Alma Mater Studiorum – Università di Bologna

DOTTORATO RICERCA IN

ECONOMICS

Ciclo 32°

**Settore Concorsuale: 13/A1**
**Settore Scientifico Disciplinare: SECS-P/01**

# Essays in Applied Economics: new empirical approaches to study individual behavior and improve policy targeting

**Presentata da: Michela Boldrini**

**Coordinatore Dottorato**
**Prof. Maria Bigoni**

**Supervisor:**
**Prof. Maria Bigoni**

**Esame Finale Anno 2020**

# Contents

# *Abstract*

This PhD thesis is composed of three, seemingly almost unrelated, chapters.

The first chapter, titled *"Twice Losers: How the shadow of cheating affects tax behavior and norms"* relies on a lab experiment to study whether the way income is generated in a society can impact individuals' willingness to pay taxes and judgements on the acceptability of tax evasion. In particular, I focus on whether the introduction of the suspicion that some individuals in the society could have got their income by cheating at the expenses of others alters individuals' behavior and acceptability ratings on tax evasion. This chapter is based on the first project I developed during my PhD and the experimental data collection was made possible by a generous grant received from International Foundation for Research in Experimental Economics (IFREE) under the Small Grants Program 2018.

The second chapter, titled *"Machine learning in the service of policy targeting: The case of Public Credit Guarantees"* originates from a joint work I developed with some colleagues at the Research Unit of the Bank of Italy, where I spent a few months during the Summer of 2017 and 2018 as a research intern. This project relies on a combination of tools from Machine Learning and casual inference in the attempt to propose an alternative targeting rule for Italy's main public guarantee program, named 'Fondo di Garanzia', which is a nation-wide scheme aimed to ease small and medium enterprises' access to bank credit through publicly funded collaterals.

The third and last chapter, titled *"Social preferences and strategic incentives for cooperation in infinitely repeated Prisoner Dilemmas"*, which I first started working on while visiting the Economics Department of the University of California Santa Barbara in early 2019, bridges my interests for applied econometrics and experimental economics. This paper investigates the role of structural game parameters and of social preferences in shaping cooperation in infinitely repeated Prisoner Dilemmas: in the first part, I collect data from previous experiments to run a meta-analysis aimed to test, using simple supervised learning algorithms, the predictive power of structural game parameters. In the second part, I develop a novel experimental design to collect data on both individuals' social preferences and cooperative attitude in infinitely repeated Prisoner Dilemmas, in order to further address my main research question.

## Chapter 1

# Twice Losers: How the shadow of cheating affects tax behaviors and norms [1]

[1] The author thanks the 'International Foundation for Research in Experimental Economics' (IFREE) for the financial support received under the Small Grant Program 2018.

## 1.1 Introduction

Both tax evasion and income inequality are extremely relevant issues in policy-makers agenda world-wide. Although at different speeds, income inequality has increased in nearly all countries in recent decades, with an increasingly uneven distribution of gains from the global income growth among global citizens, with the top 1% earners capturing twice as much the 50% poorest individuals (Alvaredo et al. (2018)).

This exacerbating trend has revived a lively debate on the extent to which these inequalities can be deemed fair, challenging the notions of deservingess in a context where, in an increasing number of cases, immoral or illegal behavior is found to be the source of the underserved wealth of a few.

The economic literature has explored how the degree of perceived fairness/deservingness at the basis of income inequality can impact individuals' redistribution preferences, but no evidence has been provided on whether and how this could also impact individuals' predisposition towards tax evasion.

We study whether tax evasion prevalence and tolerance are different in a society where income inequality may result from cheating rather than from sole differences in effort or ability, in a lab-controlled environment. In our experimental manipulation, the 'shadow of cheating' introduces an alteration in the degree of fairness attached to the process which generates incomes, opening up to the suspicion that the top positions in the pre-tax income distribution might have been reached not only through honest means, but also by taking advantage of available cheating opportunities at the expenses of others.

Working on previous literature contributions, which highlighted the importance of some income inequality features in shaping individuals' redistribution preferences (Alesina and Angeletos (2005); Durante et al. (2014); Bortolotti et al. (2017); Cappelen et al. (2018)), we design a novel experimental protocol to contribute to the study of how the characteristics of the income-generating process in a society can shape (i) individuals' behavior when they are personally involved in a redistribution action as taxpayers and not only consulted as un-involved third-party actors, and (ii) individuals' judgements on the social acceptability of tax evasion.

Our experimental design, based on a two-prong approach, allows to manipulate subjects' perceptions on the presence of the *shadow of cheating* on the income-generation process and to observe - relying on two different non-overlapping sets of participants - both subjects' actual contribution behavior in a context where complete anonymity is guaranteed (Experiment 1), and subjects' evaluations on the social acceptability of under-contribution choices (Experiment 2).

We find that the introduction of the shadow of cheating in the income-generating process does not alter neither subjetcs' evasive behavior nor subjects' judgements on the social acceptability of evasion. On average, participants evade approximentaly 60% of their due contributions and both the means and the distributions of percentage evasion measured in Experiment 1 are the same in the *Cheating* and in the *No Cheating* treatments. Likewise, the distributions of social acceptability ratings measured in Experiment 2 are the same across the two treatments for almost all possible scenarios. Based on the results reported by the literature that will be discussed below, this is an interesting null result, given that the absence of a significant effect is neither driven by an ineffective experimental manipulation nor by the presence of excessive noise.

However, we find that social acceptability norms, which result to be highly sensitive to the size of evasion and the income level of the evader, are a relevant driver of subjects' contribution choices. This suggests that injunctive social norms enter in the utility maximization process solved by taxpayers when deciding whether and/or how much to evade.

## 1.2 Motivation and Theoretical Framework

Income inequality and tax evasion represent two major issues for our modern societies but the relationship between these two phenomena is under-explored over a series of dimensions. With a handful of exceptions, the distributional effects of tax evasion have received little attention in the literature. The available evidence (Slemrod and Johns (2010), Matsaganis and Flevotomou (2010), Bishop et al. (2000), Benedek and Lelkes (2011)), which relies on microdata from different countries to address this issue, pointed out at the detrimental distributional effects of the distortions induced by tax evasion, both in terms of tax burden redistribution and horizontal income inequality. Another recent work by Nygard et al. (2018) confirms these findings for Norway as well, where they find that accounting for tax evasion leads to an estimate of income inequality that is higher than the official estimate, and to a level of actual income tax progressivity that is lower than what declared in the official figures.

Little is known, on the other side, on whether and how inequality - and in particular its severity and its origin - can impact tax evasion. A recent contribution by Alstadsaeter et al. (2019), which relies on a unique dataset on Scandinavia, estimates tax evasion rates at different points of the wealth distribution, finding a much larger percentage of tax evaded at the top ($\sim 25\%$) as compared to the other wealth groups ($< 5\%$). There is no evidence, however, on whether and how the origin or the degree of inequality in a society could impact tax compliance, for example, through behavioral channels.

In the study of tax evasion, behavioral determinants are increasingly gaining momentum as they can both allow for a better modeling of how individuals actually make their choices and represent a relevant instrument from a policy perspective, especially in those contexts where tax evasion is sizeable (as it is the case in Italy for example, where the amount of yearly evaded taxes has been recently estimated to be around 124,5-132,1 billions Euros [2]) but resources to fight evasion are limited and legal enforcement measures lack efficacy.

The literature on behavioral tax compliance determinants marginally focused on how individuals' attitude towards tax evasion could be impacted by how fair the individuals perceive the tax system and the distribution of the tax burden among taxpayers to be (see Luttmer and Singhal (2014) for a recent review and Mascagni (2017) for a review on related experimental applications). In this literature, alterations in individual perceptions of the degree of fairness in the tax system have been related to the way individuals value the quantity and the quality of the public good provided by the Government in exchange for tax payments (Spicer and Lundstedt (1976), Alm et al. (1992), Alm et al. (1993)), and the way individuals judge the equity of the tax burden they bear, given the tax structure or how much other taxpayers actually contribute (Spicer and Lundstedt (1976), Spicer and Becker (1980), Bordignon

[2] http://www.senato.it/application/xmanager/projects/leg17/attachments/documento/...pdf

(1993),Fortin et al. (2007)). In this framework, alterations in what individuals perceive to be the 'fair' amount of taxes to be paid could ultimately lead to changes in individuals' tax evasion behavior given that, at least in the short run, they cannot change what the actual tax rate is: individuals would opt for tax behaviors that allow them to minimize the distance between the due amount and what they consider the 'fair' amount of taxes to be paid, using evasion as a tool to restore equity in the system. The possibility that also the way income is earned by subjects could alter individuals' perceptions of what the 'fair' tax rates to be applied would be, has been insofar overlooked. Yet, the economic literature showing how fairness perceptions and redistribution preferences are affected by the way individuals earn their incomes, which defines the way income inequalities are generated, is abundant.

The economic literature that relates income inequality origins to redistributive preferences, indeed, proved that individuals do exhibit different redistribution preferences - which translates into different preferred tax rates - depending on what is the origin of income inequality (Alesina and Angeletos (2005); Cappelen et al. (2010); Durante et al. (2014)). Alesina and Angeletos (2005) develop a theoretical model to show how the interaction between social beliefs and welfare policies can lead to multiple equilibria: in this context, societies that believe individual effort is the only determinant of income prefer low redistribution and lower tax rates, while societies that believe the fundamental determinants of income and wealth are luck, birth, connections and corruption, prefer high redistribution and higher taxes rates. The experimental evidence brought by Durante et al. (2014), who study how subjects' preferred tax rates vary according to the way income is generated, also supports the idea that taste for redistribution depends on social preferences: eliciting subjects' tax preferences under two different scenarios - that of a disinterested decision maker and of an involved decision maker under the veil of ignorance about his position in the income distribution - they find subjects tend to support more redistribution when earnings are 'arbitrary' rather than when they are 'earned'. Interestingly, however, this difference vanishes when subjects are asked about their preferred tax rates from the perspective of an involved decision maker that is informed of his actual position in the income distribution.

Two recent experimental contributions by Bortolotti et al. (2017) and Cappelen et al. (2018) further strenghtened this evidence by focusing on the role that the introduction of *cheating opportunities* in the income-generation process might play. Bortolotti et al. (2017)

look at how third-party spectators are willing to redistribute income within pairs of players who either had or had not the opportunity to maximize their income, which is based on a gamble, by means of cheating: even if spectators are not in the position to detect cheating with certainty, the presence of the *shadow of cheating* is sufficient to shift their fairness views leading to a large increase in share of 'Egalitarians', who are willing to implement perfect equality through redistribution. Cappelen et al. (2018) look at how third-party spectators are willing to redistribute income within groups of players who worked on a real effort task and either had or had not the opportunity to falsely report to have completed their assignment. Spectators have limited information on cheating, since they are informed of the number of *cheaters* in the group but are forced to treat all group members equally when deciding on the level of redistribution, trading-off between false positives (giving some more than they deserve) and false negatives (giving some less than they deserve). They find that, despite a high degree of heterogeneity, spectators are generally more concerned with avoiding false negatives rather than false positives.

Our aim is to expand this branch of the experimental literature, by investigating whether the presence of this suspicion – namely, of the *shadow of cheating* – can also have an impact on tax evasion behavior, in a situation which mimics more closely the reality, where the outcome of the redistributive process depends on the actions of the same actors involved in the income production phase and not on the decisions of an external un-involved third-party spectators. Our experimental design is tailored to the identification of the causal effect of the presence of cheating opportunities in the income-generating process on tax evasion, through the use of a novel experimental procedure that allows us to introduce cheating opportunities and to manipulate subjects' perceptions of the intensity of cheating, while minimizing the effects of other factors that may confound the effect of our treatment (experimenter demand effect, observability concerns, etc.). We are further interested in studying whether the presence of the *shadow of cheating* can also impact the social norms on the acceptability of evasion, and in testing whether those norms are relevant drivers of actual evasion behavior, as a growing body of recent economic research suggests to be the case across a wide range of contexts (Burks and Krupka (2012), Gȧchter et al. (2013); Krupka and Weber (2013); Banerjee (2016); Gȧchter et al. (2017); Krupka et al. (2017)).

The motivation to focus on this question comes from the observation that in the real world, and in some countries more than others, the suspicion that some individuals in the society got their income at least partially my means of cheating is far from being rare, especially with respect to individuals who hold very top positions in the income distribution.

As reported by the World Values Survey data [3] (see Figure 1.1 people tend to attach a high weight to factors such as luck and connections when asked to report whether they believe that in the long-run life success depends on hard work and effort rather than by luck and connections. Overall, 31% of respondents report that luck and connections outweigh effort as long-run life success's driver, but there's a good degree of heterogeneity across countries and in Italy, for example, this share is equal to approximately 46%.

The same scenario is reported by the International Social Survey Programme data [4] where individuals are asked how important they would rate knowing the right people and having political connections to get ahead in life (see Figure 1.2). Overall, 55% and 25.5% of the respondents, respectively, claimed that knowing the right people and having connections is either a very important or essential factor: still there's a good degree of heterogeneity across countries and Italy, again, is positioned above the overall average in both cases with amost 59% and 41% of the respondents recognizing the two channels as at least very important to succeed in life.

From the World Values Survey data, we can also observe a positive and significant correlation between the perception on how justifiable is tax evasion in a country, and the average weight attached to luck and connections, as opposed to own effort, as the main determinant of life success (see Figure 1.3). This evidence, however, has two main limitations: first, it does not distinguish the role of luck from the role of connections, although these two factor are likely to bring about different considerations in terms of fairness; second, this analysis is able to

[3] WVS Wave 5: 2005-2009 - Countries included in the survey: Andorra, Argentina, Australia, Brazil, Bulgaria, Canada, Chile, China, Taiwan, Colombia, Cyprus, Ethiopia, Finland, France, Georgia, Germany, Ghana, Hungary, India, Indonesia, Iran, Italy, Japan, Jordan, South Korea, Malaysia, Mali, Mexico, Moldova, Morocco, Netherlands, New Zealand, Norway, Peru, Poland, Romania, Russia, Rwanda, Vietnam, Slovenia, South Africa, Spain, Sweden, Switzerland, Thailand, Trinidad and Tobago, Turkey, Ukraine, Egypt, Great Britain, United States, Burkina Faso, Uruguay, Serbia and Montenegro, and Zambia.

[4] International Social Survey Programme "Social Inequality" data: 2009 - Countries included in the survey: Argentina, Australia, Austria, Bulgaria, Chile, China, Croatia, Cyprus, Czech Republic, Denmark, East-Germany, Estonia, Finland, Flanders, France, Great Britain, Hungary, Iceland, Israel, Italy, Japan, Latvia, Lithuania, New Zealand, Norway, Philippines, Poland, Portugal, Russia, Slovakia, Slovenia, South Africa, South Korea, Spain, Sweden, Switzerland, Taiwan, Turkey, Ukraine, United States, Venezuela, West-Germany.

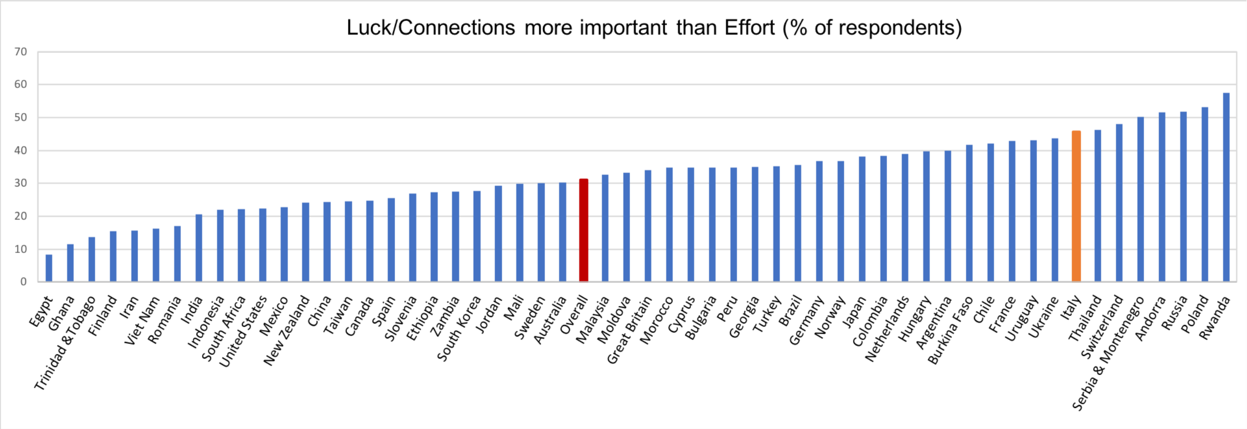


Figure 1.1: World Values Survey data (2005-2009)

show only a simple correlation given that it is not possible to draw any casual statement on the impact of the role of luck and connections on tax evasion acceptability perceptions.

Our experimental approach aims to overcome these limitations to study in isolation what is the impact of cheating, which mimics the opportunity to exploit personal connections and other non-regular or transparent channels to maximize own benefits in the real life, on tax evasion.

We do not aim to directly test a specific theoretical model, but rather to advance experimental knowledge on tax evasion behavioral mechanisms, incorporating the new insights coming from the recent experimental evidence on the impact of fairness concerns on redistribution preferences cited above. Given that individuals' preferences for redistribution have been shown to be sensitive to how incomes – and related inequalities – are generated, we aim to test experimentally whether the presence of the 'shadow of cheating' in the income generating process, altering individuals' preferences for redistribution and thus their perception on the degree of fairness of the tax burden, can also have an impact on their decision to engage in tax evasion and on the degree to which they consider tax evasion to be socially tolerable.

We can formulate some predictions which will guide us through the analysis of the experimental evidence by relying on a simple theoretical framework that is obtained starting from a simplified version of the model for individual tax evasion decision developed by Bordignon

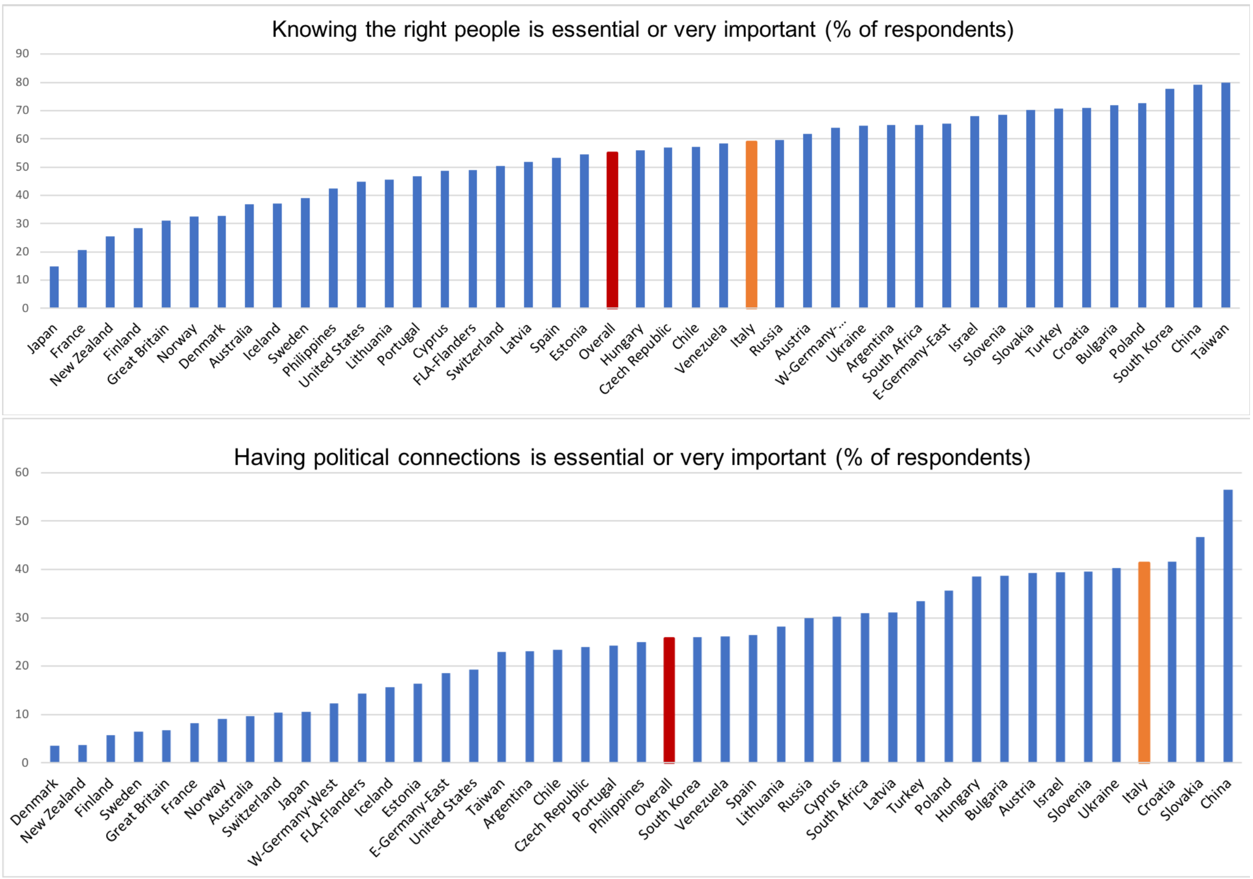


Figure 1.2: International Social Survey Programme data (2009)

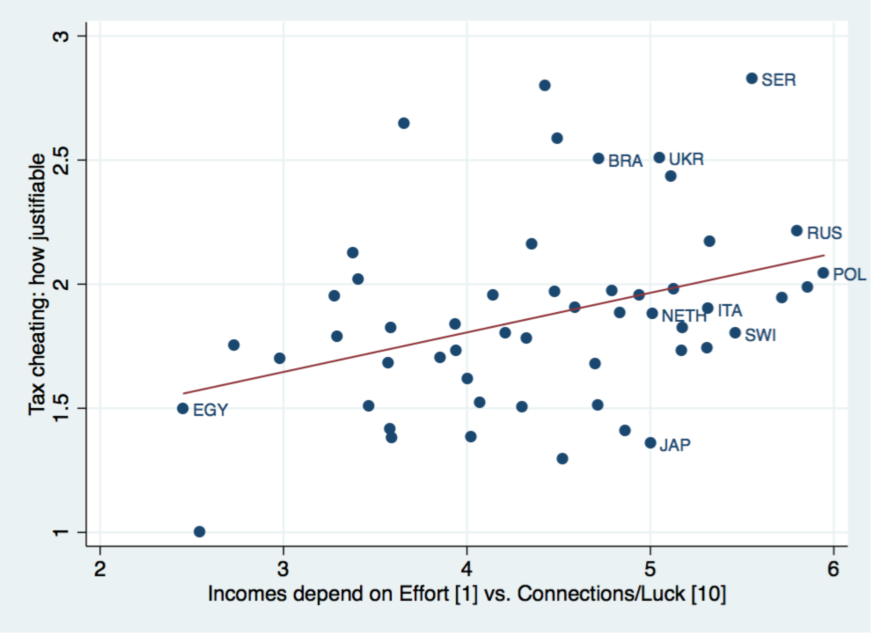


Figure 1.3: World Values Survey data (2005-2009)

(1993).

We consider a situation where a population of size $N$ is composed by $N$ types of identical individuals indexed by $i = 1, .., N$, where each type is endowed with an exogenously given income $I_i$. Each type's perferences can be represented by a strictly concave, twice differentiable utility function denoted by $U^i = U^i(C_i, G)$ that is defined over two types of goods: private consumption $C_i$ and a public good $G$. Taxpayer behavior is modeled as the result of a constrained maximization problem, where the taxpayer attempts to maximize his own utility subject to a costraint given by his desired amount of (feasible) tax evasion, in a simplified context where there's a zero-detection probability that makes evasion a risk-less activity.

The taxpayer problem can be described as follows:

$$\begin{aligned} max \quad & U^i[(1-t)\cdot I_i + x_i; G] \\ s.t. \quad & 0 \leq x_i \leq \bar{x_i} \end{aligned}$$

where:

$t$ is the proportional tax rate;

$x_i$ is the choice variable, which represents the amount of tax evaded by individual i;

$\bar{x_i}$ is the *fairness constraint*, which is determined endogenously as a function of fiscal parameters $t$ and $G$, evasion by other taxpayers $x_j$ and the intensity of cheating $\lambda$.

In absence of any fairness constraint, the solution of the standard utility maximization problem would be $x_i^0$, which, in absence of any probability to be caught and punished for evading (zero probability of detection), would be trivially equal to $x_i^{MAX} = t \cdot I_i$.

In presence of a fairness concern, the solution of the maximization problem would be written as $\hat{x_i} = min\{x_i^0, \bar{x_i}\}$: given that $x_i^0 = x_i^{MAX}$ we would have $\bar{x_i} \leq x_i^0$ and the constrained maximization problem would be solved for $\hat{x_i} = \bar{x_i}$.

The magnitude of the fairness constraint $\bar{x_i}$ depends on the desired level of tax evasion $z_i$, which is given by the difference between what should be paid $(t \cdot I_i)$ and what the taxpayer considers to be 'the fair tax' $(q_i^F)$: $z_i = t \cdot I_i - q_i^F$.

My definition of $q_i^F$ departs from the model by Bordignon in one important way. As in the original model, $q_i^F$ depends on actual fiscal parameters (mainly the actual tax rate and quantity/quality of the public good provided), on the level of tax evasion by other taxpayers [5] and the desired tax rates each individual would like to apply to himself and on others.

Yet, compared to the original version of the model, where an individual of type $i$ has a unique desired tax rate for each individual of type $j$ ($t_j^i$) given fiscal parameters, we introduce two levels of desired tax rates for each individual of type $j$, in order to account for the possibility that some individuals in the population could have got their income by cheating: $t_{j,C}^i$ is the tax rate individual $i$ would apply to type $j$ in the case type $j$ earned his income by cheating, and $t_{j,NC}^i$ is the tax rate individual $i$ would apply to type $j$ in the case type $j$ earned his income by honestly, without exploiting cheating opportunities. The actual tax rate an individual of type $i$ would apply to an individual of type $j$, given that he is not in the position to distinguish with certainty whether type $j$ actually cheated or not, would then be the weighted average of these two tax rates, where weights depend on the perceived intensity of cheating occurrences in the income-generating process, which is measured by $\lambda$ ($0 \leq \lambda \leq 1$).

$$t_j^{i*} = \lambda \cdot t_{j,C}^i + (1 - \lambda) \cdot t_{j,NC}^i$$

If $\lambda = 0$ then $t_j^{i*} = t_{j,NC}^i$, bringing us back to the standard case, while when $\lambda > 0$ we would have $t_j^{i*} > t_{j,NC}^i$, as suggested by the literature (Bortolotti et al. (2017), Cappelen et al. (2018)).

