Behind the Veil of Ignorance:
Choosing in [Harsanyi's] Original Position

Shachar Kariv (Berkeley)

Bill Zame (UCLA)

To be presented at Academia Sinica
Mar 3, 2009

The original position

Harsanyi and Rawls argue for theories of *social justice* based on the choices that agents would make for society in the *original position*, behind a *veil of ignorance*.

> *. . . without knowing their own social and economic positions, their own special interests in the society, or even their own personal talents and abilities (or their lack of them).* – Harsanyi (1975) –

Harsanyi and Rawls come to quite different conclusions, not because they view the original position differently, but because they treat uncertainty quite differently.

Harsanyi's (1953, 1955) model for moral value judgments

Suppose an agent wants to make a moral value judgment about the relative merits of two alternative social systems.

> *. . . act in such a way as if he assigned the same probability to his occupying each social position under either system. . .*
>
> *. . . then, he would clearly satisfy the impartiality and impersonality requirements to the fullest possible degree.* – Harsanyi (1978) –

The agent has two different sets of preferences: *personal preferences* and *moral preferences*.

Our point of departure from the work of Harsanyi and Rawls – and the enormous literature they spawned – comes from two observations:

[1] Choice behavior/preferences *behind* the veil of ignorance can be decomposed into choice behavior/preferences *in front of* the veil of ignorance:

> Choices that involve only personal consumption under uncertainty and choices that involve personal and social consumption – but no uncertainty.

[2] Choices behind the veil of ignorance *can* be presented – and choices in the other two environments as well – in a controlled *laboratory* setting.

Template for analysis

- Consider choice behavior by a single agent in each of three environments.

- Each choice has consequences for *self* (the agent) and for an (unknown) *other*.

- We consider only environments that involve binary choices and equiprobable lotteries.

- The results extend to more general choices and lotteries, and to unknown probabilities as well.

Consider lotteries over outcomes $[a, b]$, where a is consumption for *other* and b is consumption for *self*.

For our purposes, it suffices to consider binary lotteries with equal probabilities:

$$(.5)[a, b] + (.5)[c, d]$$

where $a, b, c, d \geq 0$. Write $\mathcal{L}$ for the space of all such lotteries, and identify $\mathcal{L}$ with the convex cone $\mathbb{R}^4_+$.

Define closed convex subcones of $\mathcal{L}$:

$$\begin{aligned}
\mathcal{R} &= \{(.5)[0,b]+(.5)[0,d]\},\\
\mathcal{S} &= \{(.5)[a,b]+(.5)[a,b]\},\\
\mathcal{V} &= \{(.5)[a,b]+(.5)[b,a]\}.
\end{aligned}$$

We can interpret choice in each of the environments as choice in one of the corresponding cones by making an obvious identification:

– Risk: identify $\mathbb{R}^2_+$ with $\mathcal{R}$ by

$$(x, y) \mapsto (.5)[0, x] + (.5)[0, y].$$

– Social Choice: identify $\mathbb{R}^2_+$ with $\mathcal{S}$ by

$$(x, y) \mapsto (.5)[x, y] + (.5)[x, y].$$

– Veil of Ignorance: identify $\mathbb{R}^2_+$ with $\mathcal{V}$ by

$$(x, y) \mapsto (.5)[x, y] + (.5)[y, x].$$

Research questions

[1] What is the relationship between moral preferences and personal/altruistic preferences?

[2] How can behavior behind [Harsanyi's] veil of ignorance be characterized experimentally?

[3] Is behavior behind a veil of ignorance consistent with the utility maximization model?

[4] Can the underlying moral preferences be recovered from observed choices?

Assumptions

Given a preference relation $\succeq$ on $\mathcal{L}$, write $\succeq_{\mathcal{R}}$, $\succeq_{\mathcal{S}}$, $\succeq_{\mathcal{V}}$ for its restrictions to $\mathcal{R}$, $\mathcal{S}$, $\mathcal{V}$, respectively.

$[i]$ $\succeq$ satisfies the usual requirements: completeness, transitivity, reflexivity, continuity, and the Sure Thing Principle.

$[ii]$ $\succeq$ satisfies (weak) *independence*:

$$\begin{aligned} [a,b] \succeq_{\mathcal{S}} [a',b'] \quad \text{and} \quad & [c,d] \succeq_{\mathcal{S}} [c',d'] \\ \Rightarrow \quad & (.5)[a,b] + (.5)[c,d] \succeq (.5)[a',b'] + (.5)[c',d'] \end{aligned}$$

(*not* the usual independence axiom and does not have the usual consequences).

$[iii]$ Zero is the worst outcome: $[a, b] \succeq_{\mathcal{S}} [0, 0]$ for every $[a, b] \in \mathcal{S}$

$[iv]$ $\succeq_{\mathcal{S}}$ is *self-regarding*: for each outcome $[a, b]$ there is an outcome $[0, s]$ such that $[0, s] \succeq_{\mathcal{S}} [a, b]$.

$[i]$ and $[ii]$ are rationality requirements (should not necessarily be given any philosophical interpretation).

$[iii]$ and $[iv]$ limit the extent to which the *self* is (respectively) spiteful or altruistic toward *others*; they seem very natural requirements but they are not entirely innocuous.

Result I: Every preference relation $\succeq$ on $\mathcal{L}$ that satisfies $[i]$-$[iv]$ is determined by its restrictions $\succeq_{\mathcal{R}}$and $\succeq_{\mathcal{S}}$.

Proof: Fix an outcome $[x, y]$. Because $\succeq_{\mathcal{S}}$ is self-regarding, there is some s such that $[0, s] \succeq_{\mathcal{S}} [x, y]$.

Define the *selfish equivalent* of $[x, y]$ by

$$\sigma[x, y] = \inf\{s : [0, s] \succeq_{\mathcal{S}} [x, y]\}.$$

Continuity and worse outcome guarantee that $[0, \sigma[x, y]] \sim_{\mathcal{S}} [x, y]$, and by construction,

$$[a, b] \sim_{\mathcal{S}} [0, \sigma[a, b]] \text{ and } [c, d] \sim_{\mathcal{S}} [0, \sigma[c, d]].$$

independence guarantees that

$$(.5)[a,b] + (.5)[c,d] \sim (.5)[0,\sigma[a,b]] + (.5)[0,\sigma[0,\sigma[c,d]].$$

Hence

$$\begin{array}{rcl} (.5)[a,b] + (.5)[c,d] & \succeq & (.5)[a',b'] + (.5)[c',d'] \\ & \Updownarrow & \\ (.5)[0,\sigma[a,b]] + (.5)[0,\sigma[c,d]] & \succeq & {}_{\mathcal{R}}(.5)[0,\sigma[a',b']] + (.5)[0,\sigma[c',d']] \end{array}$$

which decomposes preferences over $\mathcal{L}$ into preferences over $\mathcal{S}$ (selfish equivalents) and preferences over $\mathcal{R}$, as desired.

Given a linear budget constraint, we identify choice behavior in the Social Choice environment as

- *selfish* if the choice subject to every budget constraint is of the form $[0, y]$ – giving nothing to *other*.

- *symmetric* if (a, b) is chosen subject to $px+qy \leq w$ iff (b, a) is chosen subject to the mirror-image budget constraint $qx + py \leq w$.

Result II: If the preference relation $\succeq$ satisfies $[i]$ and $[ii]$ and choice behavior in the Social Choice environment is selfish then choice behavior in the Risk environment coincides with choice behavior in the Veil of Ignorance environment.

Result III: If the preference relation $\succeq$ satisfies $[i]$ and $[ii]$ and choice behavior in the Social Choice environment is symmetric, then choice behavior in the Social Choice environment coincides with choice behavior in the Veil of Ignorance environment.

Experimental analysis

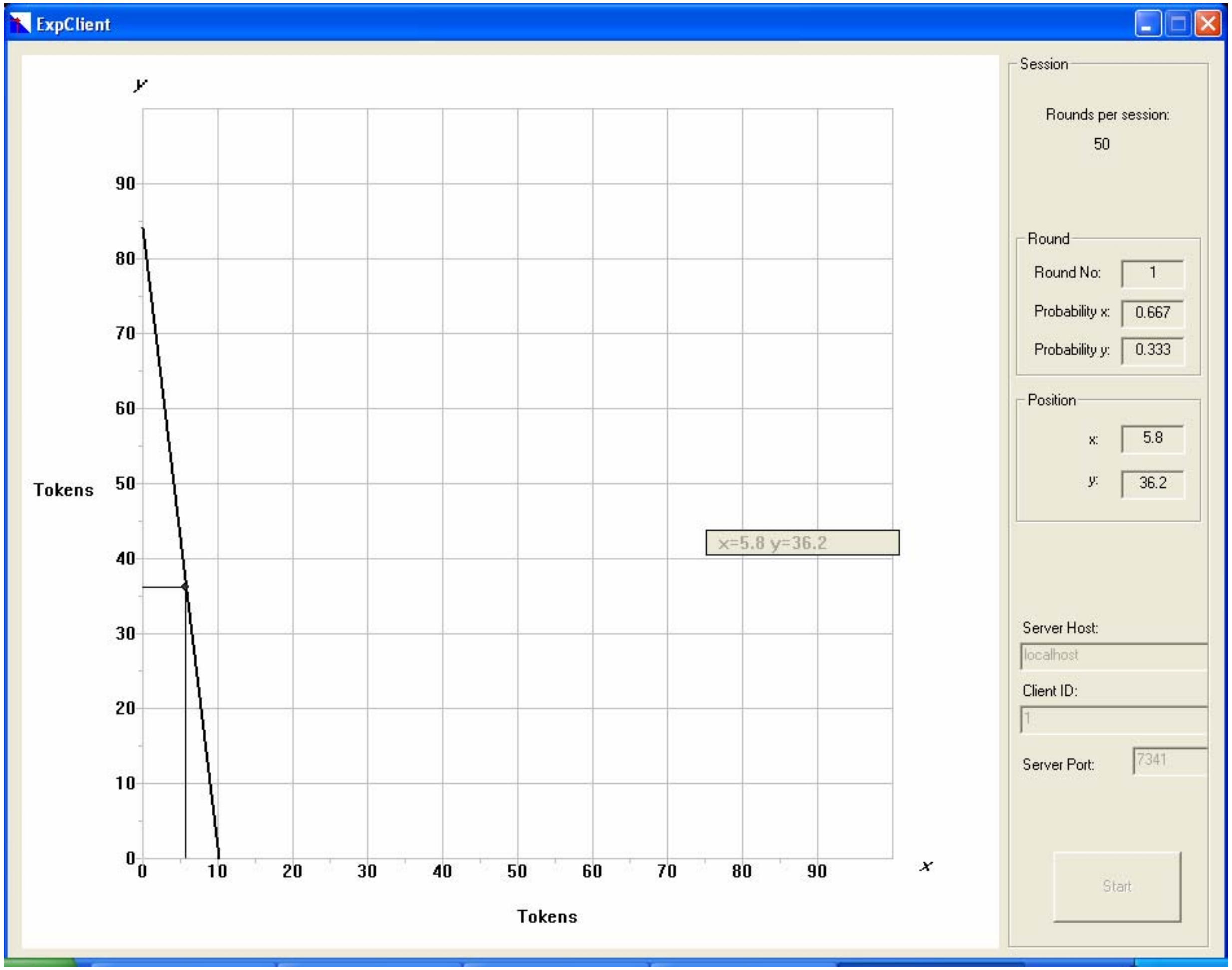
- Subjects in the experiments were recruited from all classes at UCLA and Yale Law School.

- Each decision problem is presented as a choice from a two-dimensional budget line.

- A choice (x, y) from the budget line represents an allocation between accounts x, y (corresponding to the horizontal and vertical axes).

- Choices are made through a simple point-and-click design using a graphical computer interface.

The computer program dialog window



Session

Rounds per session:
50

Round

Round No: 1

Probability x: 0.667

Probability y: 0.333

Position

x: 5.8

y: 36.2

Server Host: localhost

Client ID: 1

Server Port: 7341

The actual payoffs of a particular choice in a particular environment/treatment are determined by the allocation to the x and y accounts:

- Risk: involves only pure risk; it is identical to the (symmetric) risk experiment of Choi, Fisman, Gale & Kariv (*AER*, 2007).

- Social Choice: involves only altruism; it is identical to the (linear) two-person dictator experiment of Fisman, Kariv & Markovits (*AER*, 2007).

- Veil of Ignorance: involves equiprobable binary lotteries over symmetric pairs of consumption for *self* and for *other*.

The advantages of this experimental design are several:

- The choice of an allocation subject to a budget constraint provides more information than a binary choice.

- Quick and efficient elicitation of many decisions per subject under a wide range of budget sets.

- Apply statistical models to estimate preferences at the level of the individual subject rather than assuming homogeneity across subjects.

Testing rationality

Let $\{(p^i, x^i)\}_{i=1}^{50}$ be some observed individual data (p^i denotes the i-th observation of the price vector and x^i denotes the associated allocation).

A utility function $u(x)$ *rationalizes* the observed behavior if it achieves the maximum on the budget set at the chosen allocation

$$u(x^i) \geq u(x) \text{ for all } x \text{ s.t. } p^i \cdot x^i \geq p^i \cdot x.$$

Generalized Axiom of Revealed Preference (GARP): If x^i is indirectly revealed preferred to x^j, then x^j is not strictly directly revealed preferred (i.e. $p^j \cdot x^j \leq p^j \cdot x^i$) to x^i.

GARP is tied to utility representation through a theorem, which was first proved by Afriat (1967).

Afriat's (1967) Theorem: The following conditions are equivalent:

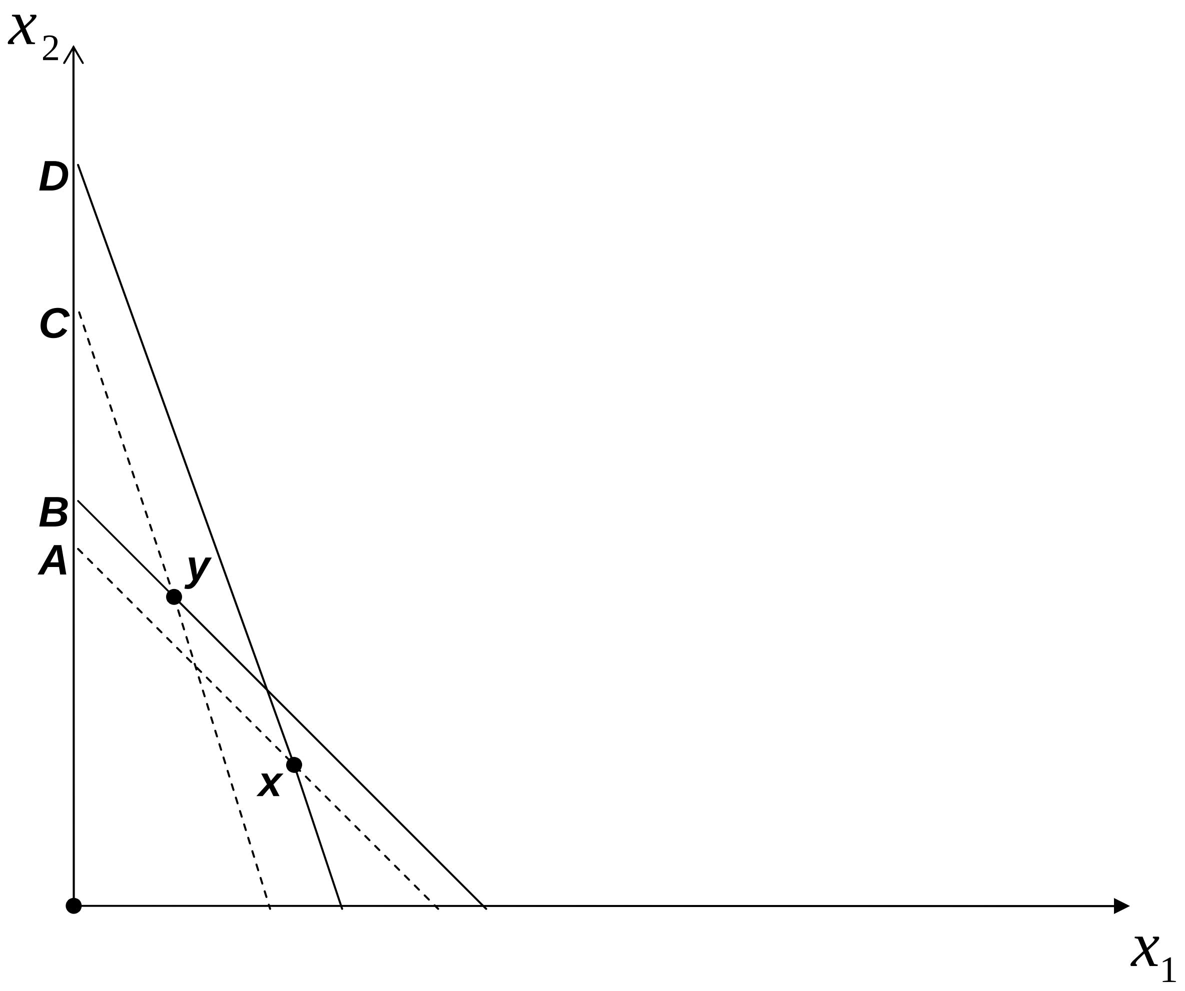
- The data satisfy GARP.
- There exists a non-satiated utility function that rationalizes the data.
- There exists a concave, monotonic, continuous, non-satiated utility function that rationalizes the data.

Since GARP offers an exact test, it is necessary to measure the extent of GARP violations.

Afriat's (1972) critical cost efficiency index (CCEI): The amount by which each budget constraint must be "relaxed" in order to remove all violations of GARP.

The CCEI is bounded between zero and one. The closer it is to one, the smaller the perturbation required to remove all violations and thus the closer the data are to satisfying GARP.

The construction of the CCEI for a simple violation of GARP



The agent is ‘wasting' as much as *A/B*<*C/D* of his income by making inefficient choices.

To provide a benchmark level of consistency, consider random sample of hypothetical subjects who implement the power utility function

$$u(x) = \frac{x^{1-\rho}}{1-\rho},$$

commonly employed in the empirical analysis of choice under risk, with error.

The likelihood of error is assumed to be a decreasing function of the utility cost of an error.

