University of Zurich[UZH]

Department of Economics

The Many Faces of Human Sociality

Uncovering the Distribution and Stability of Social Preferences

Journal of the European Economic Association

Volume 17, Issue 4, August 2019, Pages 1025–1069,

https://doi.org/10.1093/jeea/jvy018

Adrian Bruhin
Univ. of Lausanne

Ernst Fehr
Univ. of Zurich

Daniel Schunk
Univ. of Mainz

Motivation

Social preferences appear to be not only important drivers of behavior but also vastly heterogeneous across subjects

- This heterogeneity is important for aggregate outcomes & interacts with the institutional environment
  - Selfish or non-selfish types may be decisive for aggregate outcomes depending on the institutional set-up
- Examples
  - Public good games with and w/o sanctions
  - Competitive markets with complete or incomplete contracts
  - Effectiveness of various incentive mechanisms
  - Foundations for incomplete contracts (failure of subgame perfect implementation)

$\Rightarrow$ We need a parsimonious characterization of social preference heterogeneity that is stable over time and across contexts

Goal of this study

- Develop an experimental design that identifies simultaneously consequentialist and reciprocity-based social preferences
- Provide a parsimonious characterization of heterogeneity in social preferences
- Examine how stable the distribution of types is over time
- Examine the out-of-sample predictive power ("stability") of the empirical model
- **Stability across time and games is decisive criterion for the extent to which the model is capturing the key motivational forces**

Outline

1. Related Literature
2. Experimental Design
3. Empirical Analysis
4. Results
5. Conclusion

Related Literature: Identification of Social Preferences

Studies on the identification of social preferences using a similar

- Preference model: Fehr & Schmidt (1999); Charness & Rabin (2002); Bellemare et al. (2008)
- Experimental design: Kerschbamer (2015)

Studies applying finite mixture models to take social preference heterogeneity into account

- Iriberri & Rey-Biel (2011, 2013); Breitmoser (2013)
- Bardsley & Moffatt (2007); Conte & Moffatt (2014); Conte & Levati (2014)

Main Contributions: Identification of Social Preferences

- Identification of distribution- and reciprocity-based preferences
- Endogenous instead of predefined preference types

Related Literature: Stability of Social Preferences

Temporally correlated contributions to public goods suggest that social preferences may be stable over time

- Lab: Volk et al. (2012)
- Field: Carlson et al. (2014)

Studies investigating correlations between lab and field behavior indicate that social preferences may be stable across contexts

- Trust: Karlan (2005); Fehr & Leibbrandt (2011)
- Donations: Benz & Meier (2008)
- Contributions to public goods: Laury & Taylor (2008)

Main Contributions: Stability of Social Preferences

- Estimated structural model allows for predictions across games
- Accounting for preference heterogeneity

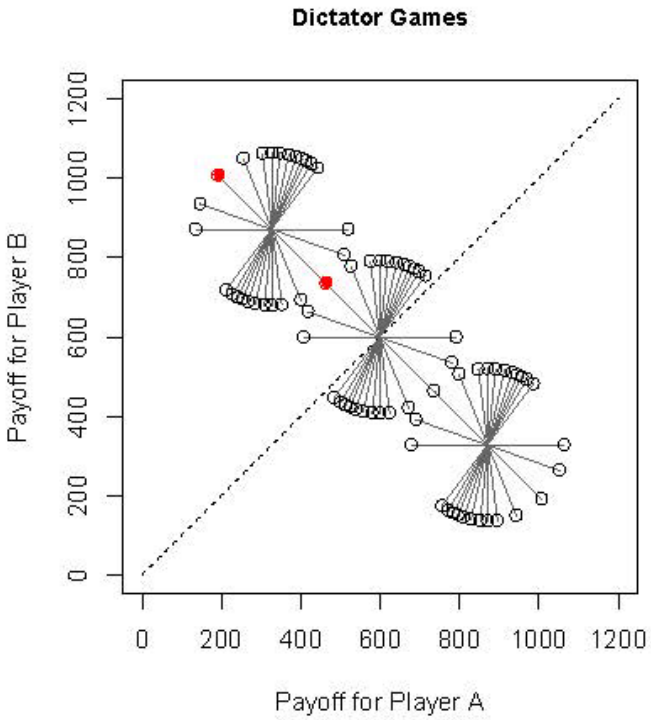
Experimental Design

The experimental design features 39 binary dictator games and 78 positive and negative reciprocity games

Dictator games:

- Player A chooses between two allocations X and Y
- Costs of altering player B's payoff vary systematically