In presence of fairness and reciprocity considerations, assuming that an individual of type $i$ would be concerned only about the tax behavior of individuals of other types $j \neq i$ we can define - selecting a linear specification for simplicity - the fair tax $q_i^F(t, G, x_j, \lambda)$ as:

$$q_i^F = t_i^i I_i - \sum_{j=1}^{N-1} \phi_j^i [t_j^i I_j - (tI_j - x_j)]$$

where $x_j$ is the average level of tax evaded by individuals of type j. If we compare the fair tax we would have in absolute absence of cheating, $q_{i,\lambda=0}^F$, and the fair tax we would have when cheating opportunities are present, $q_{i,\lambda>0}^F$, assuming only judgements on desired tax rates for other types are affected by the presence of cheating opportunities ($t_{i,C}^i = t_{i,NC}^i$) and average tax evasion by other types is constant across the two scenarios ($x_j = x_{j,C} = x_{j,NC}$),

[5] In a population composed of N types of individuals with different income levels and one individual of each type, we assume that each individual only cares about evasion by other types $x_j$, where $j \neq i$.

we obtain that the difference in fair tax brought by the presence of cheating with respect to the baseline without cheating opportunities would depend on:

- reciprocity weights on others' contributions $\phi_j^i$ , where $j = 1,..,N-1$ and $0 \leq \sum_{j=1}^{N-1} \phi_j^i \leq 1$
- the difference between the two desired tax rates in presence and absence of cheating $[t_{j,\lambda>0}^i - t_{j,\lambda=0}^i]$

$$\begin{aligned} q_{i,\lambda>0}^F &= t_i^i I_i - \textstyle\sum_{j=1}^{N-1} \phi_j^i [t_{j,\lambda>0}^i I_j - (tI_j - x_j)] \\ q_{i,\lambda=0}^F &= t_i^i I_i - \textstyle\sum_{j=1}^{N-1} \phi_j^i [t_{j,\lambda=0}^i I_j - (tI_j - x_j)] \\ \Delta q_i^F &= -\textstyle\sum_{j=1}^{N-1} \phi_j^i \cdot I_j \cdot [t_{j,\lambda>0}^i - t_{j,\lambda=0}^i] \end{aligned}$$

In presence of reciprocity concerns ($\sum_{j=1}^{N-1} \phi_j^i > 0$), if, as suggested by experimental evidence, the difference between the two desired tax rates $[t_{j,\lambda>0}^i - t_{j,\lambda=0}^i]$ is positive, we would have that the difference in fair tax induced by cheating is *negative*. This would, in turn, lead to a higher level of desired tax evasion in presence of cheating: $z_{i,\lambda>0} = t \cdot I_i - q_{i,\lambda>0}^F > z_{i,\lambda=0} = t \cdot I_i - q_{i,\lambda=0}^F$, and therefore a higher fairness constraint $\bar{x_i}$, which, under general conditions, would be equal to the desired level of evasion $z_i$ [6].

Within this framework, working on available experimental evidence on how the source of inequality can influence individual redistribution preferences (Bortolotti et al. (2017); Cappelen et al. (2018)), and on how individuals react to fairness violations (Houser et al. (2012); Spicer and Becker (1980)), our goal is to investigate three main questions: first, we test whether in presence of an opaque income-generation process - where opportunities to cheat and maximize own revenues at the expenses of others are available - individuals are more likely to engage in under-contribution (RQ1). We further test whether the introduction of cheating opportunities in the income-generation process affects individuals' average tolerance towards evasion, looking at whether and how the introduction of cheating opportunities impacts injunctive social norms on the acceptability of evasion (RQ2). Lastly, to verify whether individuals' fairness constraint $\bar{x_i}$ is also a function of the injunctive social norm on the acceptability of evasion $\bar{x}_i(t, G, x_j, \lambda, \bar{N})$, we test whether the social norms we elicit are good

[6] Under general conditions we have that $0 \leq z_i \leq 1$ and we exclude both the possibilities that the taxpayer can evade a negative amount or expect a subsidy from the Government.

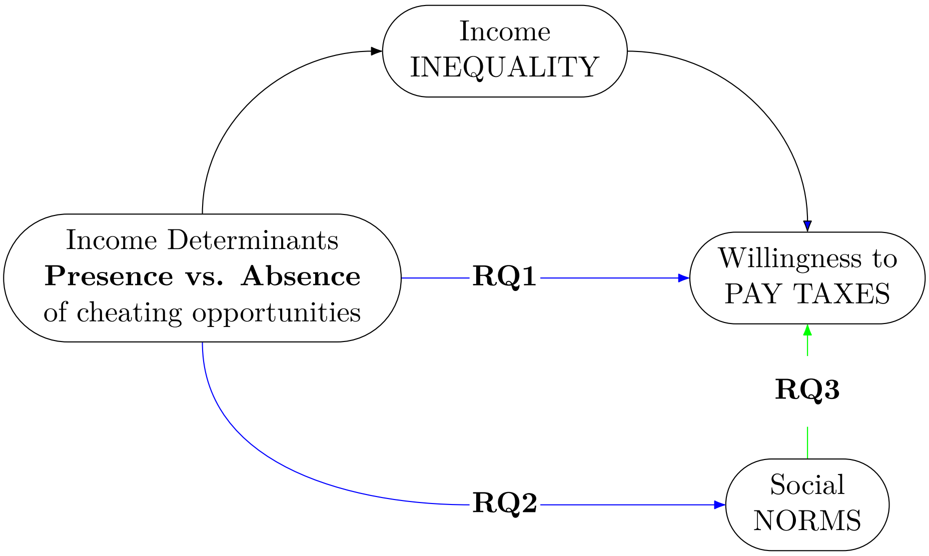


Figure 1.4: Research Questions

predictors of subjects' actual evasion behavior (RQ3).

We will therefore test whether the introduction of the shadow of cheating has both a direct (RQ1) or an indirect effect on tax evasion through the social norms channel (RQ2 & RQ3), see Figure 1.4.

While the literature on tax compliance has already explored the role of *descriptive norms* (or empirical expectations) on tax behavior, finding some evidence that tax compliance behavior is indeed influenced by what individuals perceive to be the prevalent behavior in the society (Doerrenberg and Peichl (2018)), no attention has been devoted insofar to the study of what factors shape *injunctive norms* (or normative expectations), which reflect collective perceptions about the appropriateness and acceptability of a certain behavior, and to the extent to which this type of norms can influence individual behavior. Our experimental design would allow us both to study what are the main determinants of injunctive social norms on tax evasion and to verify whether also this type of social norms can be relevant for tax compliance.

## 1.3 Experimental Design

The objective of the experiment is to re-create in the lab a small-scale economic system where individuals, as it happens in the real-life, first earn their income based on a combination of their own effort and luck, and later are asked to pay taxes contributing with a share of the income they earned to finance publicly shared services. The main manipulation of the experiment refers to the presence, and the role, of *cheating* in the income-generating process. The experimental design is specifically tailored to introduce cheating in the lab in a controlled way while minimizing subjects' scrutiny and observability concerns and the risks of a strong experimental demand effect.

It would have been hard to collect clean evidence on this phenomenon from observational data since it would have proved difficult to measure whether and to what extent individuals in real-world situations ascribe their peers' earnings to their effort and ability rather than to some forms of cheating (e.g. personal or political connections or small bribes), while controlling for the multiplicity of other factors that may influence individuals' decisions to engage in tax evasion in real life (e.g. different perceptions of the risk to be caught evading, different judgements in terms of the quality/quantity of publicly provided services provided in exchange for tax payment, personal attitude towards the Government etc.). Our experimental protocol is designed to overcome these limitations and offers a unique opportunity to identify the causal effect of the introduction of the shadow of cheating on tax evasion behavior and norms, by exogenously varying the characteristics of the process through which income is generated while keeping all other factors constant across treatments.

Our experimental design is based on a two-prong approach with two non-overlapping sets of participants: Experiment 1 is used to trace subjects' actual contribution behavior, while Experiment 2 serves to elicit (injunctive) social norms from an un-involved sample of participants.

The experiments were conducted in the Bologna Laboratory for Experiments in Social Sciences (BLESS) between October 2018 and May 2019. It was programmed using z-Tree (Fischbacher (2007)) and subjects were recruited via ORSEE (Greiner (2004)). A total of 270 subjects took part in the two experiments [7]: 180 subjects participated in Experiment 1 (Behaviors) and 90 subjects participated in Experiment 2 (Norms), see Figure 1.5. Each session involved 15 subjects and lasted approximately 75-80 minutes with participants earning on average 13.5 Euros in both experiments.

[7] Additional 64 participants took part in the first four pilot sessions of the two experiments, run in June 2018. Data from these sessions are discarded from the analysis due to a slight difference in experimental procedures.

| | **EXP 1 - Behaviors** | **EXP 2 - Norms** |
|---|---|---|
| **T1 (*Cheating*)** | 90 obs. | 45 obs. |
| **T2 (*No Cheating*)** | 90 obs. | 45 obs. |

Figure 1.5: Two-Prong Experimental Design

### 1.3.1 Experiment 1: Behaviors

The experiment is divided into three phases, see Table 1.1. The design of the second and the third phase is identical across treatments, while the design of the first phase varies across the two *Cheating* and *No Cheating* treatments.

| *Phase* | *Activity* | *T1: Cheating* | *T2: No Cheating* |
|---|---|---|---|
| | Subjects enter the lab and have a randomly assigned seat | Subjects sign the attendance sheet and randomly extract secret IDs | Subjects sign the attendance sheet and randomly extract secret IDs |
| 1st Phase | Slider Task | Slider trial n.1<br>Slider trial n.2<br>▷ *Non-dominant* hand rule<br>.<br>.<br>Slider task<br>[**No Monitoring**] | Slider trial n.1<br>Slider trial n.2<br>▷ *Non-dominant* hand rule<br>**Colored Gloves**<br>.<br>Slider task<br>[**Perfect Monitoring**] |
| 2nd Phase | Contribution Task | Paper-based contribution task using secret IDs | Paper-based contribution task using secret IDs |
| 3rd Phase | Personality Test | Paper-based HEXACO test using secret IDs | Paper-based HEXACO test using secret IDs |

Table 1.1: **Timing of Experiment 1**

In the $1^{st}$ phase, individuals are grouped in 5-people groups and have the opportunity to earn money based on their performance in the *slider task* (Gill and Prowse (2012)).

In the slider task, participants see a series of 48 sliders ranging from 0 to 100 on their computer screen and have to adjust each slider to exactly the middle position (50) within the given 120 seconds (see the Additional Figures section in the Appendix A.1, Figure A1.1): subjects are allowed to use only the touch-pad to drag and adjust sliders' position [8] and earn one point for each correctly positioned slider. Subjects' objective is to maximize the number of correctly positioned sliders before the time is over: when the time runs out subjects are ranked based on their performance in the task with respect to their other group-mates, according to a rank-order tournament payment scheme with multiple prizes (Freeman and Gelber (2010), Moldovanu and Sela (2001)), see Figure 1.6.

The earnings scheme is fixed across sessions and does not depend on the actual (absolute) performance of participants: this implies the level of income inequality in the society (group) at the end of the income-generation task is always the same in all sessions. Participants' earnings are expressed in ECUs (Experimental Currency Units), which are then converted in Euros at the exchange rate 1 ECU = 0.5 Euros.

| | |
|---|---|
| $1^{st}$ Best performance | 20 ECUs |
| $2^{nd}$ Best performance | 16 ECUs |
| $3^{rd}$ Best performance | 12 ECUs |
| $4^{th}$ Best performance | 8 ECUs |
| $5^{th}$ Best performance | 4 ECUs |

Figure 1.6: Earnings Scheme - Slider Task

Subjects face the slider task three times: the first two times represent trial sessions [9], the last session is only one relevant for the determination of subjects' earnings. When the first two trial sessions are over subjects receive instructions on the main execution rule [10]: subjects are instructed to use their *non-dominant hand* only when executing the task in the last payoff-relevant round. Our treatment manipulation relates to the monitoring over the implementation of the non-dominant hand rule and on the availability of cheating opportunities, which subjects can choose to exploit to boost their performance in order to get to higher rank positions and obtain higher earnings. In this context, exploiting cheating opportunities goes 'at the expenses' of other group members: cheating, in fact, cannot increase the amount of total surplus in the group but can only change the way in which it is distributed among

[8] Throughout the whole duration of the experiment we used a keyboard locker software ('KeyTweak' ) to prevent subjects from using the arrow keys or the mouse wheel.

[9] The first trial session lasts 200 seconds, the second one 120 seconds just like the final incentivized session.

[10] This design allows us to use data on the second trial session lasting 120 seconds as a control for subjects' individual ability in the task.

members, favoring cheaters at the expenses of the best-performing honest members.

In the *No Cheating* treatment cheating is impossible because we can can implement perfect monitoring on the respect of the non-dominant hand rule: when subjects arrive to the lab we are able to identify left- from right-hand users by checking which hand they use to sign the lab attendance sheet [11]. We assign each subject textile glove, which has a different color based on whether the subject is classified as a left- (gray glove) or right-hand user (white glove). After subjects learn of the non-dominant hand rule, in the *No Cheating* treatment they are invited to wear the textile glove over their dominant hand and to place the hand wearing the glove on their desk in plain sight: these textile gloves prevent subjects from using the touch-pad with their dominant hand and, given the two different colors, allow us to check whether the subjects are wearing the glove on the right hand during the execution of the payoff-relevant task.

Conversely, no form of monitoring is implemented in the *Cheating* treatment: subjects are simply instructed to use their non-dominant hand during the payoff-relevant task but we are not able to check whether the rule is respected as we do not mark subjects' dominant hands at the beginning of the experiment and therefore we cannot implement any form of monitoring.

When the last session of the slider task is over, subjects are informed of their absolute performance, of their relative performance in the group with respect to other group members and of their realized earnings [12].

The $2^{nd}$ phase is the same irrespective of what treatment subjects have been exposed to in the $1^{st}$ phase.

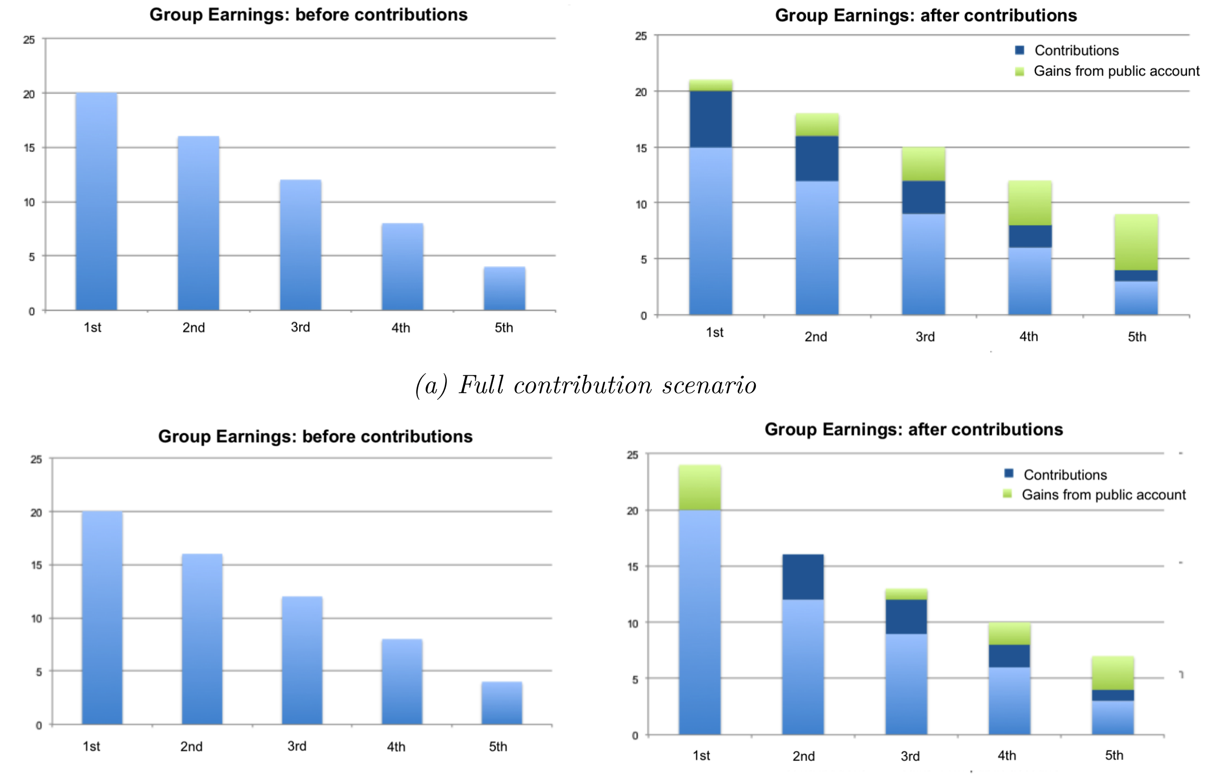
All subjects, irrespective of what is their position in the income distribution, are required to contribute with the same share of their realized earnings - equal to 25% - to a project that is common to all their group mates: subjects know that all contributions made by group members will be summed up, multiplied by 2 and then equally divided among participants, as in

[11] We double check by explicitly asking them which hand they usually use as their dominant-hand in daily tasks.

[12] Before providing this information we elicit subjects' guesses on their positions in the group ranking and subjects' perception on the frequency of cheating occurrences in the payoff-relevant slider task to check whether our manipulation is effective in altering subjects' perception on the presence and intensity of cheating across treatments.

a standard Public Good Game, where the marginal per capita return (MPCR) of each ECU contributed to the group project is equal to 0.4 ECUs for every group member, irrespective of their own actual contributions. This design mimics a proportional income-tax scenario, where tax contributions have moderate redistributive effects. Returns from contributions to the public account are decreasing in the level of individuals' pre-contribution income and, under a full-contribution scenario, we observe the highest efficiency and redistribution gains: global aggregate earnings increase by 25% and inequality decreases, with a 10pp decline in the Gini index, from 0.26 to 0.16. When some individuals decide to under-contribute or not to contribute at all, however, the distribution of gains changes and the effects on inequality can also be negative: if, for example, all individuals but top-earners fully contribute, we will have higher inequality than in the pre-tax scenario, with an ex-post Gini of 0.28, and lower efficiency, with a lower increase in overall aggregate earnings (see Figure 1.7).

Figure 1.7: Income distribution before and after contributions



*(a) Full contribution scenario*

*(b) Full contribution by all but top-earners*

In order to minimize demand effects and to guarantee the highest level of anonymity, tax contributions are collected through a paper-based procedure that is largely inspired by the procedure developed by Barmettler et al. (2012) for their the ‘double-anonymous’ treatment, which allows to implement a high degree of subject-experimenter anonymity. Barmettler et al. (2012), at the very beginning of the experiment, ask their subjects to draw a small envelope out of a nontransparent fabric bag: all envelopes contain a set of small identity cards with an identity number printed on them, and since envelopes are randomly drawn the identity number is known only to the subject himself/herself. Subjects use those identity cards to mark all the decisions they take throughout the experiment and when the experiment is over the experimenter calculates the payoffs and places the corresponding amounts of money in stuffed envelopes labeled with the identity numbers of the players: players are then called to the payout table by an assistant who was not present while payoffs were calculated, and are given the payout envelope that corresponds to the number on their identity cards. This ‘double-anonymous’ treatment ensures nobody has the possibility to link names or faces to payoffs, guaranteeing a high degree of subjects’ anonymity with respect not only to other subjects but also to the experimenter.

We implement a similar procedure to allow our subjects to be free to contribute whatever amount of ECUs they prefer, by choosing a quantity between zero and what has been explicitly requested by the instructions (25% of Earnings from the slider task), knowing that full anonymity towards both the experimenter and the other subjects in their group is guaranteed. This sets to zero the detection probability for under-contributers, allowing us to isolate the effect of our interest from its interactions with subjects’ own predisposition towards risk[13]. By design, before the experiment is over there’s no way to detect under-contribution at the individual level thus detection probability for evaders is zero. At the end of the contribution phase subjects are paid in cash privately and the payment procedures ensure that choices taken by subjects throughout the experiment cannot be matched with subjects’ real identities.

In the 3rd and last phase of the experiment subjects are asked to answer to the 60-items version of the HEXACO personality test (Ashton and Lee (2009)) [14].

[13] For details, refer to the instructions reported in the Appendix A.3

[14] More information available at: http://hexaco.org/

### 1.3.2 Experiment 2: Norms

The experiment is divided into four phases, see Table 1.2. The design of all but the second phase is identical across treatments.

| *Phase* | *Activity* | *T1: Cheating* | *T2: No Cheating* |
|---|---|---|---|
| 1st Phase | Slider Task | Trial 1: dominant hand<br>Trial 2: non dominant hand | Trial 1: dominant hand<br>Trial 2: non dominant hand |
| 2nd Phase | Norms Elicitation | Subjects are read aloud the instructions of<br>**EXP1 - Cheating**<br>Subjects rate acceptability of contribution choices | Subjects are read aloud the instructions of<br>**EXP1 - NO Cheating**<br>Subjects rate acceptability of contribution choices |
| 3rd Phase | Beliefs Elicitation | Subjects guess actual contribution choices | Subjects guess actual contribution choices |
| 4th Phase | Personality Test | Computer-based HEXACO test | Computer-based HEXACO test |

Table 1.2: **Timing of Experiment 2**

In the $1^{st}$ phase of this experiment, a group subjects who never did or will actually participate in Experiment 1, is asked to perform the slider task twice: first using their dominant hand and later using their non-dominant hand [15]. Subjects are not paid based on their performance and receive a fixed payment of 2 Euros for their participation in this phase.

The design of the $2^{nd}$ phase of the experiment heavily relies on the contributions by Krupka and Weber (2013) and Krupka et al. (2017), who designed an incentived experimental procedure to measure the 'injunctive social norms', defined as the jointly recognized beliefs on what members of a society would consider the right thing to do in a given context. Applications of this procedure to strategic contexts are very rare, an exception is represented by Gerxhani and Breemen (2019), who employ an adapted version of the norm-elicitation procedure originally designed by Krupka and Weber (2013) in a PGG context, and - to the best of my knowledge - has never been employed to study injunctive social norms on tax evasion.

[15] To make the difference between the use of the dominant vs. non-dominant rule more salient we implement the use of textile gloves as in the *No Cheating* treatment of Experiment 1.

In the $2^{nd}$ phase of the experiment, subjects, after having experienced themselves the slider task, are exposed to the same decision environment faced by participants of Experiment 1: they are exposed to the same instructions that are read aloud to subjects who take part in Experiment 1 and are later asked to rate the *"degree of social acceptability"* of each contribution choice available to subjects in the contribution-phase, knowing that under-contribution is made possible by the implementation of the paper-based double-anonymous procedure. For each level of earnings (20, 16, 12 8 and 4 ECUs), subjects are asked to rate the social acceptability of each and all contribution choices available, using a 6-points scale that ranges from "Very socially unacceptable" to "Very socially acceptable" (see Figure 1.8). Mirroring the between-subjects structure of Experiment 1, subjects are exposed to a single treatment only, and are therefore either exposed to the *Cheating* or to the *No Cheating* environment, with full information on the level of monitoring implemented during the slider task in their treatment.

| **Earnings = 4 ECUs ⇒ Expected contribution = 1** | | | | | | |
|---|---|---|---|---|---|---|
| | Very Socially Unacceptable | Socially Unacceptable | Somewhat Socially Unacceptable | Somewhat Socially Acceptable | Socially Acceptable | Very Socially Acceptable |
| C = 1 | □ | □ | □ | □ | □ | □ |
| C = 0 | □ | □ | □ | □ | □ | □ |

Figure 1.8: Social Acceptability Evaluation (Earnings = 4 scenario)

We provide respondents with incentives to match their ratings to the responses of other subjects in the session, rather than to provide their personal opinions. Subjects are asked to provide what they believe would be the "average answer" provided by other participants in the session and are informed that at the end of this phase one of the contribution choices they rated will be randomly selected and they will be randomly paired to another participant in the room: their payoff depends on how similar their response is to the answer provided by the matched participant, as shown by Figure 1.9, according to an incentive-compatible quadratic scoring rule [16].

[16] According to the quadratic scoring rule, the payoff corresponds to the maximum score possible $\alpha$ minus the inaccuracy of the forecast, computed as the sum of squared deviations: given subject's $i$ guess p on choice k, the payoff of subject $i$ will depend on how similar her guess will be from the guess reported by her randomly-selected partner j $P_i(p_{i,k}) = \alpha - \beta(p_{j,k} - p_{i,k})$, where $\alpha = 7$ and $\beta = 1/2$.

| **If the response..** | **The payoff is equal to:** |
|---|---|
| Exactly matches partner’s response | 7 Euros |
| Differs from partner’s response by 1 category | 6.5 Euros |
| Differs from partner’s response by 2 categories | 5 Euros |
| Differs from partner’s response by 3 categories | 2.5 Euros |
| Differs from partner’s response by 4 categories | - 1 Euro |
| Differs from partner’s response by 5 categories | - 5.5 Euros |

Figure 1.9: Financial incentives in the Norms’ Rating phase

In the $3^{rd}$ phase of the experiment, we elicit subjects’ beliefs on actual contribution behavior by subjects who actually took part in Experiment 1 in first person. We elicit beliefs on what would be the selected contribution option under all different earnings’ levels. The elicitation of beliefs is also incentivized. Subjects are informed that their guesses will be compared to the actual choice taken by a randomly-selected participant who took part in one of the previous sessions of Experiment 1: if their guess matches the randomly-selected actual choice, they receive a fixed prize of 3 Euros.

In the $4^{th}$ and last phase of the experiment subjects answer the 60-items version of the HEXACO personality test.

## 1.4 Results

It emerges that our manipulation is effective in altering subjects’ perception of the intensity of cheating across treatments. Both in Experiment 1 (Behaviors) and Experiment 2 (Norms) subjects have significantly different perceptions of the concentration of cheating occurrences in the pre-tax income-generating task across treatments (Figure 5): subjects perceive that violations of the non-dominant hand protocol for the execution of the income-generating task, which result in a form of cheating at the expenses of others, are more frequent in ‘Cheating’ treatments than in ‘No-Cheating’ treatments.

The distributions of subjects’ perceptions are significantly different in both Experiments, see Figure 1.10 the Rank-sum z statistic is equal to -9.346 (p-value 0.000) in Experiment 1 and to -3.987 (p-value 0.000) in Experiment 2. The data used to perform these tests count one observation per subject and all observations are independent since subjects’ perceptions of the intensity of cheating are elicited before subjects receive any feedback on their results from their strategic interactions with other subjects.