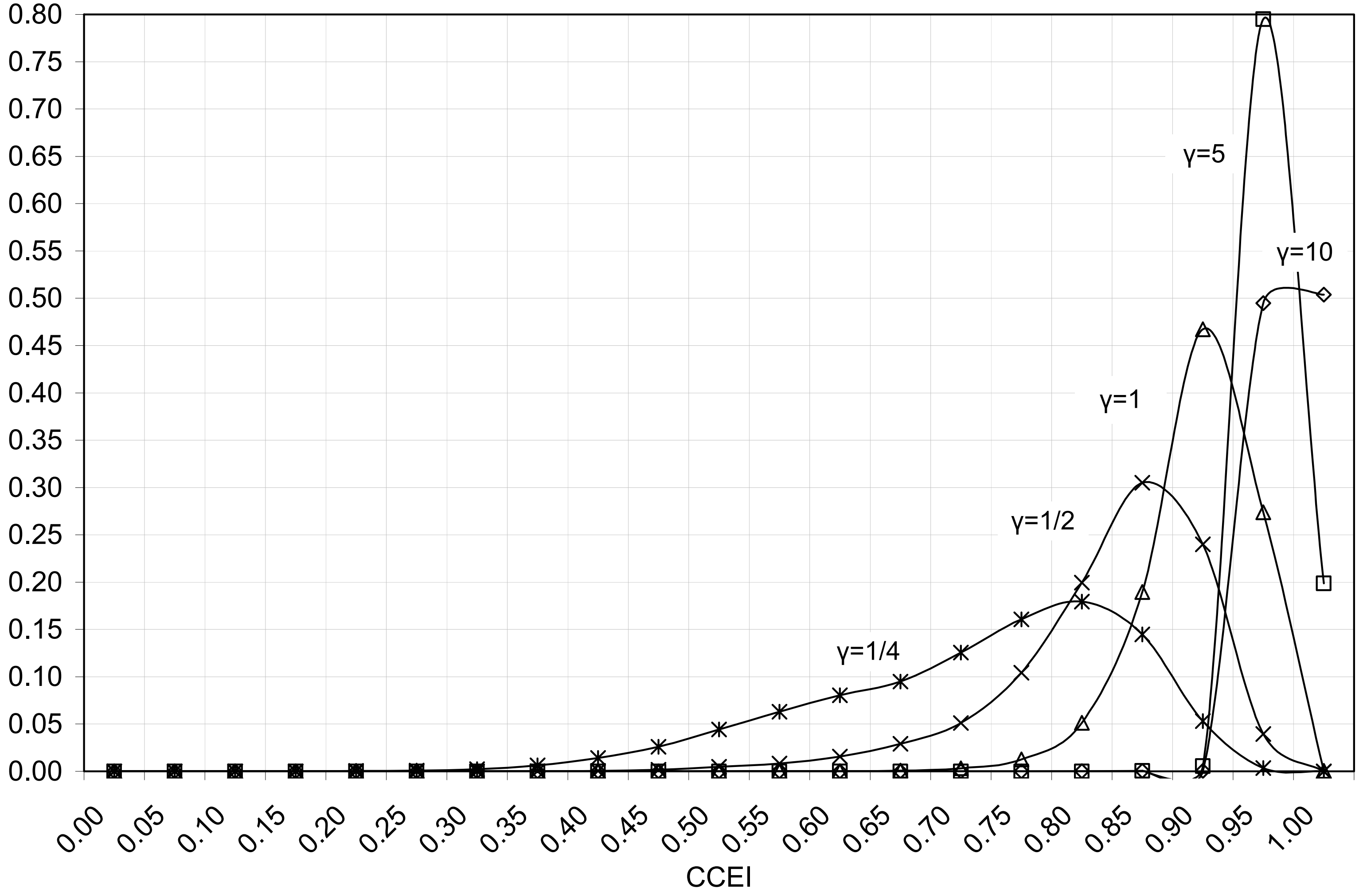
More precisely, we assume an idiosyncratic preference shock that has a logistic distribution

$$\text{Pr}(x^*) = \frac{e^{\gamma \cdot u(x^*)}}{\int\limits_{x:p\cdot x=1} e^{\gamma \cdot u(x)}},$$

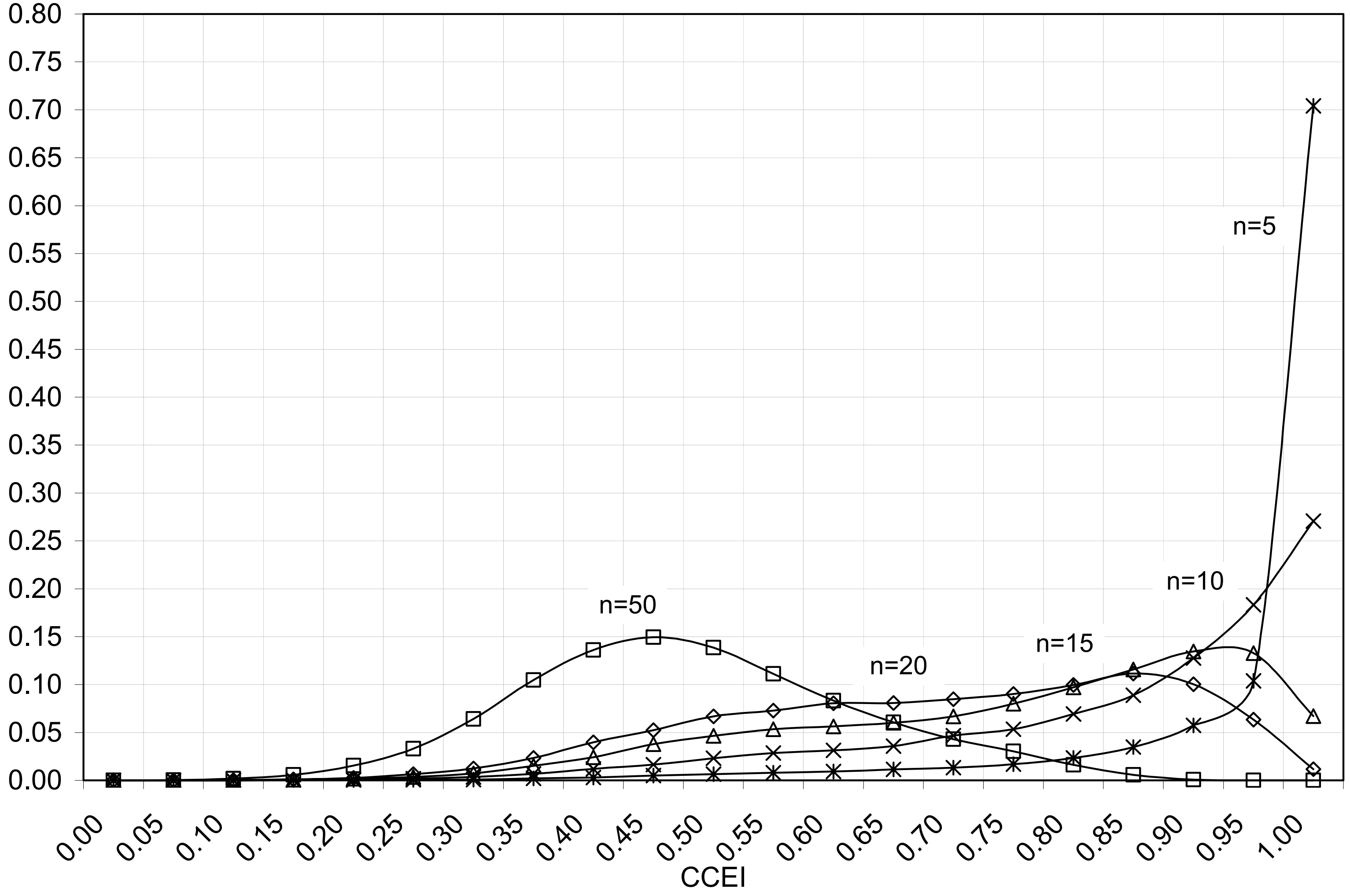
where the precision parameter γ reflects sensitivity to differences in utility.

If utility maximization is not the correct model, is our experiment sufficiently powerful to detect it?

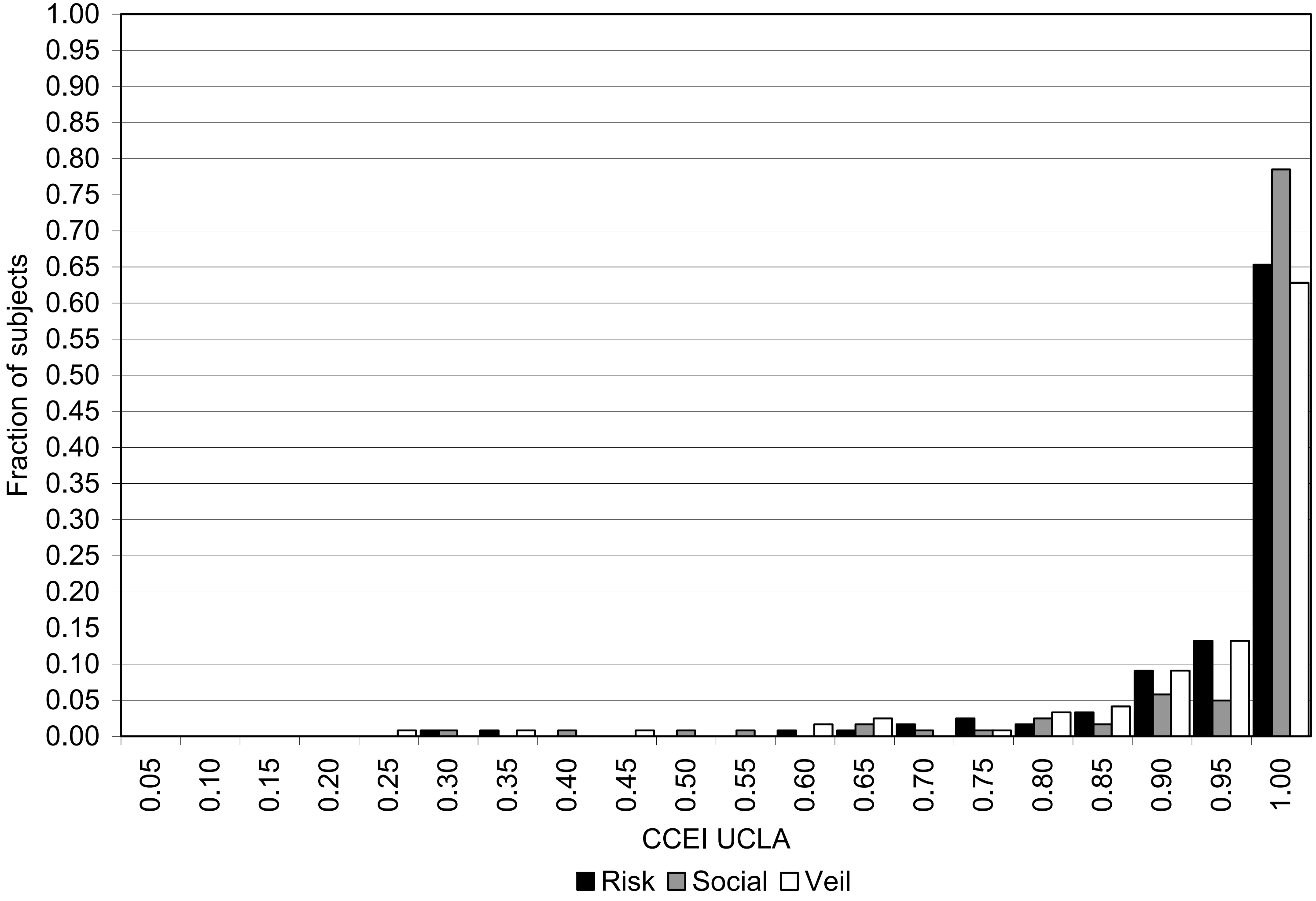
The distributions of GARP violations – ρ=1/2 and different γ

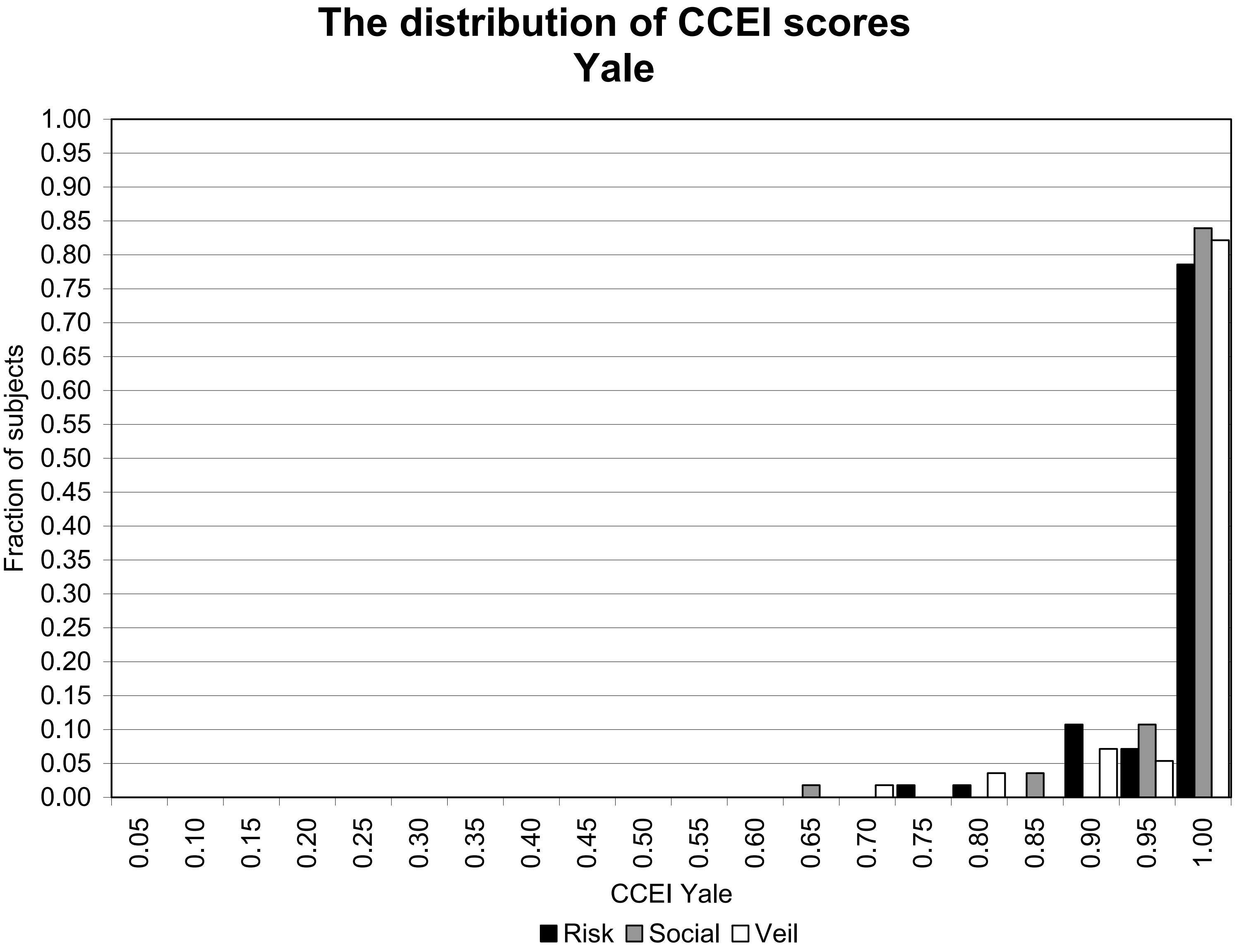


Bronnars’ (1987) test (γ=0)



The distributions of CCEI scores
UCLA





The distribution of CCEI scores
Yale

Recovering preferences

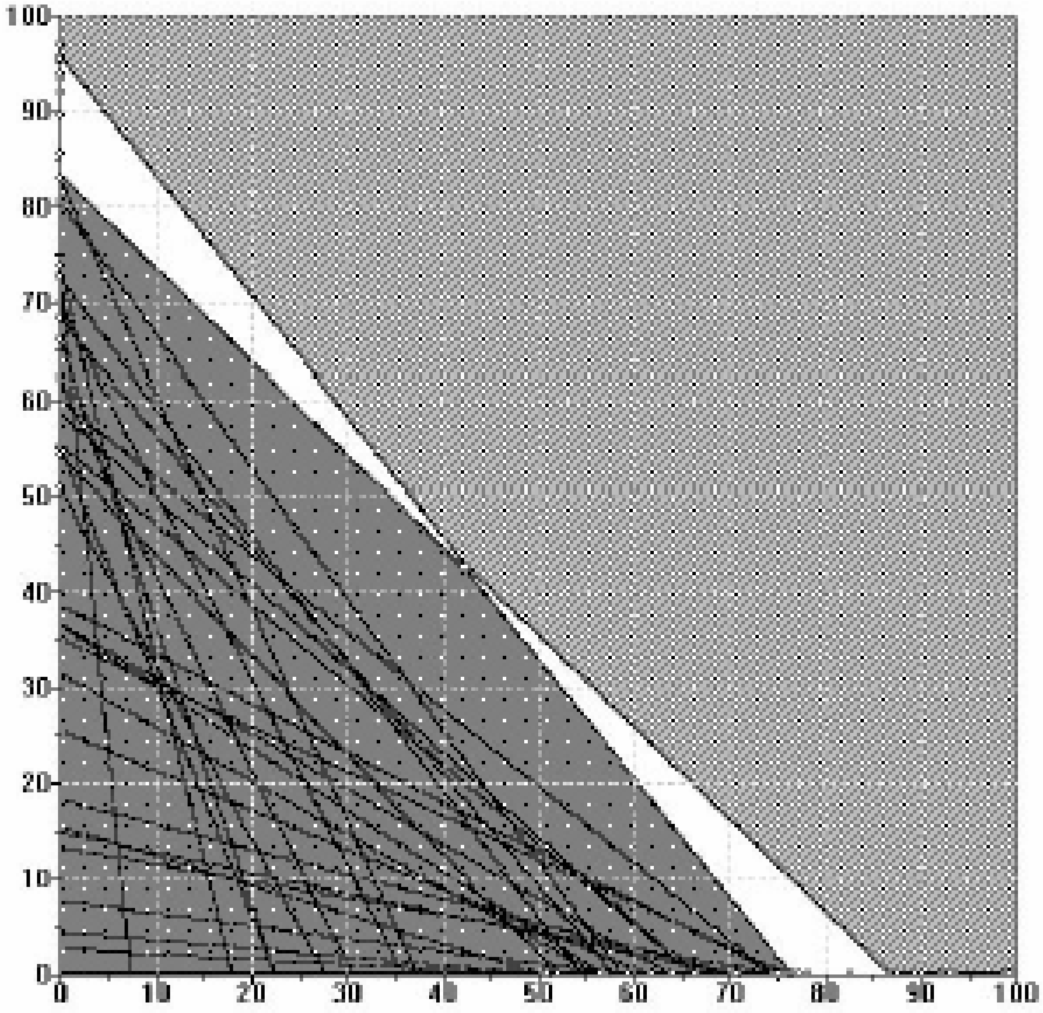
- GARP imposes on the data the complete set of conditions implied by utility-maximization.

