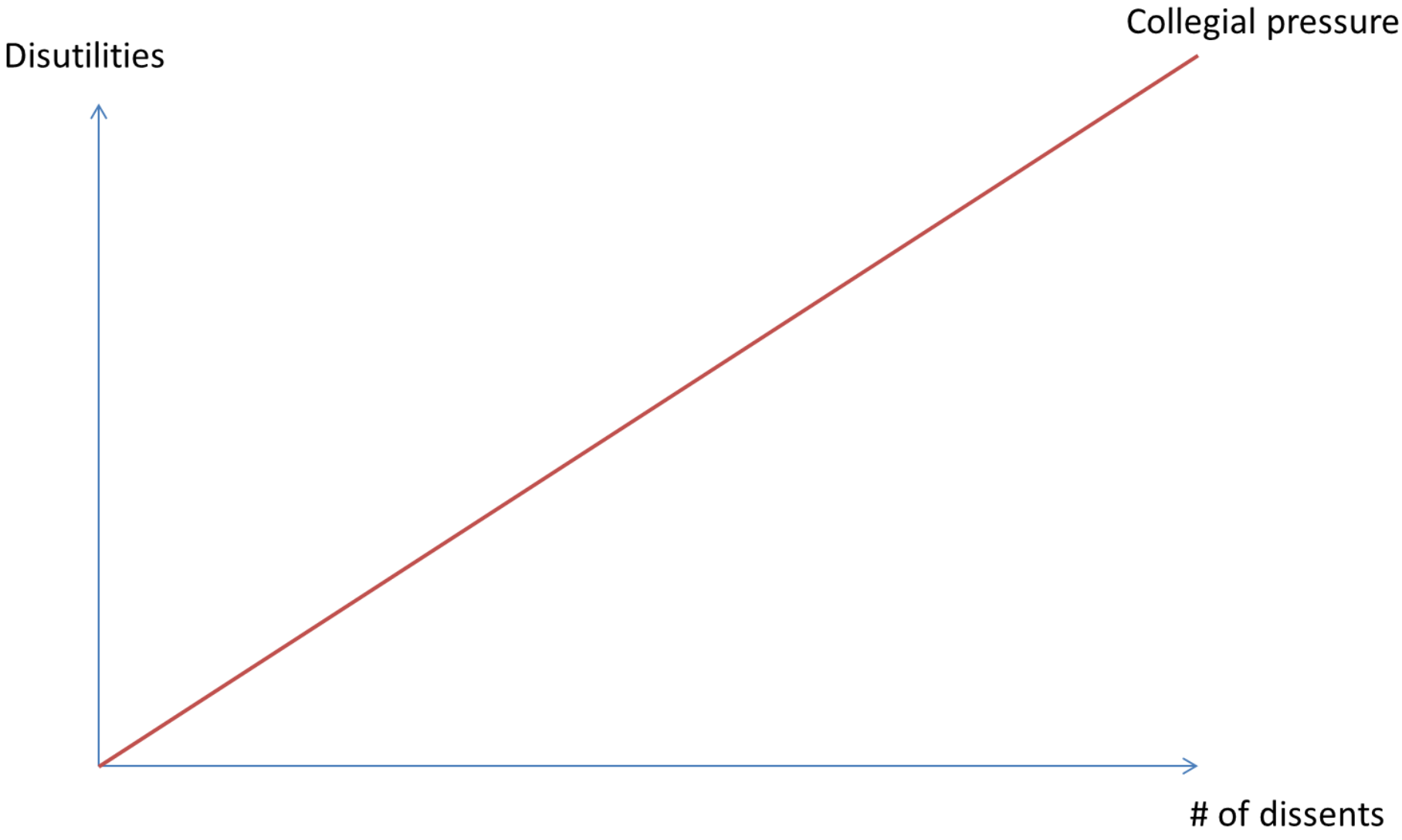
Dissent More When a Judge's Ideology Far From Panel Median

Table: Dissents and Concurrences vs. Distance to Center of Judge Panel (1950-2007)

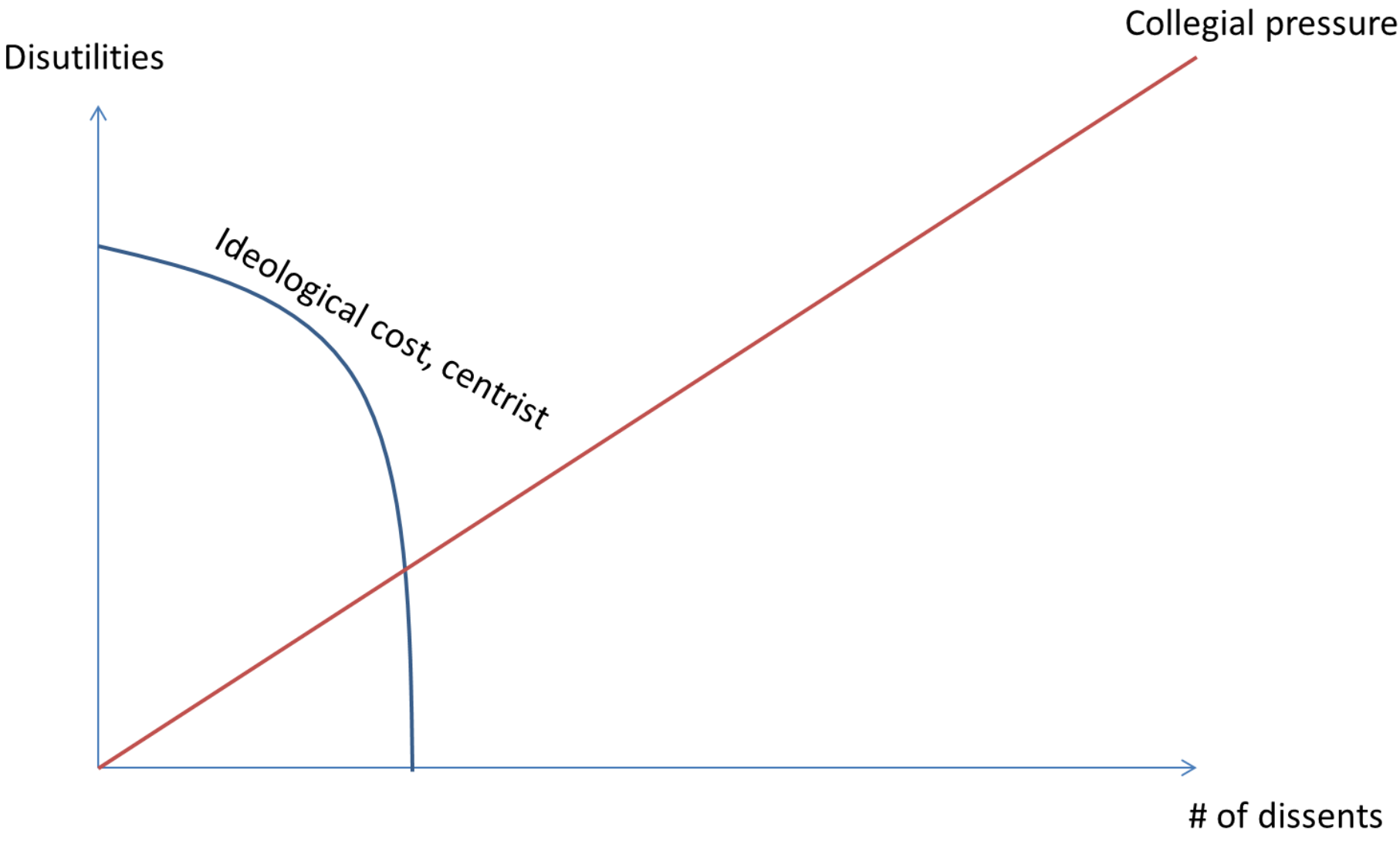
	Dissents or Concurs
Distance to Center of Panel	0.0399***
	(0.00580)
Circuit Fixed Effects	Y
Year Fixed Effects	Y
N	541182
R-sq	0.008

Extremists more often distant to panel center, should dissent more often, yet dissent less according to spider