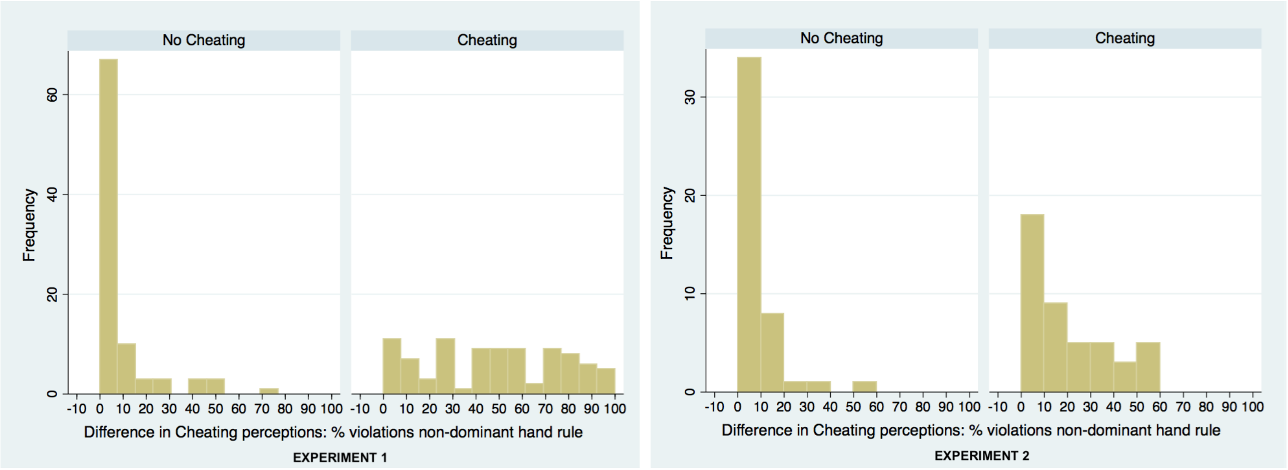
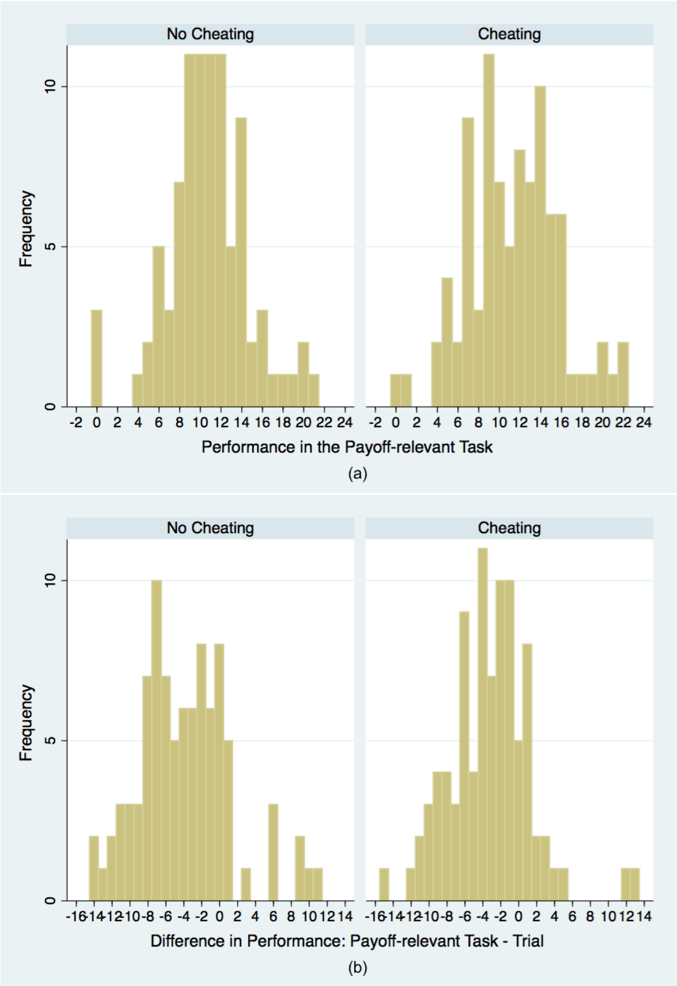


Figure 1.10: Manipulation check - Experiment 1 & 2

With respect to actual cheating occurrence, our design prevents us from detecting cheating at the individual level. If we look at the distribution of slider task's scores collected by subjects in the payoff-relevant round, we can observe that, on aggregate, subjects exposed to the ’Cheating’ treatment collect slightly higher scores but the two distributions are not statistically different (the Rank-sum z statistic is equal to -0.949 with a p-value=0.3427, see Figure 1.11, panel (a)). The same picture emerges if we control for subjects' individual ability in the task, looking at the difference in performance between the payoff-relevant round and the last trial round (both lasting 120 seconds). As predictable, the implementation of the non-dominant hand rule in the payoff-relevant round leads subjects to accrue, on average, lower scores with respect to the trial round: the difference in performance with respect to the trial round is lower in the ’Cheating’ treatment, however, the difference between the two distributions is not statistically significant (the Rank-sum z statistic is equal to -1.325 with a p-value=0.1851, see Figure 1.11 panel (b)). This evidence suggests that on average subjects followed the non-dominant hand restriction even in absence of monitoring, although the presence of a high degree of heterogeneity in subjects' learning in the task could also partially explain this attenuated result.

**Result 1**: *Our novel experimental procedure is effective in altering subjects' perceptions on the occurrence of cheating across treatments.*



(a)

(b)

Figure 1.11: Experiment 1 - Slider Task data

### 1.4.1 Experiment 1: Behaviors

With respect to tax evasion behaviors, our main conjecture is that in presence of an opaque income-generation process - where opportunities to cheat and maximize own revenues at the expenses of others are available - individuals would be more likely to engage in undercontribution behaviors. We are particularly interested in the behavior of individuals in the upper-middle part of the income distribution (those with pre-contribution earnings equal to 16 and 12 ECUs): these individuals are in the uncomfortable situation of getting smaller gains from redistribution compared to low-income individuals in the full-contribution scenario, and to be severely damaged in case of full or partial undercontribution by top-income individuals. These subjects risk to end up being "twice losers" in the Cheating treatment: being cheated twice by top-income individuals, who could first cheat in the income-generation task to boost their performance at their expenses and undeservedly get to the top position of the distribution, and then cheat again in the contribution phase, contributing less than required to the public account causing a distortion of the relative weight of the contribution burden on the shoulders of upper-middle income individuals.

Our analysis reveals there is no significant difference in subjects' aggregate contribution behavior across treatments (the Rank-sum z statistic is equal to 0.050, p-value 0.9599, see Figure 1.12 panel a and b), although our treatment manipulation proved to be effective and our test has a sufficient power to detect economically relevant effect sizes [17]. In both treatments we are clearly far from a full contribution scenario, with an average contribution equal to 59.5% of the amount due and a non-negligible quota of full under-contribution occurrences (slightly less than 30% of cases). If we look at subjects' contribution behavior, conditioning on pre-contribution earnings level, we observe some differences across earnings levels and a slight difference across treatments for top earners. However, a series of non-parametric Mann-Whitney tests reveal that contribution behavior is not statistically different across treatments even after we break down our observations by pre-contribution earnings' levels, see Figure 1.12 panel c [18].

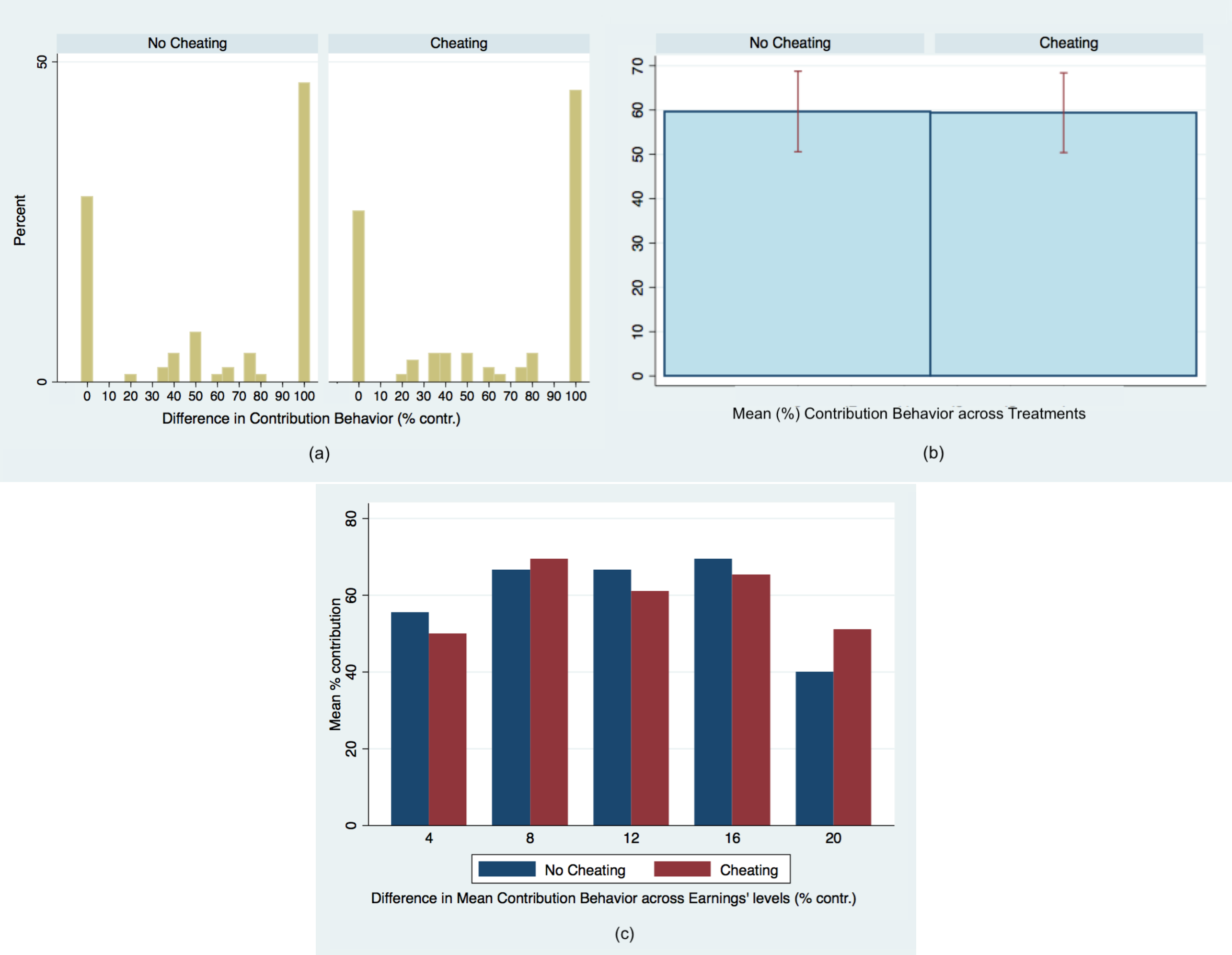
Pooling observations across treatments, a non-parametric Kruskal-Wallis equality-of-populations rank test reveals there's a marginal evidence in favor of heterogeneous distributions of contribution choices across pre-contribution earnings' levels ($\chi^2$ with ties=8.371, p-value=0.0789): a parametric analysis shows that this result is mainly driven by the contribution behavior of top-earners as this is the only group whose contribution behavior is, marginally, statistically lower [19].

As it emerges from Figure 1.12 (panel a), the distribution of percentage contribution

[17] See Appendix A.3 for further details on the power analysis.

[18] Two-sample Wilcoxon rank-sum (Mann-Whitney) tests: (a) Earnings=20, z=-0.860 p-value=0.3898; (b) Earnings=16, z=0.067 p-value=0.9463; (c) Earnings=12,(a) z=0.346 p-value=0.7295; (d) Earnings=8, z=-0.250 p-value=0.8025; (e) Earnings=4, z=0.329 p-value=0.7420

[19] Results of the estimation are reported in the Additional Tables section in Appendix A.2, Table **??**



(a)

(b)

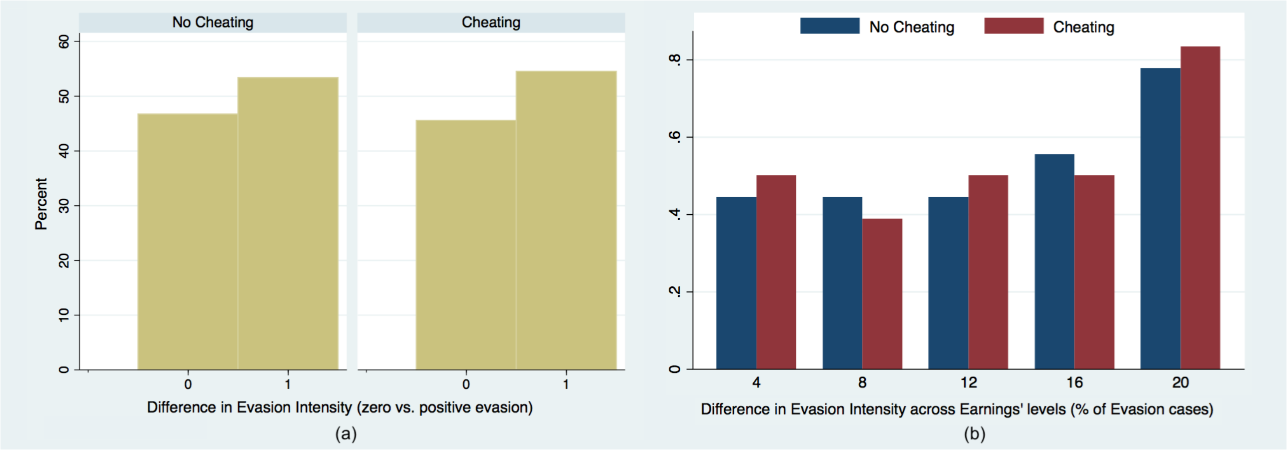
(c)

Figure 1.12: Experiment 1 - Contribution data (A)

choices has almost a bimodal shape with frequency peaks at 0% and 100% contribution choices, which correspond to the two extreme cases of evasion-to-the-full-extent and no-evasion [20]. We therefore check whether there's a significant difference in evasion intensity across treatments, looking at a dichotomous version of our variable of interest, which is coded as 0 when subjects fully contribute and as 1 when there's at least some undercontribution, irrespective of its magnitude: no significant difference is detected across treatments (the Pearson $\chi^2$ statistic is equal to 0.0224, p-value 0.881), with an average frequency of cheating equal to, respectively, the 53% and 54% of the cases in the *No Cheating* and *Cheating* treatments, see Figure 1.13, panel a. The result is stable even after we condition evasion frequency on

[20] This is partially due to the fact that subjects face different menus of available contribution choices based on what is their position in the pre-contribbution earnings' distribution, but in all cases subjects have the full-evasion and full-contribution choices available.

the pre-contribution earnings' level [21], see Figure 1.13, panel b.



(a)

(b)

Figure 1.13: Experiment 1 - Contribution data (B)

Since our design for Experiment 1, in order to ensure subjects' complete anonymity in the contribution task, doesn't allow us to observe cheating at the individual level, we are not in the position to disentangle the effects of (the perceptions of) *cheating by others* and of *own cheating* - which might induce some "moral cleansing" concerns - on contribution behavior. However, we can observe that perceptions of the intensity of cheating by others in the slider task are decreasing in the level of realized earnings (see Figure 1.14), and at the same time, we can reasonably assume that if cheating occurred, it more likely occurred at the top positions of the distribution rather than at the bottom. Looking at the results on subjects' contribution behavior, we can infer that any of these two effects - which would have called for higher(lower)-than-average contribution by top(bottom) earners - seems to be strong. The lower-than-average contribution observed for top earners, instead, is compatible with both: (a) a stronger entitlement effect driven by lower perceptions of cheating intensity, especially in the No-Cheating scenario, and (b) a mere "magnitude of stakes effect", given that top earners by design have the highest direct benefit from evasion in absolute terms.

**Result 2**: *Our treatment manipulation on the intensity of cheating does not lead to difference in subjects' contribution behavior across treatments.*

[21] Pearson $\chi^2$ tests: (a) Earnings 20, stat=0.1773 and p-value=0.674 (b) Earnings 16, stat=0.1115 and p-value=0.738 (c) Earnings 12, stat=0.1115 and p-value=0.738 (d) Earnings 8, stat=0.1143 and p-value=0.735 (e) Earnings 4, stat=0.1115 and p-value=0.738

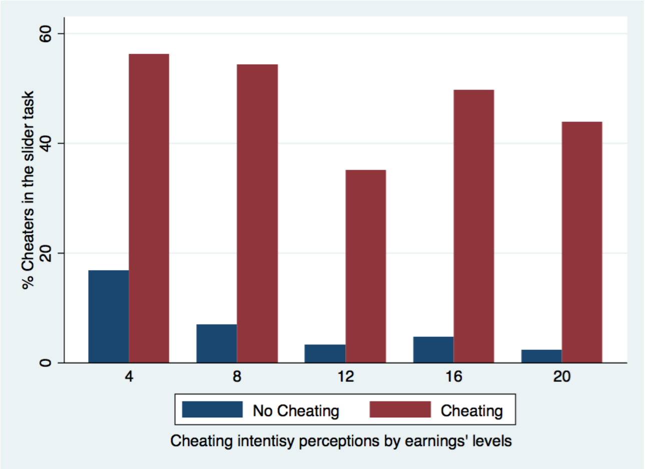


Figure 1.14: Experiment 1 - Cheating intensity perceptions

### 1.4.2 Experiment 2 - Social Norms

We explore the determinants of norms on evasion acceptability, investigating whether norms are sensitive to (i) the presence of cheating opportunities in the pre-tax income-generation process, (ii) the size of evasion and (iii) individual beliefs on the actual extent of under-contribution in each contribution context (empirical expectations).

We find that the general pattern of social acceptability ratings is the same across treatments, see Table 1.3. This is another interesting null result since our manipulation on the intensity of cheating proved to be effective also among Experiment 2 participants and our tests have a sufficient power to detect an economically relevant effect size [22]

Norms on tax evasion acceptability appear to be highly sensitive to the size of evasion: average acceptability ratings decrease as the size of evasion increases and the modal evaluations on the two "extreme" actions (full contribution and full under-contribution) are always polarized at the two respective extremes of the social acceptability spectrum: full-contribution is generally regarded as a "very socially acceptable" action while full under-contribution is recognized as a "very socially unacceptable" action. Ratings on the acceptability of "extreme" contribution actions, however, seem to be sensitive to the level of pre-contribution earnings' levels as well: full-contribution choices tend to be more highly rated when selected by high/top earners, while full under-contribution choices tend to be more highly rated when associated with low/bottom levels of pre-contribution earnings. This suggests that, overall, evasion by low-income earners is well tolerated and considered as somewhat socially accept-

[22] See Appendix A.3 for further details on the power analysis.

able irrespective of the characteristics of the process that generated income in the first place.

Table 1.3: Experiment 2 - Social acceptability ratings

| Action | No Cheating (n=45) | | | | | | | Cheating (n=45) | | | | | | | *Rank-sum z test* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Mean* | --- | -- | - | + | ++ | +++ | *Mean* | --- | -- | - | + | ++ | +++ | |
| **Earnings = 4 \| C=1** | *5.26* | 2.22% | 0.00% | 2.22% | 17.78% | 20.00% | 57.78% | *5.26* | 0.00% | 2.22% | 2.22% | 15.56% | 26.67% | 53.33% | *0.233* |
| **Earnings = 4 \| C=0** | *2.86* | 17.78% | 20.00% | 31.11% | 20.00% | 11.11% | 0.00% | *2.73* | 15.56% | 26.67% | 31.11% | 22.22% | 4.44% | 0.00% | *0.486* |
| **Earnings = 8 \| C=2** | *5.42* | 2.22% | 0.00% | 2.22% | 6.67% | 26.67% | 62.22% | *5.51* | 0.00% | 0.00% | 0.00% | 8.89% | 31.11% | 60.00% | *0.944* |
| **Earnings = 8 \| C=1** | *3.82* | 0.00% | 11.11% | 13.33% | 57.78% | 17.78% | 0.00% | *3.77* | 0.00% | 11.11% | 17.78% | 53.33% | 17.78% | 0.00% | *0.286* |
| **Earnings = 8 \| C=0** | *2.11* | 33.33% | 31.11% | 28.89% | 4.44% | 2.22% | 0.00% | *2.17* | 28.89% | 33.33% | 28.89% | 8.89% | 0.00% | 0.00% | *-0.410* |
| **Earnings = 12 \| C=3** | *5.44* | 0.00% | 2.22% | 2.22% | 4.44% | 31.11% | 60.00% | *5.55* | 0.00% | 0.00% | 0.00% | 4.44% | 35.56% | 60.00% | *-0.187* |
| **Earnings = 12 \| C=2** | *4.22* | 0.00% | 2.22% | 8.89% | 53.33% | 35.56% | 0.00% | *4.31* | 0.00% | 0.00% | 6.67% | 57.78% | 33.33% | 2.22% | *-0.350* |
| **Earnings = 12 \| C=1** | *2.8* | 2.22% | 33.33% | 46.67% | 17.78% | 0.00% | 0.01% | *2.84* | 0.00% | 33.33% | 51.11% | 13.33% | 2.22% | 0.00% | *-0.136* |
| **Earnings = 12 \| C=0** | *1.73* | 53.33% | 28.89% | 11.11% | 4.44% | 2.22% | 0.00% | *1.71* | 44.44% | 42.22% | 11.11% | 2.22% | 0.00% | 0.00% | *-0.419* |
| **Earnings = 16 \| C=4** | *5.4* | 0.00% | 4.44% | 2.22% | 4.44% | 26.67% | 62.22% | *5.51* | 0.00% | 0.00% | 0.00% | 8.89% | 31.11% | 60.00% | *0.051* |
| **Earnings = 16 \| C=3** | *4.31* | 0.00% | 4.44% | 8.89% | 37.78% | 48.89% | 0.00% | *4.48* | 0.00% | 0.00% | 2.22% | 46.67% | 51.11% | 0.00% | *-0.704* |
| **Earnings = 16 \| C=2** | *3.26* | 2.22% | 11.11% | 48.89% | 33.33% | 4.44% | 0.00% | *3.46* | 0.00% | 4.44% | 46.67% | 46.67% | 2.22% | 0.00% | *-1.252* |
| **Earnings = 16 \| C=1** | *2.33* | 13.33% | 48.89% | 28.89% | 8.89% | 0.00% | 0.00% | *2.42* | 0.00% | 66.67% | 26.67% | 4.44% | 2.22% | 0.00% | *-0.387* |
| **Earnings = 16 \| C=0** | *1.46* | 68.89% | 24.44% | 2.22% | 2.22% | | 2.22% | *1.51* | 62.22% | 28.89% | 4.44% | 4.44% | | 0.00% | *-0.656* |
| **Earnings = 20 \| C=5** | *5.64* | 0.00% | 0.00% | 6.67% | 0.00% | 15.56% | 77.78% | *5.64* | 0.00% | 0.00% | 0.00% | 6.67% | 22.22% | 71.11% | *0.620* |
| **Earnings = 20 \| C=4** | *4.64* | 0.00% | 2.22% | 4.44% | 22.22% | 68.89% | 2.22% | *4.75* | 0.00% | 0.00% | 2.22% | 22.22% | 73.33% | | *-0.554* |
| **Earnings = 20 \| C=3** | *3.73* | 2.22% | 4.44% | 20.00% | 64.44% | 8.89% | 0.00% | *3.97* | 0.00% | 2.22% | 8.89% | 77.78% | 11.11% | 0.00% | *-1.639* |
| **Earnings = 20 \| C=2** | *2.66* | 8.89% | 26.67% | 55.56% | 6.67% | 2.22% | 0.00% | *3.02* | 0.00% | 11.11% | 77.78% | 8.89% | 2.22% | 0.00% | *-2.413* ** |
| **Earnings = 20 \| C=1** | *1.97* | 22.22% | 64.44% | 8.89% | 2.22% | 2.22% | | *2.22* | 6.67% | 68.89% | 20.00% | 4.44% | 0.00% | 0.00% | *-2.106* ** |
| **Earnings = 20 \| C=0** | *1.35* | 80.00% | 13.33% | 2.22% | 2.22% | 0.00% | 2.22% | *1.35* | 75.56% | 15.56% | 6.67% | 2.22% | 0.00% | 0.00% | *-0.484* |

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01; all two-tailed.
Responses are: "Very socially acceptable" (+++), "Socially acceptable" (++), "Somewhat socially acceptable" (+), "Somewhat socially unacceptable" (-), "Socially unacceptable" (--), "Very socially unacceptable" (+++). Responses were converted into numerical scores: "Very socially acceptable" = 6, "Socially acceptable" = 5, "Somewhat socially acceptable" = 4, "Somewhat socially unacceptable = 3, "Socially unacceptable" = 2, "Very socially unacceptable" = 1. Modal responses are shaded.

These findings are confirmed by our parametric analysis, see Table 1.4: the size of evasion has a negative and sizeable significant effect on acceptability ratings, while the presence of the shadow of cheating does not lead to any significantly impact. It also emerges that the negative effect of evasion size is larger the higher the income level of the evader, although the interaction effect between income level and evasion size is not dramatic in size. These results are stable even after we control for individual beliefs on the actual level of contribution for each level of earnings, which appear to negatively affect per se acceptability norms: the higher subjects' beliefs on actual contribution behavior, the lower subjects' tolerance for evasion in those contexts. This result suggests that injucntive and descriptive social norms, in the tax evasion context, are not in conflict but positively related: when evasion is perceived as more widesprad (*descriptive norm*) the degree of social acceptability attached to evasion is higher (*injunctive norm*).

**Result 3**: *Norms on tax evasion acceptability are highly sensitive to the size of the evasion and respond to differences in pre-contribution levels of earnings but do not statistically differ across treatments.*

Table 1.4: Experiment 2 - Regressions on acceptability ratings

| | [1a] | [1b] | [2a] | [2b] |
|---|---|---|---|---|
| *Cheating* | -0.0640 | -0.0732 | -0.0886 | -0.0983 |
| | (0.225) | (0.229) | (0.227) | (0.231) |
| Size of Evasion | -2.271*** | -2.271*** | -2.271*** | -2.271*** |
| | (0.281) | (0.281) | (0.281) | (0.282) |
| Level of Earnings | 0.0103 | 0.0103 | 0.0198*** | 0.0200*** |
| | (0.00764) | (0.00766) | (0.00655) | (0.00662) |
| Size of Evasion x *Cheating* | -0.0543 | -0.0543 | -0.0543 | -0.0543 |
| | (0.291) | (0.292) | (0.292) | (0.292) |
| Level of Earnings x *Cheating* | 0.0129 | 0.0129 | 0.0208* | 0.0203* |
| | (0.00987) | (0.00988) | (0.0115) | (0.0115) |
| Size of Evasion x Level of Earnings | -0.107*** | -0.107*** | -0.107*** | -0.107*** |
| | (0.00995) | (0.00997) | (0.00996) | (0.00998) |
| Beliefs | | | -0.0637*** | -0.0652** |
| | | | (0.0244) | (0.0262) |
| Beliefs x *Cheating* | | | -0.0280 | -0.0250 |
| | | | (0.0508) | (0.0501) |
| Constant | 5.265*** | 5.276*** | 5.278*** | 5.304*** |
| | (0.183) | (0.205) | (0.185) | (0.208) |
| Observations | 1,800 | 1,800 | 1,800 | 1,800 |
| Number of id | 90 | 90 | 90 | 90 |
| Controls | No | Yes | No | Yes |

*Notes.* The dependent variable is the norm rating [1-6] for each contribution action available to subjects in all Earnings' scenarios. Controls include indicators on whether subjects show levels of Honesty-Humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, Opennes to Experience (as measured by the HEXACO questionnaire) above the sample median.
Robust standard errors in parentheses. Sign. levels: *** p<0.01, ** p<0.05, * p<0.1

### 1.4.3 Experiment 1 & 2: Tax evasion behavior and norms

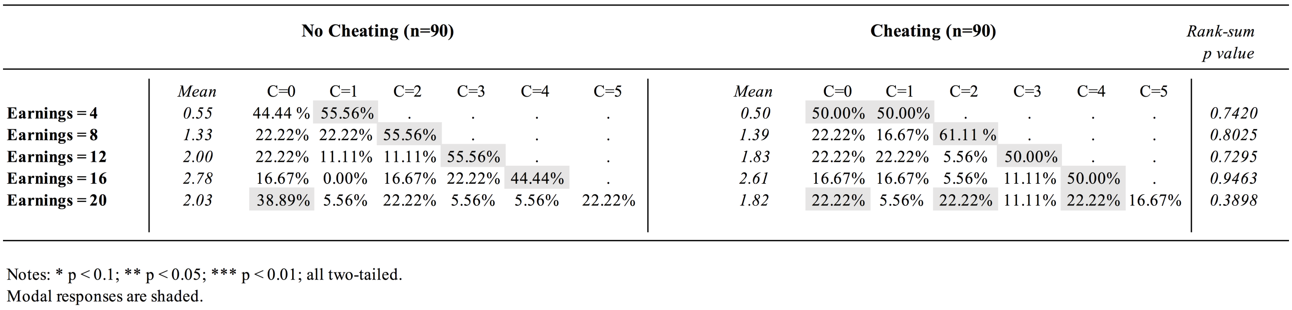
Can social norms on evasion predict subjects' behavior in different tax contribution contexts? Our empirical results show that neither norms nor behaviors, which are elicited on two different groups of subjects, are deeply affected by the characteristics of the process that generates subjects' incomes, and thus, by the origin of economic inequality within the two groups. It emerges, however, that social acceptability ratings are fairly sensitive to some of the characteristics of evasive actions, such as the size of evasion, and, to a smaller extent, the level of evaders' earnings.