- Revealed preference relations in the data thus contain the information that is necessary for recovering preferences.

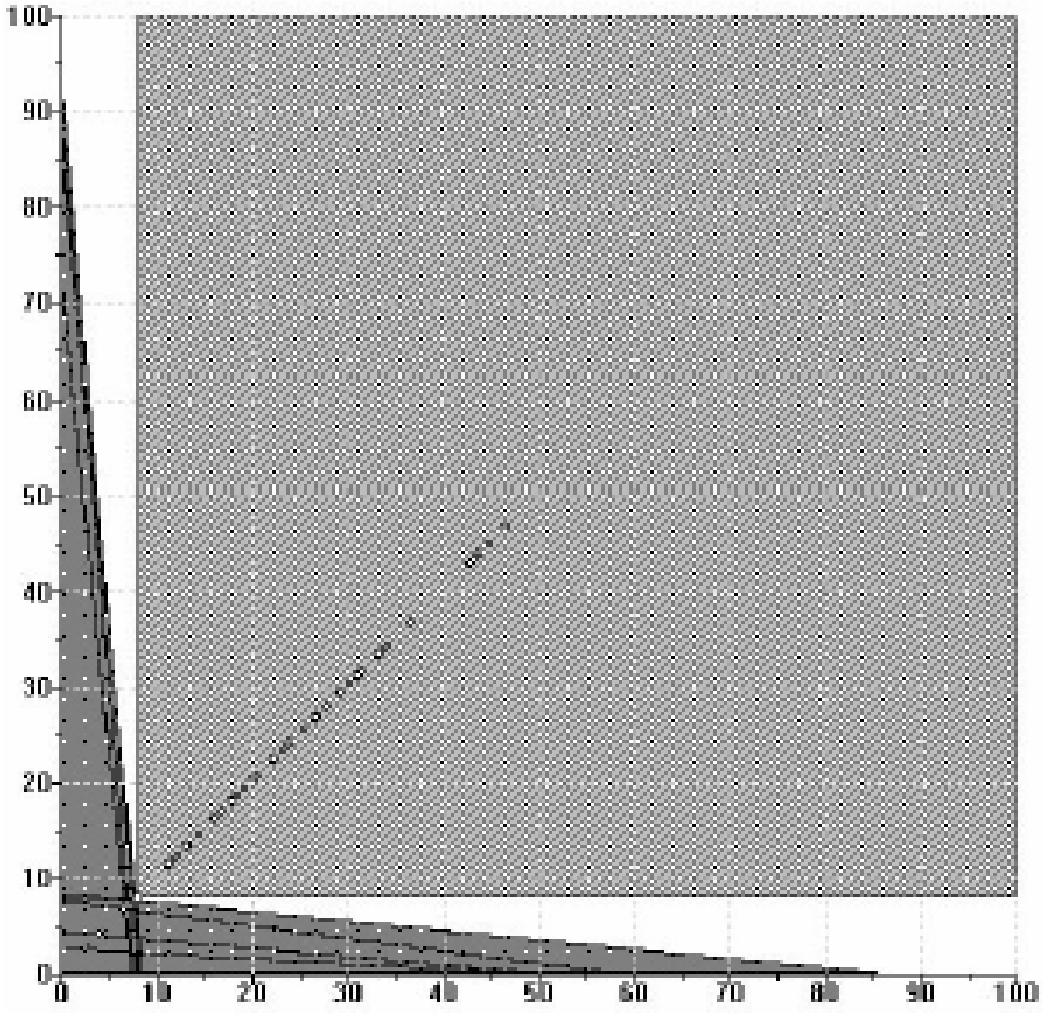
- Varian's (1982) algorithm serves as a partial solution to this so-called *recoverability problem.*

- Case studies of subjects who serve to illustrate ideal types whose choices fit with prototypical preferences.

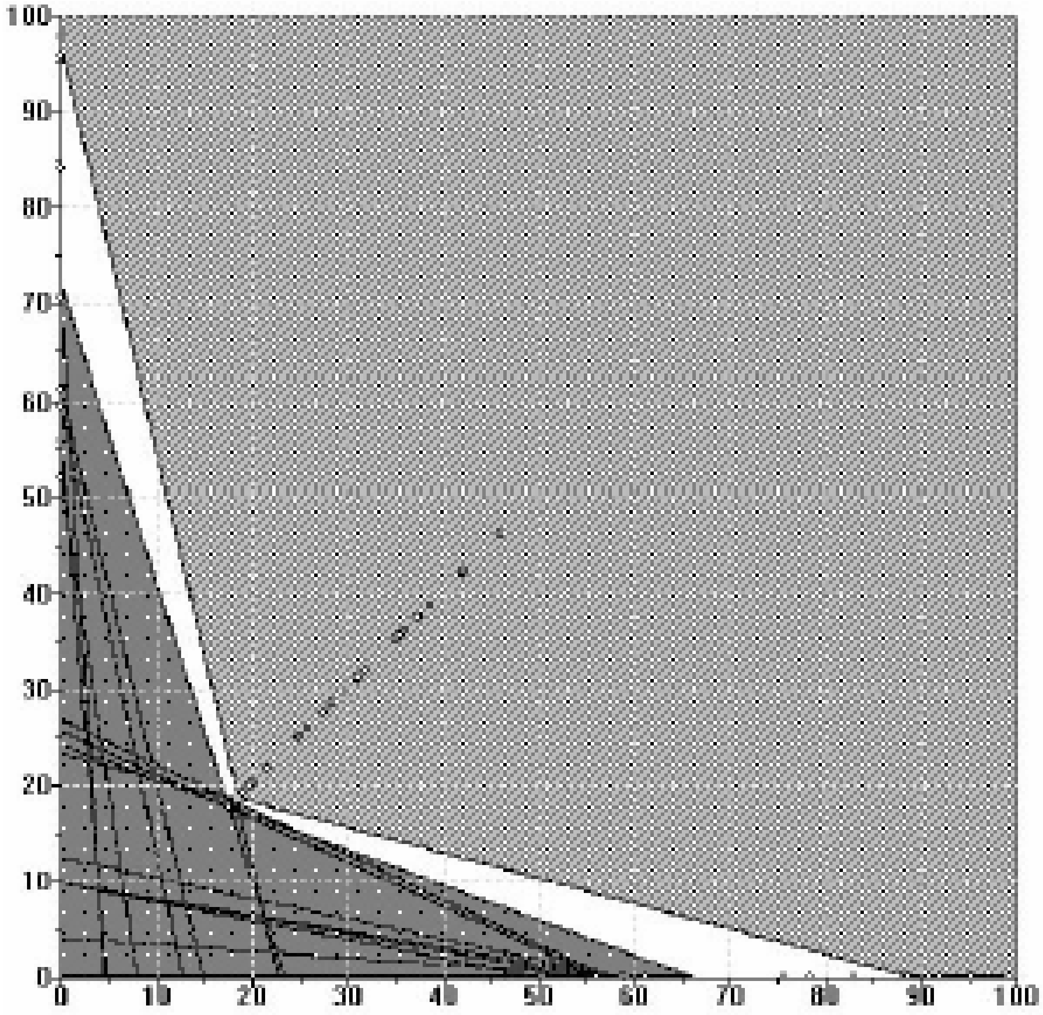
Risk neutrality



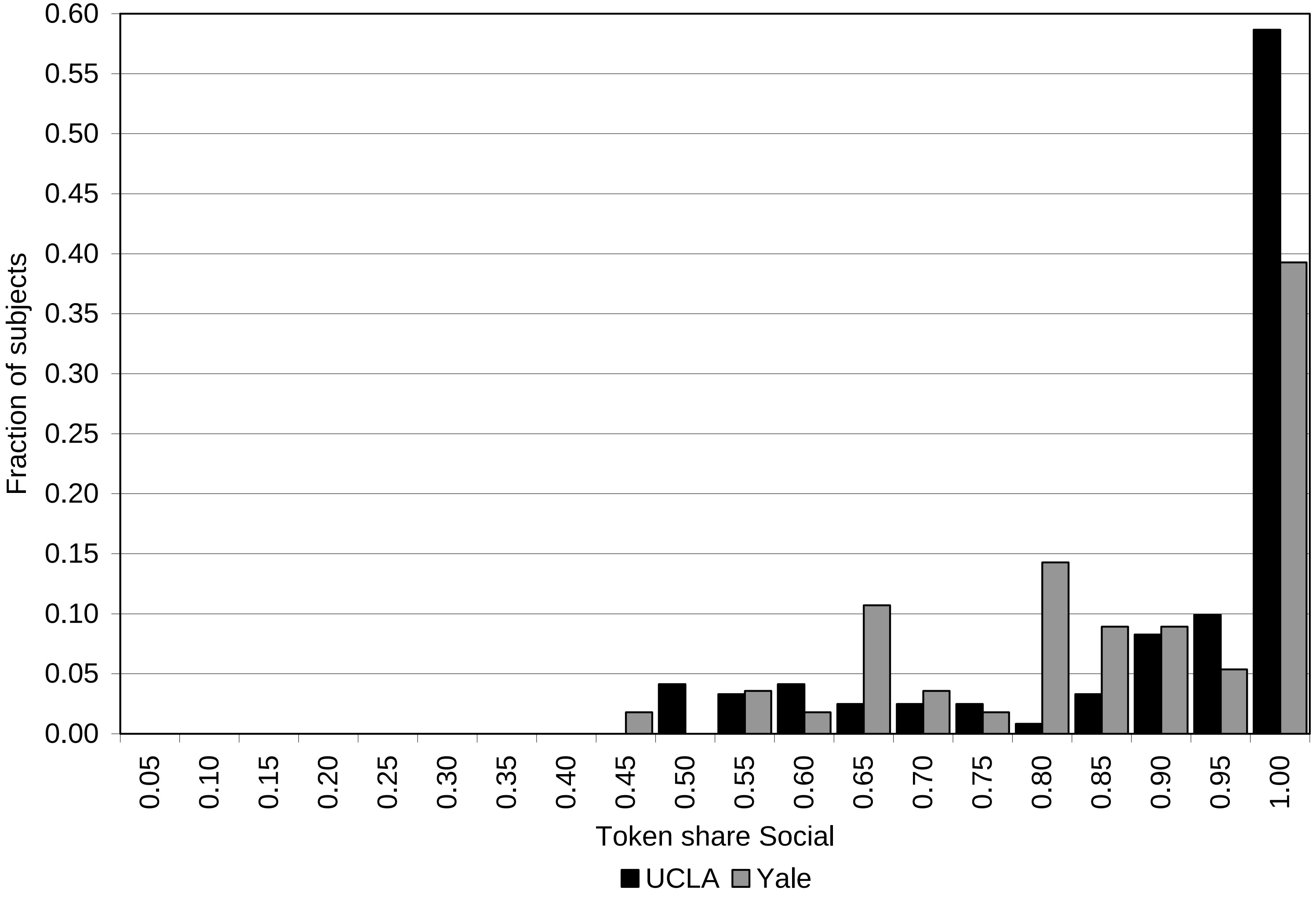
Infinite risk aversion



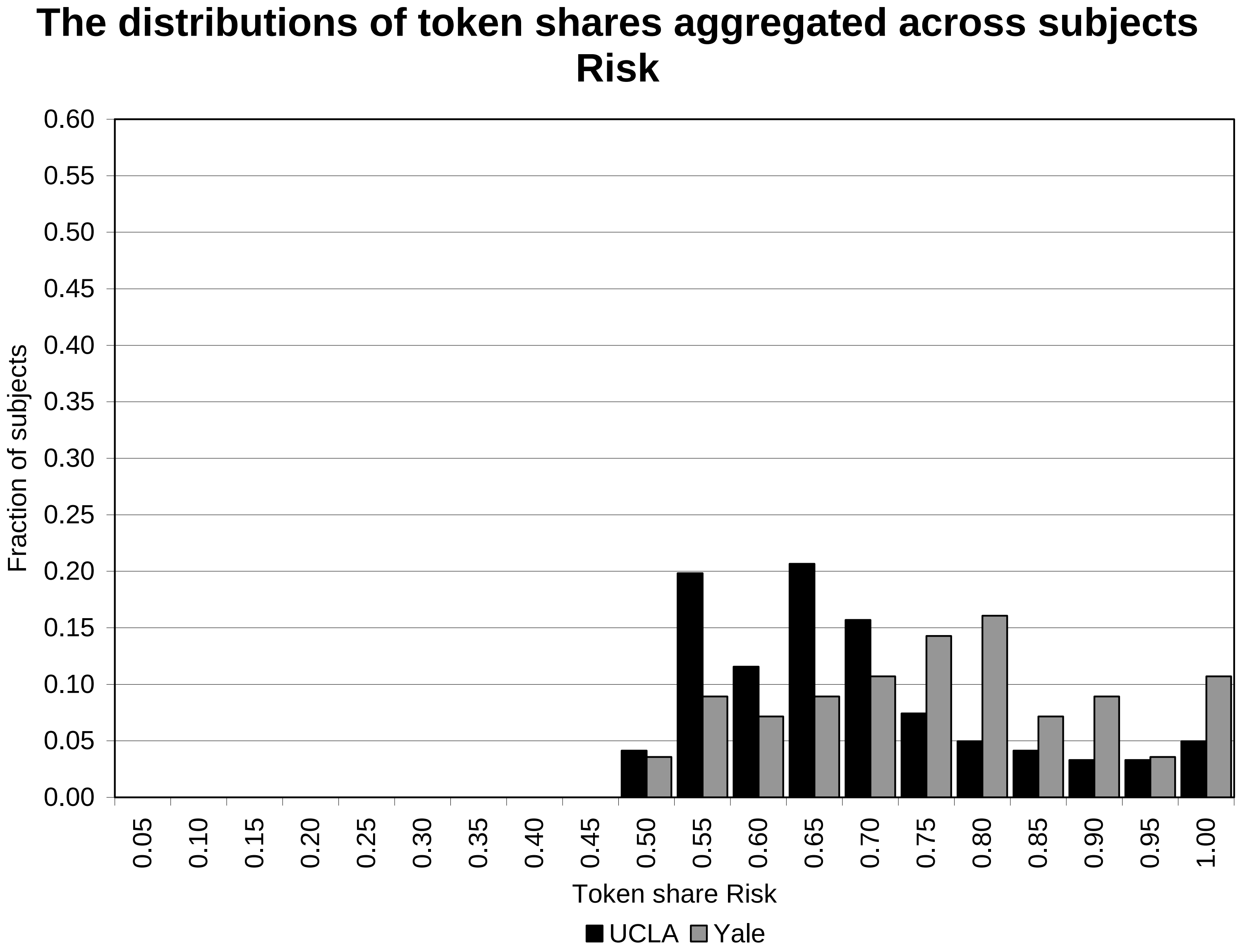
Loss / disappointment aversion



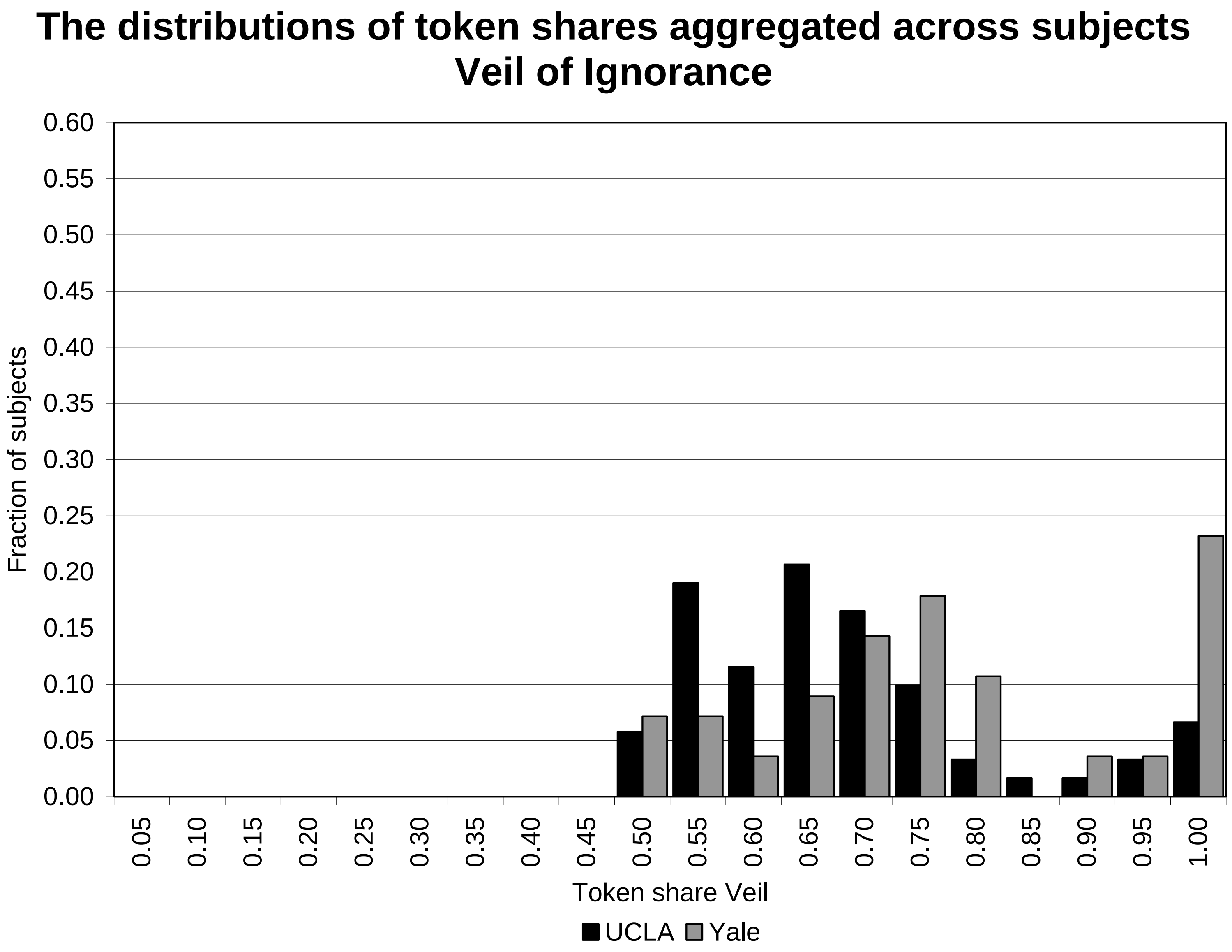
The distributions of token shares aggregated across subjects
Social Choice



The tokens kept as a fraction of the sum of the tokens kept and given to other.