$\Rightarrow$ Identify player A's distributional preferences



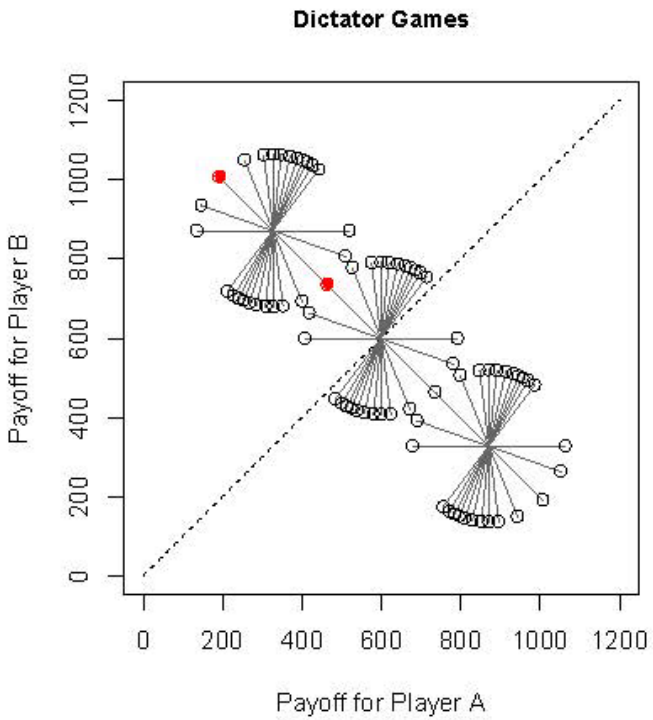
Experimental Design

The experimental design uses 39 binary dictator games and 78 positive and negative reciprocity games

Reciprocity games:

- Player B makes a prior move: She either implements allocation Z or lets A choose between allocations X and Y
- Depending on Z, letting player A choose between X and Y is either kind or unkind

$\Longrightarrow$ Differences in A's choices between the dictator and reciprocity games are due to positive or negative reciprocity



Experimental Design

We invited 200 student subjects to participate in two experimental sessions that were three months apart

- Both sessions comprised all 117 dictator and reciprocity games

$\Longrightarrow$ Allows testing the stability of social preferences over time

- The second session additionally included
  - Ten trust games with varying costs of being trustworthy
  - Two reward and punishment games

$\Longrightarrow$ Allows testing the stability of social preferences across games

- The first session additionally featured a cognitive ability test and a short version of the Big 5 personality questionnaire
- 174 subjects showed up in the second session, corresponding to a retention rate of 87%

Empirical Analysis: Preference Model

A piecewise-linear utility function represents the subjects' social preferences (Fehr & Schmidt, 1999; Charness & Rabin, 2002)

$$U^A = (1 - \alpha s - \beta r - \gamma q - \delta v)\Pi^A + (\alpha s + \beta r + \gamma q + \delta v)\Pi^B ,$$

where

- Π^A and Π^B correspond to the payoffs of players A and B
- α and β denote the weight of the other player's payoff under disadvantageous and advantageous inequality, respectively
- γ and δ indicate how this weight changes if the other player behaved kindly and unkindly, respectively
- $s, r, q,$ and v are the corresponding indicator variables

Empirical Analysis: Preference Model

- We assume a random utility model with an EV1 distributed error term (McFadden, 1981)
- Subject i in the role of player A chooses allocation X_g at game g with probability

$$\Pr(C_{ig} = X; \theta, \sigma) = \frac{\exp\left(\sigma\, U^A(X_g; \theta)\right)}{\exp\left(\sigma\, U^A(X_g; \theta)\right) + \exp\left(\sigma\, U^A(Y_g; \theta)\right)}$$

- $\theta = (\alpha, \beta, \gamma, \delta)'$ contains the behavioral parameters
- σ denotes the choice sensitivity

Empirical Analysis: Levels of Aggregation

We estimate the model at three different levels of aggregation

1. At the aggregate level assuming a representative individual
2. At the individual level
3. At the level of distinct preference types using finite mixture models

The Finite Mixture Model

- Assume K types. Then the model gives you
- Preference parameters $\theta = (\alpha, \beta, \gamma, \delta)'$ for each type
- A posterior probability τ_{ik} that assigns each individual i to a type k (clean assignment important for quality of classification)
- The proportion of subjects π_k that belong to each type
- No assumptions are made with regard to the existing types (except that they are from the broad class of feasible social preferences)
  - Could be the case, for example, that selfish types are completely absent
  - Any combination of outcome-based and reciprocal social preferences possible

Finite Mixture Model – optimal number of types

An important aspect when applying finite mixture models is to determine the optimal number of types K

- If K is too low, the model fits the data poorly, as it is not flexible enough to cope with the behavioral heterogeneity
- If K is too large, the model overfits the data, as it captures noise besides the existing preference types

Problems:

- There are no statistical tests for K that are generally applicable and exhibit a test statistic with a known distribution.
- Model selection criteria like the AIC or BIC often favor too many types as they do not penalize ambiguous classifications
- We use normalized entropy criterion plus whether the type characterization in terms of preference parameters and size is relatively stable over time

Finite Mixture Model – 0ptimal number of types

Potential solutions:

- Use prior knowledge and predefine the plausible types
- Rely on model selection criteria that penalize for entropy
- Simulate the test statistics of likelihood ratio tests
- Use cross-validation (Smyth, 2000)