Table 1.5 shows the distribution of actions selected by subjects who took part in Experiment 1 (Behaviors), conditioning on subjects' pre-contribution earnings' level: the pattern of modal actions across all income levels suggests the presence of a tension between immediate monetary gains from evasion on one side, and norms' adherence on the other side. Immediate monetary gains from evasion are obviously higher the lower the individual contribution to the tax-collection pool, and reach their peak with full under-contribution choices; on the side of norms, instead, we can see that higher contribution choices are always associated to higher acceptability ratings.

The effect of norms seems to be particularly relevant for subjects who are not at the top of the income distribution: the modal contribution action for these subjects always corresponds to the action regarded as the most socially acceptable. The same is not true for top-earners, whose modal action, although different across treatments, seems to be driven more intensely by immediate monetary benefits rather than by the desire to comply with social prescriptions on what is considered to be acceptable.

This evidence could be consistent with the hypothesis that acceptability norms play a role in subjects' decision process, with subjects exhibiting a desire for norms' compliance on top of their interest for monetary payoff maximization.

Table 1.5: Experiment 1 - Contribution choices data



| | No Cheating (n=90) | | | | | | | Cheating (n=90) | | | | | | | *Rank-sum p value* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Mean* | C=0 | C=1 | C=2 | C=3 | C=4 | C=5 | *Mean* | C=0 | C=1 | C=2 | C=3 | C=4 | C=5 | |
| **Earnings = 4** | *0.55* | 44.44 % | 55.56% | . | . | . | . | *0.50* | 50.00% | 50.00% | . | . | . | . | *0.7420* |
| **Earnings = 8** | *1.33* | 22.22% | 22.22% | 55.56% | . | . | . | *1.39* | 22.22% | 16.67% | 61.11 % | . | . | . | *0.8025* |
| **Earnings = 12** | *2.00* | 22.22% | 11.11% | 11.11% | 55.56% | . | . | *1.83* | 22.22% | 22.22% | 5.56% | 50.00% | . | . | *0.7295* |
| **Earnings = 16** | *2.78* | 16.67% | 0.00% | 16.67% | 22.22% | 44.44% | . | *2.61* | 16.67% | 16.67% | 5.56% | 11.11% | 50.00% | . | *0.9463* |
| **Earnings = 20** | *2.03* | 38.89% | 5.56% | 22.22% | 5.56% | 5.56% | 22.22% | *1.82* | 22.22% | 5.56% | 22.22% | 11.11% | 22.22% | 16.67% | *0.3898* |

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01; all two-tailed.
Modal responses are shaded.

Following Krupka and Weber (2013) and Krupka et al. (2017), we examine whether subjects' choices are guided by a desire for norms' compliance by fitting individual utility functions to choice data collected through Experiment 1, while relying on the ratings on the social acceptability of evasion that have been separately identified through Experiment 2. We assume that subjects, when facing the contribution decision, follow a logistic choice rule, where the likelihood of selecting any action $a$ out of $n$ alternatives depends on the relative utility attached to that action, compared to the other alternatives available:

$$P(a = a_k) = \frac{exp(U_k)}{\sum_{i=1}^{n} exp(U_i)} \tag{1.1}$$

We model two different utility specifications: the first specification assumes subjects' utility depends only on their monetary payoff (Selfish model - Eq. 2) [23] and allows us to estimate the weight subjects place on the monetary benefit ($mp$) they get from a particular choice ($\beta$). The second specification (Norms model - Eq. 3) assumes subjects care both about the monetary payoff and the degree to which a particular choice is perceived to be social acceptable, where the social norm $N(a_k)$ is the empirically observed judgement on social acceptability, which should reflect the existing norm in the relevant group; this specification allows us to estimate the degree to which the subjects really care about adhering to a particular norm ($\gamma$).

$$U_i(mp, a_k) = \beta \cdot mp(a_i, k) \tag{1.2}$$

$$U_i(mp, a_k) = \beta \cdot mp(a_i, k) + \gamma \cdot N(a_k) \tag{1.3}$$

The estimation is obtained through a conditional logit regression (McFadden (1974)). In the model, the dependent variable is a binary indicator that identifies whether a particular choice, out of all those available in each subject's menu, is selected. The independent variables capture the characteristics of the possible contribution choices:

- each choice's immediate monetary payoff, which depends on the pre-contribution income level and the contribution choice selected;

- and each choice's degree of social acceptability, obtained as the average social acceptability rating attached to each alternative ($N(a_k)$).

[23] We impose a linear restriction on the effect of the monetary payoff on utility.

Table 1.6 reports the estimation results: the coefficient on monetary payoffs in the Selfish model in column (1) is negative and statistically significant, which would imply that the higher the monetary payoffs the lower the probability that subjects will select that specific action. This result is driven by the negative correlation between monetary payoffs and acceptability ratings, which are omitted from the Selfish model. Once we include also social norms as an explanatory variable in the model, the coefficient on monetary payoff turns positive, although not significant, as shown by the Norms model in column (3), while the coefficient on social appropriateness is positive and slightly significant. Norms appear to be the only relevant driver of subjects' choices and moving from the Selfish to the Norms specification results in an improvement in in terms of model's predictive fit. A further improvement is reached when estimating a model where subjects' utility depends only on choices' social appropriateness and subjects' utility is non-linear in the degree of choices' social appropriateness (column 4). To get a sense of how well the different models can qualitatively account for the choice data collected through Experiment 1, we graphically compare the predictive performance of the models with actual choice frequencies observed, see Figure **??**. Except for the Earnings=4 scenario, where the Selfish model outperforms the others predicting almost an equal split between the two available options – as it is the case in the data, in all other scenarios the models that account for the role of norms clearly outperform the Selfish model.

We further test how a model where individuals gets a positive utility from conforming to *descriptive* norms ($N^d(a_k)$) on tax contributions, rather than from abiding the existing *injunctive* norms, performs in predicting subjects' actual contribution choices observed in Experiment 2. We estimate this alternative Norms model (Eq. 4) by estimating the weight attached to conformity to the descriptive norm $\gamma^d$, which is measured as the distance of each possible contribution choice from what would be considered the 'expected' contribution choice under that Earnings' scenario, that is measured as the average Belief on actual contributions elicited in Experiment 2: this distance ranges from negative to positive values where negative values indicate that the choice considered is below the 'expected' contribution, while positive values indicate that the choice considered is above the 'expected' contribution with a null value indicating perfect adherence.

$$U_i(mp, a_k) = \beta mp(a_i, k) + \gamma^d N^d(a_k) \tag{1.4}$$

Although we are not in the position to test and distinguish the relative weight attached to each of the two types of norms due to the high correlation between the two measures in our data (Pearson's correlation coefficient: 0,9518), our estimates suggets that when taken separately, injunctive norms can account for subjects contribution choices better than descriptive norms.

**Result 4**: *Social norms on the acceptability of evasion elicited in Experiment 2 are relevant predictors for subjects' actual choices observed in Experiment 1.*

## 1.5 Conclusion

This study tests the presence of the shadow of cheating on the process that generates incomes in a society as potential motivation for tax evasion.

We find that the shadow of cheating has no direct first-order causal effect neither on tax evasion behavior nor on the social acceptability of evasion, with only a weak second-order effect on top-earners. Given the effectiveness of our manipulation in altering subjects' perceptions on the presence of cheating across treatments, and the sufficient power, this is an interesting null result, which suggest that the correlational evidence on the relationship between the intensity of cheating and tax evasion does not match a causal relationship.

Our results on the factors shaping injunctive social norms on evasion acceptability, given the relevance this type of norms appear to have on subjects behavior offer some interesting insights, which could also be relevant from a policy perspective, although elicited in a simplified framework with a proportional tax system and limited stakes: if moderate amounts of evasion by low-income earners are largely socially justified, we could have that the presence of inequality itself in a society, keeping fixed the overall amount of production/wealth, has a depressive effect on tax revenues.

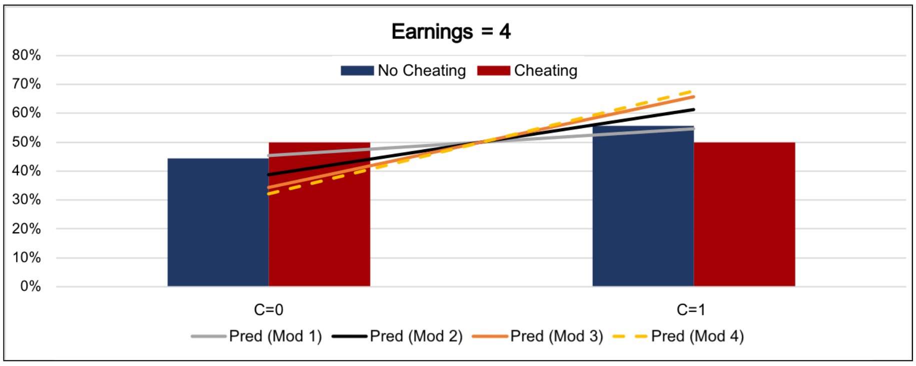
It would be interesting to further study and test whether our results are stable if we alter the

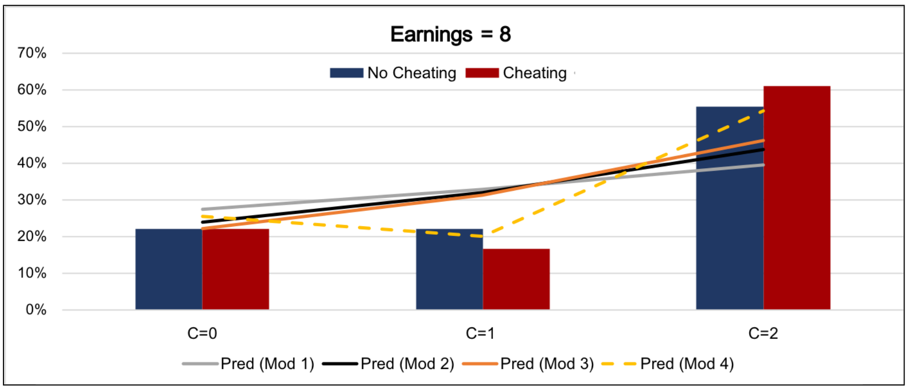
Table 1.6: Experiment 1 & 2 - Conditional Logit estimation

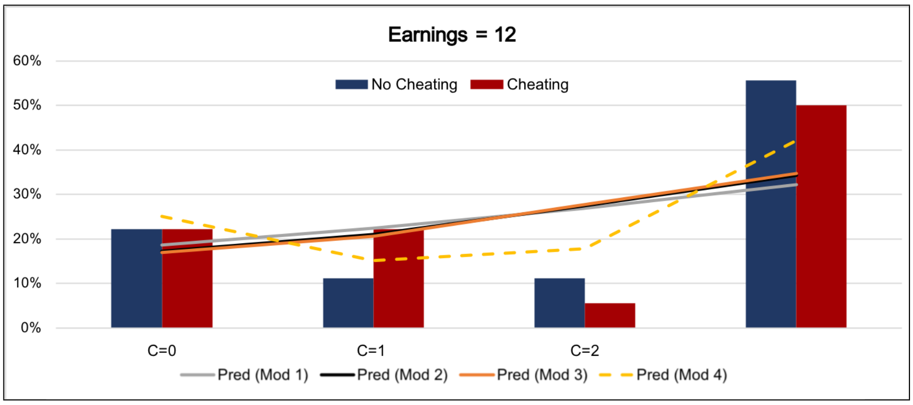
| | [1] | [2] | [3] | [4] | [5] | [6] |
|---|---|---|---|---|---|---|
| *Monetary Benefit* | -0.882** | | 1.046 | | | |
| | (0.360) | | (1.143) | | | |
| *Norm [Injunctive]* | | 0.261*** | 0.515* | 0.171** | | |
| | | (0.0874) | (0.288) | (0.0741) | | |
| $Norm\ [Injunctive]^2$ | | | | 0.426*** | | |
| | | | | (0.106) | | |
| *Norm Distance [Descriptive]* | | | | | 0.249** | 0.246*** |
| | | | | | (0.101) | (0.0817) |
| $Norm\ Distance\ [Descriptive]^2$ | | | | | | 0.330*** |
| | | | | | | (0.0879) |
| Observations | 720 | 720 | 720 | 720 | 720 | 720 |
| Log Likelihood | -232.7 | -231.2 | -230.6 | -222.8 | -232.7 | -225.3 |
| BIC | 471.93 | 468.95 | 474.45 | 458.67 | 471.93 | 463.83 |
| Pseudo R2 | 0.0176 | 0.0239 | 0.0262 | 0.0595 | 0.0176 | 0.0486 |

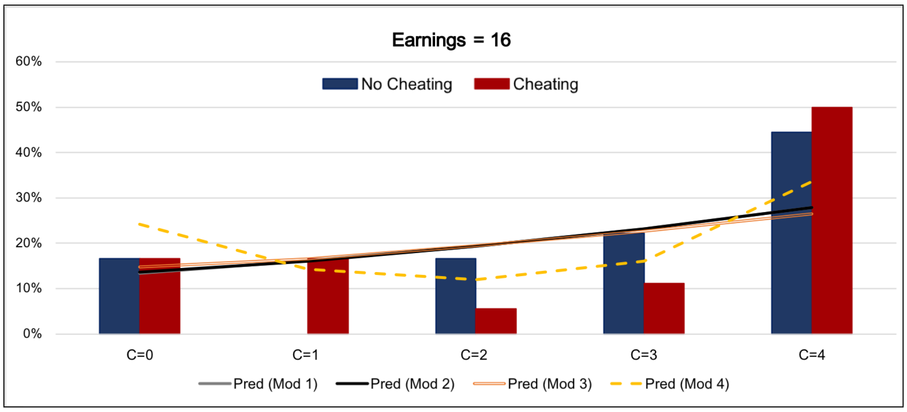
*Notes.* The Dependent variable is the chosen contribution action in the PGG (Experiment 1). The variable *Norm [Injunctive]* converts subjects' responses (Experiment 2) to numerical scores: 1 ="very socially unacceptable", 2 ="socially acceptable", 3 ="somewhat socially acceptable", 4 ="somewhat socially acceptable", 5 ="socially acceptable" and 6 ="very socially acceptable". The variable *Norm Distance [Descriptive]* measures how distant the contribution choice is from the descriptive norm, defined as the average belief on actual contributions (Experiment 2): this distance takes negative/positive values if the contribution choice is lower/higher than what is expected to be the most common option under that Earnings' scenario. To account for differences in scales, we use a standardized version of the explanatory variables. Standard errors are clustered at the subject level and reported in parentheses.
Sign. levels: *** p<0.01, ** p<0.05, * p<0.1

level of progressivity of the tax system or the severity of income inequality, or if we introduce cheating "with certainty", as to compare two scenarios where cheating is either impossible or possible and fully exploited in 100% of the cases by top-earners when available.


Earnings = 4


Earnings = 8


Earnings = 12


Earnings = 16

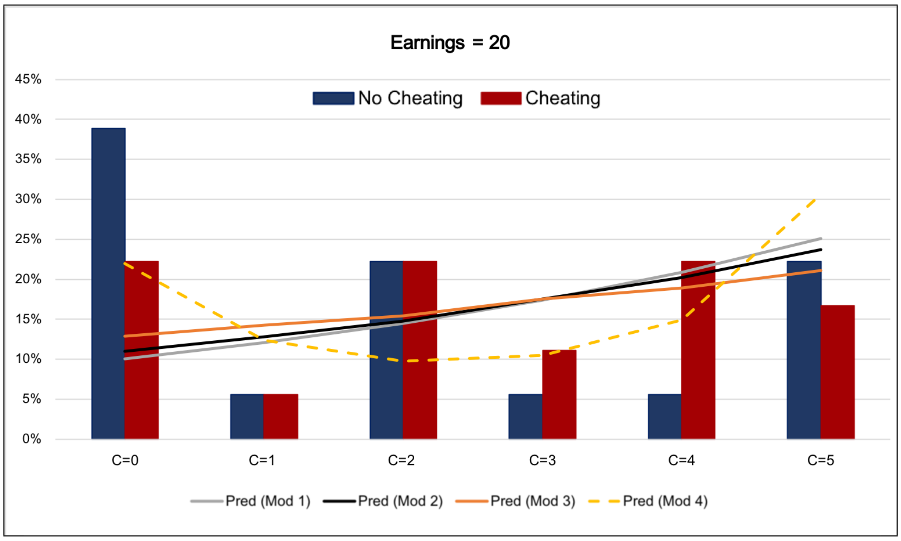


Figure 1.15: Distributions of predicted and observed choices in Experiment 1

# Bibliography

Alesina, A. and Angeletos, G.-M. (2005). Fairness and redistribution. *American Economic Review*, 95(4):960–980.

Allingham, M. G. and Sandmo, A. (1972). Income tax evasion: a theoretical analysis. *Journal of Public Economics*, 1(3-4):323–338.

Alm, J., Jackson, B., and McKee, M. (1992). Estimating the determinants of taxpayer compliance with experimental data. *National Tax Journal*, pages 107–114.

Alm, J., Jackson, B. R., and McKee, M. (1993). Fiscal exchange, collective decision institutions, and tax compliance. *Journal of Economic Behavior & Organization*, 22(3):285–303.

Alstadsaeter, A., Johannesen, N., and Zucman, G. (2019). Tax evasion and inequality. *American Economic Review*, 109(6):2073–2103.

Alvaredo, F., Chancel, L., Piketty, T., Saez, E., and Zucman, G. (2018). World inequality report. *WDI*.

Ashton, M. C. and Lee, K. (2009). The hexaco-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, 91:340–345.

Banerjee, R. (2016). On the interpretation of bribery in a laboratory corruption game: moral frames and social norms. *Experimental Economics*, 19(1):240–267.

Barmettler, F., Fehr, E., and Zehnder, C. (2012). Big experimenter is watching you! anonymity and prosocial behavior in the laboratory. *Games and Economic Behavior*, 75(1):17–34.

Barr, A., Lane, T., and Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, 164:153–164.

Benedek, D. and Lelkes, O. (2011). The distributional implications of income under-reporting in hungary. *Fiscal Studies*, 32(4):539–560.

Bishop, J. A., Formby, J. P., and Lambert, P. (2000). Redistribution through the income tax: The vertical and horizontal effects of noncompliance and tax evasion. *Public Finance Review*, 28(4).

Bordignon, M. (1993). A fairness approach to income tax evasion. *Journal of Public Economics*, 52(3):345–362.

Bortolotti, S., Soraperra, I., Sutter, M., and Zoller, C. (2017). Too lucky to be true: Fairness views under the shadow of cheating. *IZA DP No. 10877.*

Burks, S. V. and Krupka, E. L. (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Science*, 58(1):203–217.

Cappelen, A. W., Cappelen, C., and Tungodden, B. (2018). Second-best fairness under limited information: The trade-off between false positives and false negatives. *NHH Dept. of Economics Discussion Paper N.18.*

Cappelen, A. W., Sorensen, E. O., and Tungodden, B. (2010). Responsibility for what? fairness and individual responsibility. *European Econoha believe believingtt ter being informed mic Review*, 54:429–441.

Doerrenberg, P. and Peichl, A. (2018). Tax morale and the role of so- cial norms and reciprocity. evidence from a randomized survey experiment. *CESifo Working Papers*, 7149.

Durante, R., Putterman, L., and Weele, J. (2014). Preferences for redistribution and perception of fairness: An experimental study. *Journal of the European Economic Association*, 12(4):1059–1086.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Fortin, B., Lacroix, G., and Villeval, M. C. (2007). Tax evasion and social interactions. *Journal of Public Economics*, 91(11):2089–2112.

Freeman, R. B. and Gelber, A. M. (2010). Prize structure and information in tournaments: Experimental evidence. *American Economic Journal: Applied Economics*, 2(1):149–164.

Gächter, S., Gerhards, L., and Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, 97:72–86.

Gächter, S., Nosenzo, D., and Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences?. *Journal of the European Economic Association*, 11(3):548–573.

Gerxhani, K. and Breemen, J. V. (2019). Social values and institutional change: an experimental study. *Journal of Institutional Economics*, 15(2):259–280.

Gill, D. and Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American economic review*, 102(1):469–503.

Greiner, B. (2004). The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics*, 10(23):63–104.

Houser, D., Vetter, S., and Winter, J. (2012). Fairness and cheating. *European Economic Review*, 56(8):1645–1655.

Kajackite, A. (2018). Lying about luck versus lying about performance. *Journal of Economic Behavior & Organization*, 153:194–99.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A meeting of the minds: Informal agreements and social norms. *Management Science*, 63(6):1657–2048.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

List, J. A., Sadoff, S., and Wagner, M. (2011). So you want to run an experiment, now what? some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14:439–457.

Luttmer, E. F. P. and Singhal, M. (2014). Tax morale. *Journal of Economic Perspectives*, 28(4):149–168.

Mascagni, G. (2017). From the lab to the field: a review of tax experiments. *Journal of Economic Surveys*, 32:273–301.

Matsaganis, M. and Flevotomou, M. (2010). Distributional implications of tax evasion in greece. *LSE GreeSE Paper*, 31.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Zarembka P, ed. Frontiers in Econometrics (Academic Press, New York)*, page 105–142.

Moldovanu, B. and Sela, A. (2001). The optimal allocation of prizes in contests. *American Economic Review*, 91(3):542–558.

Nygard, O. E., Slemrod, J., and Thoresen, T. O. (2018). Distributional implications of joint tax evasion. *The Economic Journal*, 129(620):1894–1923.

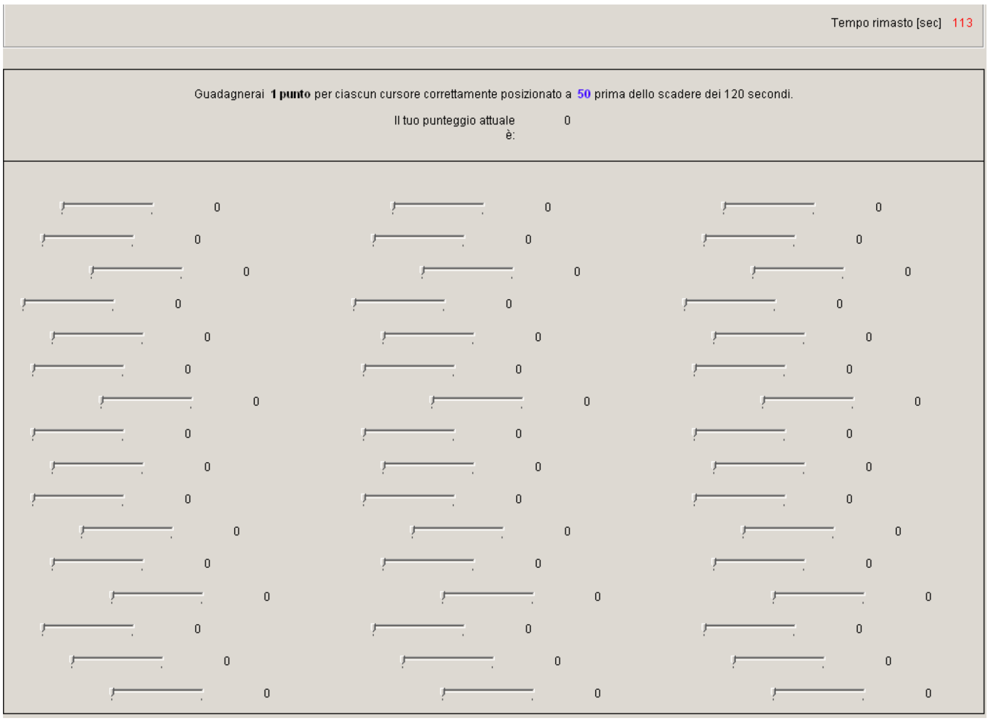
Slemrod, J. and Johns, A. (2010). The distribution of income tax noncompliance. *National Tax Journal*, 63(3):397.

Spicer, M. W. and Becker, L. A. (1980). Fiscal inequity and tax evasion: An experimental approach. *National Tax Journal*, pages 171–175.

Spicer, M. W. and Lundstedt, S. B. (1976). Understanding tax evasion. *Public Finance*, 31(2):295–305.

## 1.6 Appendix

### 1.A.1 Additional Figures



Tempo rimasto [sec] 113

Guadagnerai **1 punto** per ciascun cursore correttamente posizionato a 50 prima dello scadere dei 120 secondi.

Il tuo punteggio attuale è: 0

Figure A1.1: Slider Task screen

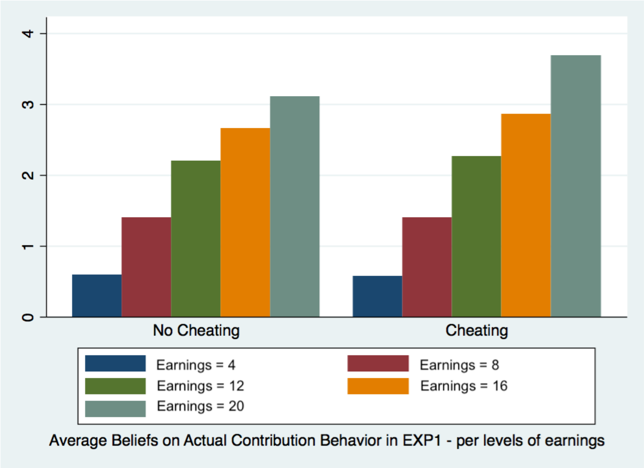


Figure A1.2: Average Beliefs elicited in EXP2 on Actual Contribution behavior in EXP1

### 1.A.2 Additional Tables

Table A2.1: Experiment 1: Contribution choices

| | [1] | [2] | [3] |
|---|---|---|---|
| *Cheating* | -5.556 | -0.278 | |
| | (14.36) | (6.368) | |
| Earnings=4 | -11.11 | -11.11 | -11.11 |
| | (14.36) | (10.07) | (10.04) |
| Earnings=8 | 4.24e-07 | 4.167 | 4.167 |
| | (14.36) | (10.07) | (10.04) |
| Earnings=16 | 2.778 | 3.472 | 3.472 |
| | (14.36) | (10.07) | (10.04) |
| Earnings=20 | -26.67* | -18.33* | -18.33* |
| | (14.36) | (10.07) | (10.04) |
| Earnings=4 x *Cheating* | -0 | | |
| | (20.31) | | |
| Earnings=8 x *Cheating* | 8.333 | | |
| | (20.31) | | |
| Earnings=16 x *Cheating* | 1.389 | | |
| | (20.31) | | |
| Earnings=20 x *Cheating* | 16.67 | | |
| | (20.31) | | |
| Constant | 66.67*** | 64.03*** | 63.89*** |
| | (10.16) | (7.799) | (7.099) |
| Observations | 180 | 180 | 180 |
| R-squared | 0.0485 | 0.0428 | 0.0428 |

*Notes.* OLS Regression: the dependent variable is individual contribution (expressed as a percentage of due contribution, equal to 25% of realized earnings). Std errors in parentheses; Sign. levels:*** p<0.01,** p<0.05,* p<0.1

Table A2.2: Beliefs on Actual Contribution Behavior in No-Cheating vs. Cheating treatments

| | No Cheating (n=90) | | | | | | | Cheating (n=90) | | | | | | | *Rank-sum p value* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Mean* | C=0 | C=1 | C=2 | C=3 | C=4 | C=5 | *Mean* | C=0 | C=1 | C=2 | C=3 | C=4 | C=5 | |
| **Earnings = 4** | *0.6* | 40% | 60% | . | . | . | . | *0.58* | 42.22% | 57.78% | . | . | . | . | *0.8313* |
| **Earnings = 8** | *1.4* | 2.22% | 55.56% | 42.22% | . | . | . | *1.4* | 11.11% | 37.78% | 51.11% | . | . | . | *0.7448* |
| **Earnings = 12** | *2.2* | 2.22% | 6.67% | 60% | 31.11% | . | . | *2.26* | 4.44% | 4.44% | 51.11% | 40% | . | . | *0.4658* |
| **Earnings = 16** | *2.67* | 8.89% | 4.44% | 24.44 % | 35.56 % | 26.67% | . | *2.87* | 8.89% | 2.22% | 17.78% | 35.56% | 35.56% | . | *0.3257* |
| **Earnings = 20** | *3.11* | 15.56 | 6.67% | 2.22% | 24.44% | 28.89% | 22.22% | *3.69* | 4.44% | 0.00% | 6.67% | 33.33% | 22.22% | 33.33% | *0.1863* |

Notes: * p < 0.1; ** p < 0.05; *** p < 0.01; all two-tailed.
Modal responses are shaded.