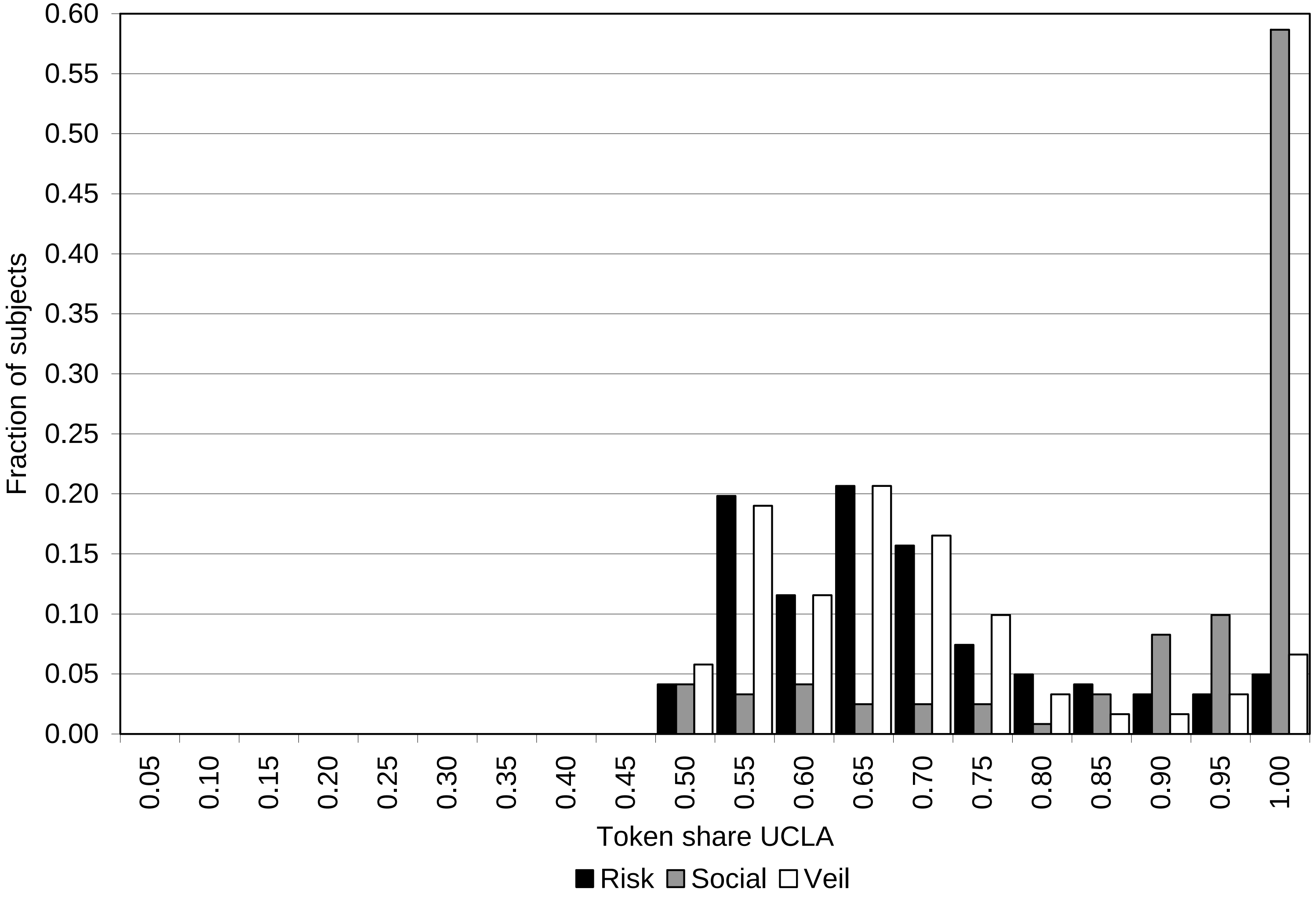
The fraction of tokens allocated to the cheaper account.



The distributions of token shares aggregated across subjects
Veil of Ignorance

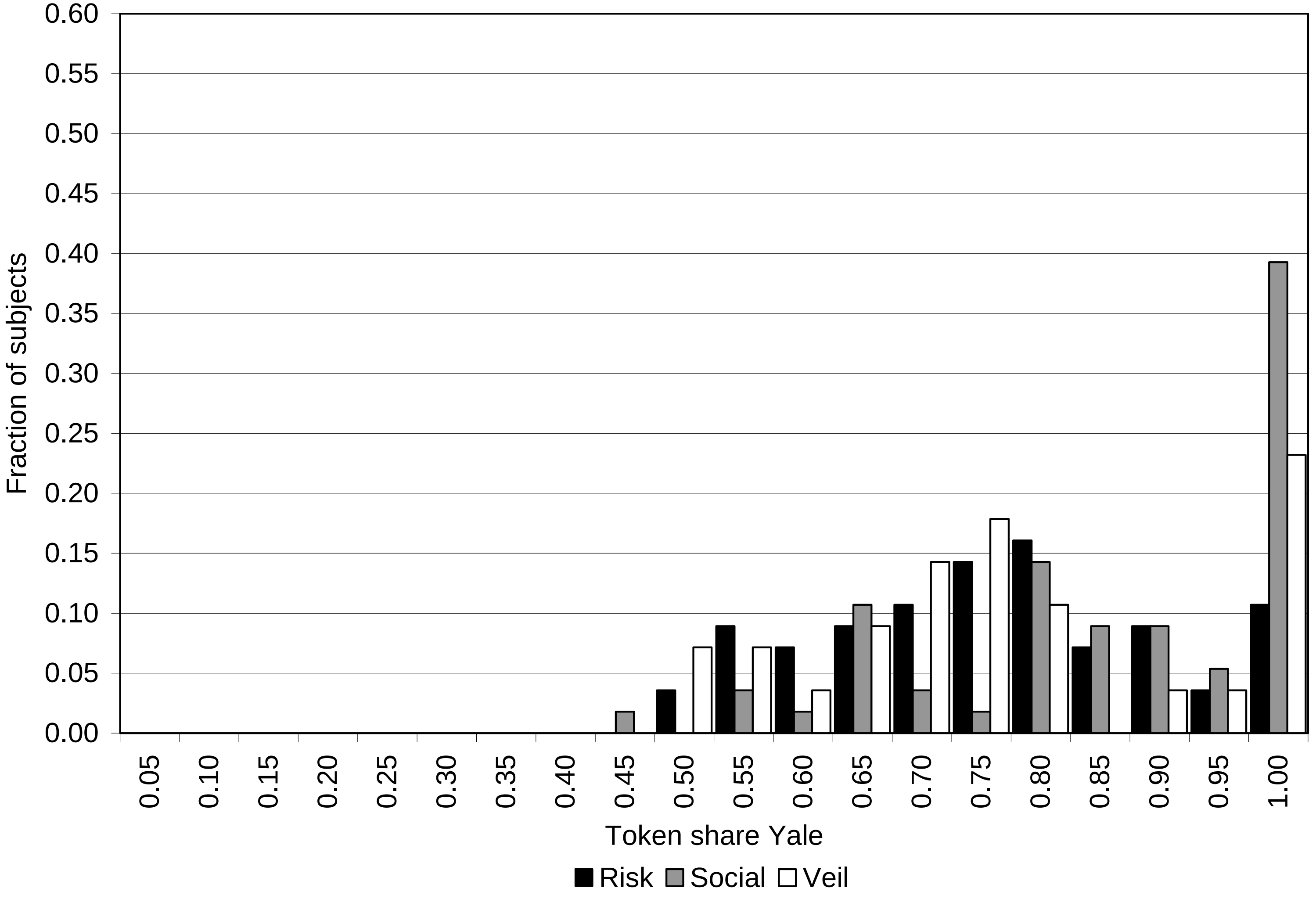
The fraction of tokens allocated to the cheaper account.

The distributions of token shares aggregated across subjects
UCLA



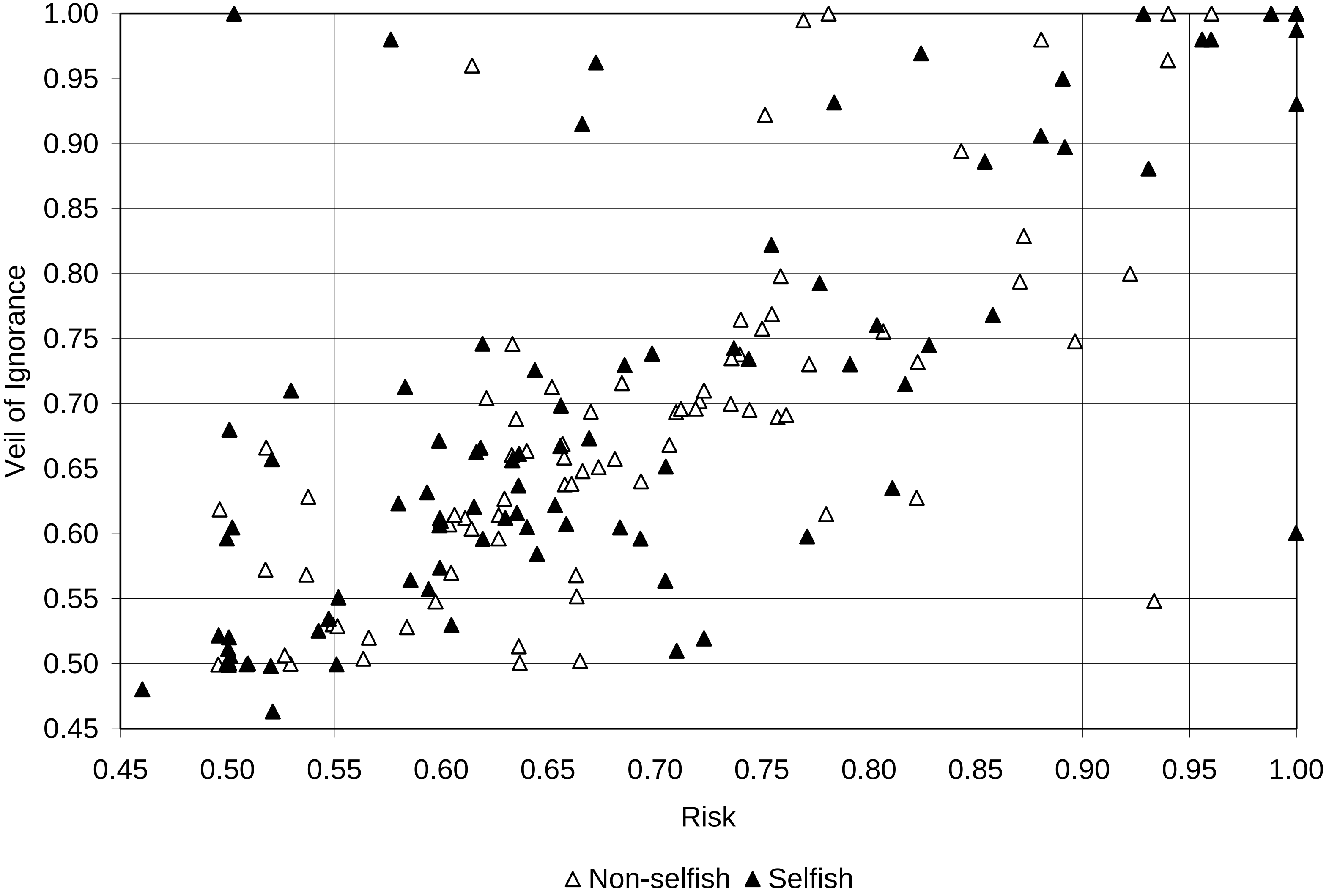
Social: fraction of tokens kept by self. Risk and Veil: fraction of tokens allocated to the cheaper account.

The distributions of token shares aggregated across subjects
Yale



Social: fraction of tokens kept by self. Risk and Veil: fraction of tokens allocated to the cheaper account.

**The average fraction of tokens allocated to the cheaper account
Risk and Veil of Ignorance**



Individual behavior

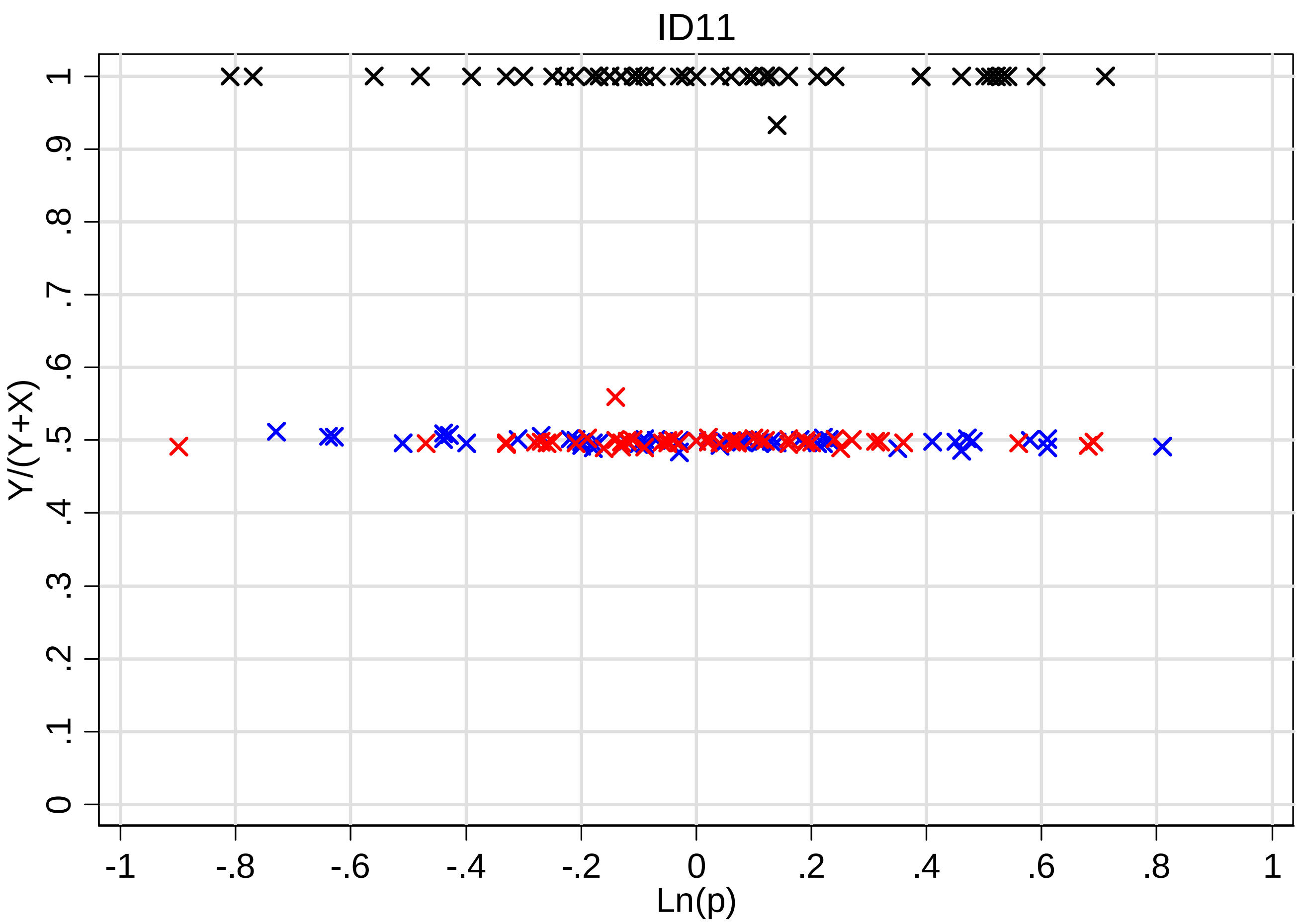
- The aggregate data tell us little about the choice behavior of individual subjects.

- Scatterplots of all choices of illustrative subjects – each entry plots $y/(x+y)$ as a function of $\log(p_x/p_y)$ in a particular treatment.