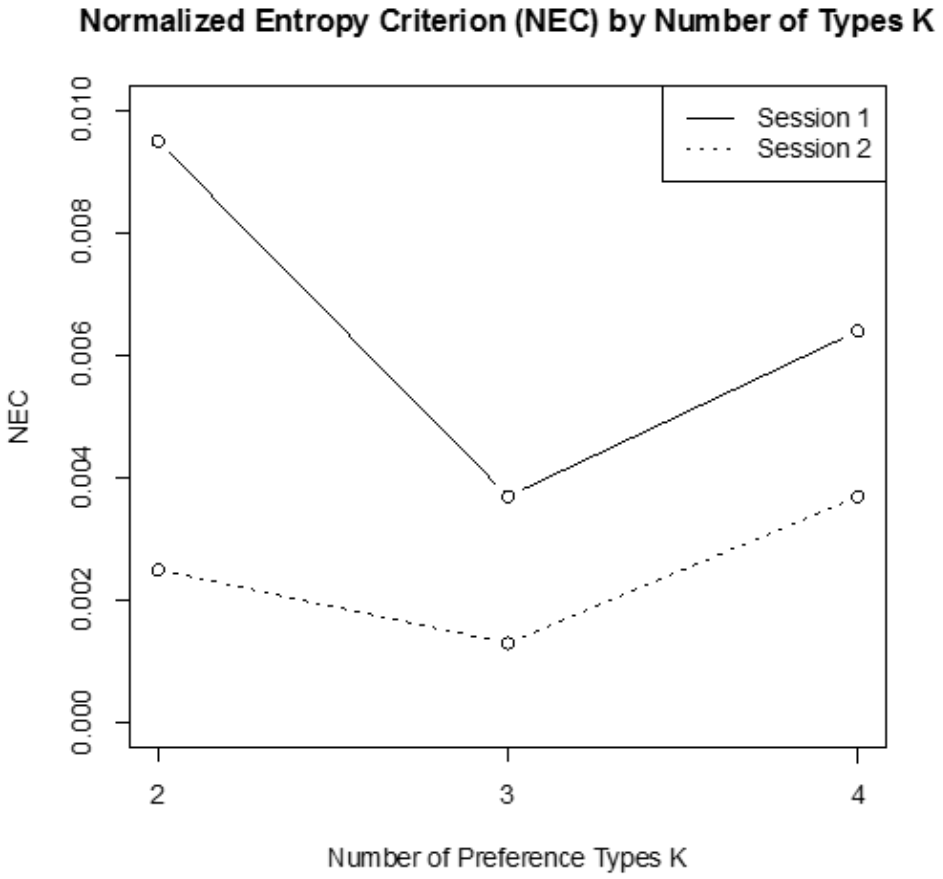
⇒ Here: We choose K so that they types remain stable over time

Social preferences of the representative agent

	Estimates of Session 1	*Estimates of Session 2*
α	0.083***	0.098***
	(0.015)	(0.013)
β	0.261***	0.245***
	(0.019)	(0.019)
γ	0.072***	0.029***
	(0.014)	(0.010)
δ	-0.042***	-0.043***
	(0.011)	(0.008)
σ	0.016***	0.019***
	(0.001)	(0.001)

- Subjects are altruistic on average
- Preferences are stable over time
- Weight of the other's payoff is lower under disadvantageous inequality than under advantageous inequality
- Distributional preferences are more important than reciprocity
- Positive reciprocity equally important as negative reciprocity

Type-specific characterization – how many types?

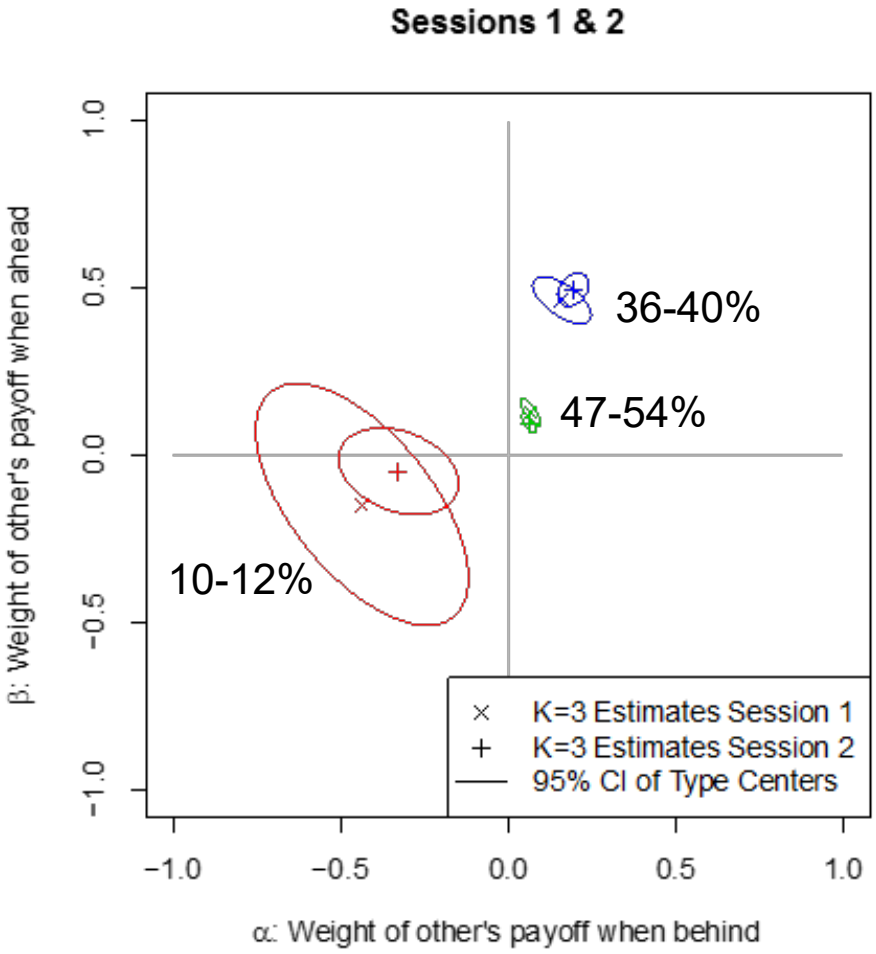


Results: Type-Specific Level ($K = 3$)