### 1.A.3 Power Analysis

After having run the first pilot sessions in June 2018 we conducted an ex-ante power analysis calculation in order to determine the sample size needed for both Experiment 1 and Experiment 2, observing that the size of the standard deviations of our outputs of interest ($Y_{EXP1}$: measured both in terms of percentage contributions and presence/absence of cheating; $Y_{EXP2}$: norm ratings, ranging from 1 to 6) was comparable in size across treatments ($\sigma^2 = \sigma^2_C = \sigma^2_{NC}$).

We can now provide ex-post power calculations for the minimum detectable effect (MDE) size, following List et al. (2011)'s approach, setting stardard levels for the significance level $\alpha = 0.05$ and the power of the test $1 - \beta = 0.8$ (which gives us $t_{\alpha/2} = 1.96$ and $t_{\alpha/2} = 0.84$ from standard normal tables. In Experiment 1 we would be able to detect a difference of approximately 18 percentage points in terms of percentage contributions, which corresponds to a 0.42 standard deviations of the mean percentage contribution, or a difference of 20 percentage points if refer to the dichotomous measure of our variable of interest, which measures intensity of evasion occurrencies irrespective of the size of evasion. In Experiment 2, we would be able to detect a difference equal to 0.59 standard deviations of the mean norm rating, which corresponds to a difference that ranges between 0.4 and 0.7 points in acceptability ratings.

| | Mean | SD | CHEATING | | NO CHEATING | | MDE |
|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD | |
| $Y_{EXP1} : Perc.\ Contribution$ | 59.53 | 43.05 | 59.39 | 42.95 | 59.67 | 43.38 | 18 |
| $Y_{EXP1} : Evasion\ Intensity$ | 53.89 | 49.98 | 54.44 | 50.08 | 53.33 | 50.16 | 20 |
| $Y_{EXP2} : NormRating$ | | | | | | | |
| $Earnings = 4\,, C = 1$ | 5.27 | 1.01 | 5.27 | .96 | 5.267 | 1.075 | 0.60 |
| $Earnings = 4\,, C = 0$ | 2.8 | 1.18 | 2.73 | 1.12 | 2.87 | 1.25 | 0.70 |
| $Earnings = 8\,, C = 2$ | 5.47 | .84 | 5.51 | .66 | 5.42 | .99 | 0.50 |
| $Earnings = 8\,, C = 1$ | 3.8 | .86 | 3.78 | .88 | 3.82 | .86 | 0.51 |
| $Earnings = 8\,, C = 0$ | 2.14 | .98 | 2.18 | .96 | 2.11 | 1.01 | 0.58 |
| $Earnings = 12\,, C = 3$ | 5.5 | .74 | 5.56 | .59 | 5.44 | .87 | 0.44 |
| $Earnings = 12\,, C = 2$ | 4.27 | .67 | 4.31 | .63 | 4.22 | .70 | 0.40 |
| $Earnings = 12\,, C = 1$ | 2.82 | .74 | 2.84 | .74 | 2.8 | .76 | 0.44 |
| $Earnings = 12\,, C = 0$ | 1.72 | .87 | 1.71 | .76 | 1.73 | .99 | 0.51 |
| $Earnings = 16\,, C = 4$ | 5.46 | .85 | 5.51 | .66 | 5.4 | 1.01 | 0.5 |
| $Earnings = 16\,, C = 3$ | 4.4 | .70 | 4.49 | .55 | 4.31 | .82 | 0.41 |
| $Earnings = 16\,, C = 2$ | 3.37 | .73 | 3.47 | .63 | 3.27 | .81 | 0.44 |
| $Earnings = 16\,, C = 1$ | 2.38 | .76 | 2.42 | .69 | 2.33 | .83 | 0.45 |
| $Earnings = 16\,, C = 0$ | 1.49 | .86 | 1.51 | .79 | 1.47 | .94 | 0.51 |
| $Earnings = 20\,, C = 5$ | 5.64 | .71 | 5.64 | .608 | 5.64 | .80 | 0.42 |
| $Earnings = 20\,, C = 4$ | 4.7 | .63 | 4.76 | .53 | 4.64 | .71 | 0.37 |
| $Earnings = 20\,, C = 3$ | 3.86 | .68 | 3.98 | .54 | 3.73 | .78 | 0.40 |
| $Earnings = 20\,, C = 2$ | 2.84 | .72 | 3.02 | .54 | 2.67 | .83 | 0.43 |
| $Earnings = 20\,, C = 1$ | 2.1 | .72 | 2.22 | .64 | 1.98 | .78 | 0.43 |
| $Earnings = 20\,, C = 0$ | 1.36 | .83 | 1.36 | .71 | 1.36 | .93 | 0.49 |

# C. Instructions

We provide an English translation (from Italian) of the Instructions used for the two experiments: Experiment 1 on Behaviors and Experiment 2 on Norms.

The text in black is the text shared by the instructions used for both treatments *Cheating* and *No Cheating*. The parts of the text in blue identify those parts that were only present in the instructions for the *No Cheating* treatment.

# EXPERIMENT 1 (BEHAVIORS)

Welcome.

You are about to participate in a study on how economic decisions are made.
All the decisions you will make during the study will be completely anonymous.

### Personal Identification Code (ID)

When you entered the lab you were asked to extract a green envelope from a container. All envelopes look the same. Each green envelope contains three small stickers with a three-digits number. Each envelope contains a different number. This number represents your personal identification code (ID), which will be used to encode your choices anonymously and to calculate your final profits. Since you randomly extracted your envelope, neither the experimenter nor the other participants know your ID. Please don't open your green envelope yet. We will tell you later, when to open the envelope.

### Duration of the study and Payments

This study is divided into **three parts**.
You will receive **3 Euros** for your participation in this study.
In addition, you will have the opportunity to earn more money and the final amount of your earnings will depend on the choices you and other participants make in the first two parts of the study (Part 1 and Part 2) and on your participation in the third part of the study (Part 3).
You will be paid privately (in a sealed envelope) and in cash at the end of the study.
Please turn off your cell phone. Any form of communication with other participants is strictly prohibited. If you violate this rule you will be excluded from all payments.
You will now receive instructions for Part 1. You will receive further instructions at the beginning of each of the following parts.

As we read the instructions we will ask you to answer some questions to verify your understanding of the instructions. There will also be several breaks in which you can ask questions: if you have a question, raise your hand and we will answer you private.

# INSTRUCTIONS FOR PART 1

### Groups

At the beginning of this part of the study you will be randomly matched to four other participants in the room.
I gruppi rimarranno fissi fino alla fine della Parte 2.
The groups will be fixed until the end of Part 2.

### Your activity

In this part of the study you and all other participants will face the same activity. Your earnings will depend on both your performance and the performance of other participants in your group.

We will repeat the activity three times: the first two times will be trial rounds for which you will not be paid, the third time will be the one determinant for your payment.

The activity consists of a screen with **48 sliders**, each of which is initially set to 0 and can be moved up to 100. Your **"score"** in the activity will be given by the number of sliders that you will be able to position exactly at 50 before the **120 seconds** are over.
Each slider has a number to its right that shows its current position, you can use the touch pad to move each slider and adjust its position as many times as you wish.

Your earnings in this phase depend on the number of sliders that you will be able to position correctly, and on the number of sliders correctly positioned by the other four members of your group.
Your earnings are expressed in tokens. At the end of the study, the tokens will be converted into Euros: each token is worth the **0.5 Euro** (fifty cents).

If you achieved the **best "score"** in your group, you will receive 20 tokens.
If you achieved the **second-best "score"** in your group, you will receive 16 tokens.
If you achieved the **third-best "score"** in your group, you will receive 12 tokens.
If you achieved the **fourth-best "score"** in your group, you will receive 8 tokens.
If you achieved the **fifth-best "score"** in your group, you will receive 4 tokens.
In the event of a tie, a random draw will determine the position of the two participants with the same score in the final ranking.

There will be two trial sessions.

The first trial session will have a longer duration (200 seconds), to allow you to practice with the activity. The second trial session will last exactly 120 seconds, just like the third session for which you will be paid.

Are there any questions?
Otherwise, click on the OK button and the first of two trial sessions will start shortly.

[ *Trial Sessions* ]

We will now proceed with the third session, which will determine your earnings for this part of the study.

To perform the activity, from now on, you will allowed to use exclusively your **non-dominant hand**, that is the one you do NOT normally use to write, sign documents, use the computer mouse, etc.

We ask you to wear the **glove** we gave you when you entered the laboratory on your dominant hand; the gloves we distributed are of two different types: to those who declared that they use their left hand as their dominant hand we gave a gray glove, while to those who declared that they use their right hand as their dominant hand we gave a white glove. This will allow us to identify those who will not follow the rule. Whoever violates the rule will earn 0 tokens for this part of the study.
We ask you to wear the glove now and to hold the hand wearing the glove on the desk, clearly visible, for the entire duration of the activity. Do not remove the glove until we request you to do so.

At the end of this round, you will see a summary screen where your total "score" and your earnings in tokens from this part of the study (Part 1) are reported.
Remember that your earnings will depend on your position in the group rankings, therefore on your performance and on the performance of the other group members. When the round is over, you will be informed of your position in the ranking and of the score you have made - that is the number of sliders you have correctly positioned - but not of the number of sliders correctly positioned by the other members of your group, who occupy the other positions in the final ranking.
In Part 2, we will ask you to contribute, with a fraction of the earnings you obtained from this part of the study (Part 1), to a project common to all the members of your group.

Are there any questions?

Before starting we remind you that:

- No one knows who his/her groupmates are and no one in your group will ever be able to link to you the decisions you will make during today's study.
- Your "score" in the activity is given by the number of sliders that you will be able to position exactly at **50** before the 120 seconds expire.
- Your earnings in this phase depend on the number of sliders that you will be able to correctly position and the number of sliders correctly positioned by the other four members of your group, according to the scheme reported in the Table.

| Final Group Ranking | Individual Earnings (in tokens) |
|---|---|
| 1° place | 20 |
| 2° place | 16 |
| 3° place | 12 |
| 4° place | 8 |
| 5° place | 4 |

- To perform the activity, you are allowed to use only your **non dominant hand** and you will have to wear the **glove** we gave you on the other hand. Hold the hand with the glove, clearly visible, on your desk.
- At the end of the session, you will know your "score" and your position in the group ranking but not the scores achieved by the other group members.

## INSTRUCTIONS FOR PART 2

During this part of the study, you and the other members of your group will make decisions that will determine your income from Part 2.

We ask you to contribute with the 25% **of the earnings** you realized in Part 1 to a COMMON PROJECT for all the members of your group. The part of your earnings that you will not contribute will be kept in your PRIVATE ACCOUNT and will represent an immediate gain for you.

The table below shows how much you should contribute to the common project (in tokens) based on how much you earned in Part 1.

| If in Part 1 you earned | You should contribute to the common project with |
|---|---|
| 20 tokens | 5 tokens |
| 16 tokens | 4 tokens |
| 12 tokens | 3 tokens |
| 8 tokens | 2 tokens |
| 4 tokens | 1 token |

Your earnings, at the end of Part 2, depend on:

- how many tokens you have in your PRIVATE ACCOUNT

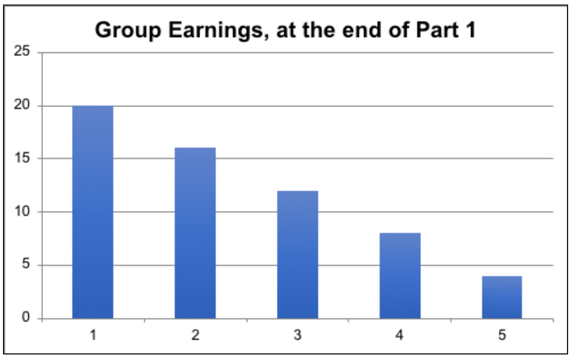
- how many tokens you and the other members of your group decide to contribute to the COMMON PROJECT

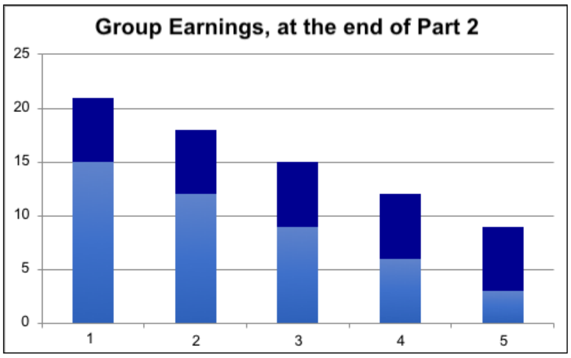
The sum of all the tokens contributed at the group level is multiplied by 2 and then divided equally among all members of the group. Each member of the group receives the same amount from the COMMON PROJECT, regardless of whether he actually contributed to it.

Your **final earnings** at the end of Part 1 and Part 2 will be then given by:

**the number of tokens you have in your PRIVATE ACCOUNT**
**+ 2*(sum of the tokens contributed by all group members to the COMMON PROJECT) /5**.

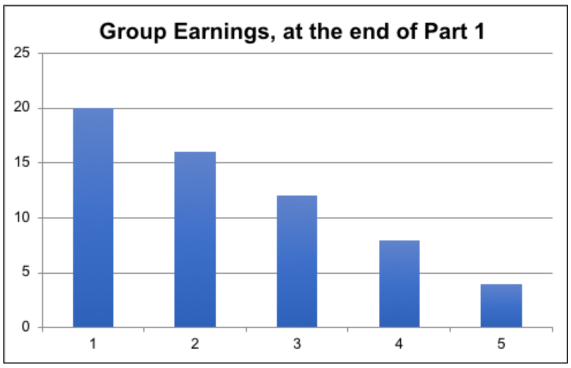
For example (1):
Imagine you earned 20 tokens from Part 1 and you contribute 5 tokens to the common project. Imagine also that the other members of your group contribute 4, 3, 2 and 1 tokens, respectively, to the common project. Your earnings, at the end of Part 2, will be given by the sum of the 15 coins in your private account and the $2 \cdot [5 + 4 + 3 + 2 + 1]/5 = 6$ tokens produced by the common project:
15 + 6 = 21 tokens.
All group members get the same benefits from the common project (6 tokens in the example) but have a different total income (having a different number of tokens in their private accounts).

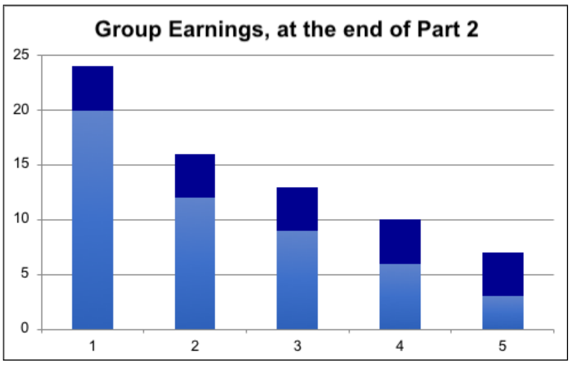




For example (2):
Imagine you earned 20 tokens from Part 1 and you contribute 0 tokens to the common project. Imagine also that the other members of your group contribute 4, 3, 2 and 1 tokens, respectively, to the common project. Your earnings, at the end of Part 2, will be given by the sum of the 20 coins in your private account and the $2 \cdot [0 + 4 + 3 + 2 + 1]/5 = 4$ tokens produced by the common project:

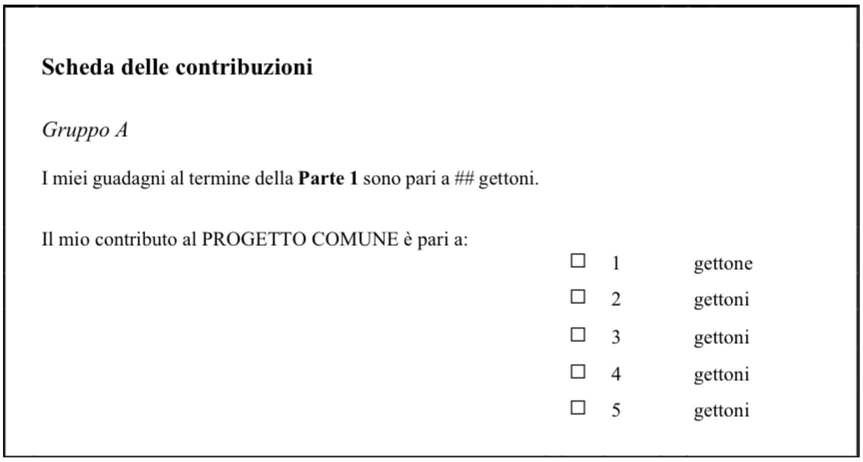
$20 + 4 = 24$ tokens.





Are there any questions?
Before Part 2 begins, you and all other participants will receive a white envelope. All white envelopes contain:

- a *'Contribution Sheet'*, where the amount of your earnings form Part 1 and the letter that identifies your group are reported.



**Scheda delle contribuzioni**

*Gruppo A*

I miei guadagni al termine della **Parte 1** sono pari a ## gettoni.

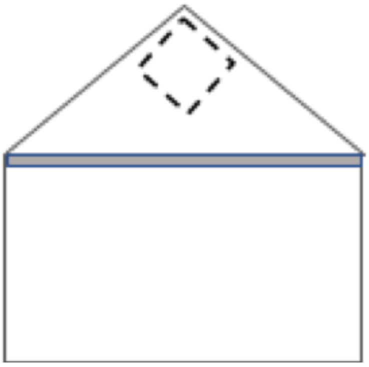
Il mio contributo al PROGETTO COMUNE è pari a:

□ 1 gettone
□ 2 gettoni
□ 3 gettoni
□ 4 gettoni
□ 5 gettoni

- a small yellow envelope.

After opening the white envelope, we ask you to fill in the *'Contribution Sheet'*.
On the *'Contribution Sheet'* you can indicate the amount of tokens you are willing to contribute to the COMMON PROJECT, by marking the corresponding option.

Once you filled in the *'Contribution Sheet'*, insert the sheet inside the small yellow envelope.
Before sealing the envelope, stick one of the stickers with your ID code (found in the green envelope extracted at the beginning of the study) on the inner surface of the upper triangle of the envelope, inside the dotted area.



When all the participants have made their own decisions and have closed their envelopes, an assistant will collect all the (sealed) yellow envelopes inside a box.

*Your Final Earnings*

The information contained in the *'Contribution Sheet'* will allow the experimenter to calculate the earnings of all participants, based on their performance in Part 1 and the result of the group decision-making process in Part 2.

As soon as the final earnings of all participants have been calculated, the experimenter will convert the earnings into Euros (1 token = 0.5 Euros) and the money earned by each participant - including the fixed amount of the participation fee equal to **3 Euros** - will be placed in a white envelope, which will be marked with the anonymous identification code (ID) that corresponds to the code printed on the sticker found inside the yellow envelope.
The experimenter has no way to connect, in any way, the identification codes (IDs) with the real identities of the participants.

Are there any questions?
Before we start, we ask you to answer a few comprehension questions to verify your understanding of the instructions.

Before we start, we remind you that:

- In this part of the study you are required to contribute to a PROJECT that is COMMON to all the members of your group with the 25% **of the earnings** you realized in Part 1, according to the scheme reported in the Table.

| If in Part 1 you earned | You should contribute to the common project with |
|---|---|
| 20 tokens | 5 tokens |
| 16 tokens | 4 tokens |
| 12 tokens | 3 tokens |
| 8 tokens | 2 tokens |
| 4 tokens | 1 token |

- The envelopes we will distribute contain the *Contribution sheet* and a small yellow envelope: once you filled in your sheet, we ask you to insert it in the yellow envelope and to stick one of your stickers with the ID code on the inside of the envelope, as indicated, before closing it.
- All the decisions you make will be completely anonymous, neither the experimenter nor the other members of your group will be able to link to you the decisions you make.

Once you are finished with filling in your sheet and sealing your yellow envelope, raise your hand: an assistant will come to your desk and will let you enter your envelope in a box.

# INSTRUCTIONS FOR PART 3

Now we ask you to fill in the questionnaire which we will distribute shortly.
You will receive another **2 Euros** for completing this questionnaire. We will directly insert this money into the payment-envelope that you will collect at the end of the study.

We ask you to remain seated and refrain from talking to the other participants.

We will distribute a paper-copy of the questionnaire to all participants, along with a large envelope. When you have finished answering all the questions, please insert the questionnaire sheets inside the envelope. Before closing the envelope, stick another of your stickers with the identification code (IDs) on the inner side of the envelope, inside the dotted area.

When all participants will be done answering the questionnaire and will have closed their envelopes, we will collect all the (closed) envelopes in a box and proceed with payments.

The questionnaire consists of 60 questions.
Remember to answer **to all the questions** before handing over your copy, and to report all your answers on the Answers' sheet.

[ Displayed on subjects' screens after all envelopes have been collected ]

Thank you for your participation.
As soon as all payment-envelopes are ready, the experimenter will place the envelopes containing your payments on a table outside the laboratory, sorted according to the identification code (ID). The experimenter will then return to the laboratory.
All envelopes will be filled in so that it will be impossible to see how much money they contain from the outside. An assistant will be standing at the payments' table and all participants will withdraw their money one by one, receiving the envelope with their payment in private.

Once at the payments' table, each participant must hand the third sticker with his/her identification code (ID) to the assistant, so that the right envelope is collected by each participant. As soon as all participants have received their envelope, the study will be over.

# EXPERIMENT 2 (NORMS)

Welcome.

You are about to participate in a study on how economic decisions are made.
All the decisions you will make during the study will be completely anonymous.

### Duration of the study and Payments

This study is divided into **four parts**.
You will receive **2 Euros** for your participation in this study.
In addition, you will have the opportunity to earn more money and the final amount of your earnings will depend on the choices you and other participants make in the two central parts of the study (Part 2 and Part 3) and on your participation in the first and the last part of the study (Part 1 and 4).
You will be paid privately (in a sealed envelope) and in cash at the end of the study.
Please turn off your cell phone. Any form of communication with other participants is strictly prohibited. If you violate this rule you will be excluded from all payments.
You will now receive instructions for Part 1. You will receive further instructions at the beginning of each of the following parts.

As we read the instructions we will ask you to answer some questions to verify your understanding of the instructions. There will also be several breaks in which you can ask questions: if you have a question, raise your hand and we will answer you private.

# INSTRUCTIONS FOR PART 1

In this part of the study we ask you to repeat the same activity twice.
You will receive **2 Euros** for your participation in the activity. We will directly insert this money into the payment-envelope that you will collect at the end of the study.

The activity consists of a screen with **48 sliders**, each of which is initially set to 0 and can be moved up to 100.
Your **"score"** in the activity will be given by the number of sliders that you will be able to position exactly at 50 before the **120 seconds** are over. Each slider has a number to its right that shows its current position, you can use the touch pad to move each slider and adjust its position as many times as you wish.
At the end of each round, you will see a summary screen where your total "score" is reported.

We will now proceed with the first round.
During this round, we ask you to perform the activity using exclusively your **dominant hand**, that is the one you do normally use to write, sign documents, use the computer mouse, etc.

Are there any questions?
Otherwise, click on the OK button and the first of two rounds will start shortly.

[ *Round n.*1 ]

We will now proceed with the second round.

During this round, we ask you to perform the activity using exclusively your **non-dominant hand**, that is the one you do NOT normally use to write, sign documents, use the computer mouse, etc.
Before we startm, we ask you to wear the **glove** we gave you when you entered the laboratory on your dominant hand; the gloves we distributed are of two different types: to those who declared that they use their left hand as their dominant hand we gave a gray glove, while to those who declared that they use their right hand as their dominant hand we gave a white glove.
This will allow us to identify those who will not follow the rule.

Are there any questions?

Otherwise, click on the OK button and the second and last of two rounds will start shortly.

[ *Round n.*2 ]

# INSTRUCTIONS FOR PART 2

Now we will read the descriptions of a series of situations. These descriptions correspond to situations in which an individual - *the individual A* - must make a decision. For each situation, we will describe the decision faced by individual A, considering all the available options.

After reading the description of each decision, we will ask you to evaluate the different choices available for individual A and decide, for each of the possible actions, whether undertaking that action would be "socially acceptable" and "consistent with adequate social behavior "or" socially unacceptable "and" inconsistent with adequate social behavior ".
By *socially acceptable* we mean a behavior that most people would agree to define the "reasonable" or "right" thing to do.

You and all other participants will face the same evaluation activity.
When assessing how much you believe that most people would consider acceptable each of the available actions, we would like you to provide us the same answers that you believe the "average" individual sitting in this room today would provide.
This is because at the end of the evaluation phase, we will randomly select one situation and one of the possible actions, and we will randomly match you with another individual in the room: your earnings will depend on the similarity between the ratings provided by you and the rating provided by other individual with whom you have been matched.

To give you an idea of how the evaluation phase will proceed, we will start with the description of an *'example situation'* to show you how to indicate your evaluations.

### Description of the 'example situation'

Individual A is in a café near the campus. While there, individual A realizes that someone has left a wallet on one of the tables. Individual A must decide what to do and has four possible choices: take the wallet, ask others nearby if the wallet belongs to them, leave the wallet where it is, or give the wallet to the manager.
Individual A can choose one among these four options.