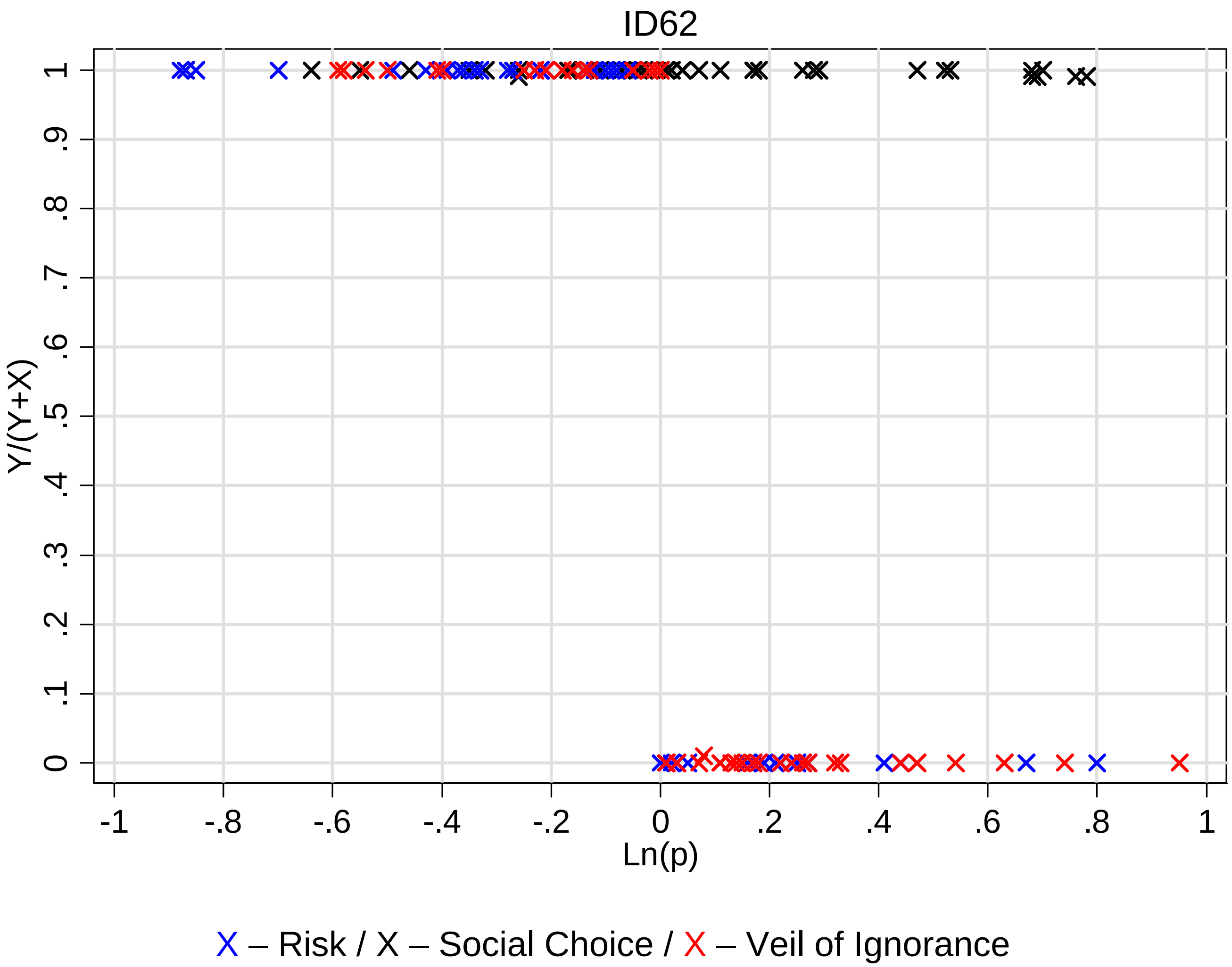
- There is no taxonomy that allows us to classify all subjects unambiguously.

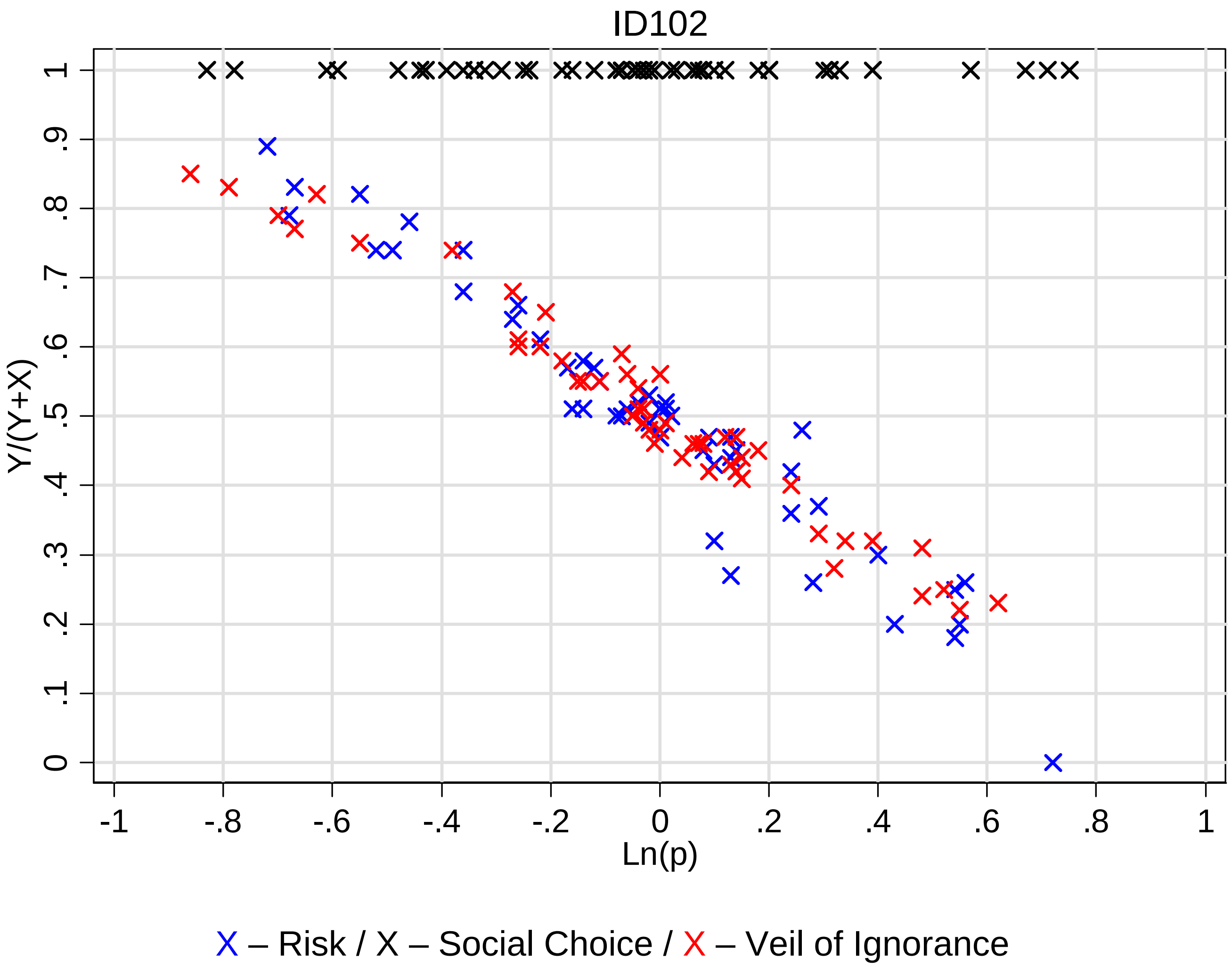
- The characteristic of all our data is striking regularity *within* subjects and heterogeneity *across* subjects.

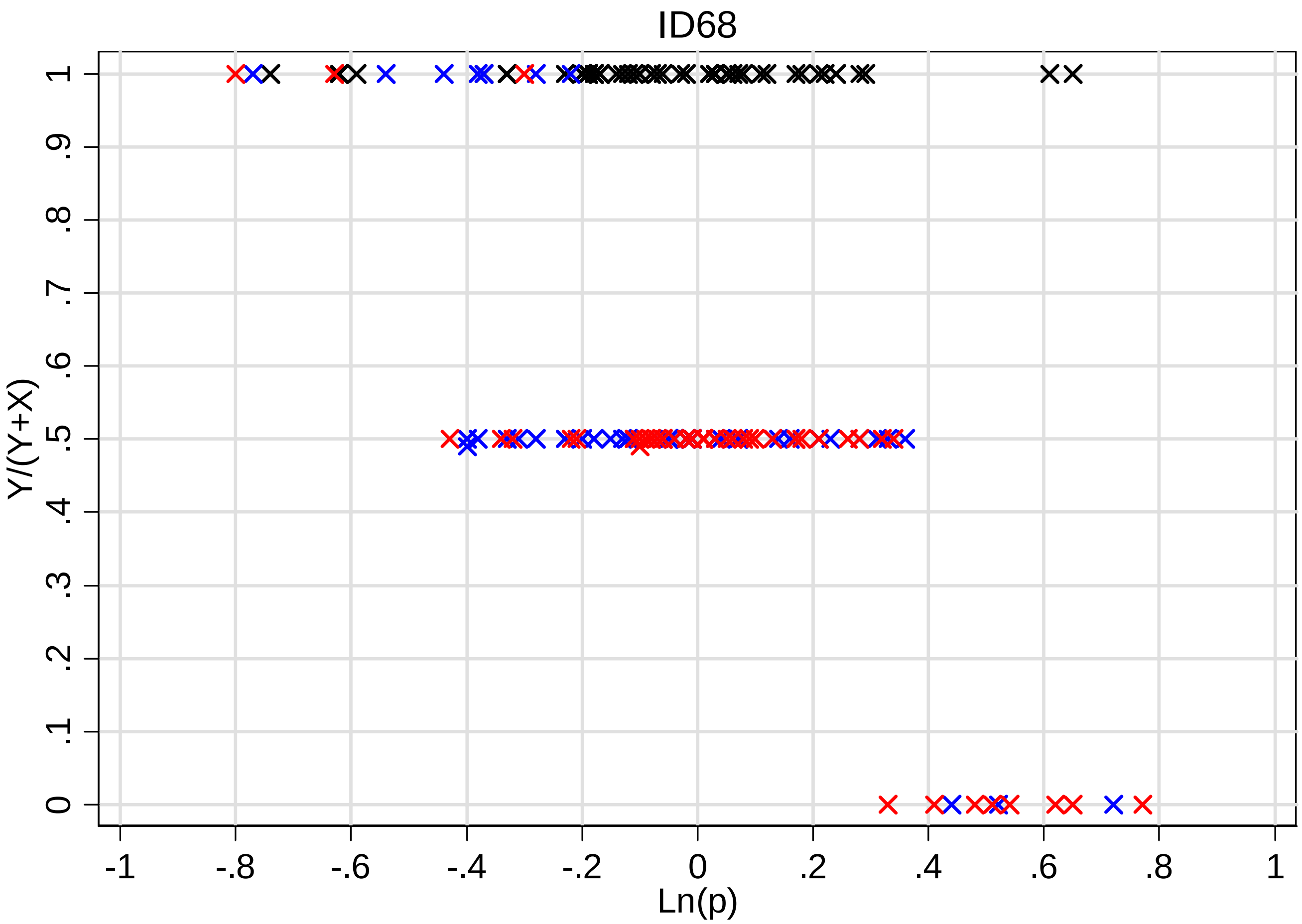
The relationship between the log-price ratio and the token share



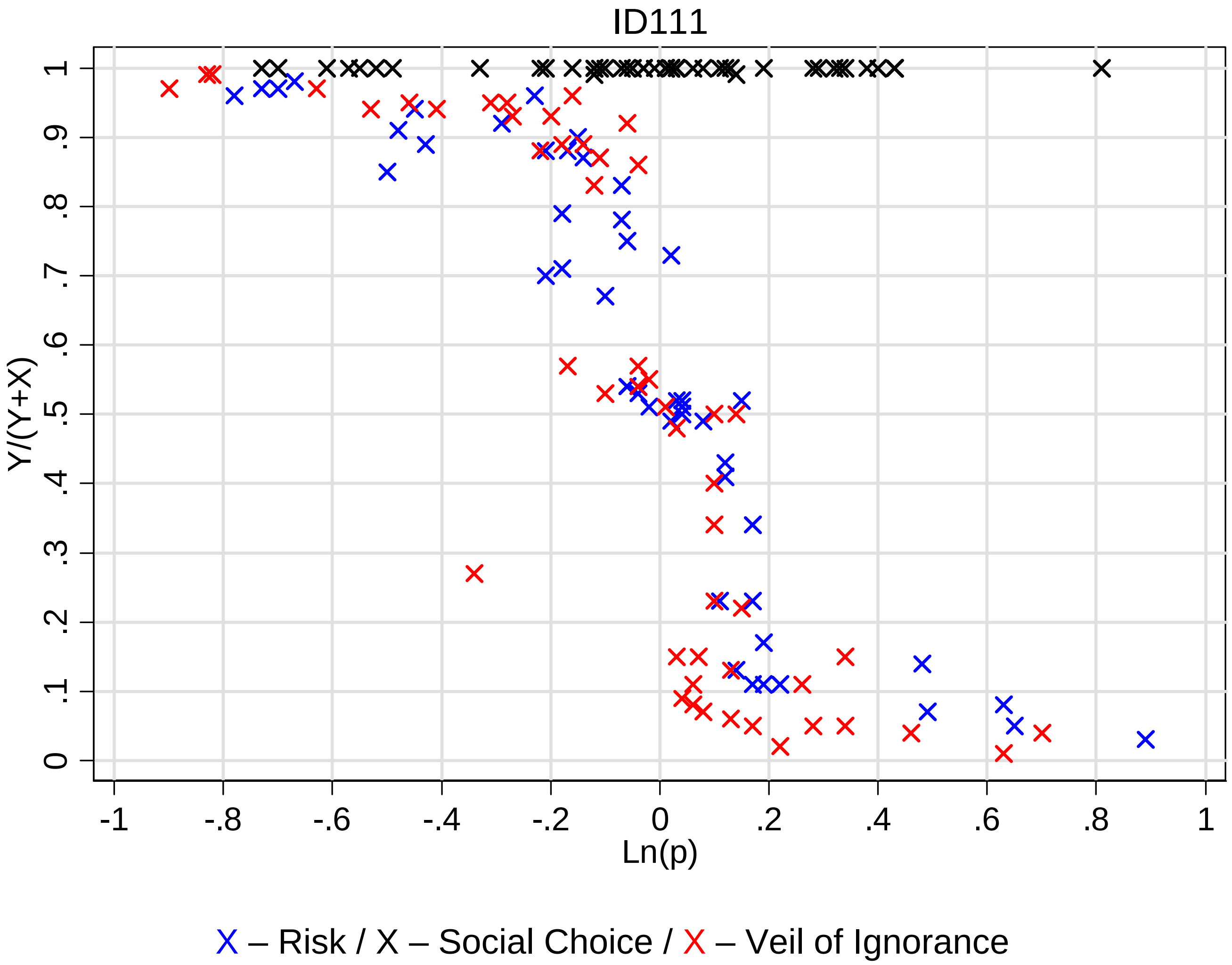
X – Risk / X – Social Choice / X – Veil of Ignorance