	Session 1			Session 2		
	Moderately Altruistic	Strongly Altruistic	Behindness Averse	Moderately Altruistic	Strongly Altruistic	Behindness Averse
π	0.474***	0.405***	0.121***	0.544***	0.356***	0.100***
	(0.042)	(0.047)	(0.039)	(0.041)	(0.039)	(0.024)
α	0.065***	0.159***	-0.437***	0.061***	0.193***	-0.328***
	(0.013)	(0.036)	(0.130)	(0.009)	(0.019)	(0.073)
β	0.130***	0.463***	-0.147	0.095***	0.494***	-0.048
	(0.017)	(0.028)	(0.147)	(0.012)	(0.020)	(0.053)
γ	-0.001	0.151***	0.170	-0.005	0.099***	-0.028
	(0.012)	(0.026)	(0.119)	(0.006)	(0.024)	(0.030)
δ	-0.027**	-0.053**	-0.077	-0.019***	-0.082***	-0.015
	(0.012)	(0.025)	(0.162)	(0.007)	(0.018)	(0.035)
σ	0.032***	0.018***	0.008***	0.049***	0.019***	0.015***
	(0.002)	(0.001)	(0.002)	(0.004)	(0.001)	(0.002)

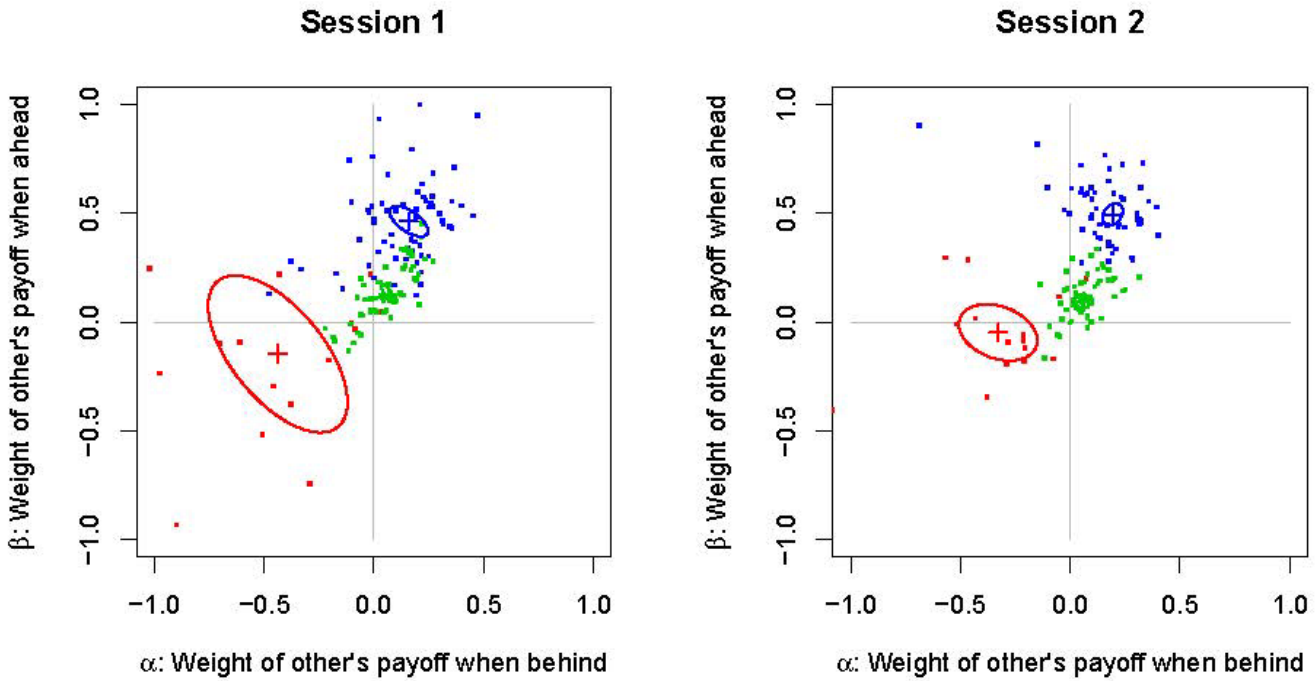
*** 1%, ** 5%, * 10%

What are the preference types?



- No purely selfish types exists
- Strongly altruistic type displays significant positive and negative reciprocity
- If anything, positive reciprocity is stronger in strong altruists
- No stable reciprocity in other types