***Evaluation of the choices available in the 'example situation'***

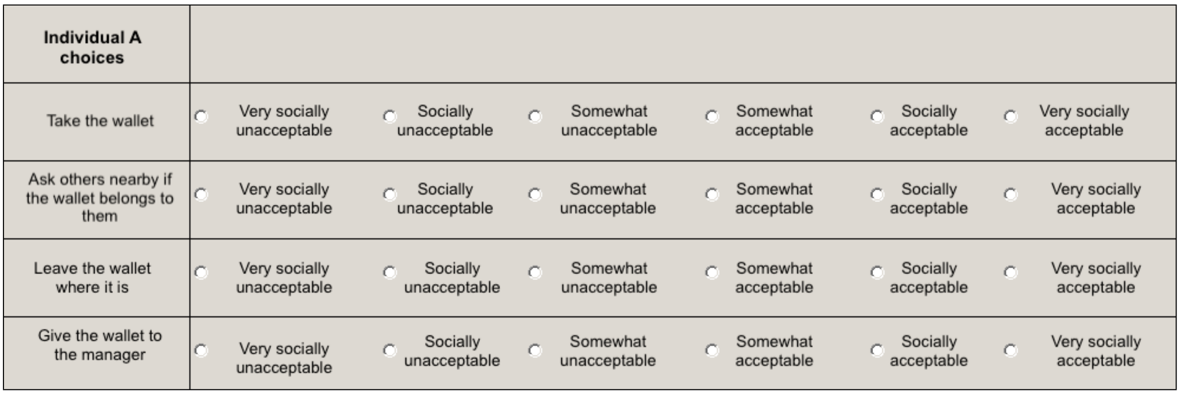
*The table below shows the list of possible choices available for individual A. For each of the choices, we ask you to indicate if you believe that the choosing this option is:*

- *Very socially unacceptable*
- *Socially unacceptable*

- *Somewhat unacceptable*
- *Somewhat acceptable*
- *Socially acceptable*
- *Very socially acceptable*

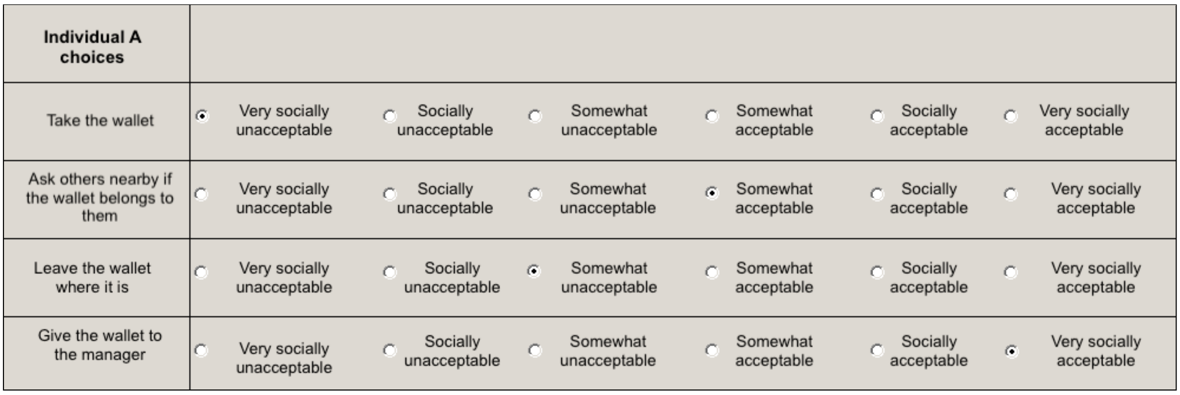
*by marking the corrisponding option.*

If this were one of the situations for this study, you should consider all the possible choices listed and, for each, indicate to what extent you believe the action is "socially acceptable" and "consistent with adequate social behavior" or "socially unacceptable" and "inconsistent with adequate social behavior".
Remember that by socially acceptable we mean a behavior that most people would agree to define a "reasonable" or "right" thing to do.



| Individual A choices | | | | | | |
|---|---|---|---|---|---|---|
| Take the wallet | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Ask others nearby if the wallet belongs to them | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Leave the wallet where it is | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Give the wallet to the manager | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |

For example, if you believe that the "average" individual in this room today believes that taking the wallet is a *versy socially unacceptable* choice, that asking others nearby if the wallet belongs to them is *somewhat acceptable*, that leaving the wallet where it is is *somewhat unacceptable* and that giving the wallet to the manager is *versy socially acceptable*, you should indicate your evaluations in this way:



| Individual A choices | | | | | | |
|---|---|---|---|---|---|---|
| Take the wallet | ● Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Ask others nearby if the wallet belongs to them | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ● Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Leave the wallet where it is | ○ Very socially unacceptable | ○ Socially unacceptable | ● Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Give the wallet to the manager | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ● Very socially acceptable |

At the end of Part 1, we will randomly select one of the situations you evaluated. For this situation, we will also randomly select one of the possible choices that individual A could have made.
Your earnings depend on the similarity between the assessment provided by you and that provided by the other individual with whom you have been matched, more precisely:

**If you provided exactly the same answer that was provided by the other participant to whom you have been matched, you will earn 7 Euros.**
For example, let's imagine we select the "example situation" above and the choice "Leave wallet where it is". Imagine that you have evaluated this choice as *"somewhat unacceptable"*: if this was also the evaluation provided by the other participant, you would receive a **7 Euros prize**.

**If your rating and the one provided by the other participant differ only by one category, you will earn 6.5 Euros.**
Imagine that you have evaluated this choice as *"somewhat unacceptable"* but that the other participant evaluated this choice as *"somewhat acceptable"*, you would receive a **6.5 Euros prize**.

**If your rating and the one provided by the other participant differ only by two categories, you will earn 5 Euros.**
Imagine that you evaluated this choice as *"somewhat unacceptable"* but that the other participant evaluated this choice as *"very socially unacceptable or socially acceptable"*, you will receive a **5 Euros prize**.
**If your rating and the one provided by the other participant differ only by three categories, you will earn 2.5 Euros.**

**If your rating and the one provided by the other participant differ only by four categories, you will earn -1 Euros.**

**If your rating and the one provided by the other participant differ only by five categories, you will earn -5.5 Euros.**

This amount will be paid to you in cash at the end of the study.

Are there any questions on this 'example situation' or on how to indicate your evaluations?

In the following pages we will describe a different situation regarding the decisions that the individual A, a generic participant in a study like this, might have to take. These decisions will be the object of the evaluation activity that you will perform in the Part 2 of today's

study.
Before we proceed, we ask you to answer some questions based on the 'Example Situation' that we have just described, in order to verify yur understanding of the instructions.

[ *Comprehension questions* ]

We will now describe the situation that will be the object of the evaluation activity that you will perform today.

## Description of the situation

Imagine that 5 individuals - A, B, C, D and E - take part in a study and are randomly assigned to the same group. The grouping is anonymous, in the sense that each individual will never know the identity of the other individuals with whom it has been matched.

In the first part of the study (Part 1) all participants have the opportunity to earn money based on their performance in the same computer activity that you performed in Part 1 of today's study.
All participants are required to perform this activity by using ff **exclusively their non-dominant hand** (the one they do NOT normally use to write, sign documents, use the computer mouse, etc.).

[*Cheating treatment*]
The experimenter, however, is not in the position to implement monitoring over the compliance with this rule.
The experimenter has no information on which is the dominant hand of each participant, so it is impossible for him to identify those who are not abiding the rule.

[*No-Cheating treatment*]
The experimenter has information on which is the dominant hand of each participant and is able to perfectly check whether all participants are abiding the rule.
During the activity, all participants are required to wear a glove on the dominant hand, like the one you have worn before, which prevents participants from using the touchpad with that hand to position the sliders and at the same time allows the experimenter to check whether the rule is respected.

Participants' earnings depend on their "score", which is equal to the number of sliders correctly positioned, and the score of the other participants in their group.

| Final Group Ranking | Individual Earnings (in tokens) |
|---|---|
| 1° place | 20 |
| 2° place | 16 |
| 3° place | 12 |
| 4° place | 8 |
| 5° place | 4 |

At the end of Part 1, individuals A, B, C, D and E are informed of their score, of their position in the ranking and therefore of their earnings but not of the score achieved by the other members of their group that occupy the other positions in the ranking. In the second part of the study (Part 2) individuals are required to contribute a share equal to the 25% of their earnings to a project common to all the members of their group.

| If in Part 1 you earned | You should contribute to the common project with |
|---|---|
| 20 tokens | 5 tokens |
| 16 tokens | 4 tokens |
| 12 tokens | 3 tokens |
| 8 tokens | 2 tokens |
| 4 tokens | 1 token |

The tokens participants decide not contribute will be kept in their PRIVATE ACCOUNTs and will represent an immediate gain for them.

The sum of all the tokens contributed at the group level is multiplied by 2 and then divided equally among all members of the group. Each member of the group receives the same amount from the COMMON PROJECT, regardless of whether he actually contributed to it.

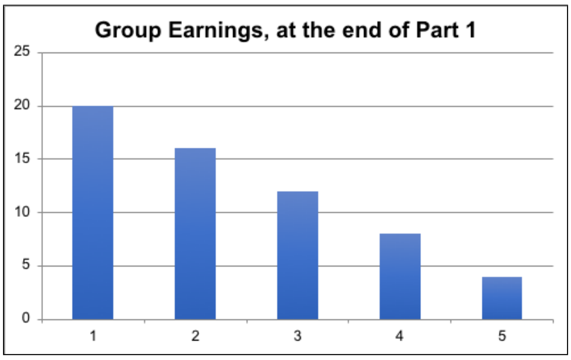
The earnings of Individual A, at the end of Part 2, depend on:

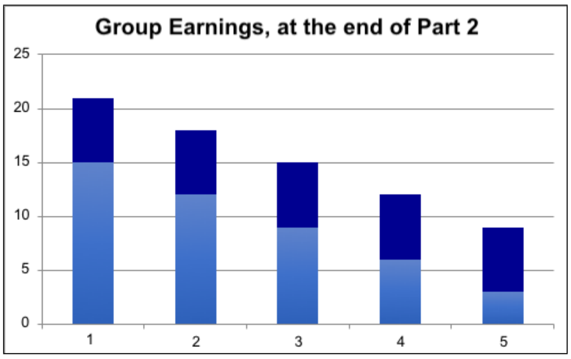
- how many tokens A has in his/her PRIVATE ACCOUNT
- how many tokens A and the other members of his/her group decide to contribute to the COMMON PROJECT

**A Earnings = the number of tokens you have in your PRIVATE ACCOUNT + 2*(sum of tokens contributed by all group members to the COMMON PROJECT)/5**.

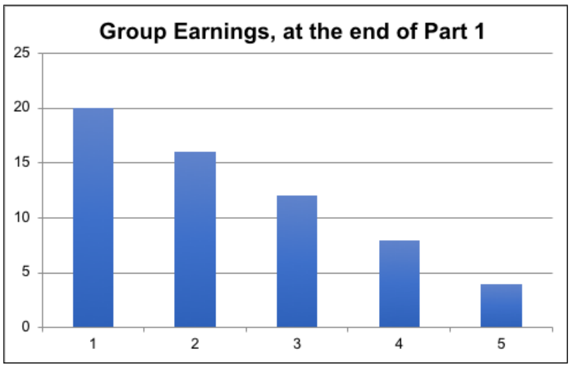
Contributions to the common project are collected in a completely anonymous way: individuals are asked to indicate on a paper sheet the quantity of tokens they wish to contribute and the sheet on which the choice is indicated is inserted in an envelope.
All the envelopes are closed by the individuals themselves and once sealed are collected in a box: neither the experimenter nor the other participants in the study have means to reconnect the contribution choices to the identity of the subjects who are present in the room.

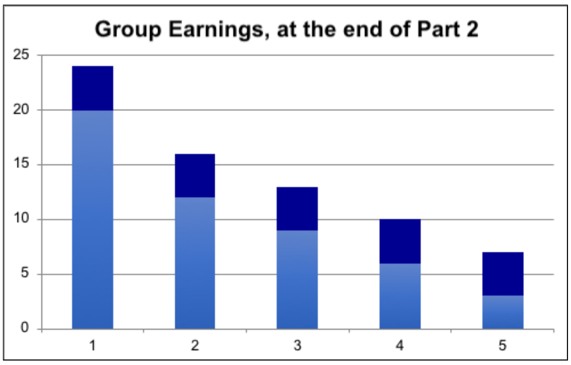
For example (1):
Imagine Individual A earned 20 tokens from Part 1 and contributes 5 tokens to the common project. Imagine also that the other members of his/her group contribute 4, 3, 2 and 1 tokens, respectively, to the common project. A's earnings, at the end of Part 2, will be given by the sum of the 15 coins in his/her private account and the $2 \cdot [5 + 4 + 3 + 2 + 1]/5 = 6$ tokens produced by the common project:
$15 + 6 = 21$ tokens.
All group members get the same benefits from the common project (6 tokens in the example) but have a different total income (having a different number of tokens in their private accounts).





For example (2):
Imagine Individual A earned 20 tokens from Part 1 and contributes 0 tokens to the common project. Imagine also that the other members of his/her group contribute 4, 3, 2 and 1 tokens, respectively, to the common project. A's earnings, at the end of Part 2, will be given by the sum of the 20 coins in his/her private account and the $2 \cdot [0 + 4 + 3 + 2 + 1]/5 = 4$ tokens produced by the common project:
$20 + 4 = 24$ tokens.





***Social Acceptability Evaluation***

*The table below lists the possible contribution choices available for individual A in a specific situation (earning from Part 1 = 12 tokens).*
*For each of the choices, we ask you to indicate if you believe that the "average" individual sitting in this room today believes that the choice is very socially unacceptable, socially unacceptable, somewhat unacceptable, somewhat acceptable, socially acceptable or very socially acceptable,*

*[Cheating treatment]*
***assuming that individual A has earned his money without violating the non-dominant hand execution rule****, but accounting for the fact that Individual A, when deciding ow much to contribute,* ***does not know whether the others have done the same****.*

*[No-Cheating treatment]*
*accounting for the fact that Individual A, when deciding ow much to contribute, knows that* ***all his/her group memebers were wearing the glove on their dominant hand as well while performing the task****.*

To indicate your answer, mark the corresponding option.
Before we start, remember that:

- If you provide exactly the same answer that is provided by the other participant to whom you will be randomly matched at the end of the evaluation phase, you will earn 7 Euros.

- If your rating and the one provided by the other participant differ only by one category, you will earn 6.5 Euros.

- If your rating and the one provided by the other participant differ only by two categories, you will earn 5 Euros.

- If your rating and the one provided by the other participant differ only by three categories, you will earn 2.5 Euros.

- If your rating and the one provided by the other participant differ only by four categories, you will earn -1 Euros.

- If your rating and the one provided by the other participant differ only by five categories, you will earn -5.5 Euros.

Are there any questions?

Before we start, we ask you to answer some other comprehension questions, in order to verify your understanding of the instructions, and we remind you that:

- In this activity we ask you to express an evaluation on the *social acceptability* of a series of choices that the individual A could take. By *social acceptability*, we mean a behavior that most people would agree to define a "reasonable" or "right" thing to do.

- If your acceptability rating corresponds to that expressed by the other participant to whom you will be randomly matched to, you will receive a prize of **7 Euros**. The greater the distance between your rating and the one expressed by the other participant matched to you, the lower your earnings will be.

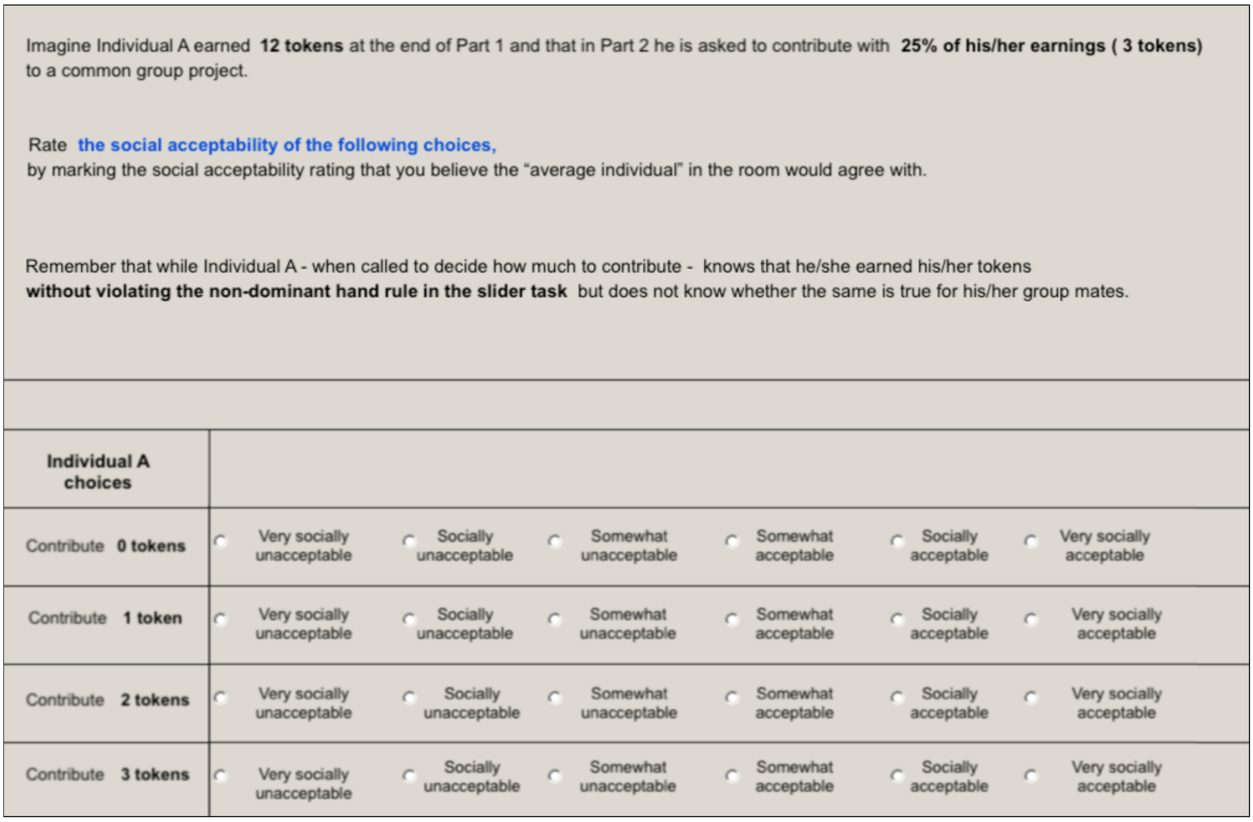
- In the situations that you will be asked to evaluate, individual A is a participant in an economic study that is divided into two parts. In this study, individual A is assigned to a group with 4 other participants and its final earnings depend on:

  - *[Cheating treatment]*
    his/her relative performance with respect to the other members of the group in the sliders task in Part 1. In this phase all the participants are asked to perform the task using their **non-dominant** hand only but the experimenter is not able to check whether this rule is actually respected. Individual A knows that he has earned his income without violating the rule but does not know if the other members of his/her group have done the same.
  - *[No-Cheating treatment]*
    his/her relative performance with respect to the other members of the group in the sliders task in Part 1. In this phase all the participants are asked to perform the task using their **non-dominant** hand only but the experimenter is able to implement monitoring and verify whether this rule is actually respected through the use of gloves. Individual A knows that all the members of his/her group wore the glove, just like him/her, on their dominant hand during the execution of the activity.
  - his/her own choice on how much to contribute to the common project and the contribution choices of the other members of his/her group in Part 2. The contribution choices are made in a completely anonymous way and each participant

can decide how much tokens to contribute to the common project. The total of all the contributions collected within the group is multiplied by two and equally distributed among all members.

- All the choices made by individual A during the study are completely anonymous, neither the experimenter nor the other participants will ever be able to reconnect the actions of the individual A to his/her identity.

*[Cheating treatment]*



Imagine Individual A earned **12 tokens** at the end of Part 1 and that in Part 2 he is asked to contribute with **25% of his/her earnings ( 3 tokens)** to a common group project.

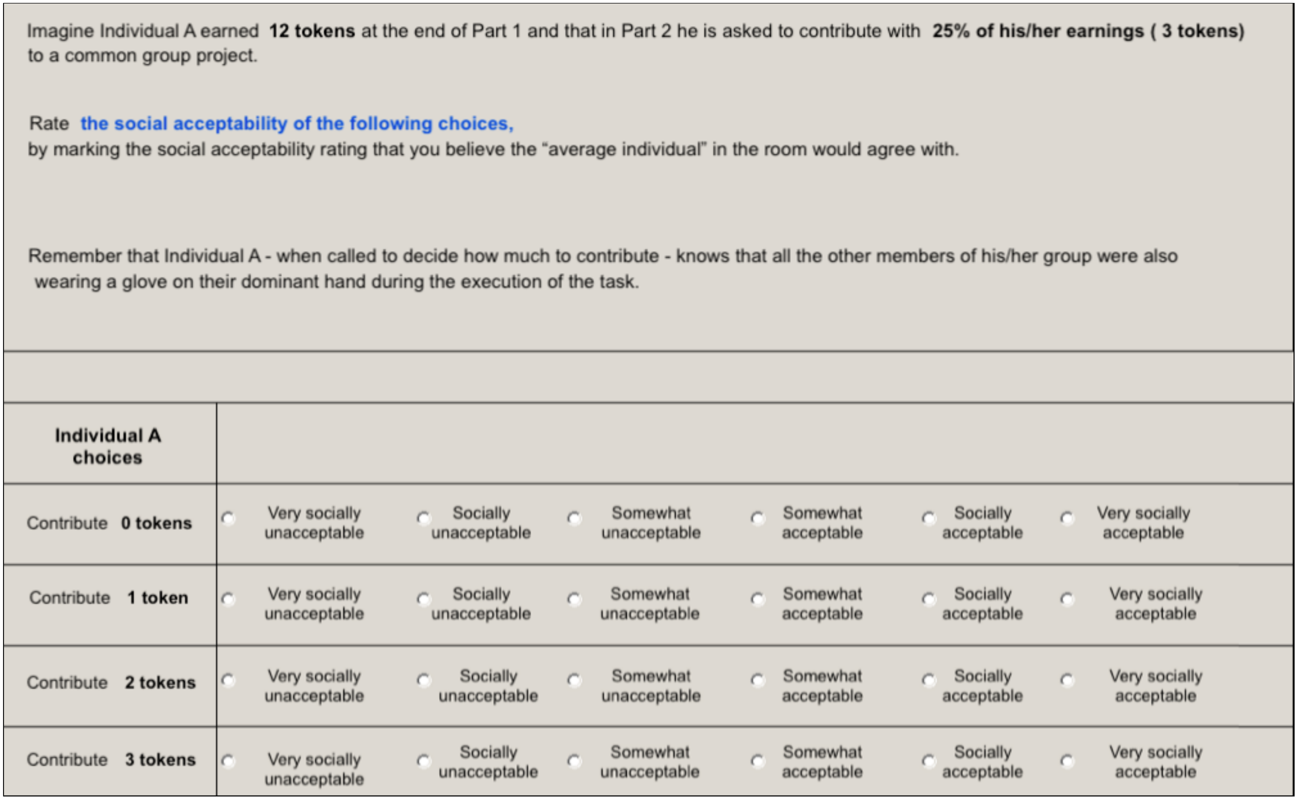
Rate **the social acceptability of the following choices,**
by marking the social acceptability rating that you believe the "average individual" in the room would agree with.

Remember that while Individual A - when called to decide how much to contribute - knows that he/she earned his/her tokens
**without violating the non-dominant hand rule in the slider task** but does not know whether the same is true for his/her group mates.

| **Individual A choices** | | | | | | |
|---|---|---|---|---|---|---|
| Contribute **0 tokens** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Contribute **1 token** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Contribute **2 tokens** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Contribute **3 tokens** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |

**Make sure you have a marked an option for each possible choice/ for each line**

*[No-Cheating treatment]*



Imagine Individual A earned **12 tokens** at the end of Part 1 and that in Part 2 he is asked to contribute with **25% of his/her earnings ( 3 tokens)** to a common group project.

Rate **the social acceptability of the following choices,**
by marking the social acceptability rating that you believe the "average individual" in the room would agree with.

Remember that Individual A - when called to decide how much to contribute - knows that all the other members of his/her group were also wearing a glove on their dominant hand during the execution of the task.

| **Individual A choices** | | | | | | |
|---|---|---|---|---|---|---|
| Contribute **0 tokens** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Contribute **1 token** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Contribute **2 tokens** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |
| Contribute **3 tokens** | ○ Very socially unacceptable | ○ Socially unacceptable | ○ Somewhat unacceptable | ○ Somewhat acceptable | ○ Socially acceptable | ○ Very socially acceptable |

**Make sure you have a marked an option for each possible choice/ for each line**

# INSTRUCTIONS FOR PART 3

We will now describe, again, the situations we referred to previously, looking at all possible contribution choices available to Individual A.

In this case, however, we ask you to guess what was the choice that Individual A, who really took part in first person in the study we described, **actually selected** under each circumstance.

Your earnings from this part of the study will depend on the accuracy of your conjecture, which will be compared to the choice actually made in those circumstances by individual A, who is randomly selected from the pool of individuals who took part in one of the previous sessions of the study. We will randomly select one of the situations: if your conjecture corresponds to the choice actually made by the individual A in that situation, you will receive a prize of **3 Euro**. This amount will be paid to you in cash at the end of the study.

For example, imagine that we select the contribution situation "Individual A has earned 16 tokens in Part 1" and that in that circumstance we observe that the choice actually selected by the individual A was "Contribute 4 tokens".
If your conjecture is that individual A has selected the "Contribute 4 tokens" option, at the end of the study you will receive a prize of **3 Euro**, in addition to the participation fee and the earnings you realized in the other parts of the study. Otherwise, you will only receive the participation fee the earnings you realized in the other parts of the study.

Are there any questions?

Before we start, we remind you that:

- We ask you now to *guess* what the choice *actually selected* by individual A in each circumstance.
- If your conjecture matches the choice actually made by the individual A - randomly selected from those who took part in one of the previous sessions of the study - you will receive a prize of **3 Euro**.

# INSTRUCTIONS FOR PART 4

We ask you now to complete the questionnaire which will shortly appear on your screen. You will receive additional **2 Euros** for completing this questionnaire. We will insert this money in the payment envelope that you will receive at the end of the study.

We ask you to remain seated and refrain from talking to the other participants.

For each of the statements, we ask you to select the number that you think best represents your opinion, according to the following scale:

5 = Strongly Agree
4 = Agree
3 = Neutral (neither agree nor disagree)
2 = Disagree
1 = Strongly Disagree

Click the OK button to start the questionnaire.

When you are done answering the questionnaire the details on payment procedures will appear on your screen.

[ Displayed on subjects' screens ]

Thank you for your participation.

We kindly ask you to remain seated, an assistant will bring the envelope containing your payment directly to your desk.

# Chapter 2

# Machine learning in the service of policy targeting: The case of Public Credit Guarantees [1]

[1] This chapter originates from a joint work with Monica Andini, Emanuele Ciani, Guido de Blasio, Alessio D'Ignazio (Bank of Italy, Structural Economic Analysis Directorate) and Andrea Paladini (University of Rome "La Sapienza", Department of Mathematics) and was partially developed while I was a research intern at the Structural Economic Analysis Directorate of the Bank of Italy in 2017 and 2018. A previous version of the work circulated as Bank of Italy Discussion Paper No. 1206, February 2019

## 2.1 Introduction

Public guarantee schemes aim to support firms' access to bank credit by providing publicly funded collateral. They typically target small and medium-sized enterprises (SMEs), which are the kind of firms most likely to suffer from credit constraints. These programs, which are widespread in both developed and developing countries, experienced a dramatic surge in popularity in the aftermath of the global financial crisis (Beck et al. (2008)). The literature has, however, highlighted that these schemes often fail to reach firms that are actually credit constrained (see, for instance, Zia (2008)): if the guarantee is provided to firms that are not credit constrained, the effectiveness of the program languishes as these firms would obtain funding anyway. One of the reasons for this misallocation is that credit rationing is difficult to gauge, while firms' creditworthiness is more easily assessed by means of balance sheet variables. As a result, the eligibility condition usually winds down into naïve rules that pinpoint financially sound borrowers, without considering indicators for credit rationing (OECD (2013)). We propose an assignment mechanism, based on Machine Learning (ML) algorithms, that explicitly accounts for both creditworthiness and credit rationing. The latter is proxied by the mismatch between firms' loan applications and actual credit granted by banks, which can be observed through the Bank of Italy Credit Register (CR). To show the advantages of the ML targeting in practice, we focus on the case of the Italian Guarantee Fund (GF). First introduced in 2000, the Fund became especially popular with the unfolding of the financial crisis, as the total amount of guarantees that were granted rose from about 1.2 billion Euros in 2008 to 11.6 billion Euros in 2016.