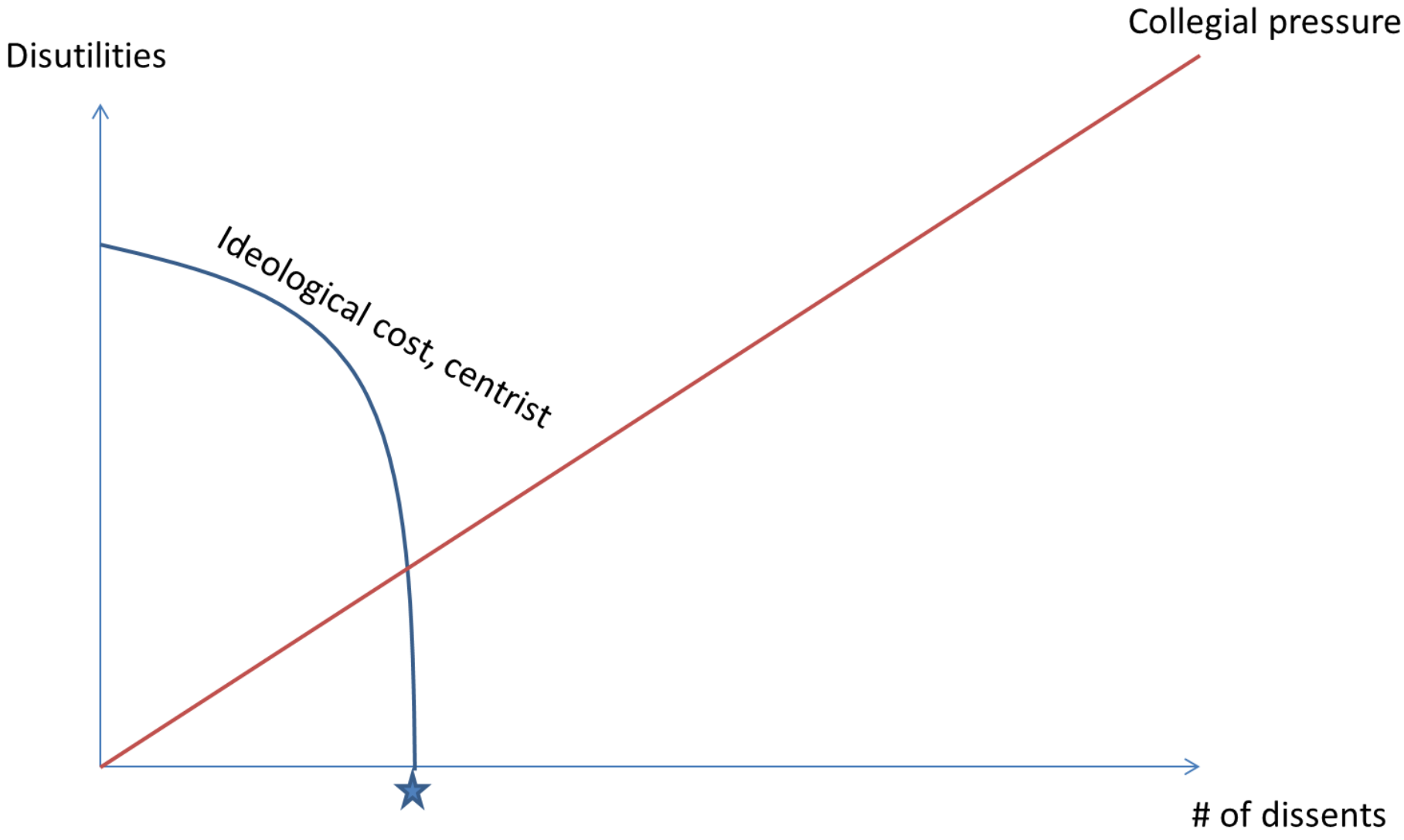
Intuition



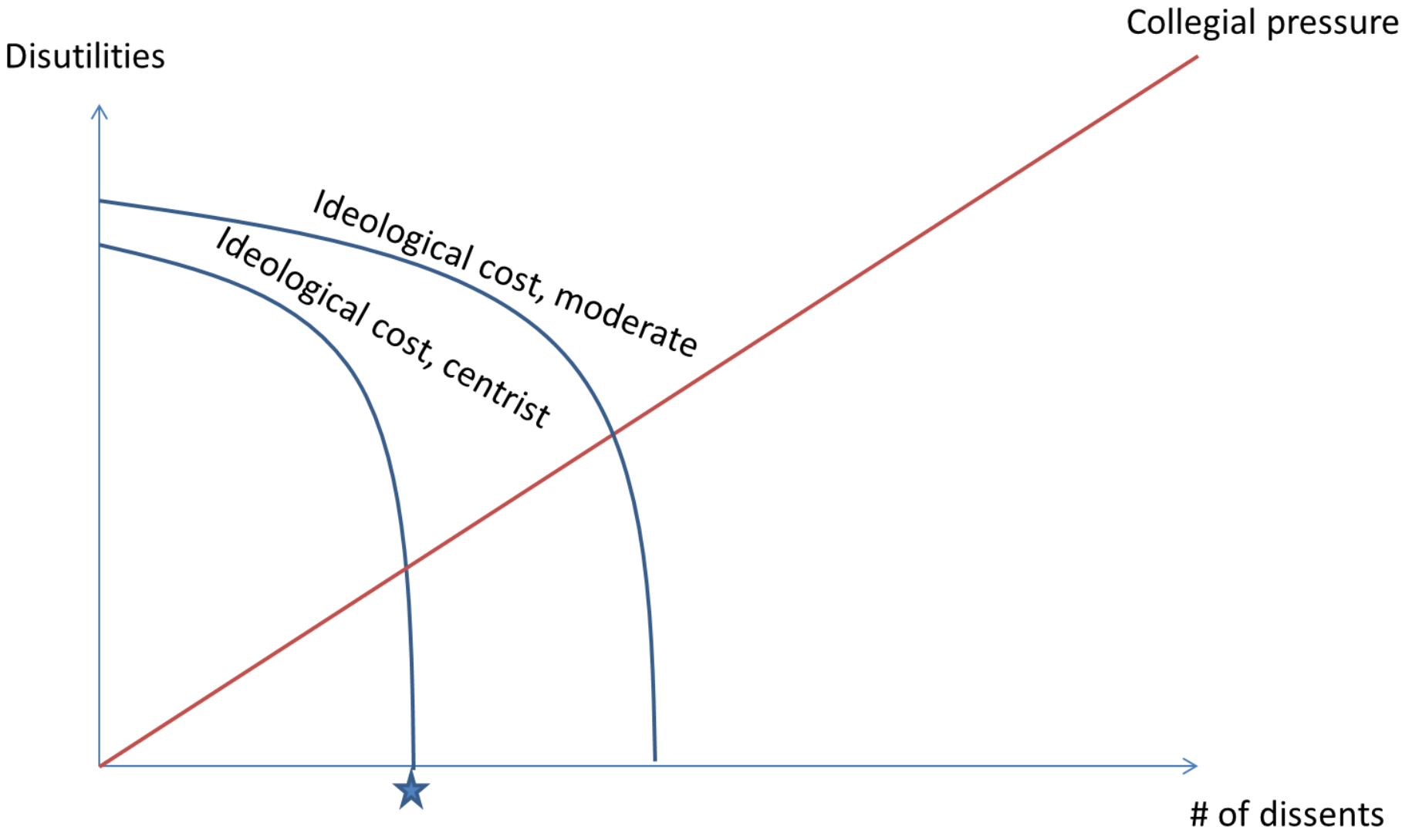
Intuition


Collegial pressure
Disutilities
Ideological cost, centrist
# of dissents

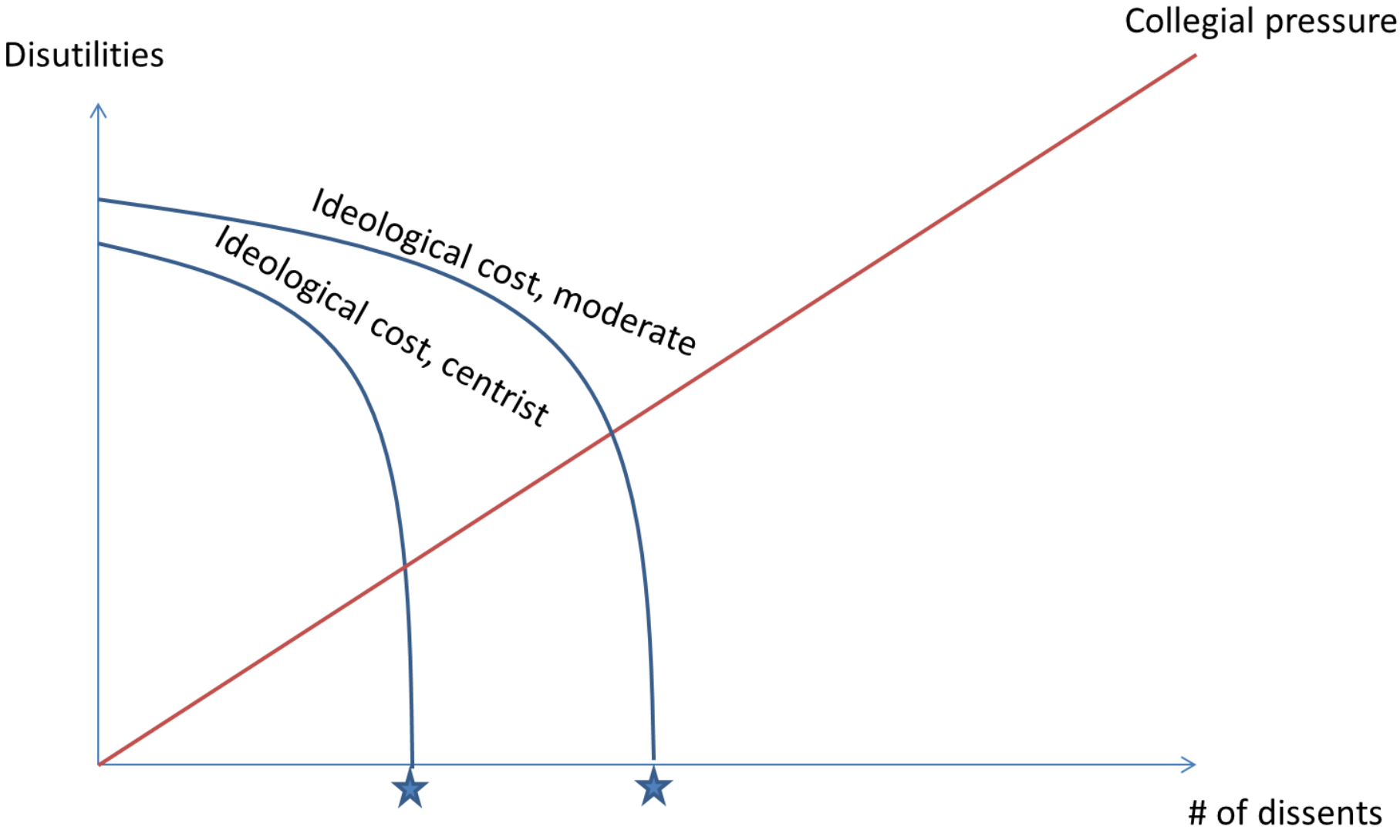
Intuition


Collegial pressure
Disutilities
Ideological cost, centrist
# of dissents

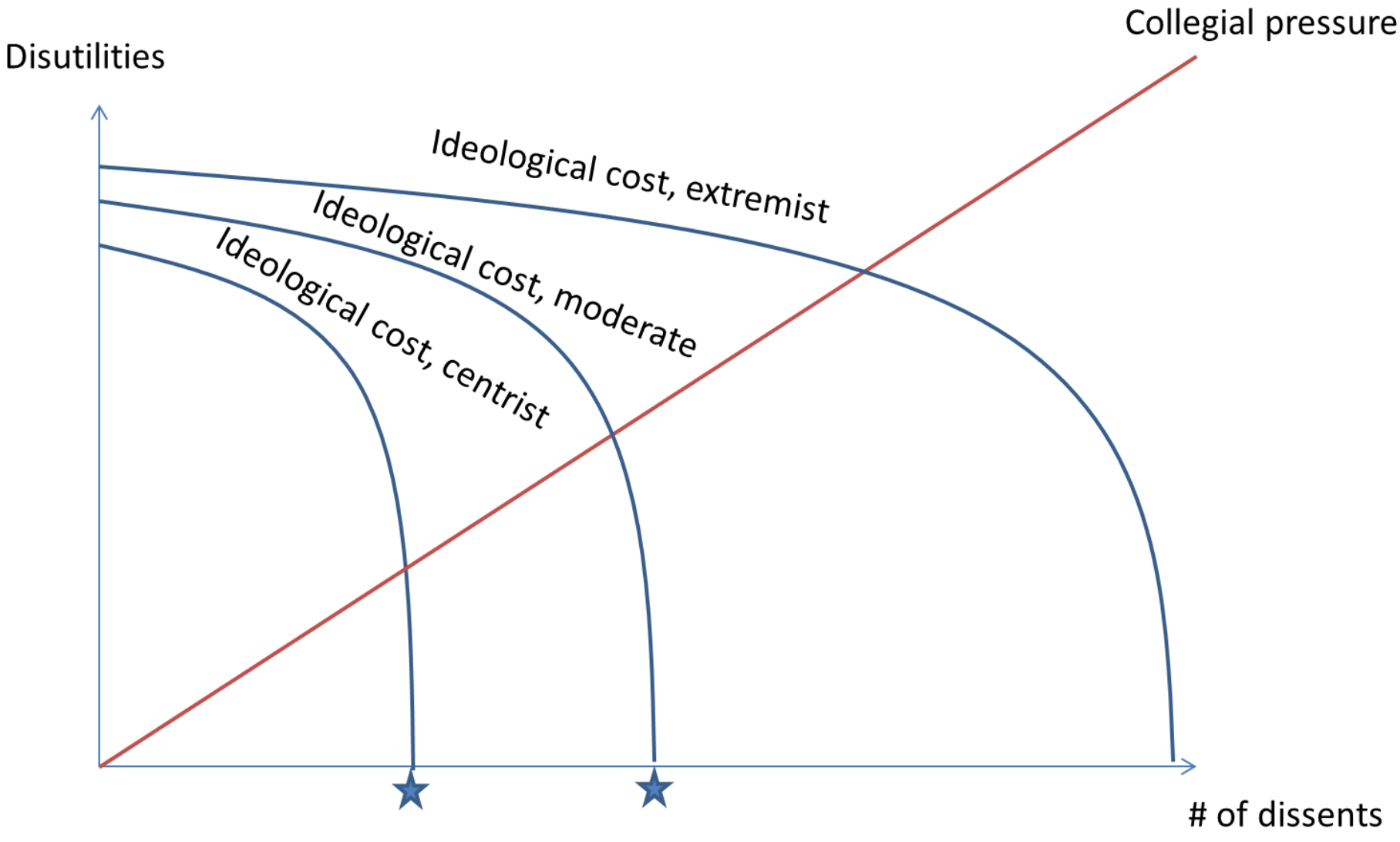
Intuition


Collegial pressure
Disutilities