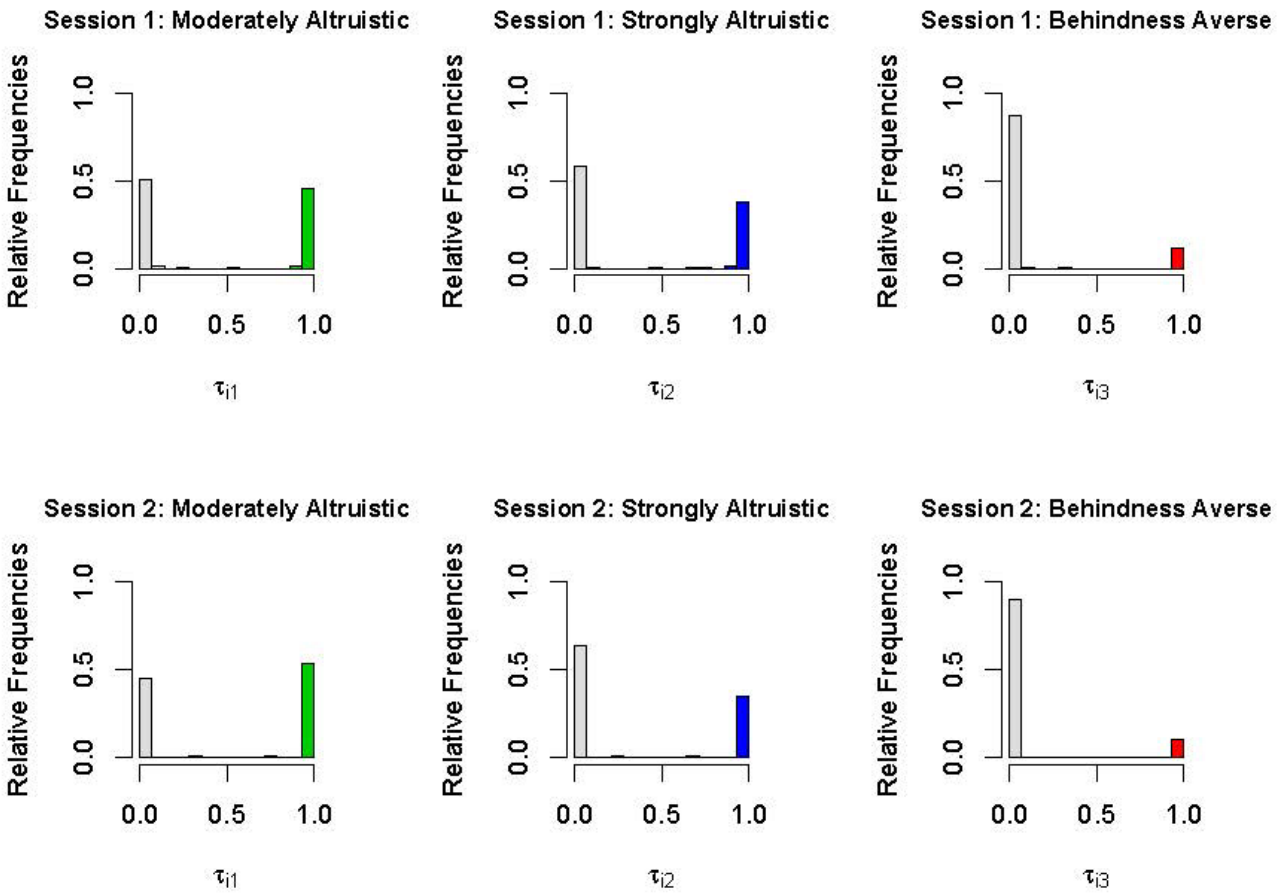
Type-Specific ($K = 3$) Preference Parameters and Subjects Individual Preference Parameters



Shares: 47%, 41%, 12%

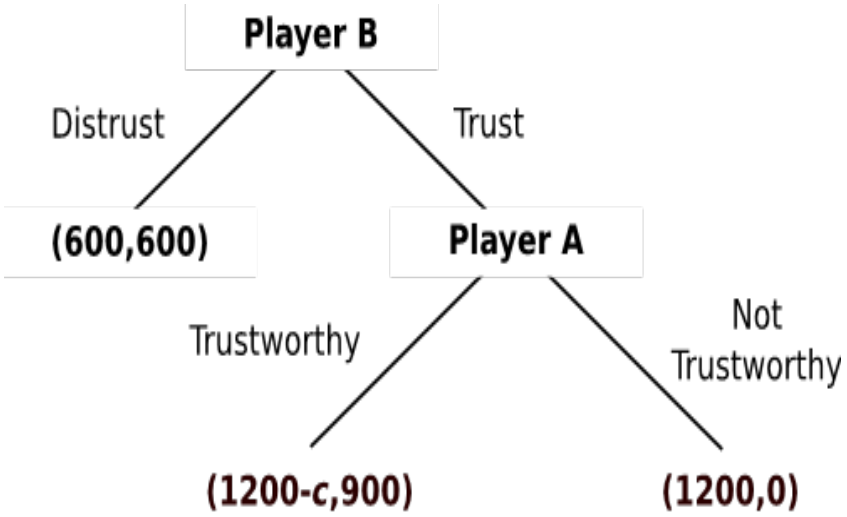
Shares: 54%, 36%, 10%

Unambiguous assignment of individuals to types?



Out of sample predictions

To test for stability across games, we predict the subjects' behavior in the additional games based on their estimated parameters



Payoffs: (Π^A, Π^B)