In the first part of the paper, we work as if we were in the ex-ante situation, in which the policymaker must design the allocation of the guarantee without prior knowledge of the intervention effectiveness and targets the hypothetical beneficiaries: firms that are both financially sound and rationed on the credit market. We make use of firm-level data from the Bank of Italy's CR, together with balance-sheet information, to develop two separate ML prediction models for firm credit constraints and firm creditworthiness, respectively. The two predictions are then combined to identify the ML hypothetical beneficiary of the GF and, by comparing the original GF assignment rule with the ML-based one, we show that the former is biased against firms that are credit constrained and we quantify the amount of resources that is misallocated.

In the second part of the paper, we substantiate the validity of our approach by looking at the ex-post dimension. As underscored by Athey (2017), nothing ensures that the ML prediction successfully identifies those for which the intervention is most beneficial. We therefore use ex-post evaluation methods to test whether the impact of the GF is stronger among the ML-targeted firms. We start by showing suggestive evidence of a greater GF effectiveness for ML-targeted firms with respect to non ML-targeted firms, among GF beneficiaries. Next, we exploit the GF assignment scheme, which is based on a scoring variable and an eligibility threshold, to run a Regression Discontinuity Design (RDD) experiment (as in de Blasio et al. (2018)), separately by ML-targeted and non ML-targeted groups of firms. We find that effectiveness of the policy is higher for the firms identified by ML as targets.

The use of ML targeting comes with several risks. One pitfall of our approach is that we train the ML algorithm by using data for a period in which the guarantee was already available. While this is a rather common situation for policymakers who try to re-design a scheme that is already in place, our ML prediction is likely to show a higher out-of-sample (forecasting) error with respect to the case in which a policy is yet to be introduced. Our results should, therefore, be taken as a conservative estimate of the benefits that could be obtained by using ML instead of the naïve rules. If the data were not contaminated by previous treatment, the prediction would have been more accurate and the gains from ML even larger. We also discuss the importance, in our case, of other issues that are typically related to the use of ML for policy decisions, such as transparency. We show that our preferred ML algorithm for this exercise – the random forest – is the one that performs worse on transparency grounds. However, it is not clear the extent to which off-the-shelf alternatives, such as the decision-tree and the Logistic LASSO, improve the transparency of the assignment process, and whether the GF rule itself, based on a scoring system that uses balance-sheet data as input, can be considered superior when accountability is at stake. In this respect, an important distinction refers to formal versus substantive transparency, where the latter includes being accountable for using public money in an effective way. We also discuss why manipulation is a less relevant issue in our set-up. Finally, we argue that additional objectives for the policymaker (the “omitted payoffs” of Kleinberg et al. (2018)) might derive from the allocation of the guarantees across banks and territories. To this aim, we contrast the distribution of collaterals across lenders and areas that would derive from ML targeting with the one based on the GF naïve rule.

While the literature on ML for policy analysis is now booming (see Athey (2019), for an updated review), the papers that deal with ML techniques to tailor the assignment of a policy are few: two exceptions are McBride and Nichols (2015), who propose to use ML to improve poverty targeting and Andini et al. (2018), who exploit ML to show how to re-target a scheme intended to boost consumption. As for the literature on credit guarantees, to the best of our knowledge, no contribution has been made on policy targeting by explicitly dealing with both the additionality and financial sustainability features of such programs. An attempt to indirectly assess the additionality of credit guarantee schemes was made by Riding et al. (2007), who use standard (non ML) econometric techniques and focus on survey data for a SMEs loan guarantee program in Canada.

Apart from the tailoring of the policy, we also make other contributions to the literature on firms' credit constraints and default risk assessment. A well-known challenge in the literature about credit constraints is posed by the availability of a measure of access to credit at firm level. Following Jimenez et al. (2012) and Jimenez et al. (2014) we measure credit constraints by elaborating on some unique features of the CR, which is collected by the Bank of Italy acting in its capacity as bank supervisor. The register records monthly information requests lodged by banks on borrowers (which are currently not borrowing from them) when they apply for a loan. As the CR database also contains detailed monthly information on bank loans, we can proxy credit constraints by means of the variation in bank credit granted to the applicant firm over the months after her loan applications. We therefore use such a proxy to provide a prediction of credit-rationing which is based on hard data from a sizable dataset. To the best of our knowledge, no previous forecasting exercise has ever been attempted for an indicator of credit market access. As for the literature on default risk assessment, after the financial crisis it became clear that traditional statistical models to predict default rates performed poorly ( Rajan et al. (2015)). When such models fail to account changes in agents' behavior, a data driven approach is deemed to be preferable. We indeed follow such an approach and use ML to predict non-performing loans using information from a very large and reliable database.

## 2.2 Credit Guarantees Schemes (CGSs)

### 2.2.1 CGSs Design: Additionality and financial sustainability

The impact of credit guarantee schemes depends on whether they actually reach firms that are credit constrained. However, providing guarantees to credit constrained firms involves a great risk taking and therefore a greater probability of incurring financial losses, which can put at risk the financial sustainability of the guarantee schemes. As stated by the WB (2015), "it is essential that credit guarantee schemes are properly designed and operated to achieve both outreach and additionality in a way that is financially sustainable". In the concrete experience of policy making, this has been a challenging task. While the financial sustainability of the schemes can be approximated by means of firm risk screening models (i.e. credit scoring models), a guidance to reach additionality based on a measure of the firms' credit constraints is largely lacking. As a result, as argued by Zia (2008), credit guarantees usually fail to reach constrained firms.

There are, however, some notable exceptions where the presence of credit constraints is explicitly addressed within the credit guarantee scheme. For instance, the U.S. Small Business Administration (SBA) requires firms applying for a guaranteed loan to demonstrate the inability to obtain credit available elsewhere on reasonable commercial terms without the SBA guarantee (so-called "credit not available elsewhere" test). This task primarily involves the potential lender, which must specify which factors prevented the financing from being accomplished without SBA support and includes the explanation in the applicant's file [2] . However, the SBA procedure is not free of drawbacks: it is vulnerable to manipulation and thus its implementation might require very high "verification" cost (Vogel and Adams (1997) ; Beck et al. (2008)) [3] .

In some cases, the pricing and the coverage ratio of the guarantee have been used to enhance the capability of the program to reach financially constrained firms. For instance, Honohan

[2] Acceptable factors include, among others: the requested loan has a longer maturity than the lender's policy permits; the requested loan exceeds either the lender's legal lending limit or policy limit regarding the amount that it can lend to one customer; the collateral does not meet the lender's policy requirements; the lender's policy normally does not allow loans to new businesses or businesses in the applicant's industry (see https://hcdc.com/credit-elsewhere-test/). In the new Standard Operating Procedure 50 10 5(J) which took effect on 1 January 2018, the agency has issued some further guidance on how to identify that credit is not available from private sources.

[3] Similar cases are those of the FAMPE (Fundo de Avail às Micro e Pequenas Empresas) in Brazil, and KGF (Kredi Garanti Fonu) in Turkey. In both cases, the guarantee fund supports SMEs that are creditworthy but proven to lack sufficient collateral.

(2010) argues that low fees may lead to the provision of guarantees to firms that are not financially constrained. When fees are high enough, only credit-constrained firms should still have an incentive to apply to the scheme. However, as argued by Saadani et al. (2011) and the OECD (2013), high fees could also lead to adverse selection, with the riskier borrowers taking part in the scheme. Setting a higher coverage ratio for riskier firms is another approach that has been adopted in order to increase additionality. This approach relies on the fact that banks generally require higher coverage to provide loans to riskier firms, typically seen as more credit constrained (Saadani et al. (2011)). However, it also displays some weaknesses. While a high coverage ratio per se does not prevent unconstrained firms from applying for the guarantee, it could lead to moral hazard behavior from both firms and banks (see, among others, Uesugi et al. (2010)).

An easier approach to reach credit constrained firms involves limiting the guarantee programs to specific categories of firms for which there is clear evidence of problems in accessing credit (i.e. firms operating in lagging areas, start-ups, female entrepreneurs). Hence, the inability to obtain market credit is assessed at an aggregate rather than individual level. Under this approach, however, there will still be credit-constrained and creditworthy firms that are left without funds (i.e. firms belonging to non-targeted sectors or areas; see Deelen and Molenaar (2004)) and firms from the chosen sectors and areas that are financially unconstrained and thus receive an undeserved benefit.

Overall, a reliable mechanism to predict which firms should be considered under a scheme of public collateral seems to be largely lacking. In this paper we propose a different, data-driven, approach: we start from micro-data, where we observe both credit default and a good proxy of credit rationing, and we let ML models assess which characteristics better predict these two conditions.

### 2.2.2 The Italian Guarantee Fund

The Italian GF started its activity in 2000. Initially the volume of bank loans with public guarantees was quite small, totaling 11 billion Euros until 2008. With the advent of the crises it experienced a boom. From 2009 to 2016, 86 billion Euros in loans to SMEs benefited from the public guarantee. The growth in volumes reflects the desire of the Italian authorities to counterbalance the effect of the credit crunch. The latter was particularly severe for SMEs that, in an environment of increased credit risk, experienced a more significant drop in credit flows and a stronger rise in interest rates with respect to larger firms (MSE (2015), CF (ears)). The provision of GF guarantees is limited to SMEs, defined according to EU criteria, active in the private sector, which includes manufacturing, construction and services. Some specific sectors, such as agriculture, automobile and financial services, are not covered by the scheme because of the limitations imposed by the EU regulation on competition. The public guarantee insures up to 80 per cent of the value of a bank loan. For each firm, however, there is a maximum amount of guarantee, which is equal to 1.5 million Euros. The GF can guarantee both short-term and long-term loans and there are no constraints in terms of the final use of the funding by the borrower. It is important to notice that, in case of default, the financing institution can immediately call on the GF to meet its obligation ("first demand guarantee"). According to the GF procedure, a SME that needs to borrow asks the bank to apply for a public guarantee (alternatively, it is the bank that proposes to the firm to apply for the guarantee). The bank has to verify the eligibility of the firm for the scheme through a scoring system (a software) provided by the GF. The scoring system is designed to minimize the likelihood that a firm defaults on its debt; no consideration is given to the actual financial constraints of the SME. The scoring system takes into account four indicators (that slightly differ according to the economic sector) of the firms' financial condition in the two financial years preceding that of the application: the soundness of firms' financial structure is measured for the industry (service) sector by the ratios of equity and long-term loans to fixed assets (short-term assets on short-term liabilities) and equity to total liabilities (short-term assets to sales); the short-term financial burden is measured by the ratio of financial expenses to sales; the cash flow is measured by the ratio of cash flow to total assets. As described in de Blasio et al. (2018), the four balance-sheet indicators are synthetized to produce a single score [4]. According to this score, the applicant firms are split in three types (0, 1, and 2). Type-0 firms are not eligible. Type-1 and Type-2 firms are both eligible but do not automatically receive the treatment. They have to go through a further assessment, which is more demanding for

[4] The sectors eligible under the GF scheme are divided into two groups: manufacturing, construction and fishing and the tradable sector, hospitality industry, transportation and other private service sectors. Firms belonging to these two groups face a slightly different screening procedure by the Guarantee Fund, based on a different set of balance-sheet indicators (and scoring thresholds).

the Type-1 firm, as they have worse scores (i.e., poorer lagged balance-sheet observables) [5]. The additional assessment concludes with final approval or rejection. Rejection, however, has been a rare event (3.8 per cent of the applicant firms were rejected over the 2011-16 period). In what follows, we will refer to the GF eligibility mechanism both in Subsection 4.4, where we compare GF-eligible firms with the ML-targeted ones, and in Subsection 5.2, where we run a RDD exercise based on the GF assignment to substantiate the ML gains in terms of effectiveness [6].

## 2.3 'Ex ante' prediction complementarity with 'Ex post' evaluation

### 2.3.1 The 'Ex ante' prediction exercise: a theoretical framework

Throughout our prediction exercise, we operate in a "ex-ante" situation, in which we assume the policymaker is called to design an allocation rule for the public guarantee scheme, without prior knowledge of the effectiveness of the intervention.

In this context, when designing the new allocation rule, our objective is to help the policy maker in reaching its desired policy-target, which, based on our discussion in Section 2.1 we identify with firms that are jointly financially sound (financial sustainability argument) and credit-constrained (additionality argument). We can easily spell out how our empirical contribution would be welfare-improving for the policy-maker, if the ML algorithms provide accurate predictions, by relying on a simple theoretical framework. In this framework, we imagine that the policy maker operates as to maximize a utility function that is increasing in the ability of the policy maker to reach its desired targets: each time a firm applies to the Guarantee Fund, the policy maker has to make an assignment decision $G \in \{0, 1\}$ and the

[5] According to the GF guidelines, the additional assessment is referred only to cash-flow requirements for Type-2 firms. As for Type-1 firms, the additional assessment is an in-depth analysis of the economic and financial situation of the firm. Again, the aspects related to credit constraints do not matter.

[6] In December 2017 the Italian Guarantee Fund was the subject of a reform, primarily aimed at: (i) enlarging the number or potential beneficiary firms, (ii) improving the screening of firms to exclude those that are not creditworthy, and (iii) increasing the support to creditworthy firms that are more exposed to the risk of credit rationing. The central point of the reform was the adoption of a new rating model to assess the creditworthiness of the firms, based on a larger set of information with respect to the mechanism described above. In order to tackle credit rationing, the new rules allow riskier (but still creditworthy) firms to benefit from a larger share of the loan covered by the guarantee. The reform has become operative since mid-March 2019. The analysis carried out in this paper is solely based on the pre-reform GF rules.

utility of the policy maker depends on firms' probability to be creditworthy ($\alpha$) and credit-rationed ($\beta$), and by the assignment decision G.

We model the policy maker utility function as increasing in:
– firm i probability to be creditworthy (financial sustainability argument);
– firm i probability to be credit-rationed (additionality argument);

$$U(\alpha_i, \beta_i, G) = G \cdot [\lambda_1 \alpha_i \beta_i - \lambda_2 \alpha(1 - \beta_i) - \lambda_3(1 - \alpha)] + (1 - G)\emptyset \tag{2.1}$$

where:
$\lambda_1 > 0$ is the weight attached to reaching firms that are both creditworthy and credit rationed;
$\lambda_2 > 0$ is the weight (cost) attached to reaching firms that are creditworthy but not rationed;
$\lambda_3 > 0$ is the weight (cost) attached to reaching firms that are not creditworthy, irrespective of their rationing status; we take the view, shared by policy-analysts, that the cost associated to financing firms that are not creditworthy [Type B error] is higher than the cost associated to financing firms that are creditworthy but are not credit-rationed [Type A error] $\lambda_3 > 0 > \lambda_2 > 0$ , since it could hinder the Fund subsistence.

$$\mathrm{U}\left(\alpha_i, \beta_i, G\right) \begin{cases} if\ G = 1 \rightarrow \lambda_1 \alpha_i \beta_i - \lambda_2 \alpha_i \left(1 - \beta_i\right) - \lambda_3 \left(1 - \alpha_i\right) \\ if\ G = 0 \rightarrow \emptyset \end{cases}$$

We can treat the utility maximization problem as unconstrained because, since the GF has been operating, there have been no cases in which firms identified as eligible by the GF authority did not receive the public guarantee due to lack or limitation of available funds.
As we already discussed in Section 2.1, properly measuring firms' financial constraints is a hard task and most of currently operating Public Guarantee schemes tend to rely on naïve approaches to quantify firms' rationing status: one of these approaches, which mirrors the strategy that is currently adopted by the Italian GF, is based on an aggregate-level rationing assessment.
In this context, we assume that the policy maker, even if able to observe, or, more precisely,

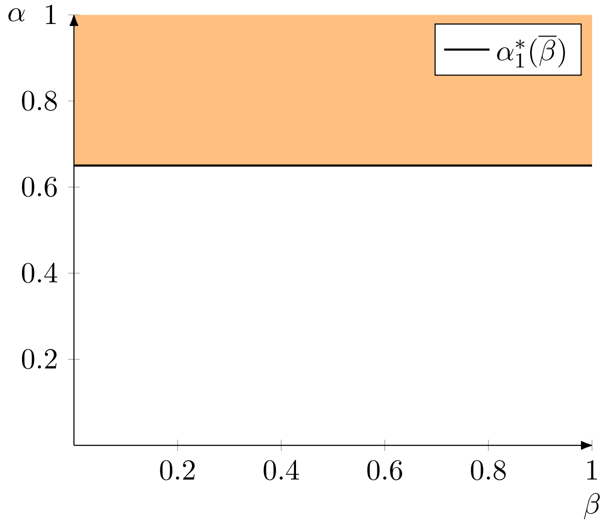
to estimate with a good degree of approximation firms' probability to be creditworthy ($\alpha$), cannot observe firms' true probability to be rationed ($\beta^T$): the policy maker is, then, forced to use an imprecise approximation for this dimension ($\hat{\beta}$) when making the assignment decision, although its utility will ultimately depend on the true probability ($\beta^T$). Under the current assignment rule (Scenario 1), the policy maker is not able to observe or to measure the 'true' rationing status of each firm, then relies on an aggregate measure at the sector level to identify target groups based on overall rationing severity: within sectors identified as targets all firms are considered eligible and no individual rationing assessment is performed. In this case the policy maker, is unable to measure the 'true' rationing status of each firm and relies on its *prior*, which is equal to the mean measured over the entire population of interest, as its best approximation. Assuming $\beta_i^T \sim (\overline{\beta},\ \sigma^2{}_B)$ the best approximation is then given by $\hat{\beta}_i = \overline{\beta}$ and the guarantee assignment rule, for firms belonging to target groups, winds down to a sole creditworthiness assessment. The policy maker defines an assignment decision rule $\gamma^1$ that will therefore depend on the optimal threshold value for $\alpha_1^*(\overline{\beta})$ [7] , which is set as the value that maximizes the utility of the policy maker given $\overline{\beta}$.

The policy maker assigns the guarantee $[\gamma^1(X) = 1]$ if firm's probability to be creditworthy, which can be expressed as a function of observable firms' features $h_i(X)$, is above the threshold $\alpha_1^*(\overline{\beta})$, as shown in Figure 2.1, where the threshold depends on the mean rationing status in the population $\overline{\beta}$ and on how the policy-maker weights the benefit of reaching the desired target ($\lambda_1$) and the costs associated to Type A and Type B errors ($\lambda_2$ and $\lambda_3$):

$$\gamma_1(X) = 1 \text{ if and only if } \hat{\alpha}_i = h_i(X) > \alpha_1^*(\overline{\beta}).$$

[7] If we substitute in $\beta_i^T$ with $\overline{\beta}$ in the equation (1): $U(\alpha_i, \beta_i^T = \overline{\beta}, G) = G \cdot [\lambda_1 \alpha_i \overline{\beta} - \lambda_2 \alpha(1-\overline{\beta}) - \lambda_3(1-\alpha)]$ and we solve the F.O.C. $\frac{\partial \mathrm{E}_\beta [\mathrm{U}(\alpha_{GF}, \overline{\beta}, G)]}{\partial G} = 0$, we obtain $\alpha_1^*(\overline{\beta}) = \frac{\lambda_3}{[(\lambda_1+\lambda_2)\overline{\beta} - \lambda_2 + \lambda_3]}$.

Figure 2.1: GF Assignment rule (Scenario 1)



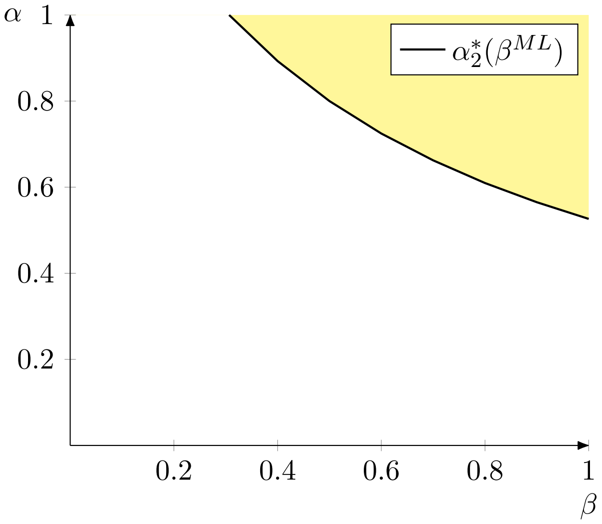
When we introduce ML techniques (Scenario 2), the policy maker is able to estimate, although with some error, the 'true' rationing status of each firm relying on prediction algorithms, which are able to provide a good *signal* of firms' true rationing situation. The ML estimate is used as the best approximation of the 'true' rationing status at the firm level $\hat{\beta}_i = \hat{\beta}_i^{ML}$.

The guarantee assignment rule will then be based on both firms' rationing and creditworthiness status. Based on observable firms' characteristics, the policy maker defines an assignment decision rule $\gamma^2$ that depends on the optimal combinations of values for $\alpha$ and $\beta$ such that the utility of the policy maker is maximized. The policy maker assigns the guarantee $[\gamma^2(X) = 1]$ if firm's propensity to be creditworthy, which can be expressed as a function of observable firms' features $\hat{\alpha}_i = m_i(X)$ (which we assume to be empirically equivalent to $h_i(X)$), and firm's probability to be rationed on the credit market, expressed as a function of observable firms' features $\hat{\beta}_i = g_i(X)$, belong to the optimal parameters' space [8], which corresponds to the area of the plane $(\beta, \alpha)$ above the function $\alpha_2^* = \alpha_2^*(\hat{\beta}_i^{ML})$, as shown in Figure 2.2,:

[8] Under the following assumptions:
**Ass.1** The true data generating model for $\beta_i^T$, given firms' characteristics $x_i$ is $\beta_i^T = f(x_i) + u_i, E[u_i] = 0$;
**Ass.2** The estimate produced by the ML model, given firms' characteristics $x_i$ is: $\hat{\beta}_i^{ML} = g(x_i) = f(x_i) + u_i + \epsilon_i$ , where $\epsilon_i$ is the ML prediction error. Therefore $\hat{\beta}_i^{ML} = \beta_i^T + \epsilon_i$;
**Ass.3** $E[\epsilon_i] = 0 \; and E[\epsilon_i^2] = \sigma_\epsilon^2$.
We substitute in equation (1) $\beta_i^T = \hat{\beta}_i^{ML}$:
$U(\alpha_i, \beta_i^T = \hat{\beta}_i^{ML}, G) = G \cdot [\lambda_1 \alpha_i \hat{\beta}_i^{ML} - \lambda_2 \alpha(1 - \hat{\beta}_i^{ML}) - \lambda_3(1 - \alpha)]$
And we solve the F.O.C. (1): $\frac{\partial \mathrm{E}_\beta\left[\mathrm{U}\left(\alpha_{GF}, \hat{\beta}_i^{ML}, G\right)\right]}{\partial G} = 0$, we obtain $\alpha_1^*(\hat{\beta}_i^{ML}) = \frac{\lambda_3}{[(\lambda_1 + \lambda_2)\hat{\beta}_i^{ML} - \lambda_2 + \lambda_3]}$.

$$\gamma_2(X) = 1 \text{ if and only if } (\hat{\alpha}_i; \hat{\beta}_i^{ML}) = (m_i(X); g_i(X)) \in [\textit{yellow area}].$$

Figure 2.2: GF Assignment rule (Scenario 2)



The difference in terms of utility of switching from $\gamma_1$ to an alternative rule based on the ML $\gamma_2$ will be driven by the sign/size of the utility associated to treating firms for which the two rules disagree. In particular, the utility of the policy-maker will depend on actual realizations of $\alpha_i$ and ${\beta_i}^T$ for:

− firms that would be treated by the old rule $\gamma_1$ but would not be treated under the new rule;

− firms that would not be treated by the old rule $\gamma_1$ but would be treated under the new rule;

In general, when substituting the old rule $\gamma_1$ with the new rule $\gamma_2$ , if we assume both the current GF rule and the ML rule can observe the probability to be creditworthy with approximately the same degree of accuracy [9] , the magnitude of utility gains from adopting the ML-empowered rule depends on the shape of the distribution of $\beta_i^T$ and on the accuracy of the ML-prediction: in general, if the distribution of $\beta_i^T$ is dispersed around the mean and the ML-prediction error $\epsilon_i$ is on average low with a dense distribution around zero, switching from $\gamma_1$ to $\gamma_2$ will actually be welfare-improving. If the distribution of $\beta_i^T$ is symmetrical and

[9] This may not be true at the time we refer to in the analysis, when the GF was using a simplified assignment rule based on a restricted set of information, but could apply to the current modus operandi of the GF (since March 2019, see footnote n.5), which relies on a more sophisticated credit scoring model to assess financial stability of applicants.

highly dense around the mean with thin long tails, the mean serves as a good approximation of $\beta_i^T$ for the whole population: this implies that under $\gamma_1$ we would not have high Type B error costs and, even if the ML-estimation is accurate enough, gains from the use of $\gamma_2$ would be limited in size. If, instead, the distribution of $\beta_i^T$ is highly skewed with a low density around the mean and a thick left- or right-tail, the mean does not serve as a good approximation of ${\beta_i}^T$ for the whole population: this implies that under $\gamma_1$ we would encounter high Type B error costs and, if the ML-estimation is accurate enough, gains from the use of $\gamma_2$ would be notable in size.

If we relax the assumption that the parameter $\alpha_i$ is accurately observed by the policy-maker ($\alpha_i = \hat{\alpha}_i$, where $\hat{\alpha}_i = \hat{\alpha}_i^{GF} = h_i(X) = \hat{\alpha}_i^{ML} = m_i(X)$) and we allow the prediction technologies used to predict $\alpha_i$ under the two rules $\gamma_1$ and $\gamma_2$ to differ ($\alpha_i \neq \hat{\alpha}_i$,where $\hat{\alpha}_i^{GF} = h_i(X) \neq \hat{\alpha}_i^{ML} = m_i(X)$), we could observe even larger utility gains switching from $\gamma_1$ to $\gamma_2$ if the prediction technology used to predict $\alpha_i$ under $\gamma_2$ outperforms the technology used by the actual GF rule $\gamma_1$ in terms of prediction accuracy $ERR_{GF} = \frac{1}{N}\sum[\hat{\alpha}_i^{GF} \neq \alpha_i] > ERR_{ML} = \frac{1}{N}\sum[\hat{\alpha}_i^{ML} \neq \alpha_i]$.

### 2.3.2 The complementarity of the ‘ex ante’ prediction and the ‘ex post’ evaluation

The main limitation of our “ex-ante” approach, which is not designed in order to identify firms to which the intervention of the policy would be the most beneficial, is that we have no guarantee that the ML-approach would lead to an improvement in the effectiveness of the policy. As underscored by Athey (2017), most of the ML applications that focus on prediction problems in the domain of public policy, especially those related to resource allocation, require some complementary statistical analyses on the causal effect of the intervention in order to ensure that the new predictive tools bring about an improvement, taking into account that the public intervention might have heterogenous effects on different groups of beneficiaries.

In our case, the prediction exercise is explicitly tailored to the objective of providing to the policy maker new and more accurate instruments to reach its desired targets, which are identified as such by the theory and the literature. By doing so, we would also contribute to increase the program effectiveness if the targets selected ex-ante by the policy maker – which we expect to be more likely reached under the ML assignment rule rather than under the current assignment mechanism – are part of the sub-population for which the treatment effect is sizeable.

In particular, we are interested in testing whether the effect of the policy is higher for firms identified as targets by our ML-empowered assignment rule because, if this is the case, channeling the resources of the Fund to ML-target firms would also contribute to increase the average effectiveness of the program. In this context, we are mainly interested in evaluating the first-round effect of the policy, that is the ability of the guarantee scheme to grant firms access to credit, which is measured in terms of credit availability, looking at the magnitude of the amount of loans actually disbursed to firms in the time window that follows the loan application.