Ideological cost, moderate
Ideological cost, centrist
# of dissents

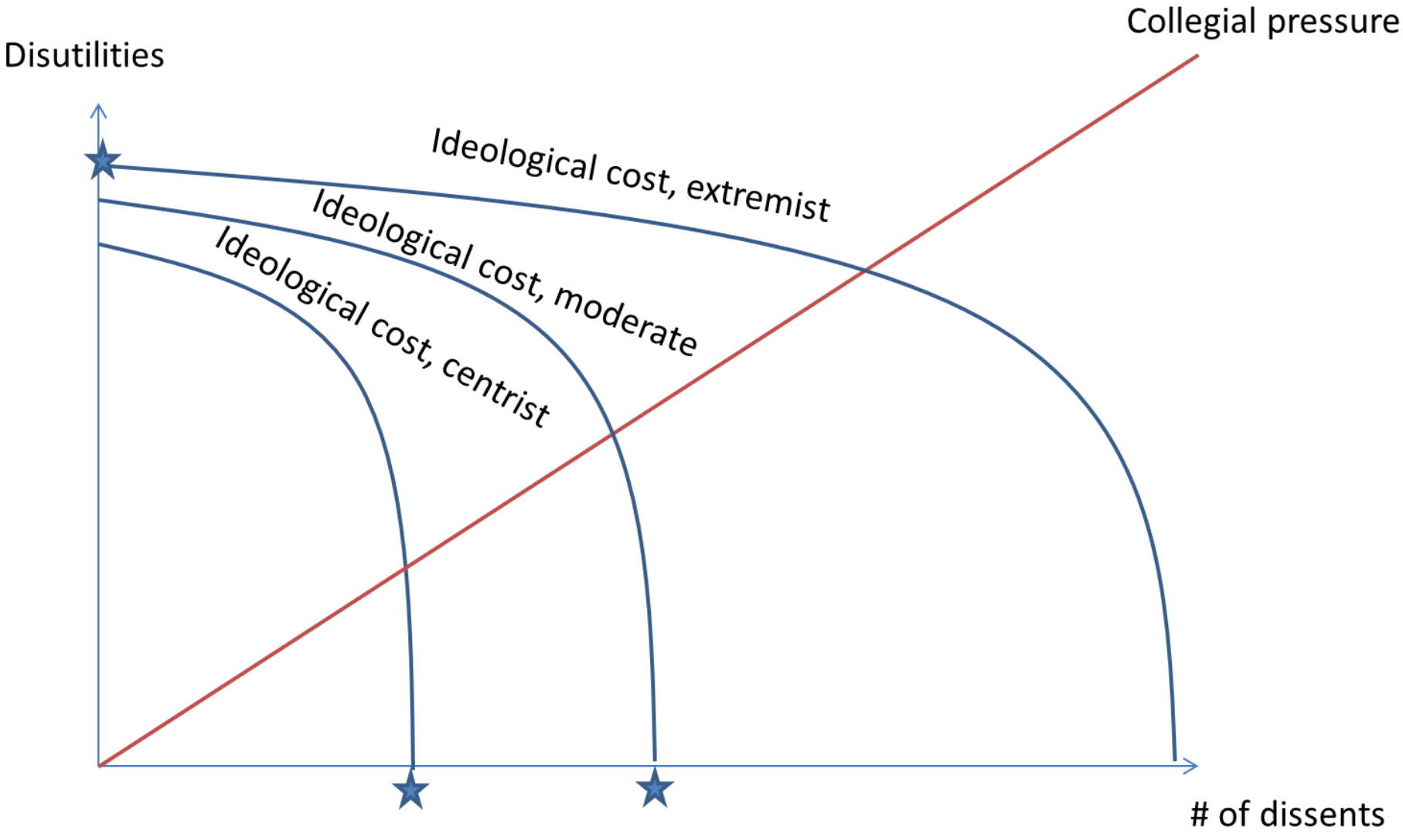
Intuition


Collegial pressure
Disutilities
Ideological cost, moderate
Ideological cost, centrist
# of dissents

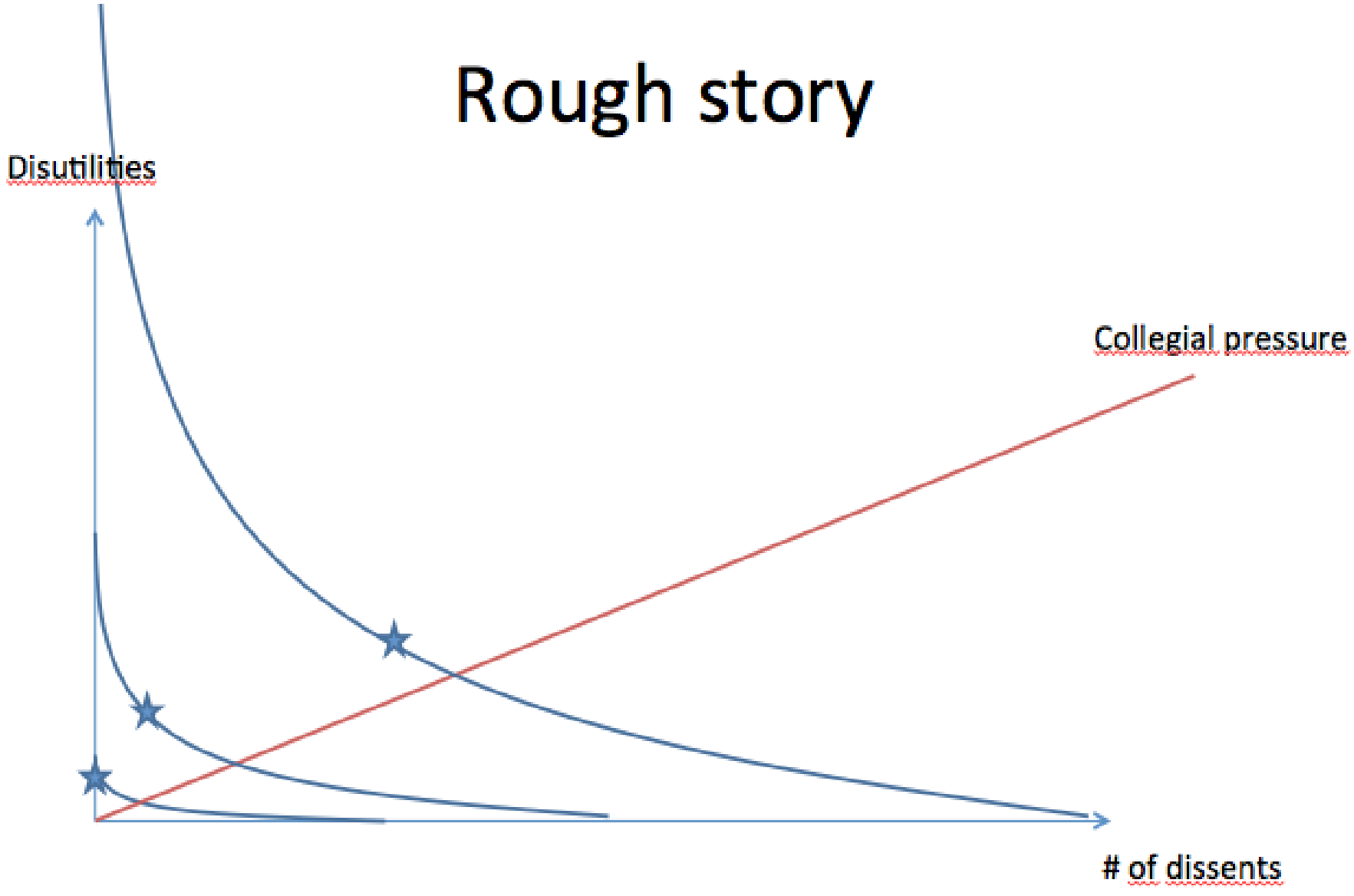
Intuition


Collegial pressure
Disutilities
Ideological cost, extremist
Ideological cost, moderate
Ideological cost, centrist
# of dissents