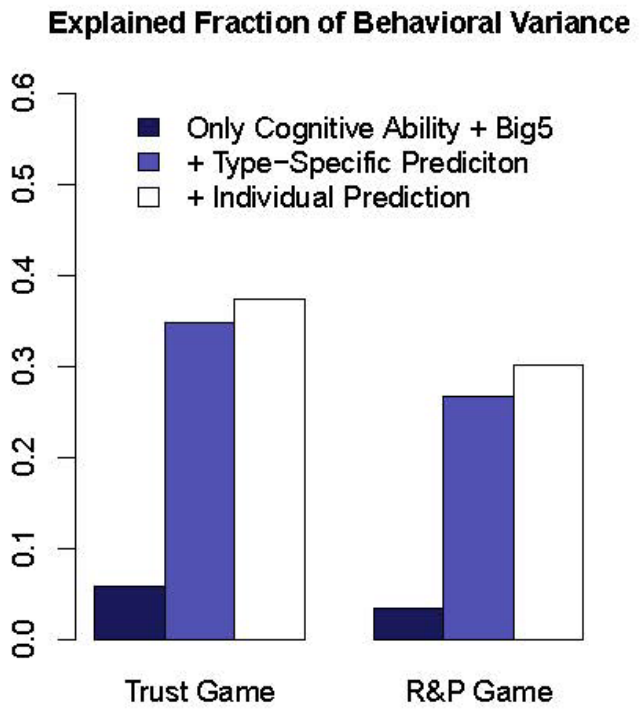
Cost of Being Turstworthy: $c \in \{0, 100, 200, ..., 900\}$

- Reward/Punishment 1
  - (600,600) vs (300, 900)
- Reward/Punishment 2
  - (700,500) vs (500, 700)
- 2nd mover could reward, do nothing or punish
  - Could pay 0, 10, 20, 30 to achieve r/p of 0, 100, 200 or 300

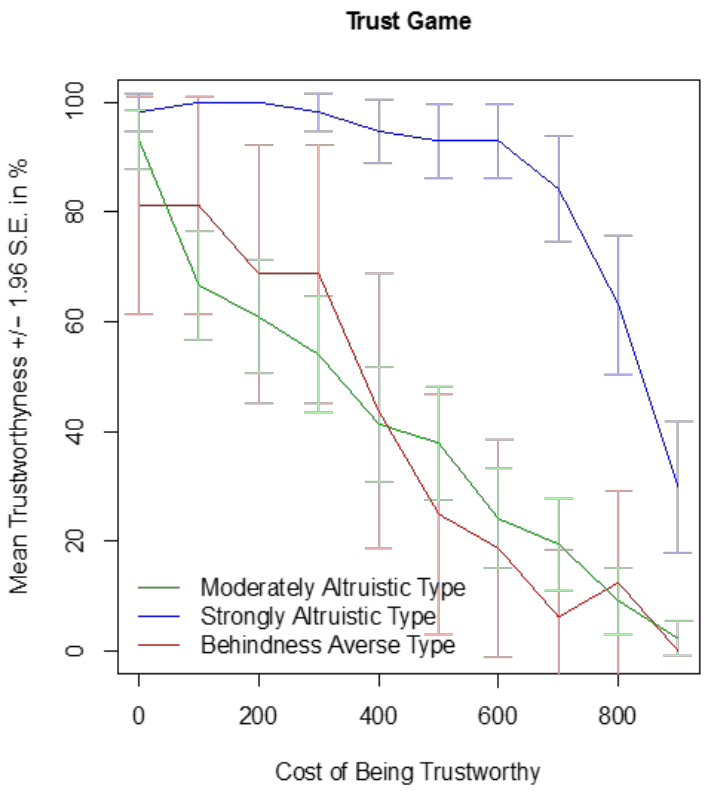
Out of sample predictions

To test for stability across games, we predict the subjects' behavior in the additional games based on their estimated parameters

- Regression of i's behavior in new games on prediction of i's behavior based on estimated preferences
  - Cog. Ability, Big5, age, gender, monthly income, field of study
  - Type-specific predictions
  - Individual-based prediction

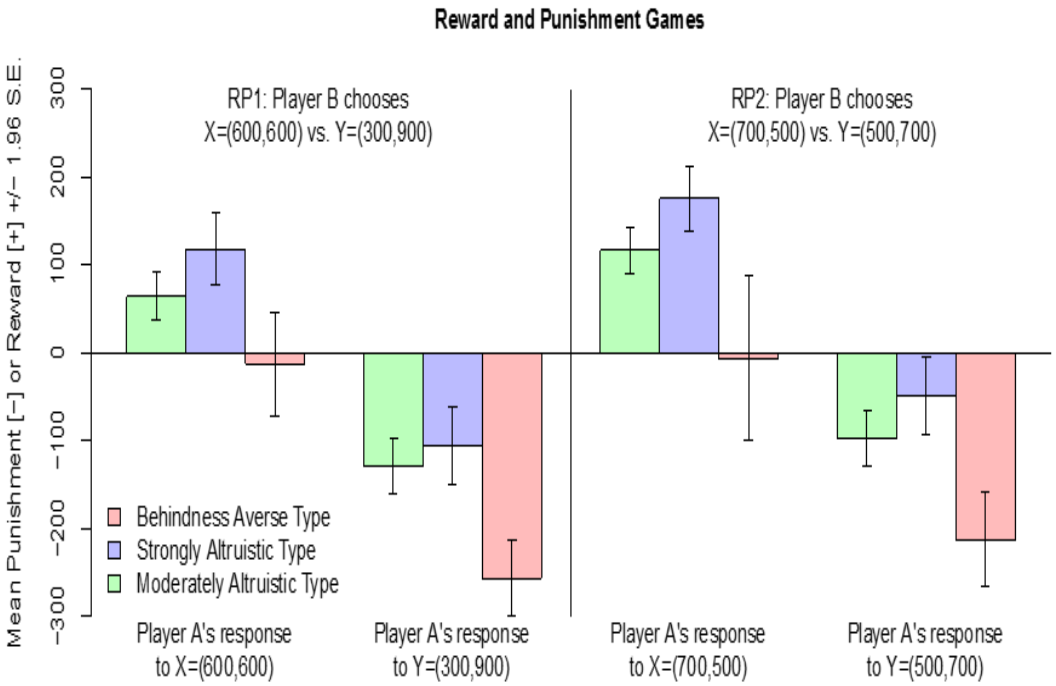


How good are our quantitative type-specific predictions?