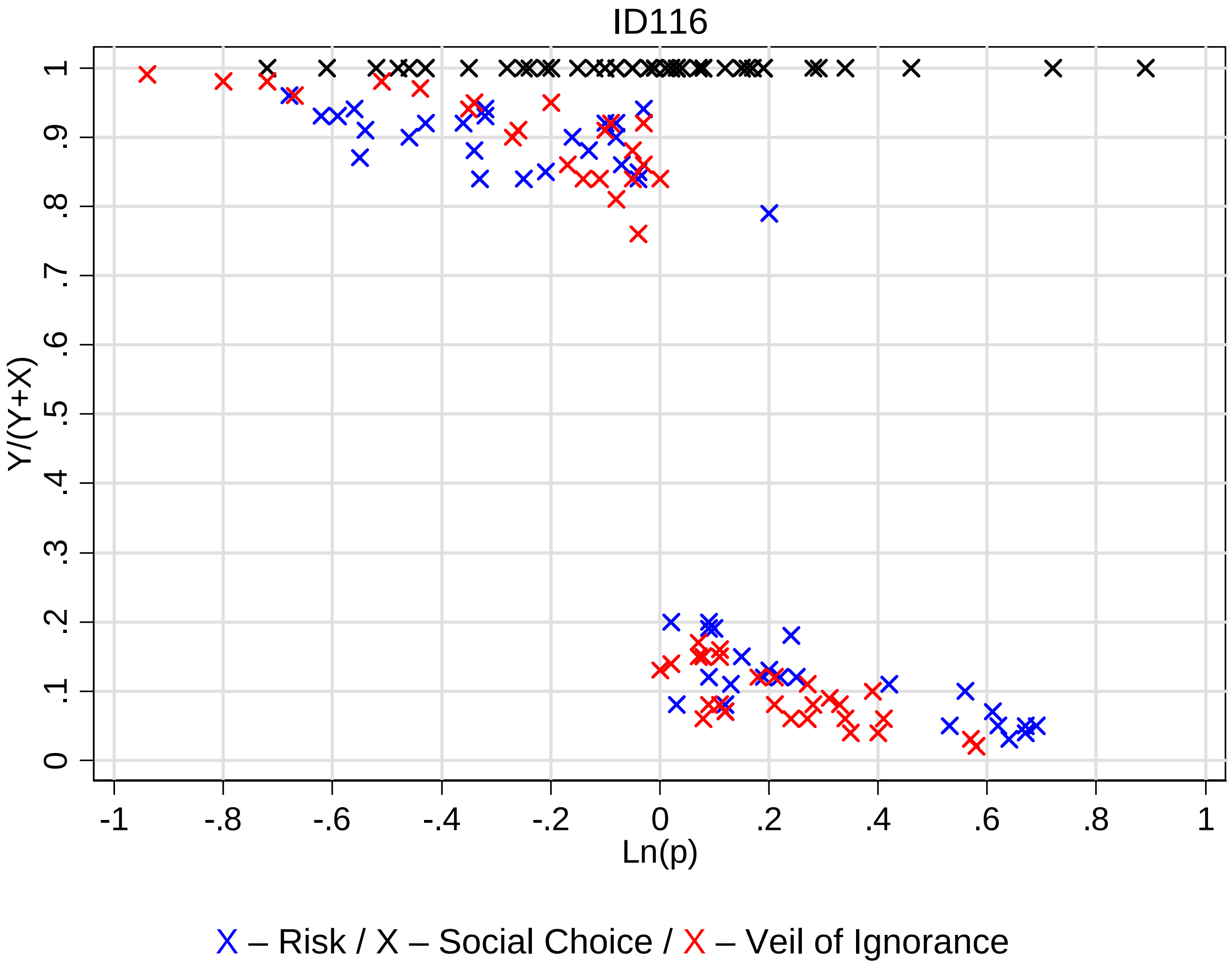

ID62
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

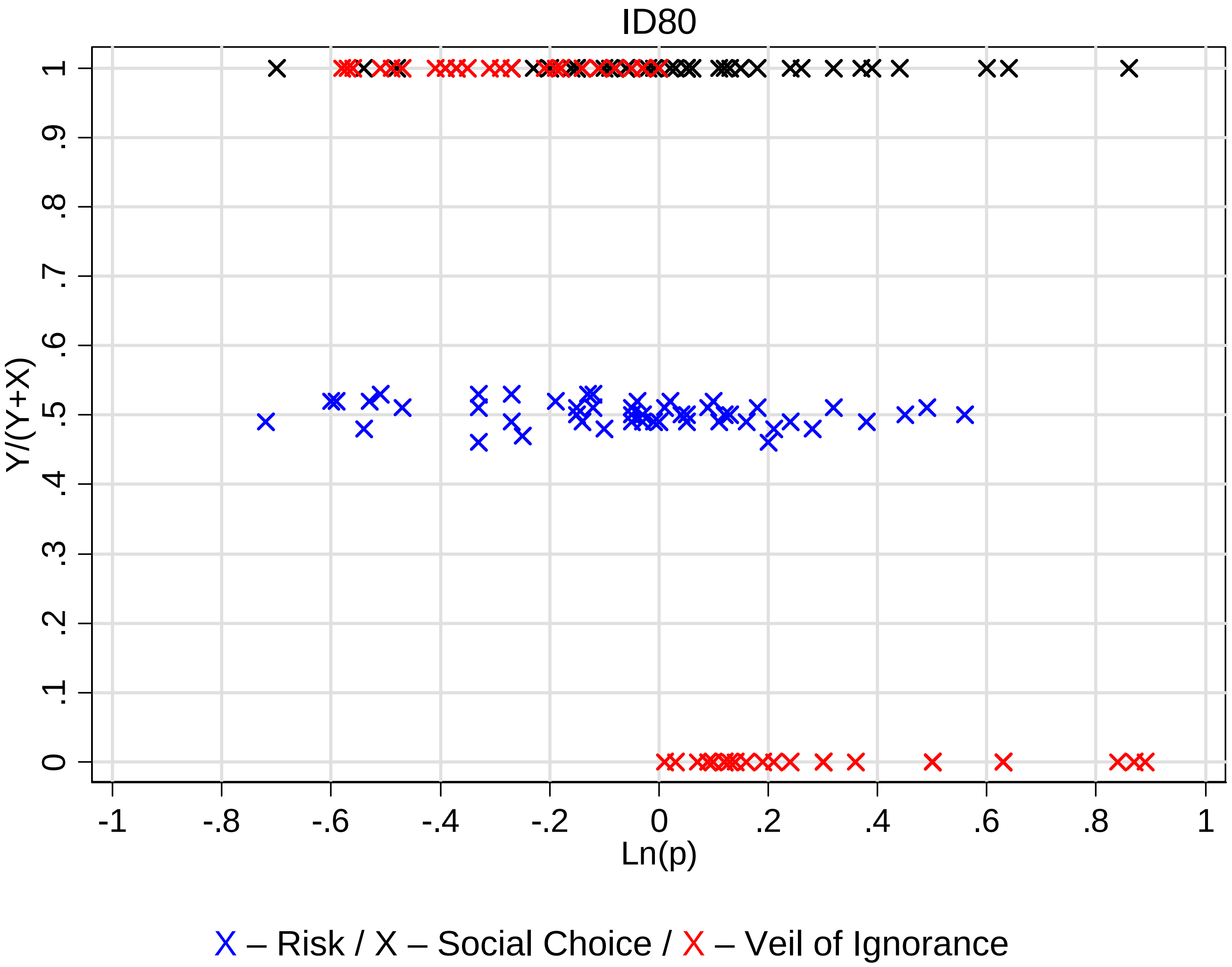

ID102
Y/(Y+X)
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

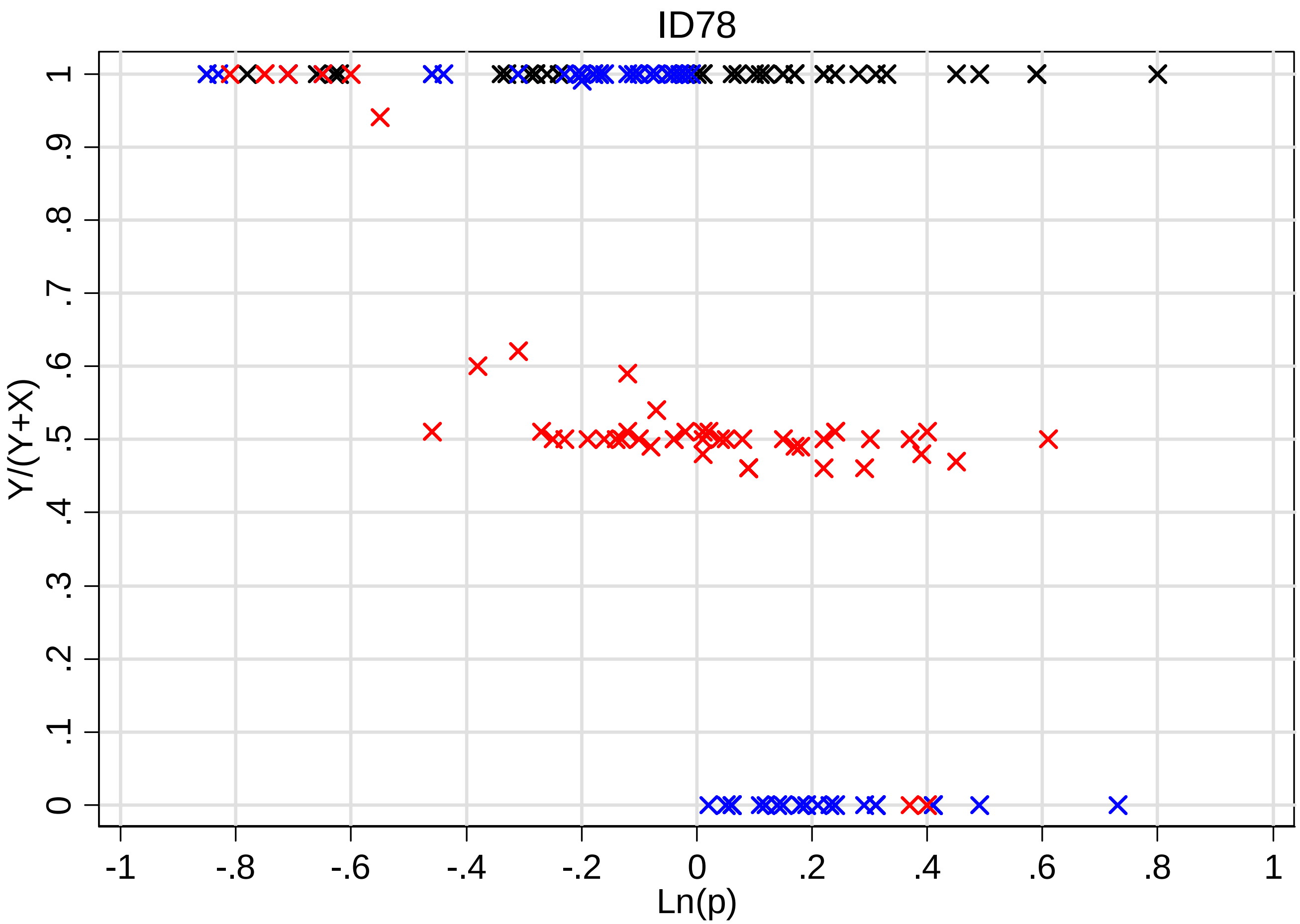


X – Risk / X – Social Choice / X – Veil of Ignorance

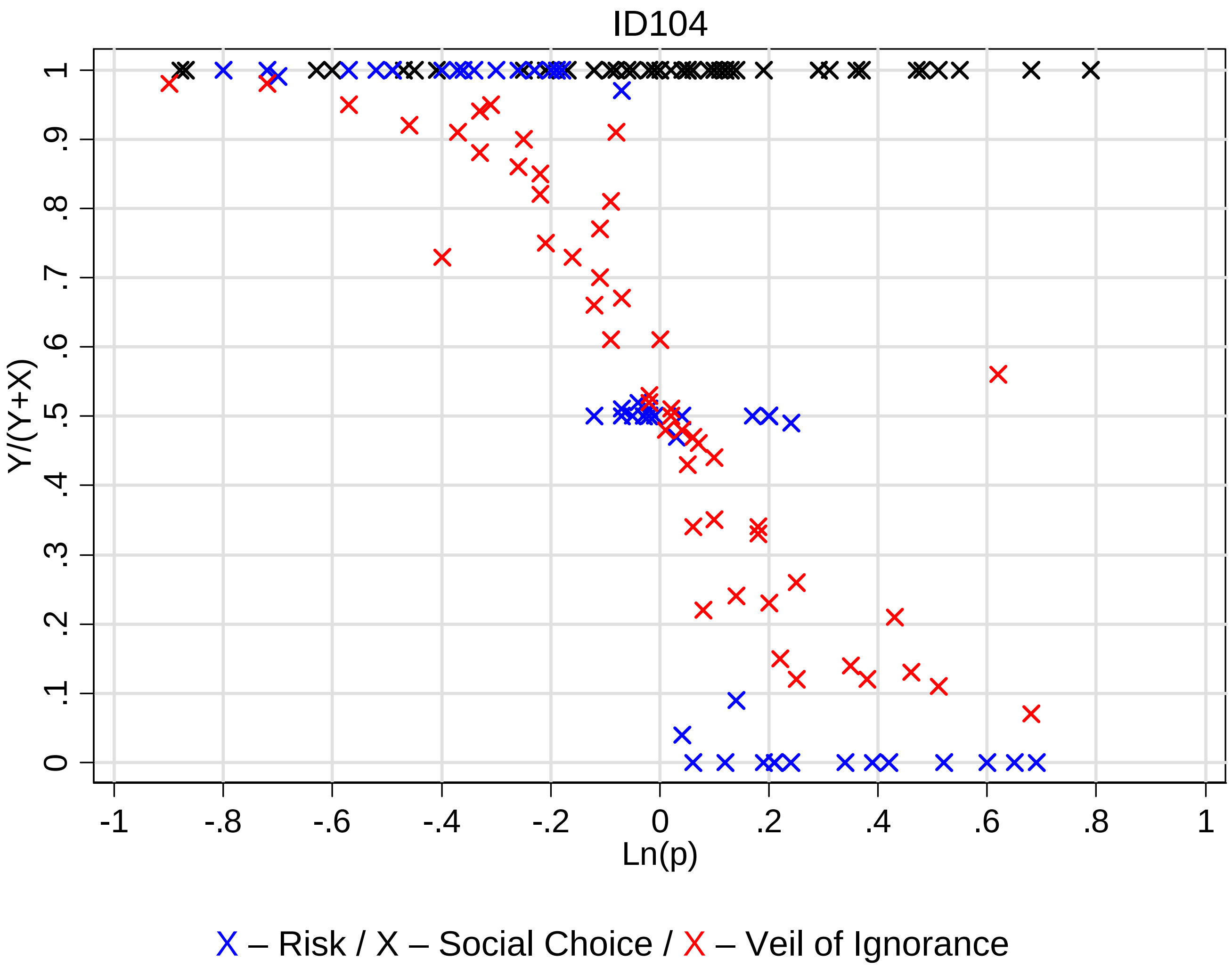

ID111
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

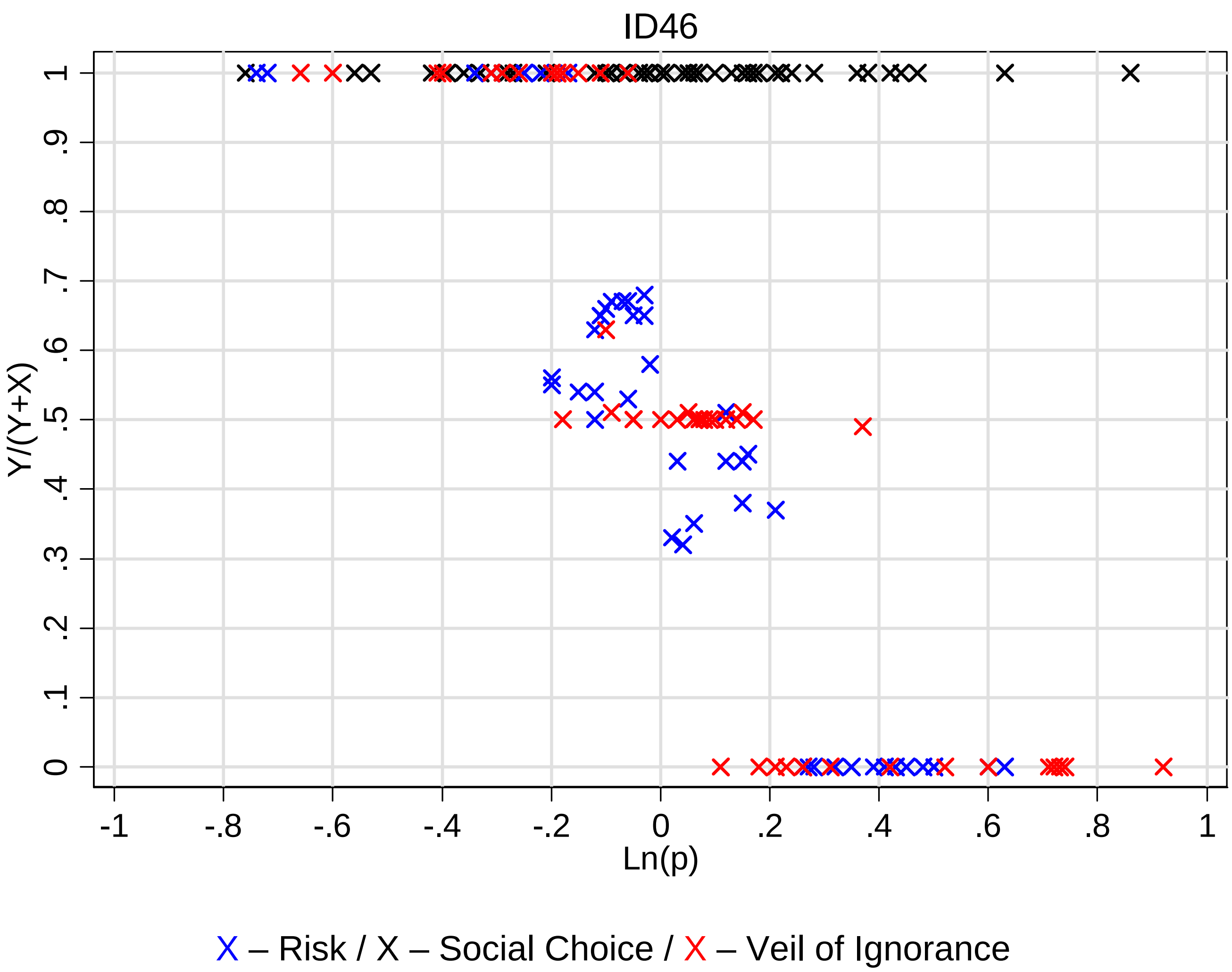

ID116
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

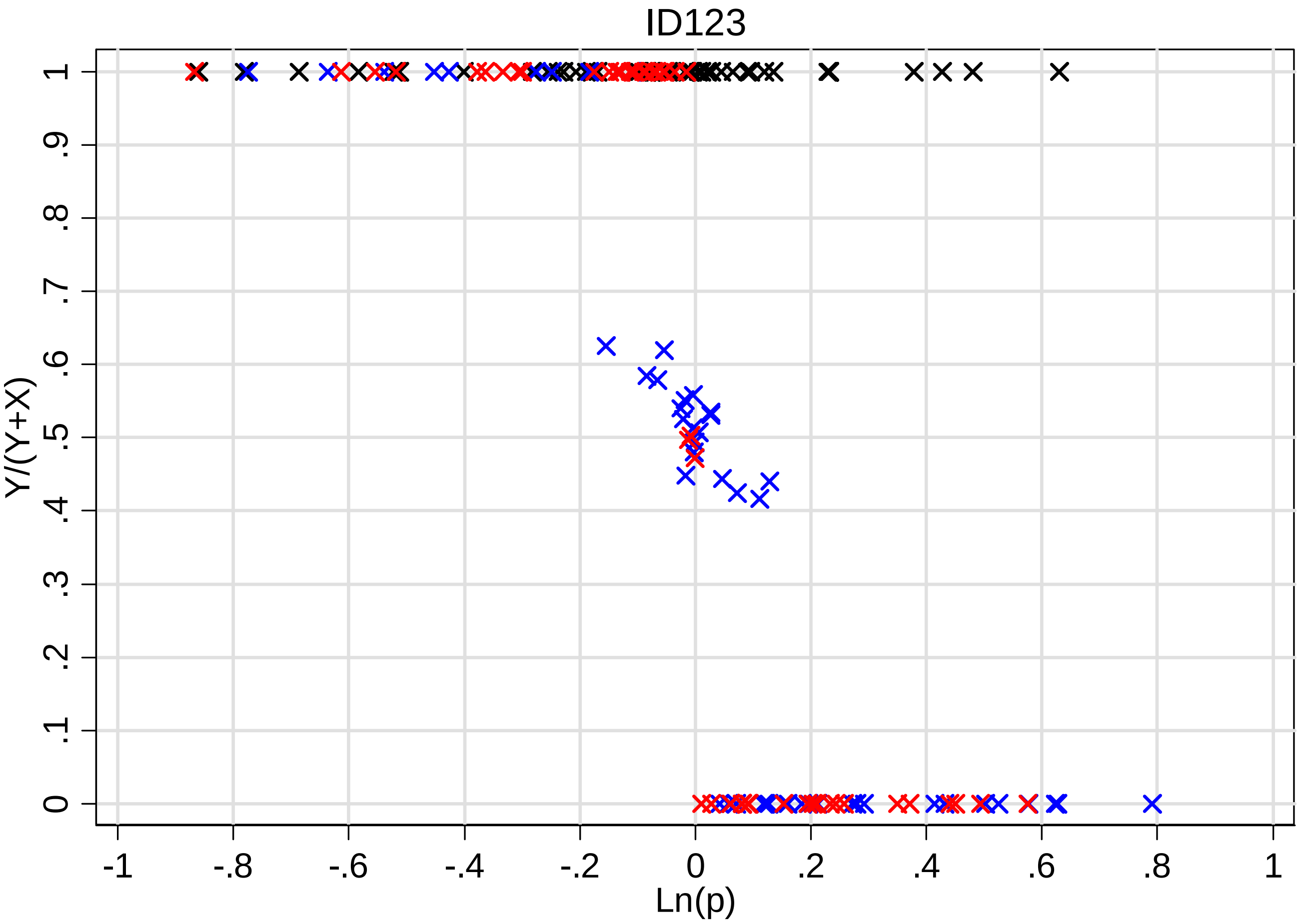