Intuition


Collegial pressure
Disutilities
Ideological cost, extremist
Ideological cost, moderate
Ideological cost, centrist
# of dissents

Theorem: Convex costs cannot explain spider



Alternative Theory

- Judges whistleblow to signal to Supreme Court. Extreme judges have no incentive to dissent since the Supreme Court will not overturn.
  - Decrease in cost should push peak outwards for "what the hell" and push peak inwards for whistleblowing
    - Caseload, 9/11 caseload surge, and Circuit 5 split proxy for increase in cost
  - For extremely low cost, "what the hell" predicts dissent positively correlated with distance to center and whistleblowing predicts negative correlation
    - Senior status is measure of extremely low cost

Conclusion

- Document non-monotonicity of dissents: extreme judges dissent less than others, moderate judges dissent the most
  - Can be explained by model of ideological perfectionism and collegial pressure
  - Test auxiliary model results
- Judges are sensitive to interaction with judges with distant ideologies
- But extremist judges get numb and give up on their ideology
- ◂ Concave ideological costs can explain backlash and legitimization

Further Predictions From Theory

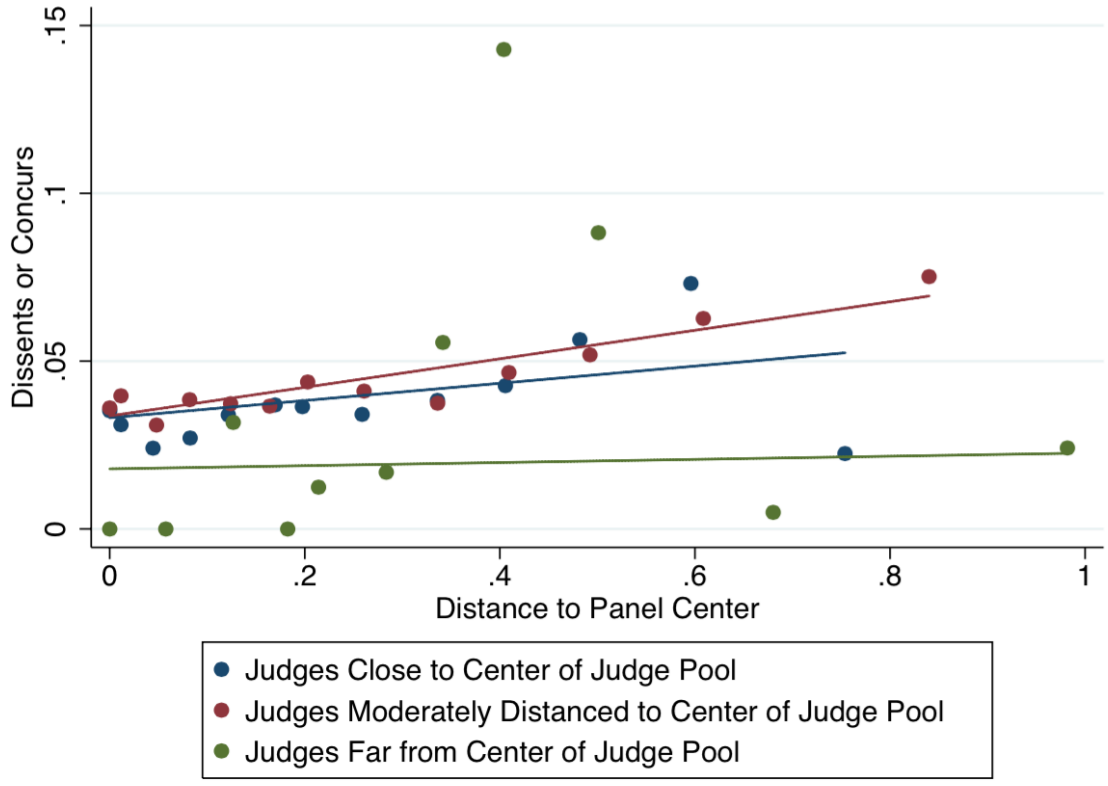
- Prediction: In all cases $v \equiv$ median judge in panel.
- Prediction: The further a judge is from panel median, the higher the probability (s)he will dissent.
- Prediction: Extreme judges sign verdicts which are more unfavorable to them than what moderate judges sign.

Does the Median Judge Decide?

	(1)
	Liberal Verdict
Score	-0.0915***
	(0.0138)
Center Judge	-0.00492***
	(0.00108)
Score * Center Judge	-0.153***
	(0.0278)
N	23031
R-sq	0.003

- Determine who in each panel has median ideology, and who among other two is closest to median and furthest from median.
- Use database of handcoded ideology (liberal=+1, conservative=-1) of each "opinion."

Extreme Judges Sign Verdicts Which Are More Unfavorable?



Polarization

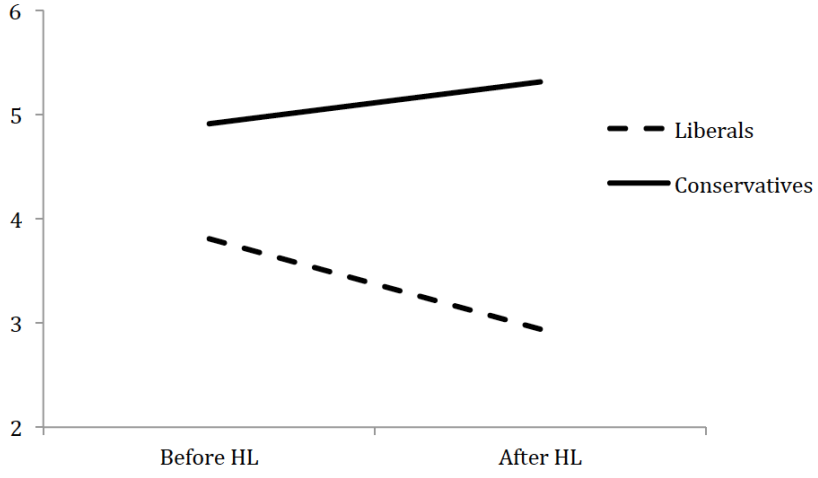


Figure 3: Liberals and conservatives views on religious liberty of closely held company before and after Hobby Lobby. The y-axis represents support for religious freedom rights: Higher scores indicate greater support for such rights

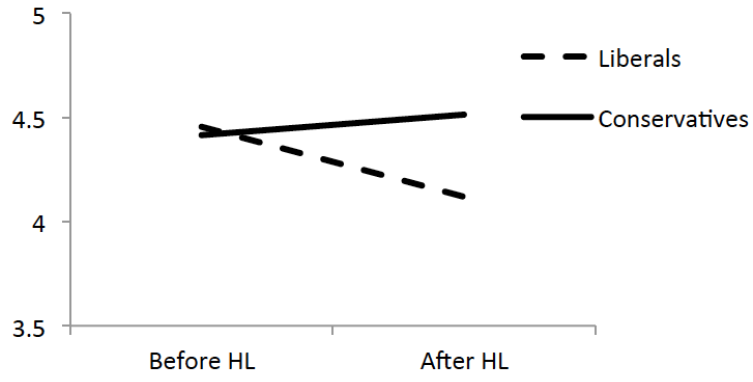
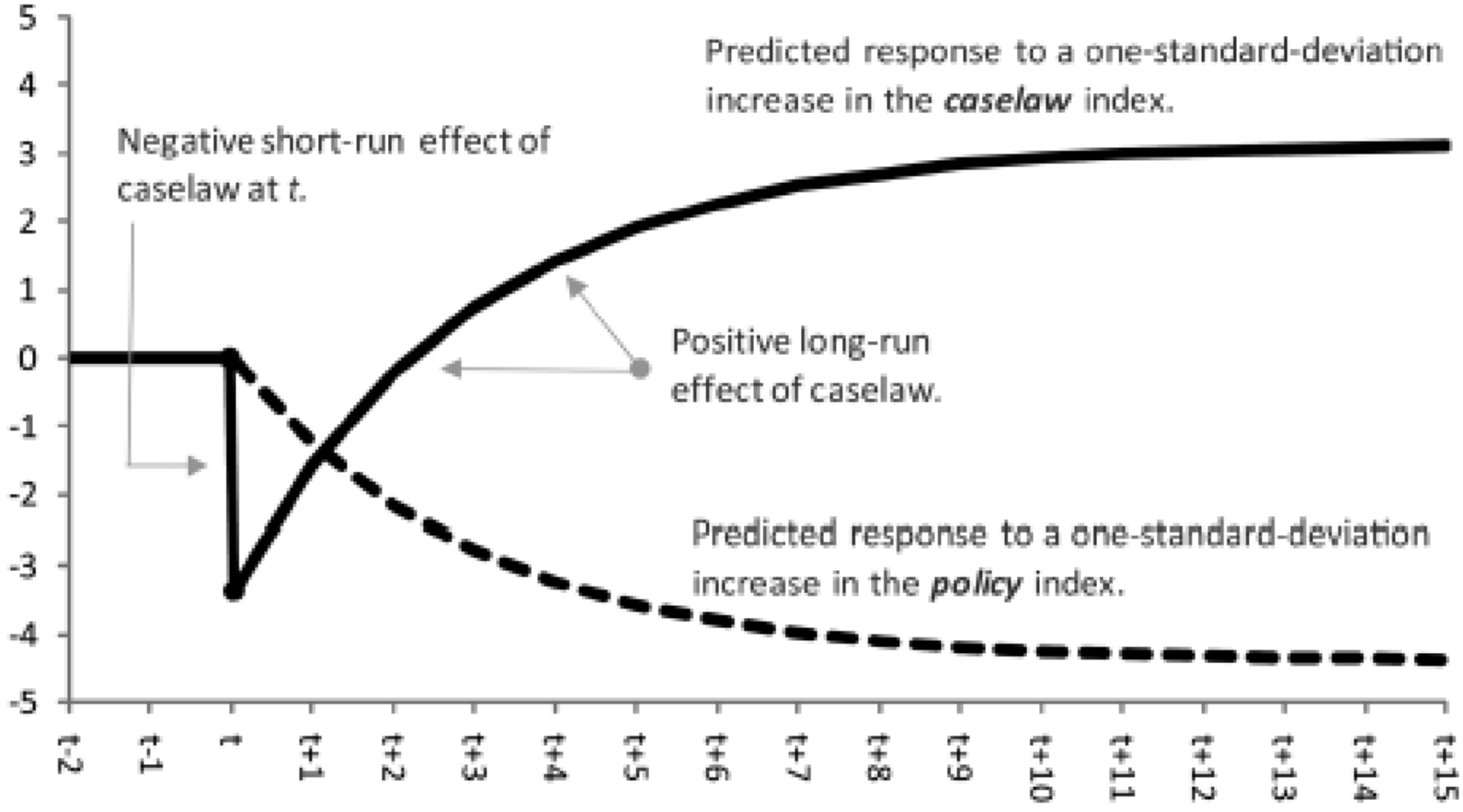