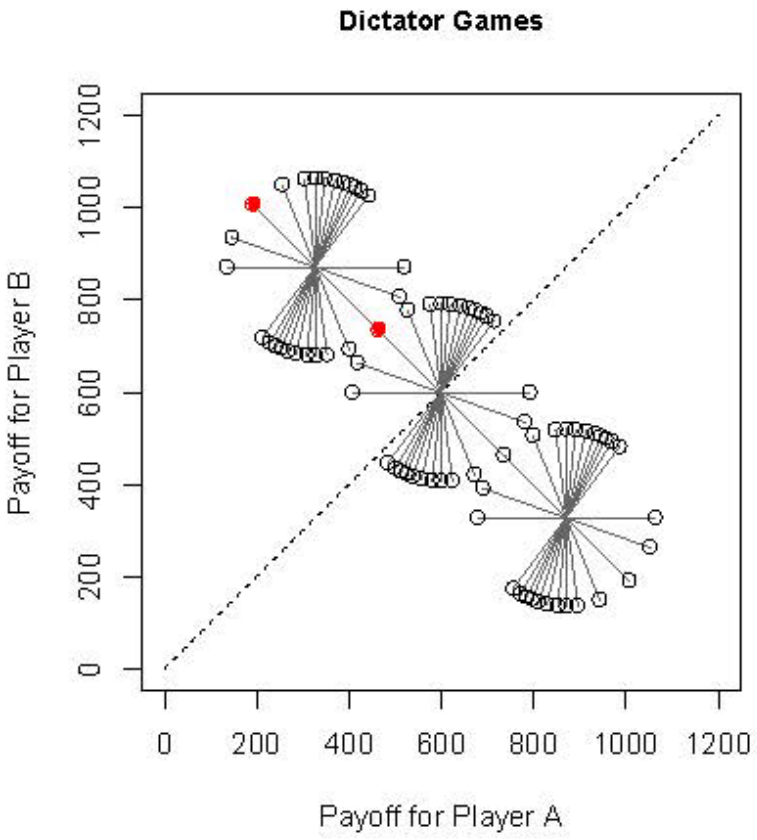
- Very good predictions for strongly altruistic types
- Too high trustworthiness for moderately altruistic types
- **Complete misprediction of behindness averse types' behavior.**
  - **They should never behave in a trustworthy manner**
- Does our preference identification capture positive reciprocity or positive inequality aversion insufficiently?

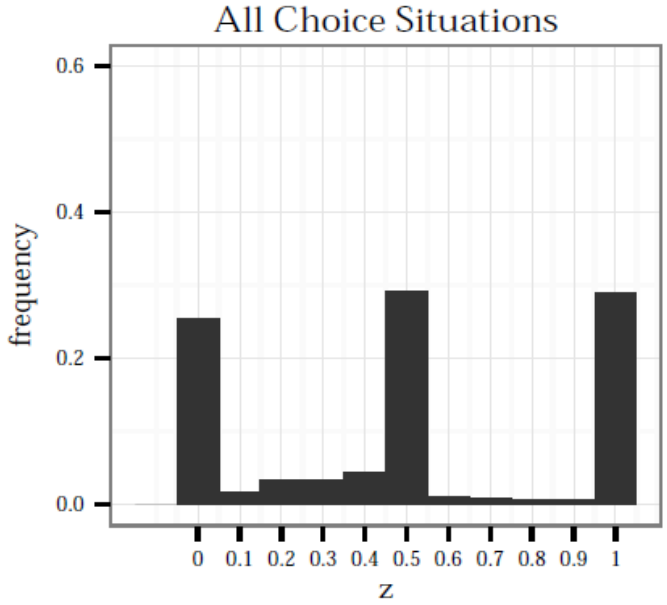
How good are our quantitative type-specific predictions?



- Very good predictions for rewarding behaviors
  - Strong A's > moderate A's > behindness averse types = 0
- **But moderate and high altruists should never punish!**
- Does our preference identification capture negative reciprocity or negative inequality aversion insufficiently?

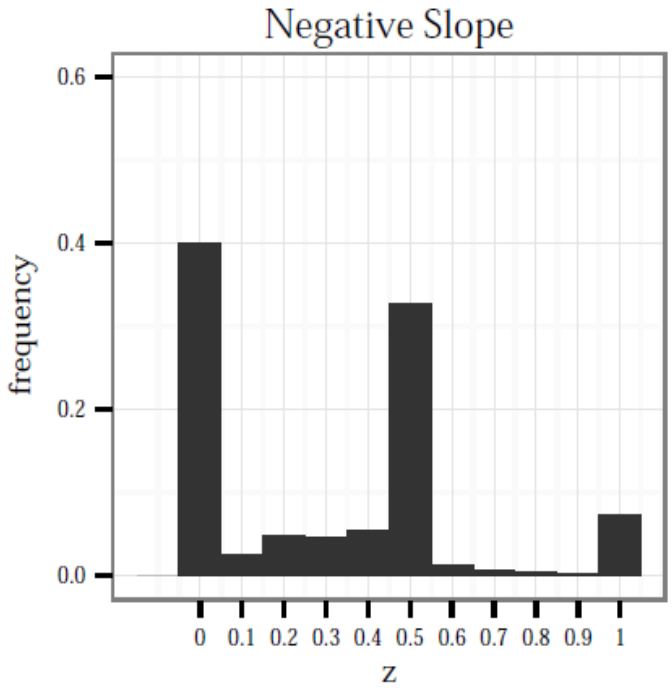
A speculation based on a large Danish sample where subjects could also choose equal payoff allocations on positively & negatively sloped budged lines