ID80
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance



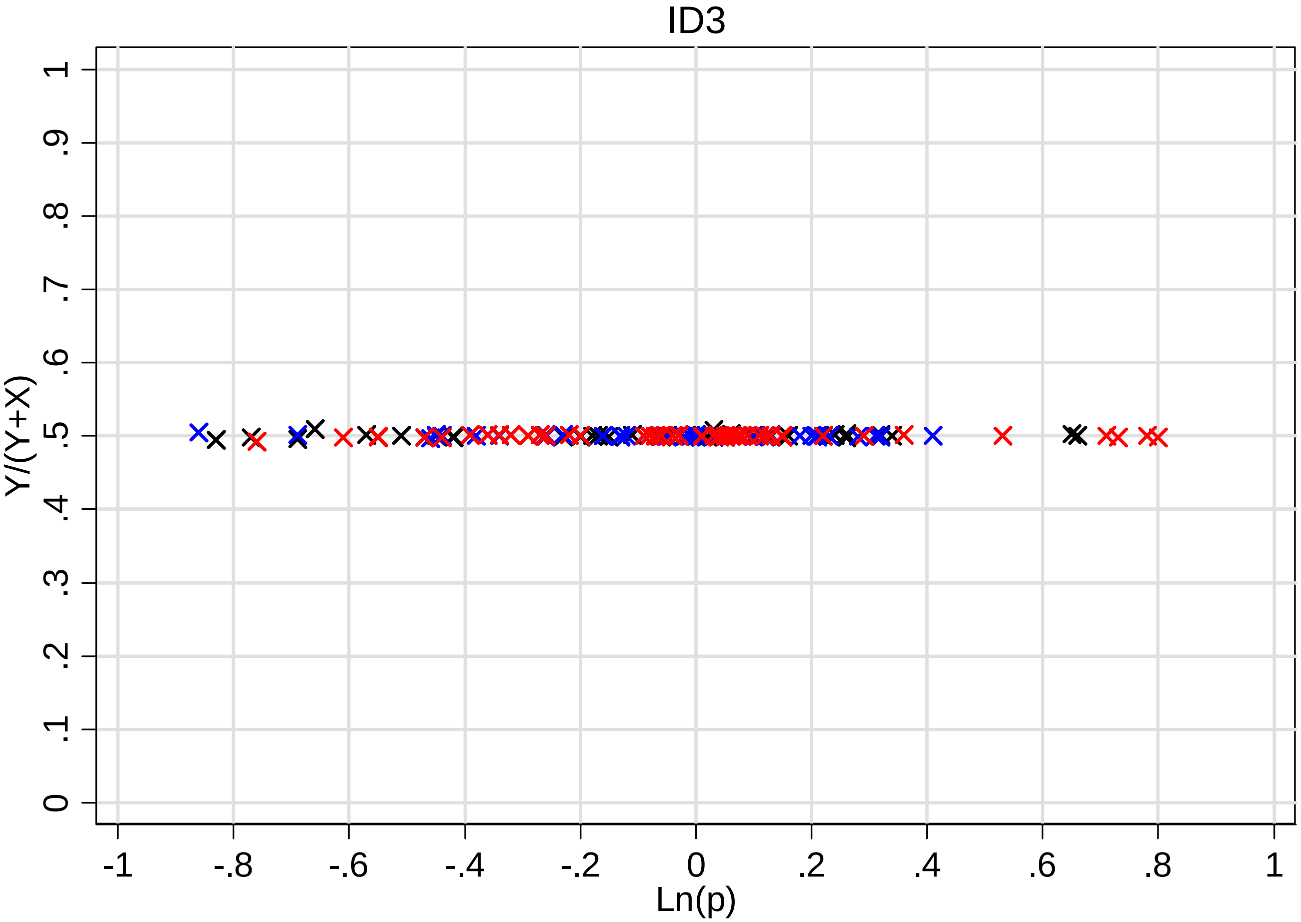
X – Risk / X – Social Choice / X – Veil of Ignorance


ID104
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

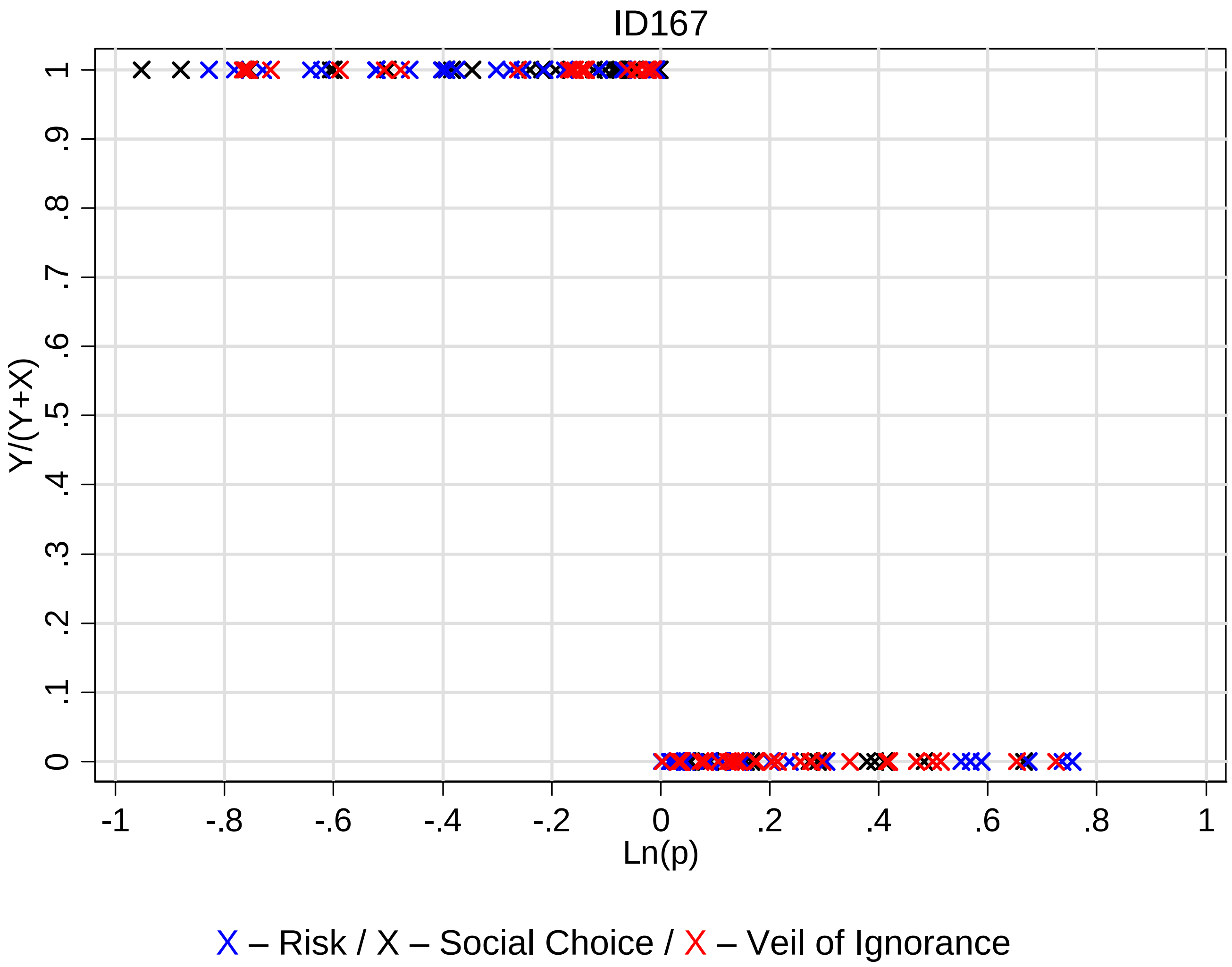

ID46
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

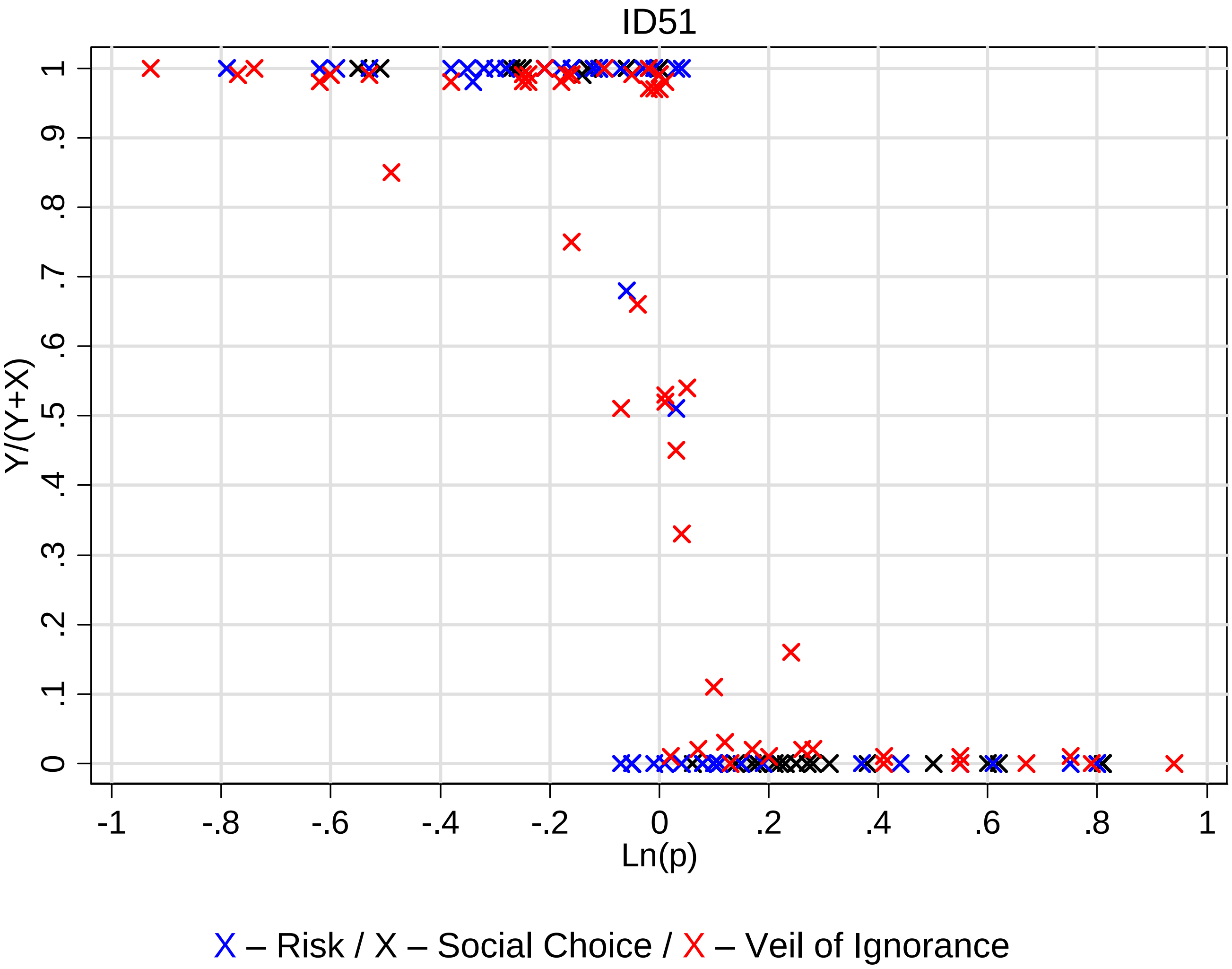


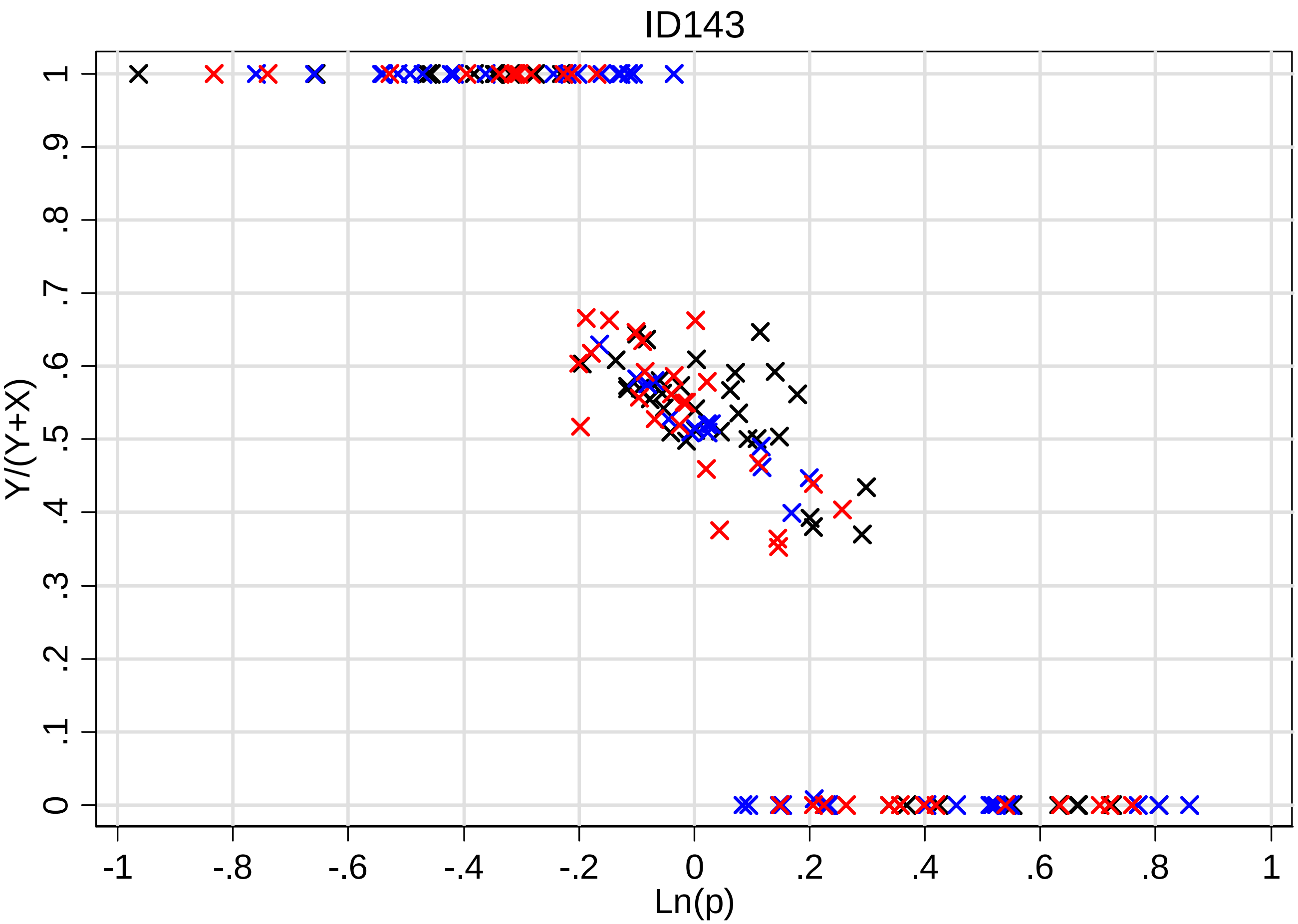
X – Risk / X – Social Choice / X – Veil of Ignorance



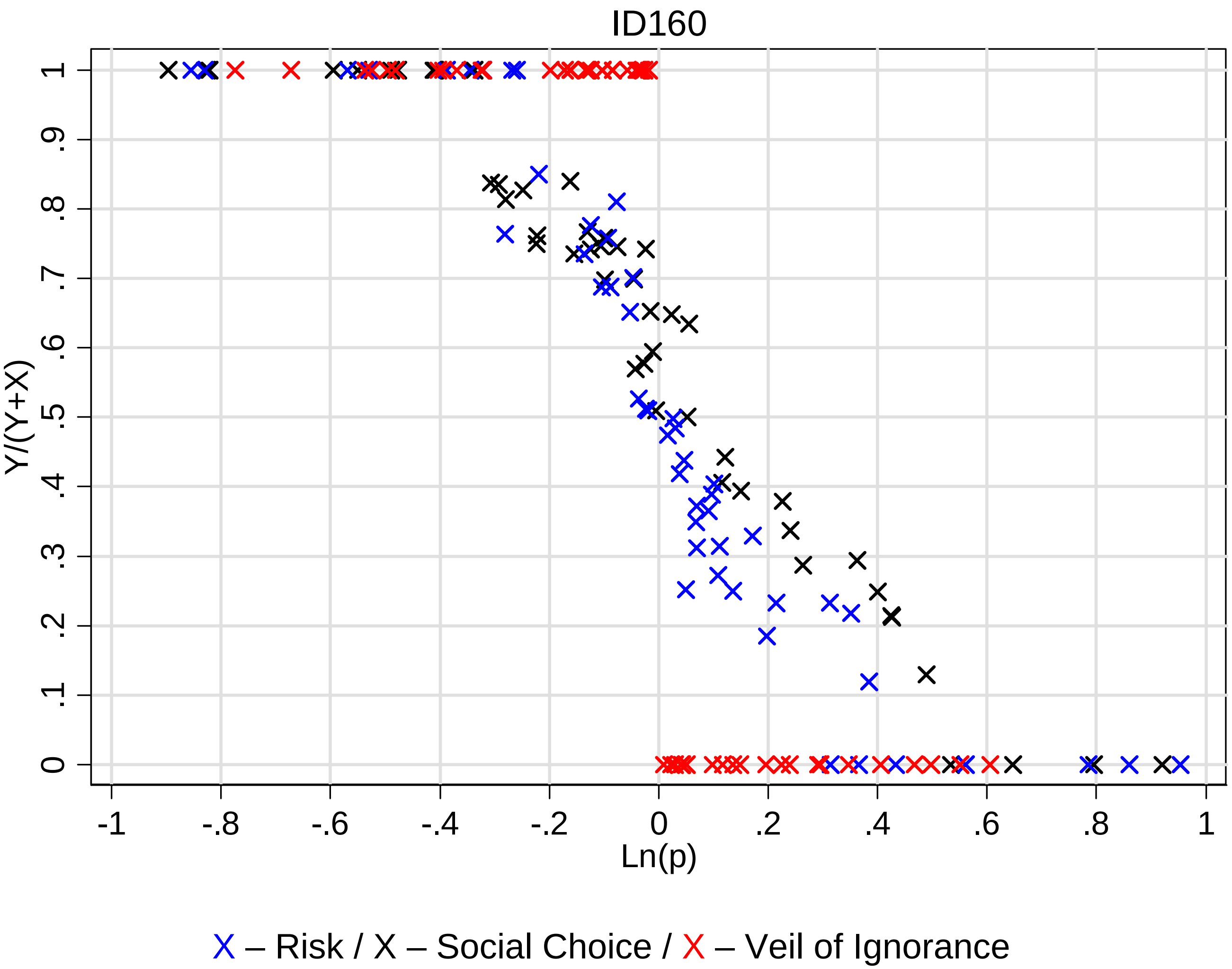
X – Risk / X – Social Choice / X – Veil of Ignorance