Figure 6b: Liberals' and conservatives' trust in the Supreme Court (before a after *Hobby Lobby*)

Morally repugnant decisions impact stated (and revealed) preferences

Backlash and Legitimization

FIGURE 2 Predicted Responses in Mood to One Standard Deviation Increases in Caselaw and Policy



Instantaneous backlash, then countervailing long-run effect that follows the law

Model

2 periods, actions at $t = 0$ that may result in abortion at $t = 1$

- Utility of no abortion: 0; an abortion yields: $-u_a < 0$
- After an abortion, no subsequent change to utility from additional abortions ("What the hell", concave cost to deviating from duty)
- q (laws, access to abortion, exogenous) -> $\uparrow$Pr(abortion)
- p (attitudes, donations, endogenous) -> $\downarrow$Pr(abortion)
  - $c(p) \geq 0,\ c' > 0, c'' > 0$
  - $P(q-p),\ P' > 0,\ P'' > 0$

$$\max_p \{(P(q-p))(-u_a) - c(p)\}$$

$$\max_p \{-P(q-p) - c(p)\}$$

Dynamics of Law and Norms

- If the agent has already had an abortion, $p^* = 0$
  - else, $P'(q-p) = c'(p)$
  - s_0 share of the population have not had an abortion
- Assume share of abortions in the society is at steady-state
  - $s = P(q-p)$ will have an abortion at $t = 1$
  - share α of new people enter; β exit
  - $s_0(1-s)(1-\beta)+\alpha$ is share without abortion at $t = 1$
  - A steady state obtains if:

$$s_0(1-s)(1-\beta)+\alpha = s_0$$

Equilibrium Effect of Laws

- Implicit Function Theorem yields:

$$\frac{\partial p^*(q)}{\partial q} = \frac{P''(q-p^*)}{P''(q-p^*)+c''(p^*)}$$

- Since $P'' > 0$, and $c'' > 0$:

$$0 < \frac{\partial p^*(q)}{\partial q} < 1$$

- Pro-choice decision at $t = 0$ stimulates p: initial backlash
  - Overall anti-abortion attitude is: $s_0 p$
- At $t = 1$, both p^* and s_0 will change

$$s_0 p^* = \frac{\alpha p^*}{s^* + \beta - s^*\beta} = \frac{\alpha p^*}{P(q-p^*)+\beta - P(q-p^*)\beta}$$

Backlash or Expressive?

- q increases both the numerator and the denominator

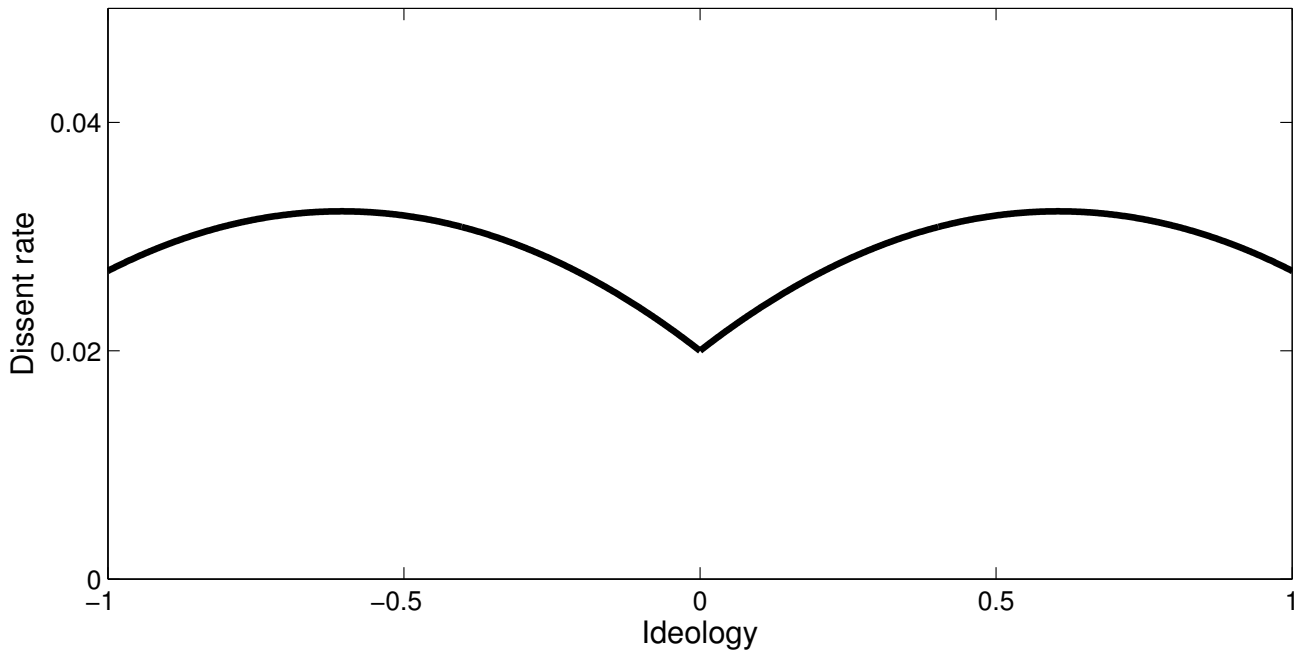
$$s_0 p^* = \frac{\alpha p^*}{P(q - p^*) + \beta - P(q - p^*)\beta}$$

  - Overall effect depends on the relative increase of p in the numerator compared to increase of $P(q - p^*)$ in the denominator
- If large increase in p^* offsets the increase in the probability of abortions, then long-term equilibrium also displays backlash
  - Otherwise, at $t = 1$, the overall effect of a pro-choice decision reduces negative attitudes, i.e. expressive
- Big q or small q?

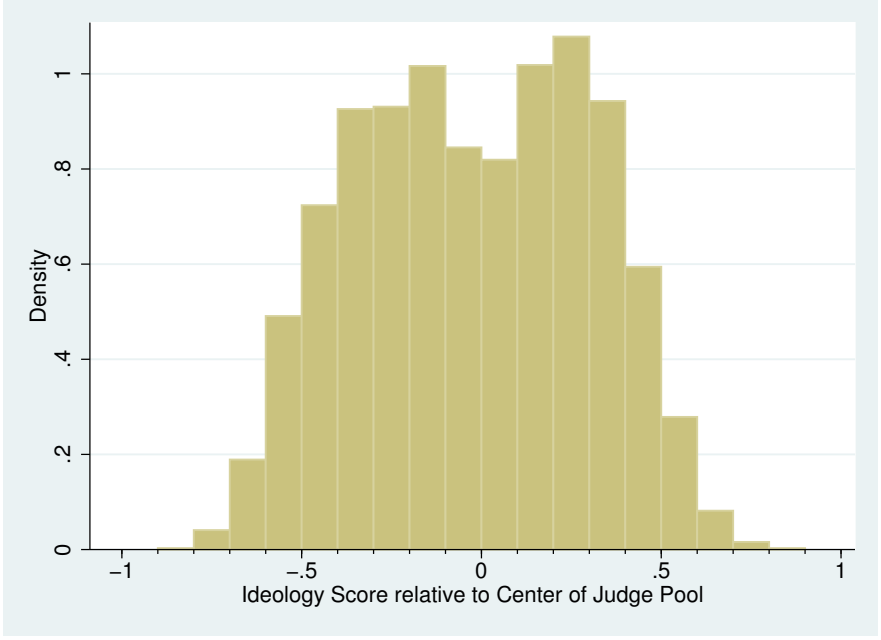
Dissent in Polynomial Distance to Expected Center

	(1)	(2)
	Dissent	Concur
Distance to Center of Judge Pool	0.0404***	0.0285***
	(0.00756)	(0.00570)
Distance^2	-0.0334***	-0.0313***
	(0.0118)	(0.00862)
Circuit Fixed Effects	Y	Y
Year Fixed Effects	Y	Y
N	10043	10043
R-sq	0.109	0.086

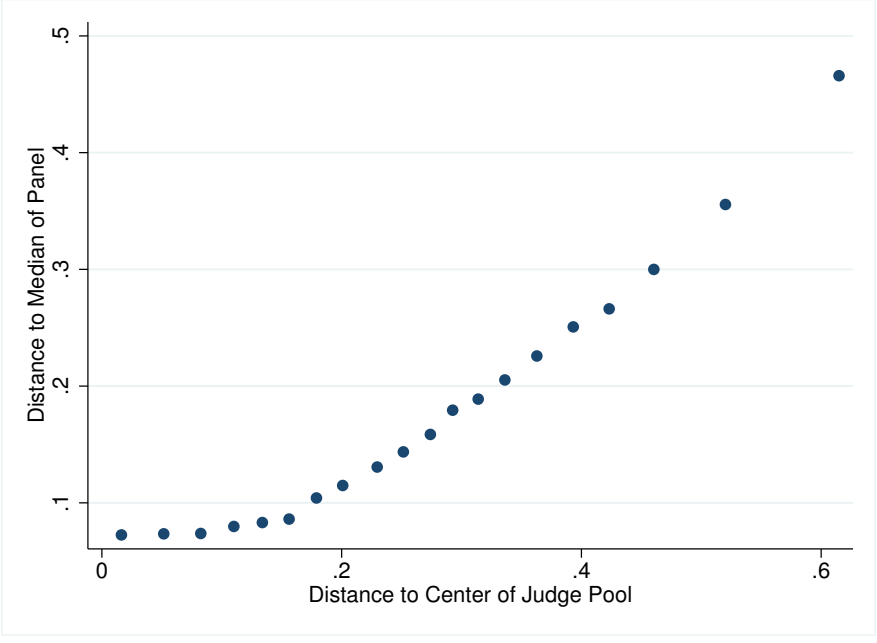
Dissent in Polynomial Distance to Expected Center