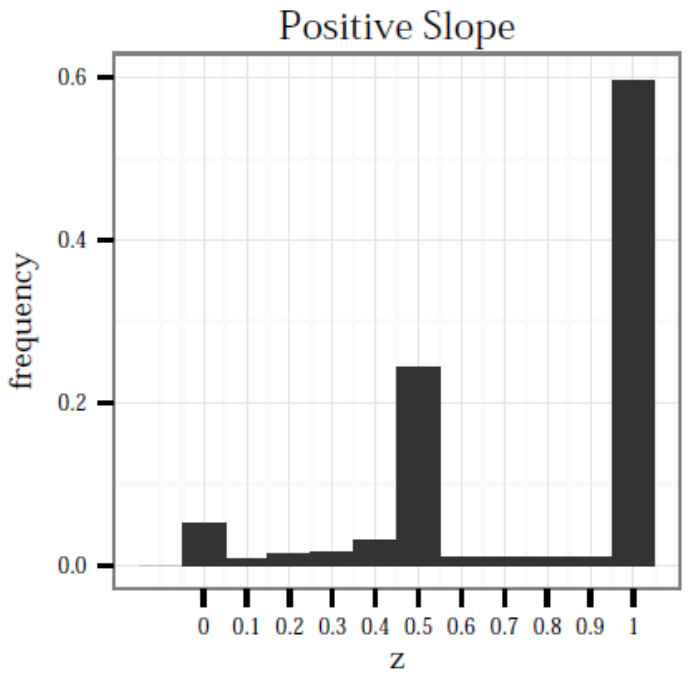




- Linear model a good approximation
- Allowing equal payoff distributions makes a big difference

z = 0 means that subjects maximize their own payoff, z = 0.5 means that subjects choose equal payoff allocation; z = 1 means that subjects maximize the other player's payoff





Many subjects are altruistic when it is costly but they don't give more than the equal split

A substantial fraction of subjects are willing to implement equality even when they would be better off maximizing the other's income

Summary I

- We can identify the relative quantitative importance of distributional and reciprocity preferences
  - Purely distributional preferences are considerably more important – at the aggregate and the type-specific level
- We provide a parsimonious characterization of the heterogeneity in terms of distinct preference types that emerge endogenously from the data :
  1. 50% moderate altruists displaying no reciprocity
  2. 40% strong altruists with significant reciprocity
  3. 10% behindness averse with no reciprocity
- No purely selfish type emerges
- Preference characteristics of the types are stable over time
- Individuals are unambiguously assigned to a type

Summary II

- Type-specific characterization is as good as individual preference estimates in out-of-sample predictions
- Type-specific estimation predicts the qualitative rankings of the intensity of various behaviors very well but
  - It underestimates the willingness to reciprocate in trust games (in moderately altruistic and behindness averse types)
  - It underestimates the willingness to punish among the altruistic types
- Our identification strategy may somewhat underestimate reciprocity and the new Danish data suggest that allowing for equal payoff allocations can be decisive
  - May be the reason that there are no inequality averse types
  - Alternatively, because the Danish sample is a broad population sample, broader population may just exhibit more inequality aversion compared to students which comprise the sample in the JEEA paper

Appendix

Empirical Analysis: Finite Mixture Model

The log likelihood of the finite mixture model is given by

$$\ln L(\Psi; C) = \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \pi_k f(\theta_k, \sigma_k; C_i),$$

where

- $f(\theta_k, \sigma_k; Ci)$ represents i's type-specific density contribution
- π_k is the mixing proportion the corresponds to type k's relative size
- $\Psi = (\theta_1, \ldots, \theta_K, \sigma_1, \ldots, \sigma_K, \pi_1, \ldots \pi_{(K-1)})'$ contains all the parameters of the model
- Maximizing the log likelihood of a finite mixture model is tricky and requires the EM algorithm (Dempster et al., 1977)

Empirical Analysis: Finite Mixture Model

After estimating the finite mixture model, we apply Bayes' rule to obtain ex-post probabilities of individual type-membership

$$\tau_{ik} = \frac{\hat{\pi}_k f(\hat{\theta}_k, \hat{\sigma}_k; C_i)}{\sum_{m=1}^{K} \hat{\pi}_m f(\hat{\theta}_m, \hat{\sigma}_m; C_i)}$$

Based on τ_{ik}, we can

1. classify each subject into the type she most likely stems from, given the fit of the model and given her data
2. Assess the quality of the classification of individuals into types.
   - If the classification is clean and the types are well separated, almost all subjects exhibit τ_{ik} close to 1 or 0
   - If the classification is ambiguous and the types overlap, many subjects exhibit $\tau_{ik} \approx 1/\mathrm{K}$

Type-specific actual average behavior versus predicted average behavior