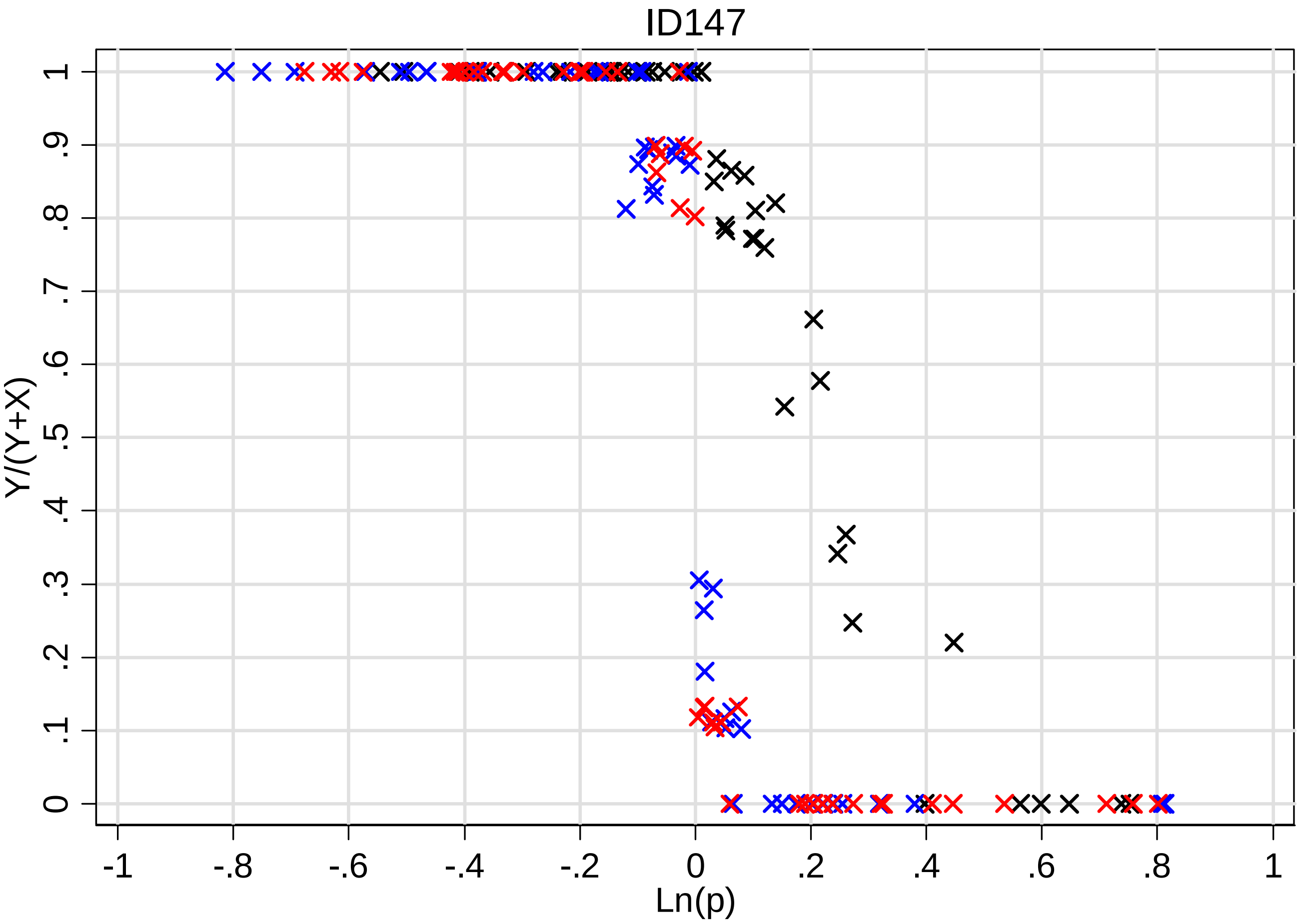

ID167
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance


ID51
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance

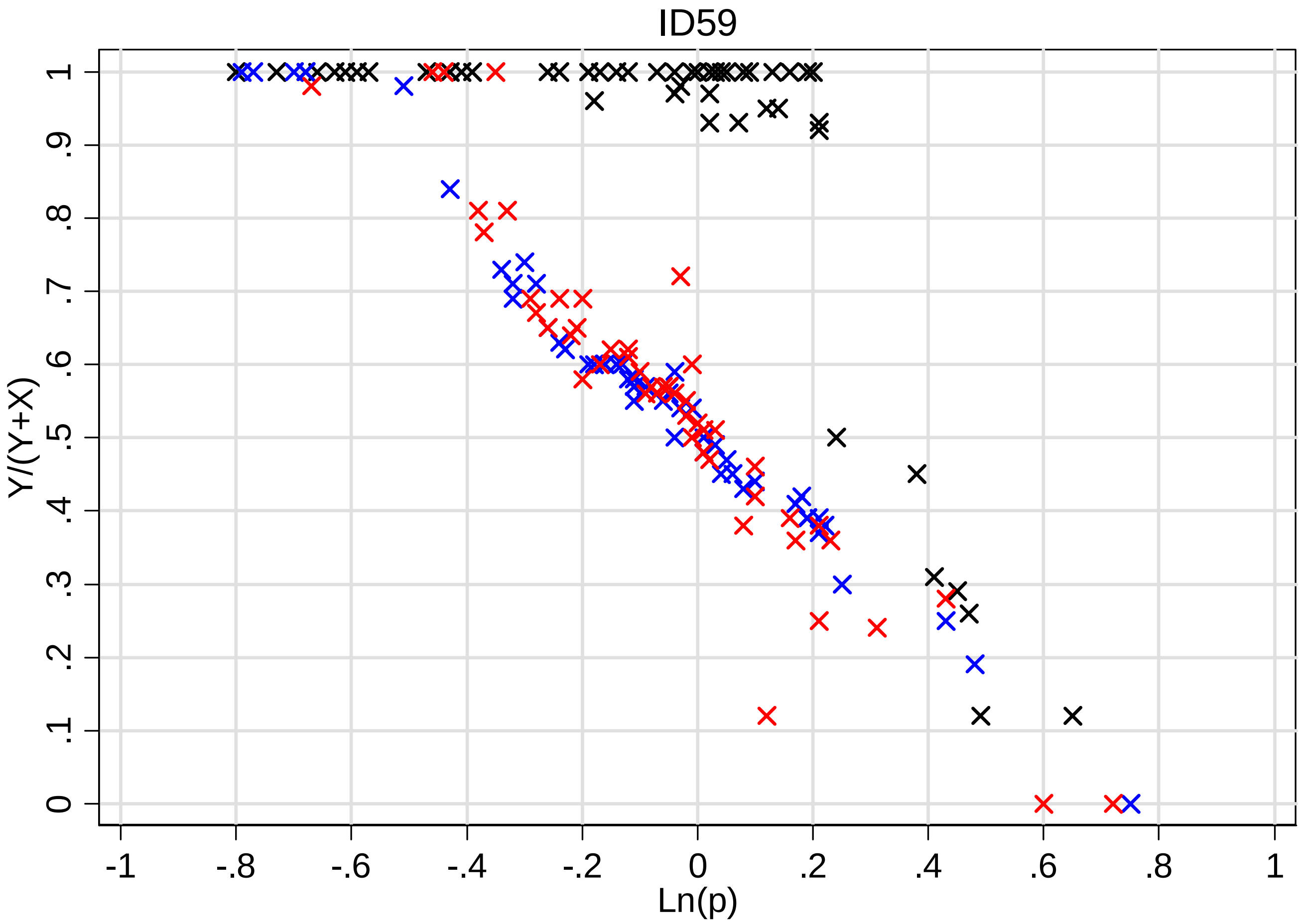


X – Risk / X – Social Choice / X – Veil of Ignorance


ID160
Y/(Y+X)
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
-1
-.8
-.6
-.4
-.2
0
.2
.4
.6
.8
1
Ln(p)
X – Risk / X – Social Choice / X – Veil of Ignorance



X – Risk / X – Social Choice / X – Veil of Ignorance



X – Risk / X – Social Choice / X – Veil of Ignorance

Testing the theory

- Many selfish subjects seem to display the same choice behaviors in the Risk and Veil of Ignorance environments, but a substantial number do not.

- Because of the nature of the data, “flexible” functional forms do not provide a plausible fit for the data.

- No satisfactory formulation to explain the “switching” between stylized behavior patterns exhibited by many subjects.

- Parametric approaches may be possible – keeping in mind that individual behaviors are extremely heterogeneous.

Non-parametric econometric approaches

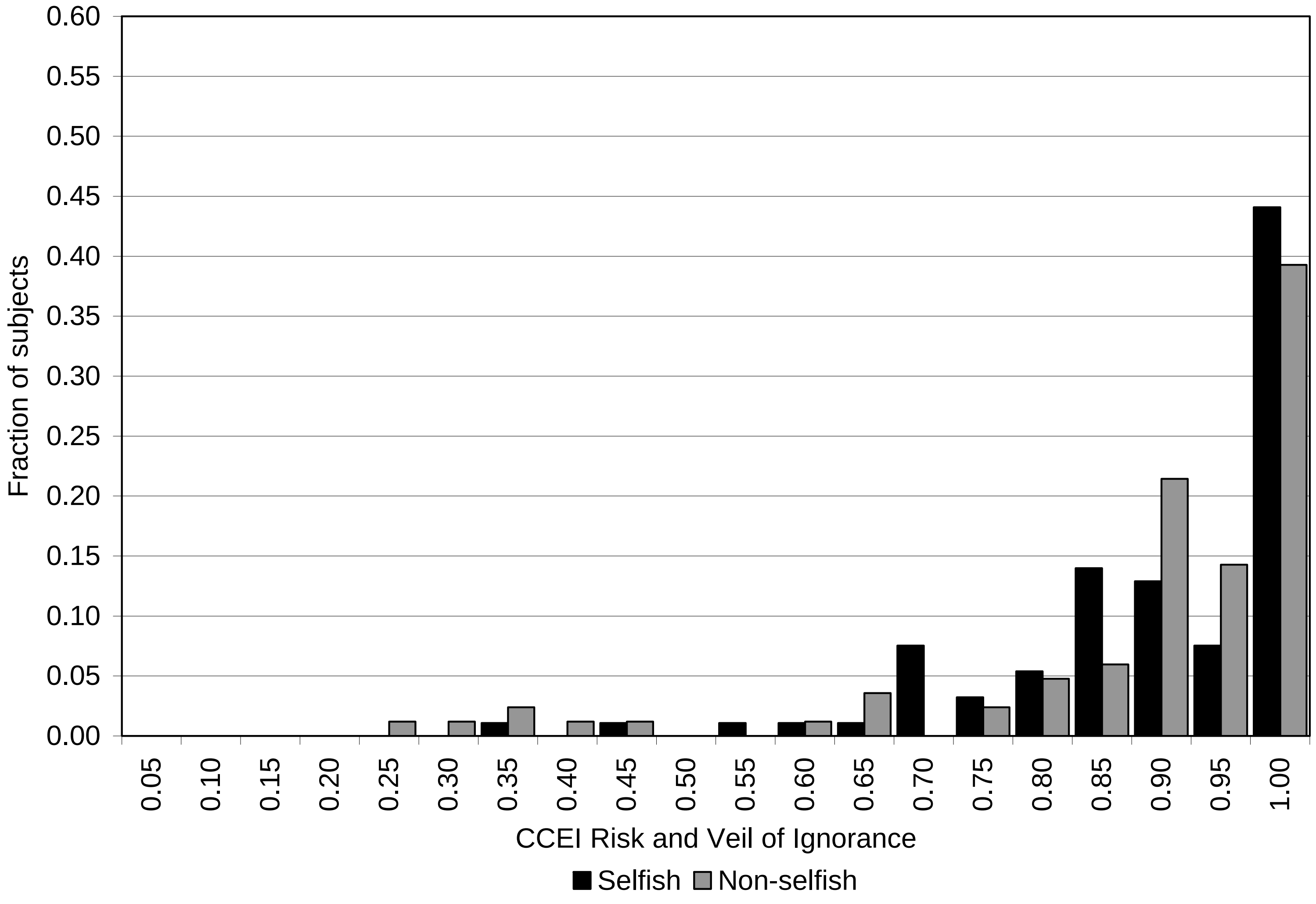
Revealed preference

- The ratio of the CCEI score for the combined data set to the *minimum* of the CCEI scores for the separate data sets.

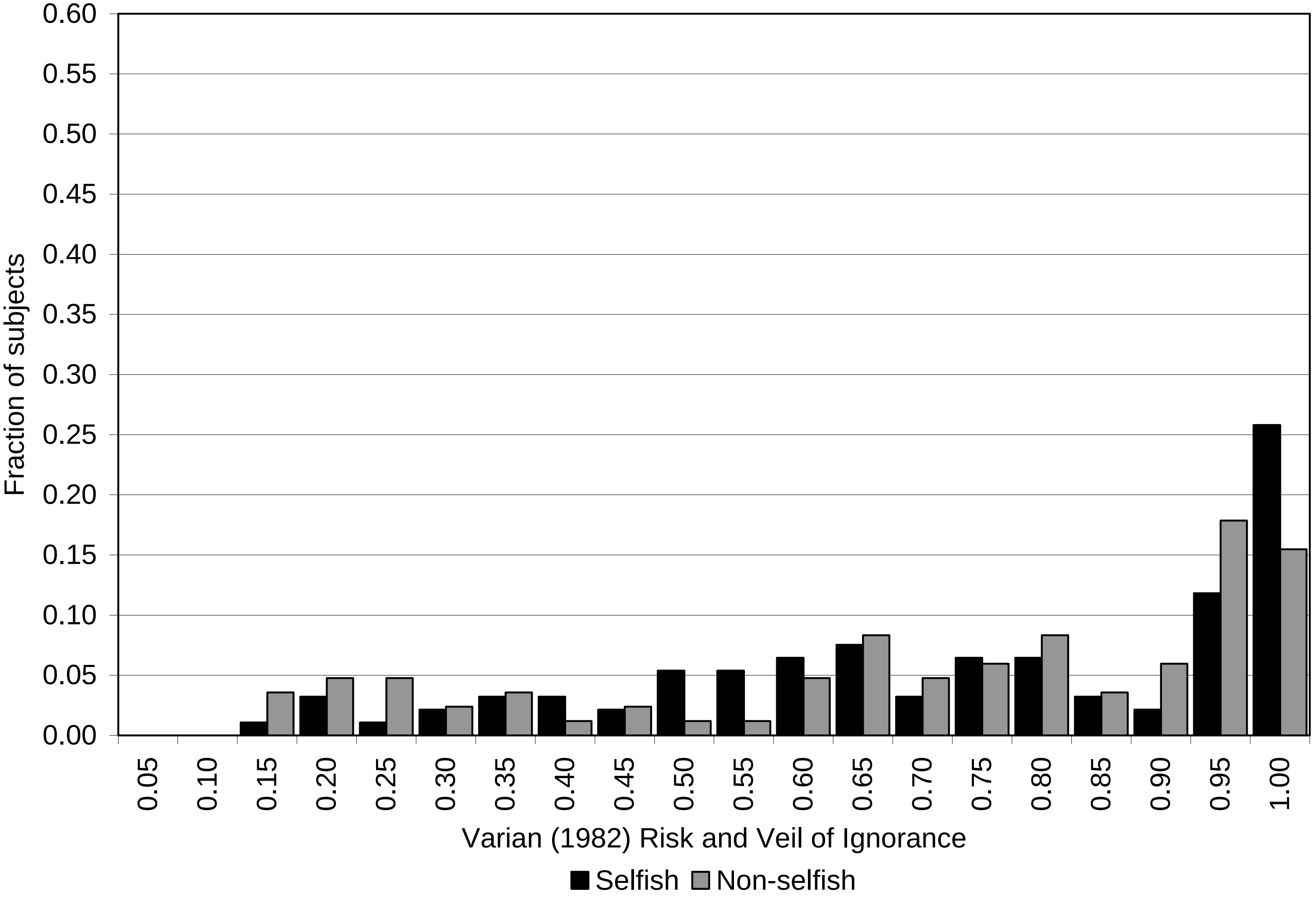
- A measure of the extent to which choice behaviors in any two environments coincide.

- Unfortunately, this test is weak – cannot discriminate between Risk and Veil of Ignorance behavior of selfish and non-selfish subjects.

The distributions of CCEI scores for the combined data set



The distributions of Varian’s (1982) scores for the combined data set



Kolmogorov-Smirnov type tests

- A two-sample Kolmogorov-Smirnov tests of the equality of distributions of token and budget shares.

- The test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

- Generalize the univariate Kolmogorov-Smirnov statistics for bivariate samples (Adler and Brown, 1986).

Conclusions I

- There are subjects who fail Result II (selfish but display different choice behaviors in the Risk and Veil of Ignorance) and others who fail Result III.

- These subjects might have preferences over $\mathcal{L}$ that do not obey independence (or might not be consistent with utility maximization).

- Testing for independence in our setting presents a substantial challenge – three-dimensional choice sets and/or non-linear choice sets.

- The potential of this data set to teach us about individual behavior has not been exhausted.

Conclusions II

- A significant majority of our subjects exhibit behavior that appears to be almost optimizing.

- Individual preferences are very heterogeneous, ranging from utilitarian to Rawlsian.

- Actual preferences "mix-and-match" behavior in ways that no extant theory would regard as justified.

- The techniques of economic analysis may be brought to bear on modeling and predicting behavior governed by moral preferences.

Takeaways

- A positive account of preferences for both personal and social consumption in rich choice environments.

- Two methodological contributions:

  - The establishment of theoretical links between preferences in various environments.

  - An experimental technique that allows for the collection of richer data about preferences.

- The experimental platform and analytical techniques are applicable to many other types of individual choice problems.