Distribution of Ideology Scores (1950-2007)



Distance to Panel Median and Distance to Center of Judge Pool



Alternative Explanations

- Do the results come from preferences?
  - No: result is driven by peer pressure
- Are extreme judges different? e.g., to signal non-bias
  - No: spider shows up mainly for relative measures of extremeness
- Do outliers explain dissent (and concurrence) spider?
  - No: would need low dissent rates for outliers, and dissent rate is bounded by zero
- Convex peer pressure and linear D?
- Is it mechanical that the presence of extreme judges requires large variance in scores?
  - No: there non-monotonicity in the spider

Alternative Explanations

- Extremists dissent less since they want to hide their private (extremist) type. Judges feel collegial pressure for their private views and not for their behavior – i.e. judges try to hide their private preferences from each other.

  - Judges know each other well; still requires a concave D

- Extremist judges think that if the verdict equals their (extremist) type then nobody will take the verdict seriously anyway – it's precedential power will be weak.

  - Requires that the ones who are supposed to cite the verdict have concave costs of deviating from it. If they had convex costs of deviating from a precedent then an extremist would always like extreme verdicts that set precedent.

- Are moderate and centrist judges who happen to sit in a panel with two extremists being backed up by others on the circuit?

  - Does the peer pressure function increase with how extreme you are?

Model

- Bell-shaped continuous distribution of judge types (t between -1 and +1)
- Continuum of cases, three judges picked randomly for each case
- Each judge foresees all cases she will sit in (alt: cases are decided upon simulatenously).
- For verdict $v \in R$, judge feels an outer disutility of $O(|v - t|)$, O is increasing fn
- Judge feels an inner disutility D which is increasing in the cumulative unfavorable verdicts s/he has signed ($s(v) = 1$):

$$D = D(\int_v |t - v| g(v) s(v) dv)$$

- For each dissent ($s(v) = 0$) judge feels collegial pressure W

Timing within Case

1. The three judges suggest and vote about verdict.
2. Each judge decides whether to sign or not.
3. Disutility is applied

$$L = \int_v O(|v - t|)g(v|t)dv$$

$$+D(\int_v (|t - v|)g(v)s(v)dv)$$

$$+W \int_v (1 - s(v))g(v|t)dv$$

Voting Procedure and Outcome

- Each of the three judges suggests a verdict.
- Condorcet winner determines final verdict v.
- Since L is increasing in $|v - t|$:

  Lemma (prediction): In all cases $v = t_m \equiv$ median type

To Sign or Not to Sign

- After v has been determined, outer disutility plays no role in decision making.

$$+D(\int_v |t-v|g(v)s(v)dv)$$

$$+W\int_v(1-s(v))g(v|t)dv$$

To Sign or Not to Sign

- Problem can be rewritten so:
- *Lemma (prediction): Each judge chooses a cutoff verdict distance, c: if verdict is beyond then dissent, if verdict is closer then sign.*
- Probability of dissent

$$P(t,c) = Pr(v < t-c) + Pr(v > t+c)$$

$$= Pr(t_m < t-c) + Pr(t_m > t+c)$$

- For given c, $P(t,c)$ increasing with extremeness $|t|$.
- (For spider we need $P(t,c)$ to decrease(!) to fall for large $|t|$.)
- Hence, necessary condition for $P(t,c(t))$ to fall in $|t|$ is for $c(t)$ to increase in $|t|$:
- **Lemma (prediction): For spider to appear, necessary that extreme judges sign verdicts which are more unfavorable to them than what moderate judges sign.**

What the Spider Needs

Lemma: If D is linear or convex then $c(t)$ is (weakly) decreasing in $|t|$ and hence $P(t, c(t))$ is increasing in $|t|$.

What the Spider Needs

- Lemma: If D is linear or convex then $c(t)$ is weakly decreasing in $|t|$ and hence $P(t, c(t))$ is increasing in $|t|$.

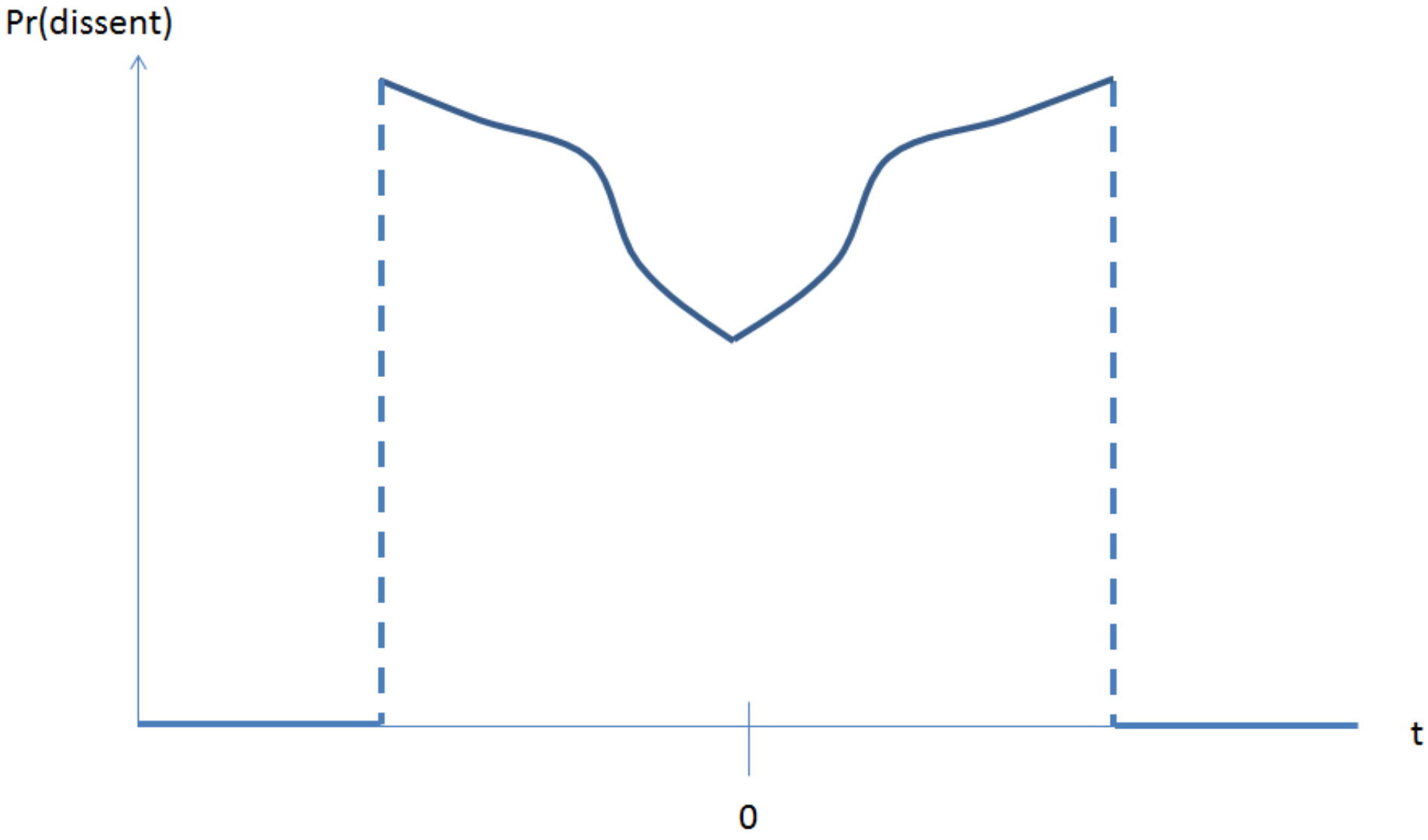
- Proposition: A necessary condition for "the spider" is that D is concave.

What Suffices for the Spider

- Suppose D is a step function
- Then signing any one $v \neq t$ gives same ideological cost as signing many $v \neq t$.
- Meanwhile, collegial cost is increasing in dissent.
- If you sign once, then sign always!
- If you dissent, then dissent any time t$\neq$tm.

$$P(t) = \begin{cases} \Pr(t \neq t_m) & \text{if } |t| < t_{\text{cutoff}} \\ 0 & \text{if } |t| \geq t_{\text{cutoff}} \end{cases}$$

The Fixed Cost Spider



Empirical Prediction: Are Extreme Judges, More Than Others, Signing Verdicts Which Are More Unfavorable?

$$\text{Prob(dissent or concur)} = a + b_1 abs(t) + b_2 abs(t)^2$$

$$+ b_3 abs(t - t_m) + b_4 abs(t - t_m) abs(t)$$

$$+ b_5 abs(t - t_m) abs(t)^2$$

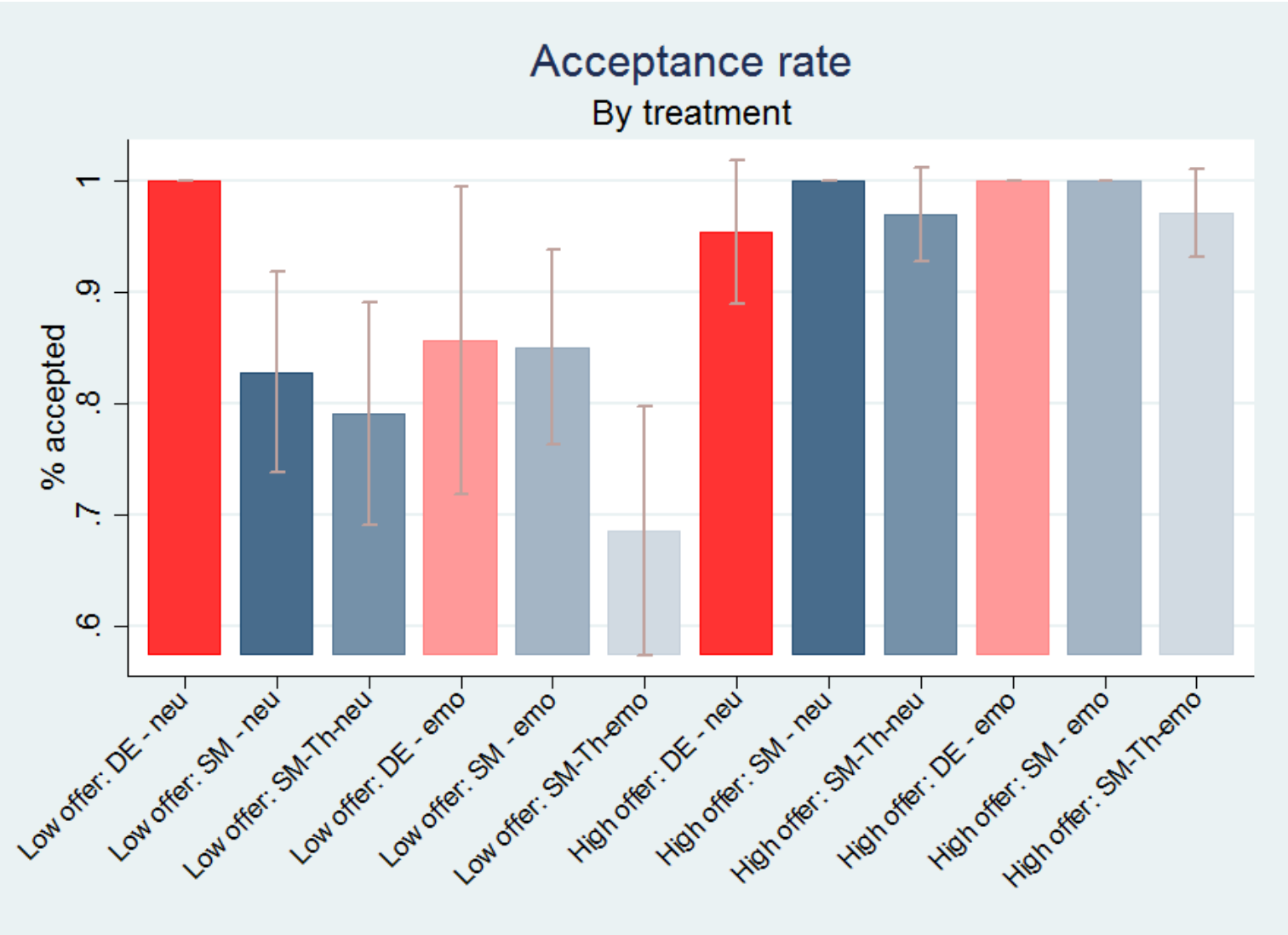
- Cutoff verdict increasing when judges become extreme: $b_5 < 0$
- b_3, b_4, b_5 together such that Pr(dissent) increases in distance from t to *panel* median: judges dissent against unfavorable verdicts

Empirical Prediction: Are Extreme Judges, More Than Others, Signing Verdicts Which Are More Unfavorable?

Dissents and Concurrences vs. Distance to Median of Judge Panel (1950 - 2007)

	(1)
	Dissents or Concurs
Distance to Center of Judge Pool	0.0180
	(0.0225)
Distance to Center of Judge Pool^2	-0.0403
	(0.0389)
Distance to Median of Panel	-0.00335
	(0.00892)
Distance to Median of Panel *	0.244***
Distance to Center of Judge Pool	(0.0572)
Distance to Median of Panel *	-0.287**
Distance to Center of Judge Pool^2	(0.103)
Circuit Fixed Effects	Y
Year Fixed Effects	Y
N	509022
R-sq	0.008

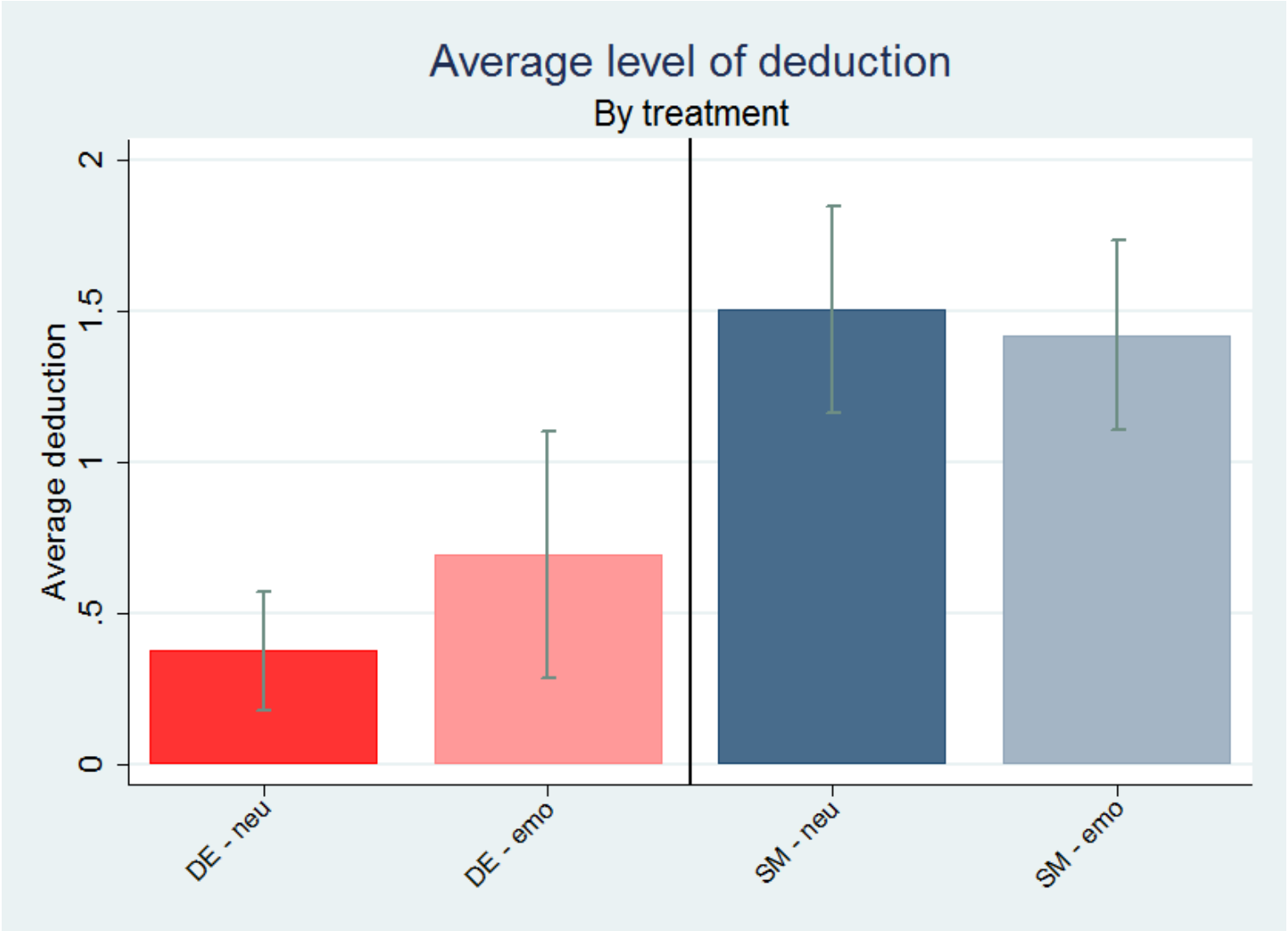
Ultimatum Game (DE vs. SM x Emo vs. Neu)



- Concern: strategy/threshold provides far more data at offer levels that are off the equilibrium or rare.
- Focusing on the sample of frequent offers that are 40 or 50% (these offers occur over 80% of the time)
- Responder acceptance rates of low offers still diverge between direct elicitation and strategy/threshold. Emotions reduce further the willingness for Responders to accept low offers in the threshold setting.

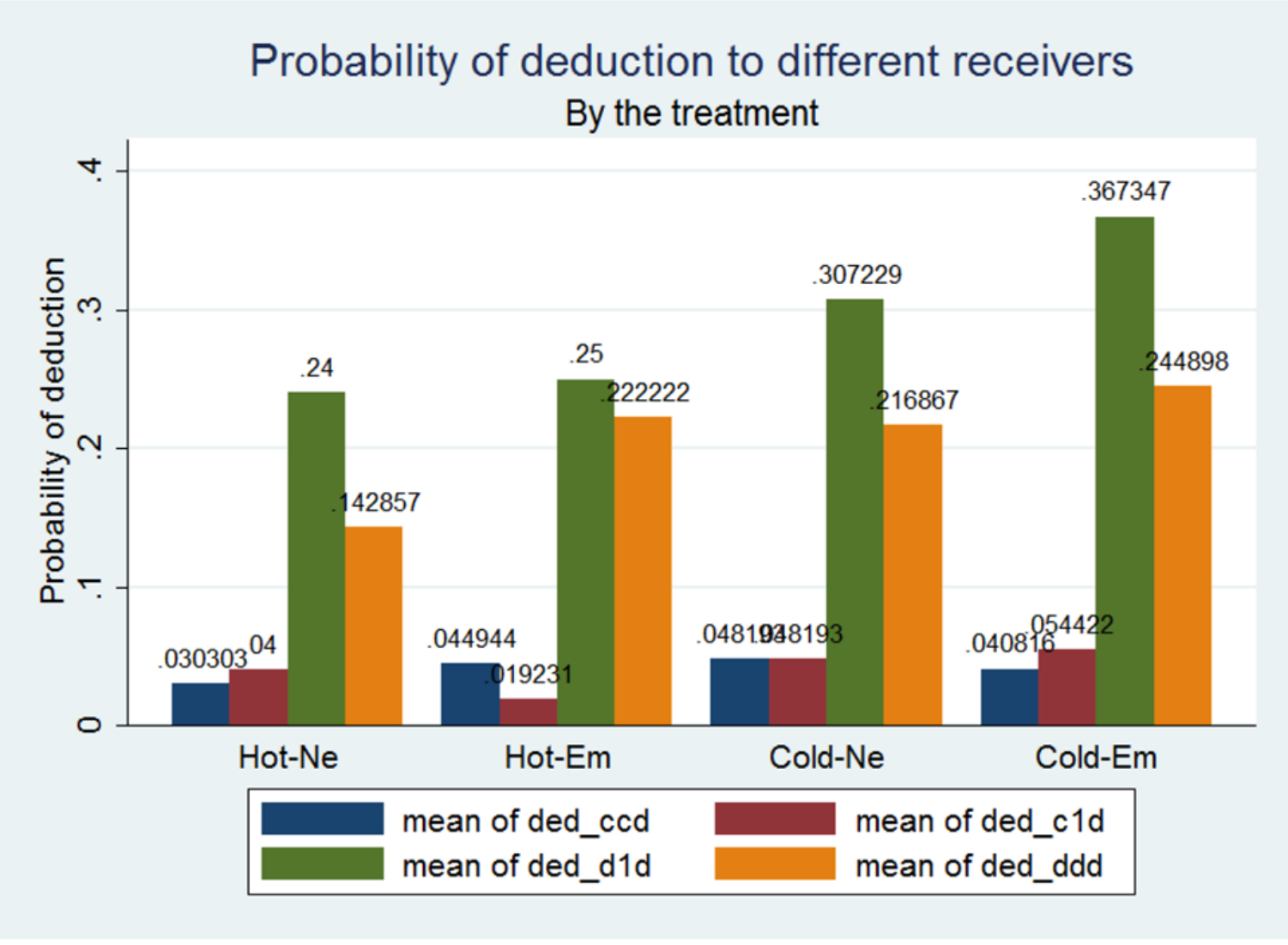
◂ UG

3-Player Prisoner's Dilemma



The strategy method increased deductions and did so primarily in the neutral setting; similar results when controlling for first stage outcome or restricting to specific first stage outcomes ◂ 3PD

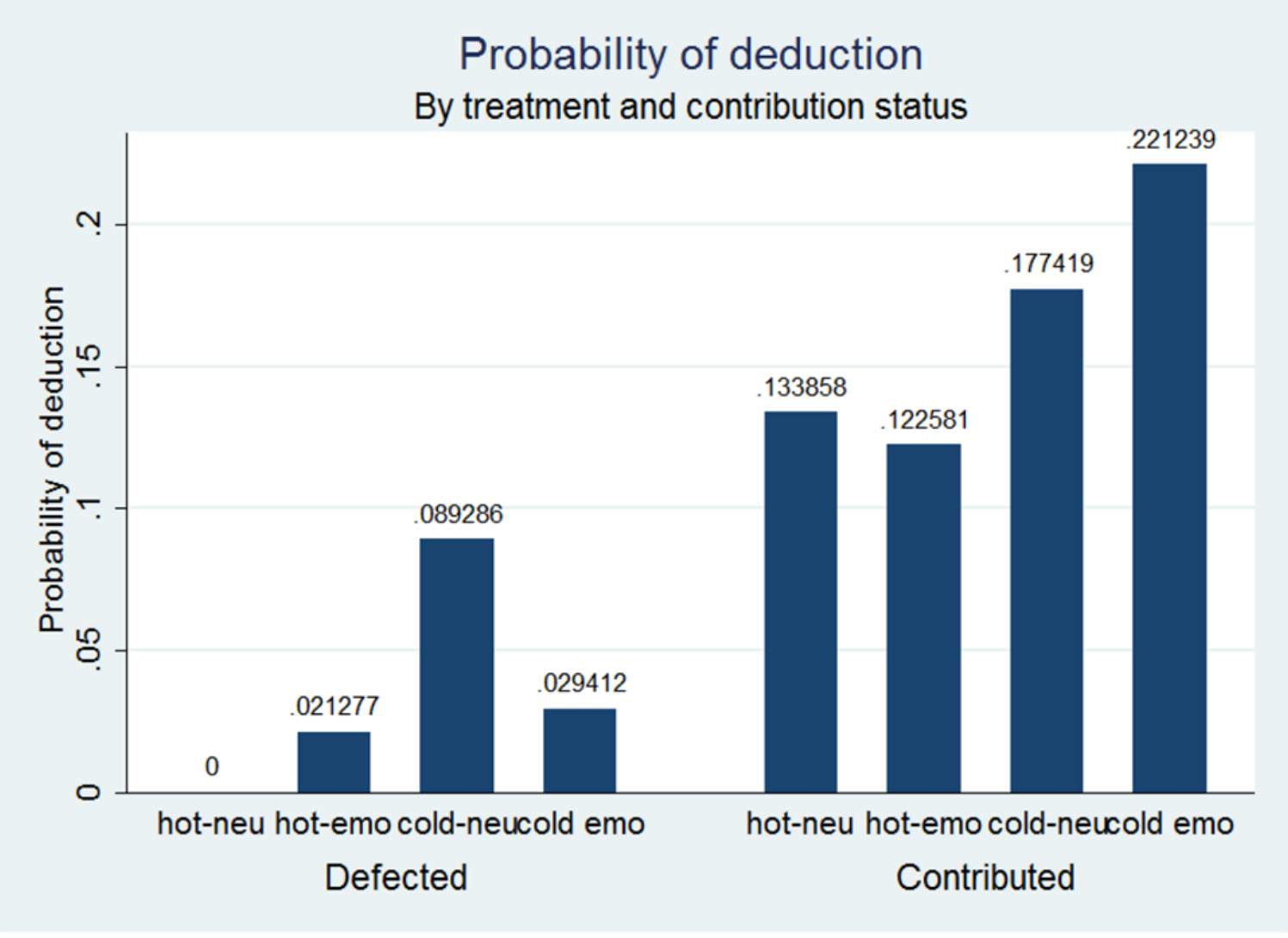
3-Player Prisoner's Dilemma



Deductions for D1 in hot and cold settings are differentially affected by emotional prime

◂ 3PD

3-Player Prisonser's Dilemma



Deductions by contributors in hot vs. cold are differentially affected by emotional prime

◂ 3PD

Sequential Contrast Effects (SCE)

Criteria for quality while judging the current case may be higher if the previous case was particularly high quality (Bhargava and Fisman, 2012)

$$Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \beta_2 Quality_{i,t-1} + Controls + \varepsilon_{it}$$

If SCE causes negatively autocorrelated decisions, we expect $\beta_2 < 0$

- Controlling for discrete decision $Y_{i,t-1}$, decision-makers should be more likely to reject the current case if the previous case was of very high quality, as measured continuously using $Quality_{i,t-1}$

Asylum Judges: Sequential Contrast Effects

	Grant Asylum Dummy	
	(1)	(2)
Lag grant	-0.0356***	-0.0352***
	(0.00788)	(0.00785)
Lag case quality	0.00691*	0.00520
	(0.00385)	(0.00360)
p-value lag case quality < 0	0.0367	0.0751
Quality Measure	1	2
N	23981	23973
R^2	0.228	0.228

- Case quality is predicted using a regression of asylum decisions on applicant characteristics

◂ Abbreviated Results

Quotas and/or Learning

Judges, loan officers (in field experiment), and umpires do not face explicit quotas or targets, but may self-impose these

We control for the fraction of the previous 2-10 decisions that were 1's

- Conditional on this fraction, the most recent decision still negatively predicts the next decision
- Unlikely to be explained by quotas/learning unless unless agents can't remember beyond the most recent decision
- Agents are highly experienced, and quality bar is given in baseball

◂ Abbreviated Results

Concerns about External Perceptions

Decision-maker is rational but judged by others who suffer from the gambler's fallacy

- This is broadly consistent with our hypothesis and would be an interesting question for follow-up research
- Not likely to be a strong factor in the loan officers experiment where they are paid for accuracy
- Asylum judges typically serve until retirement, are paid fixed salary, and can discriminate by nationality of asylum applicant
- Negative autocorrelation in umpire calls does not vary dramatically by game attendance or leverage

◂ Abbreviated Results

Fraction of Decisions Altered by Gambler's Fallacy

Simple regression

$$Y_{it} = \beta_0 + \beta_1 Y_{i,t-1} + \varepsilon_{it}$$

Base rate of affirmatives

$$\alpha \equiv P(Y = 1) = \frac{\beta_0}{1 - \beta_1}$$

Fraction of decisions altered ◂ Abbreviated Results

$$(\beta_0 - \alpha) \cdot P(Y_{i,t-1} = 0) + (\alpha - (\beta_0 + \beta_1)) \cdot P(Y_{i,t-1} = 1)$$

$$= 2\beta_1 \alpha (1 - \alpha)$$

Asylum Judges: First-in-First-Out

FIFO can be violated if asylum applicant claims work hardship, files additional applications, etc.

- Assume these violations of FIFO, which are driven by applicant behaviors, are not negatively correlated with the previous decision
- Asylum judges scheduling system usually picks the next available date

- We estimate the "quality" of each case by regressing grant decisions on case characteristics and using the predicted grant outcomes
  - Predicted case quality is positively autocorrelated
- Previous grant or deny decisions do not significantly predict whether the next case has a written decision, remote hearing, or non-decision

Asylum Judges: Ordering of Case Quality

	Lawyer Dummy	Lawyer Quality	Size of Family
	(1)	(2)	(3)
Lag grant	-0.0000772	-0.00117	-0.00927
	(0.00258)	(0.00293)	(0.0104)
N	23,990	19,737	23,990
R^2	0.0858	0.451	0.159

- A previous grant decision does not predict that the next case will be lower in observed quality measures
- ◂ Asylum Background

Asylum Judges: Summary Statistics

	Mean	Median	S.D.
Number of judges	357		
Number of courts	45		
Years since appointment	8.41	8	6.06
Daily caseload of judge	1.89	2	0.84
Family size	1.21	1	0.64
Grant indicator	0.29		
Non-extreme indicator	0.54		
Moderate indicator	0.25		
Lawyer indicator	0.939		
Defensive indicator	0.437		
Morning indicator	0.47		
Lunchtime indicator	0.38		
Afternoon indicator	0.15		