Through our ex-post analysis, we test whether the effect of the policy on credit availability and on other indicators measuring second-round effects (such as investments, sales and bad loans), is stronger for the groups of firms identified as targets by the ML rule, as compared to firms that are not identified as targets by ML. After showing some suggestive evidence of how ML target firms perform better than non-ML target firms in terms of our outcomes of interest within the sample of beneficiary firms, we rely on a RDD strategy to separately identify and quantify the causal effect of the policy for ML target and non-ML target firms, following the approach adopted by de Blasio et al. (2018).

### 2.3.3 Targeting using ML before vs. after ex-post evaluation

An alternative strategy could have been to directly employ ML to identify, within an ex-post evaluation framework, firms' characteristics associated with a stronger treatment effect, in order to target firms for which the policy intervention would have been the most beneficial/-effective. This strategy has the advantage of directly focusing on the heterogeneity of the policy effect and calls for a somehow a-theoretical identification of the treatment assignment rule, which is exclusively tailored to maximize the effectiveness of the policy. For instance, Ascarza (2018) uses a decision-tree based algorithm to identify which customers should be targeted by firms' retention programs aimed at avoiding churning. The author exploits a pilot experiment in which the treatment - the retention program - was randomly assigned among customers and uses these data to train the algorithm. In this setting, the decision tree is able to predict which characteristics are associated with stronger effects (see also Athey and Imbens (2016)) and the author can identify a targeting rule based on the identification of subjects that are most sensitive to retention interventions rather than on the identification of subjects with the highest predicted risk of churning, which proves to be profit-enhancing for the firm.

In principle, we could have proceeded as in Ascarza (2018), by using ML techniques to find which subgroups of firms showed larger effects as estimated by the RDD strategy. However, RDD estimates, compared to the estimates obtained by Ascarza (2018) through the randomized experiment, have a very local interpretation. We would therefore find it debatable to base the entire targeting of the program using exclusively the local RDD estimates as a starting point in the search for heterogeneous effects along a large set of non-pre-specified dimensions. Furthermore, our strategy could also be applied, more in general, to cases in which a policy has not yet been rolled out and it is costly to delay its introduction to run a randomized pilot, as it was the case for the GF during the recession. It is, therefore, interesting to understand the pros and cons of such a strategy.

## 2.4 The ML Prediction Exercise

### 2.4.1 The ML approach

We estimate two separate predictive models for being "credit constrained" and "creditworthy", that is:

$$credit\ constrained_i = f(X_i) + \epsilon_i \tag{2.2}$$

$$credit\ worthy_i = g(X_i) + \eta_i \tag{2.3}$$

where i indexes the loan application from a firm in a given quarter, $X_i$ is a set of P observable characteristics of the firm at the time of the application, $f(.)$ and $g(.)$ are the two functions to be learnt from the data, and $\epsilon_i$ and $\eta_i$ are noise. The outcomes are two binary variables - $credit\ constrained_i$ and $credit\ worthy_i$ - which are valued one if the application belongs to the respective status.

As we do not know the true functions $f(X_i)$ and $g(X_i)$, our aim is to estimate (or train, in ML jargon) them by using a model that has good forecasting performance out-of-sample, because the rule is meant to be used for future assessments of new requests for the GF guarantee. In this respect, ML tools are particularly useful (Mullainathan and Spiess (2017)) as they aim to minimize the out-of-sample forecasting error. In short, such tools rely on highly flexible functional forms, where greater complexity improves the in-sample fit but increases the out-of-sample error of the selected model. The complexity of the model is set through a regularization parameter, which is chosen by cross validation in order to minimize the out-of-sample error (Hastie et al. (2009)). Unlike in standard econometrics, ML models do not focus on obtaining unbiased estimates of the two functions, but rather on minimizing the out-of-sample forecasting error. Their objective function therefore allows for some bias in the estimator if this reduces the variance of the prediction.

In practice, we employ and compare three different off-the-shelf ML algorithms, the decision tree, the random forest and the logistic LASSO regression. Before fitting our models, we randomly split our sample into two subsamples, a training set and a testing set, following the 2/3 & 1/3 division rule (as suggested in Y. Zhao and Cen (2014)). We then fit our models on the training set and later test their out-of-sample predictive performance over the testing set. As a criterion for selecting the best complexity parameters and the best model across the three different alternatives, we look at the misclassification rate, which is the fraction of observations that are predicted to belong to the wrong class. In the Appendix A.3 we briefly introduce the three algorithms and we discuss the details of their implementation, including our strategy for dealing with the unbalancedness in the creditworthy status (as most of the observations have $credit\ worthy_i = 1$).

The fact that we estimate separate models for $credit\ constrained_i$ and $credit\ worthy_i$ implies that we are focusing on the two marginal probabilities, not that we are assuming that the two events are statistically independent. In Subsection 4.3 we show the relation between the two predictions and we discuss the implications for our analysis. From an econometric perspective, our purpose is to predict each status using the same set of observable characteristics. One could think of our prediction problem as a system of simultaneous equations where the probability of a given status depends on both the covariates and the (true) probability of the other status. However, we do not observe the latent probability of each status and, therefore, such an approach is unfeasible and useless for prediction purposes. By recursively substituting the unknown latent probabilities we would end up with two equations where the right-hand sides are a function of the observable $X_i$, exactly as in (1) and (2). An alternative could be to directly predict the joint status. However, in our dataset the *credit worthy*$_i$ status is highly unbalanced towards 1. A joint model would, therefore, give most of the weight to the errors related to the constrained status, where there is a larger fraction of observations valued as 0. We nevertheless tried estimating such a model, but no improvement has been reached in terms of misclassification error. Further details are provided in the Appendix A.3.

### 2.4.2 Data and sample selection

In order to gather information on firms that initiate a bank loan application, we exploit the Bank of Italy's Credit Register (CR). In particular, we use as our main data source the requests of preliminary information (PI) collected by the CR [10].

The PI request is an instrument used by banks to gain information on the reliability of new potential borrowers. Through a PI request, banks can obtain detailed information on the credit history of their loan applicants [11]. Given that obtaining information through a PI request is not free of cost, it is reasonable to assume that the decision to inspect a firm's credit history always follows a loan application by the firm to the PI-requiring bank. Throughout the paper we will therefore treat each PI request as a loan application and we will use the two terms interchangeably.

Using the PI requests we build two datasets consisting of Italian limited companies that applied for a bank loan in 2011 or in 2012. We chose 2011 and 2012 as sampling years because they leave us with a good number of follow-up years, while still allowing us to draw information on firms past history over two years. From the dataset we exclude (i) firms for which we do not have balance-sheet information on the two years preceding the PI request; (ii) firms that have never had lending relationships with banking institutions in the two years preceding the PI request. These firms, which include for instance start-ups, are likely to be different from the others, and therefore we would need to devise a separate forecasting and assignment exercise. The final sample on which we train and test our ML algorithm is composed of nearly 190,000 firms that made a bank loan request in 2011. This sample is randomly split into a 2/3 training sample (used to estimate the models) and 1/3 testing

[10] A similar register is maintained by the Bank of Spain (Jimenez et al. (2012), Jimenez et al. (2014)).

[11] The CR retains information at the loan level on all loan contracts granted to each borrower whose total debt from a bank is above 30.000 Euros.

one (used to validate and compare the models). Firms that applied for a bank loan in 2012 constitute instead our hold-out sample (see the Appendix for more details on this sample), which we will use in Subsections 4.4 and 5 to compare the current GF assignment rule with the one that we devise based on the ML algorithms.

Our sample is likely to represent only a subset of the total number of firms that request credit. In fact, it excludes those credit-requiring firms for which a PI request is not issued. This is the case when the firm that applies for a loan is already known to the bank, or the firm is outstanding and no further screen is needed by the bank. The presence of credit relations for which a PI has not been issued will likely drag our sample towards less financially sound firms. However, this is not an issue for this study, as firms that are indeed financially stable are not the primary target of the GF in first place.

Firms may issue more than one loan request within the same year. As we cannot observe the amount of each loan application, and firms may turn to more than one bank in order to finance a given project, we assume that different loan applications issued by the same firm (proxied by different PI requests issued by banks on that firm) within the same quarter refer to the same project. Different loan applications by the same firm in different quarters are instead considered as separate observations in the sample. The final 2011 dataset is therefore composed of 278,355 observations [12]. Approximately 2/3 of this dataset pertains to the training set (185,256 observations, relative to 123,276 firms), while the remaining 1/3 pertains to the test set (93,099 observations, relative to 62,052 firms).

For each firm that applied for a loan in a given quarter of 2011 we devise two outcome variables (Figure 2.3). The first is an indicator of whether the firm is credit constrained, namely if its total amount of granted loans has not increased six months after the PI request [13]. In our sample about two thirds (66.2 per cent) of loan applications refer to firms that are credit constrained, a figure in line with those obtained from the Survey on SMEs access to finance (ECB (2015)). The second is an indicator of whether the firm is creditworthy, namely if it

[12] See Figure A1 in the Appendix A7 for observed frequencies of the number of quarterly observations of PI requests issued by the same firms in the full 2011 sample: in roughly 2/3 of the cases we observe that only one PI request is issued by the same firm throughout the year.

[13] The measure considers the total amount of bank loans, and not just the loans granted by the banks that issued the PI request about the firm, in order to control for those cases where the credit is issued to the firm by banks not requiring a PI. For the details on the credit-constraints index based on PI requests, see Albertazzi et al. (2017), Carmignani et al. (2019), Galardo et al. (2017).

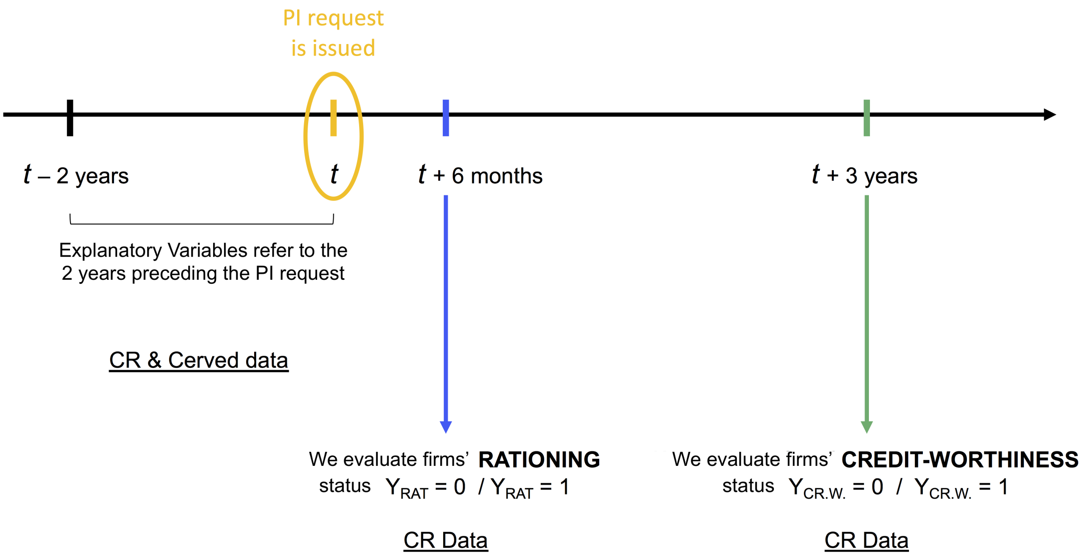

PI request is issued
t – 2 years
t
t + 6 months
t + 3 years
Explanatory Variables refer to the 2 years preceding the PI request
CR & Cerved data
We evaluate firms' RATIONING status $Y_{RAT} = 0$ / $Y_{RAT} = 1$
CR Data
We evaluate firms' CREDIT-WORTHINESS status $Y_{CR.W.} = 0$ / $Y_{CR.W.} = 1$
CR Data

Figure 2.3: Construction of the two indicators for the supervised learning exercise

does not have "adjusted bad loans" in the three-year window following the PI request [14]. About 86 per cent of the applications in our sample refer to firms that are creditworthy. The prediction makes use of a set of explanatory variables that are observable by the policymaker at the time she is required to assign the guarantee according to the decision rule in place. We focus on CR data on lending from the banking system and balance-sheet information from the Cerved database. In both cases, we include not only variables in levels, but also a measure of their change over time. We also include some additional variables capturing firm-specific characteristics: firm age, location and sector indicators. Finally, we introduce a dummy variable that takes value one for firms that have already been beneficiaries of the GF program in the years preceding the PI request (data on GF beneficiaries have been available since 2005). The complete set of covariates includes 108 variables (see Table A1 in the Appendix A8 for the complete list and a brief description; Table A2 provides summary statistics) [15]. In order to minimize information redundancy (Fan and Lv (2008)), we submit our covariates set to a pre-processing procedure before applying ML techniques in the two predictive exercises (see the Appendix).

[14] A firm has adjusted bad loans if it is reported as insolvent by a bank that accounts for at least 70 per cent of the firm's total bank loans, or if it is reported as insolvent by two or more banks that together account for at least 10 per cent of the firm's total bank loans. See "sofferenze rettificate" at: https://www.bancaditalia.it/footer/glossario/ index.html?letter=s.

[15] We recover the same set of observables for firms that applied for a loan in 2012, which form our "hold-out" sample.

### 2.4.3 Prediction results

Based on the prediction performances of the three ML models used (decision tree, random forest and logistic LASSO), our preferred model is random forest. In short, it generally provides: (i) a lower fraction of observations predicted to be in the wrong status; (ii) a higher ratio between the percentage of observations correctly predicted as credit constrained (or creditworthy) over those wrongly predicted as such; (iii) a higher number of observations that are correctly predicted to be credit constrained (or creditworthy) among those with a higher predicted probability to be in that status (see the Appendix A6 for details).

Focusing on random forest predictions, Figure 2.4 (panels a and b), shows that the predicted probability of belonging to each status (i.e. being creditworthy and credit constrained) is strongly correlated with the actual rate. Although some caveats apply, these strong correlations speak in favor of the ability of the ML assignment rule to help the policy maker improving its targeting. As we argued in Section 3.1, even if we assume the current rule has the same ability as the ML rule to identify creditworthy firms, the higher the accuracy of ML predictions on firms' probability to be credit rationed, especially in presence of a skewed distribution of actual credit rationing probability, the more useful the ML rule would be for the policy maker.

Although we do not observe actual probabilities ($\beta_i^T$) but only true binary realizations of firms' rationing status ($RAT = [0, 1]$), and although our main interest lies in correctly predicting firms' probability to be rationed in absence of the public intervention ($RAT^{G=0}$), which we can correctly measure only for the subset of firms in our data that were not actually treated, we can still find some evidence in favor of the ML rule implementation. If we plot the predicted probability to be rationed against actual rationing rate for the subset of non-treated firms (Figure 2.5, panel a), we can see that most of the observations are centered around the $45°-$line, as it happens for the bigger picture elaborated on the data relative to the entire sample (Figure 2.4, panel b): there is a slightly higher dispersion for bins with an average predicted probability between $0.4-0.6$, which is not surprising if we consider that 'being rationed' is the complement-to-one result for a firm of being identified as 'worthy' by the banking system [16].

[16] The information set available to the policy-maker when making the prediction differs from the information set available to private banks - e.g. banks, in addition to the hard information observable by the policy-maker, might dispose of some soft information on potential borrowers - and it is plausible that this type of

Furthermore, looking at the distribution of $\hat{\beta}_i^{ML}$(Figure 2.5) we can see that, although there's a good concentration of observations around the mean (0.69), the distribution exhibits a long thin tail on the left and a short thick tail on the right with approximately 60% of the observations lying outside the $[-0.1; +0.1]$ interval around the mean $(0.6 - 0.8)$.

Figure 2.4 (panel c) shows that being creditworthy is correlated with being credit constrained, but there is large dispersion around this relation. The relationship between the predicted probability of being constrained and the predicted probability of being creditworthy is negative for groups of observations that show 'extreme' probabilities to be creditworthy: both for firms with an extremely low and extremely high probability to be creditworthy, the predicted probability of being constrained strongly declines as the probability of being creditworthy increases. The interpretation is straightforward for highly risky firms, which are nevertheless associated to high levels of credit constraints, while relates to the relative easiness of access to credit for firms that are less risky and most financially sound. However, for the rest of the distribution, the relation between the two probabilities is flatter, and there is large dispersion around each point. For this reason, a measure of default probability, alone, does not seem to provide enough information to also target firms that are credit constrained.

There might be different explanations as to why firms with the same risk are not equally credit constrained. One possibility is that there is true heterogeneity in credit rationing even for firms with the same risk. For instance, the heterogeneity can be due to the availability of other guarantees or different forms of collateral (which affect the banks' loss given default) that we are not able to observe at the time of application. A different level of credit constraints for firms with the same risk might also depend on banks' policies on risk diversification and on the amount of delegation in credit management, which could impose limits on specific firms, sectors or territories. An alternative explanation is that banks have more information than we do and, therefore, assess risk better. For any given probability of default predicted by us, some firms are actually riskier (and the banks know that), hence selecting as eligible only the credit-constrained firms may lead the GF to get the lemons [17]. This issue might hinder the ability of ML-targeting to improve effectiveness, and calls for the ex-post evaluation that we

information asymmetry is more relevant for firms whose predicted probability to be rationed on the basis of hard information is around 0.5.

[17] However, the opposite might also be true if weaker banks misallocate credit towards firms on the verge of bankruptcy (Schivardi et al. (2017)) or if banks favor connected firms (Barone et al. (2016)

provide in Section 5. Finally, the dispersion in credit constraints for same-risk firms might be due to measurement error. In this case, though, we should find no improvement in terms of effectiveness, hence, once again, the issue boils down to the ex-post evaluation in Section 5.

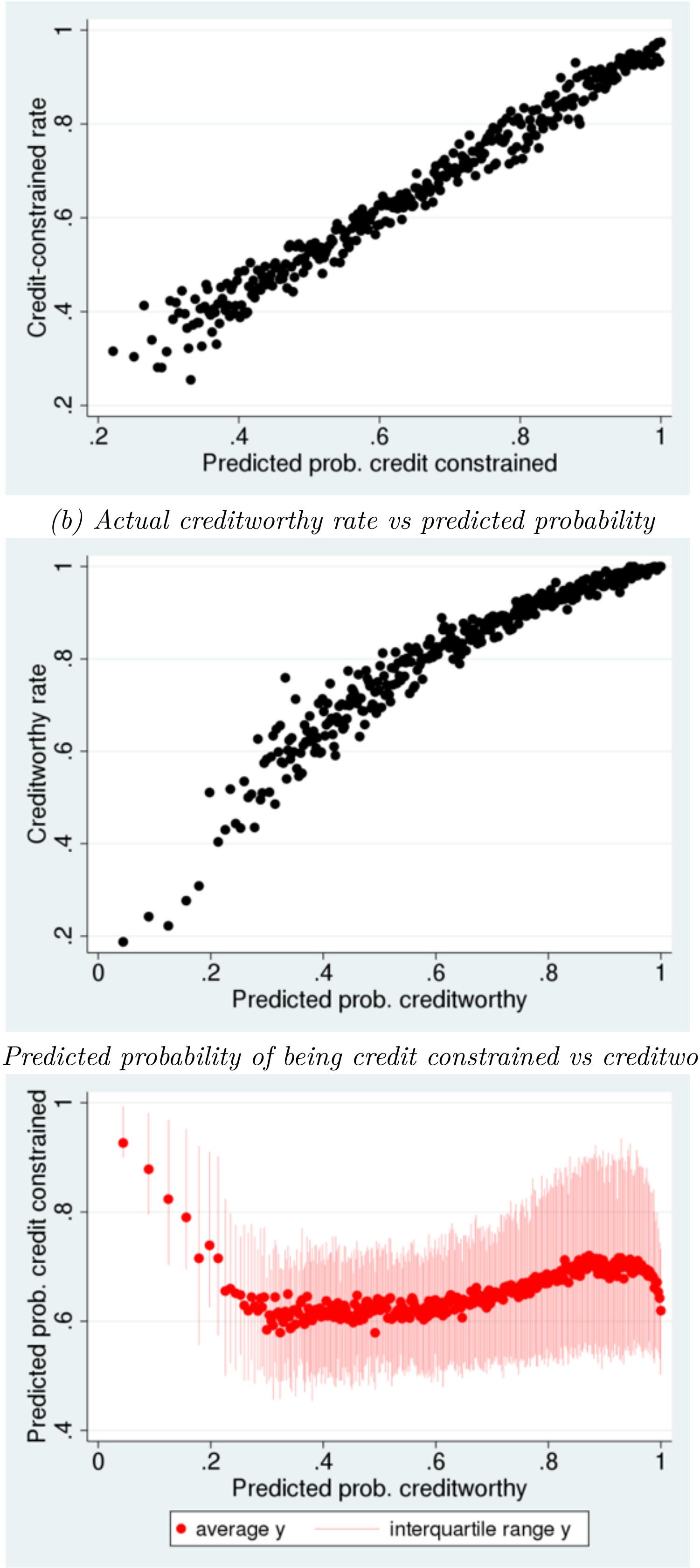
We therefore combine the two models to look at our final target: firms that are predicted to be both creditworthy and credit constrained. These are the firms whose forecasted probabilities for the two conditions are both at least equal to 0.5; for the rest of the paper, the "ML targeting rule" identifies the assignment rule that selects these firms. For this joint status, the misclassification error, reported in Table 2.1, is 36.8 per cent.

### 2.4.4 ML rule vs GF rule

We now use the 2012 hold-out sample to compare the GF rule, which evaluates whether a firm is eligible or not on the basis of the balance-sheet indicators described in Section 2.2, with the ML rule, based on random forest predictions. We consider only firms belonging to the sectors that are currently eligible for the GF scheme. We end up with a sample of about 90,000 firms (see the Appendix for details). In case of multiple loan applications from the same firm, we randomly choose only one so that the number of observations and firms coincide. In order to evaluate whether a firm is eligible to the Fund guarantee, we apply the GF scoring procedure to the firms in our dataset. It is worth noticing that we cannot perfectly mimic the GF rule as we do not have access to the firm original balance sheet data that were provided to the Fund but, instead, we observe less detailed reclassified balance sheet data (drawn from the Cerved archive; see de Blasio et al. (2018)). Notwithstanding this difficulty, we replicate the Fund eligibility mechanism fairly well (Table 2.2). Only about 2.3 per cent of the firms that received the Fund guarantee in 2012-13 are classified by us as not eligible when we replicate the GF rule on reclassified Cerved balance sheet data.

Table 2.3 compares the GF rule with the ML one. Overall, the ML rule is more selective than the GF one. Out of roughly 90.000 firms in our dataset, about 80 per cent of them would be selected by the ML targeting mechanism, while about 95 per cent are eligible according to the GF rule. In particular, the ML targeting would exclude about 20 per cent of the

Figure 2.4: Random forest predictions



*(a) Actual credit-constrained rate vs predicted probability*

*(b) Actual creditworthy rate vs predicted probability*

*(c) Predicted probability of being credit constrained vs creditworthy*

*Testing sample (2011). Each point represents one of 1,000 percentile bins of the variable on the x-axis.*

Figure 2.5: Predicted probability to be credit-rationed for non-treated firms

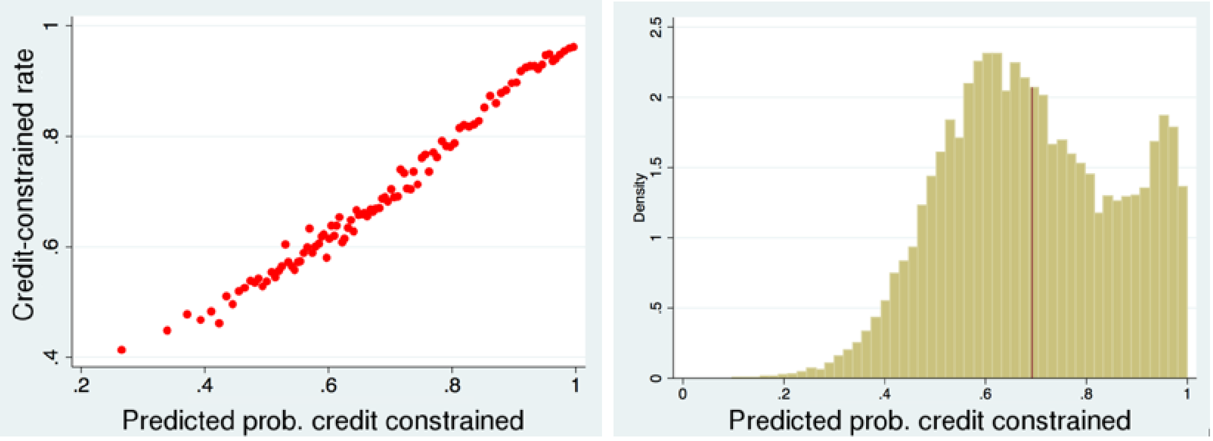


Table 2.1: Confusion matrix for the final target

| | Ypred = Not Target | Ypred = Target | Misclassification rate: 36.76% | |
|---|---|---|---|---|
| Yactual = Not Target | 17.822 | 22.779 | TN: 43.89% | FN: 21.81% |
| Yactual = Target | 11.451 | 41.047 | FP: 56.1% | TP: 78.18% |

*Notes.* Testing sample (2011). Yactual is 1 if the actual status is to be credit constrained and creditworthy, 0 otherwise; Ypred is 1 if a credit-constrained and creditworthy observation is predicted (predicted probability of each status ≥ 0.5), 0 otherwise. FP is the false positive rate computed as the percentage of observations predicted positive, but that are actually negative, over the total number of actually negative observations; TP is the true positive rate computed as the percentage of observations predicted positive, that are actually positive, over the total number of actually positive observations; FN is the false negative rate computed as the percentage of observations predicted negative, but that are actually true, over the total number of actually positive observations; TN is the true negative rate computed as the percentage of observations predicted negative, but that are actually negative, over the total number of actually negative observations.

Table 2.2: Replication of the GF screening mechanism

| | GF eligible (B) | | |
|---|---|---|---|
| GF beneficiary (A) | 0 | 1 | Total |
| 0 | 4.518 | 77.073 | 81.591 |
| 1 | 160 | 6.751 | 6.911 |
| Total | 4.678 | 83.824 | 88.502 |

*Notes.* 2012 sample, only firms belonging to the sectors that are currently eligible for the GF scheme. (A): firms that received (=1) or did not receive (=0) the Fund guarantee over the period 2012-13. (B): firms that are eligible (=1) or not eligible (=0) according to the actual Fund eligibility scoring mechanism, with the scoring procedure based on firm balance sheet data from Cerved group.

Table 2.3: GF eligibility vs ML targeting

| | Target (ML) firms (B) | | |
|---|---|---|---|
| Eligible (GF) firms (A) | 0 | 1 | Total |
| 0 | 1.174 | 3.504 | 4.678 |
| 1 | 16.860 | 66.964 | 83.824 |
| Total | 18.034 | 70.468 | 88.502 |

*Notes.* 2012 sample, only firms belonging to the sectors that are currently eligible for the GF scheme (A): firms that are eligible (=1) or not (=0) to the Fund guarantee according to the actual GF scoring mechanism. (B): firms that are selected as target (=1) or not (=0) by the ML algorithm (random forest).

Table 2.4: Characteristics of the Fund-eligible firms that are not targeted by ML

| | Constrained (B) | | |
|---|---|---|---|
| Credit-worthy (A) | 0 | 1 | Total |
| 0 | 874 | 4.395 | 5.269 |
| 1 | 11.591 | 0 | 11.591 |
| | 12.465 | 4.395 | 16.860 |

*Notes.* 2012 sample, only firms belonging to the sectors that are currently eligible for the GF scheme; subset of firms that are eligible according to the Fund rules but that are not targeted by our ML algorithm. (A): firms predicted as credit-worthy (=1) or not (=0) by the ML algorithm (random forest). (B): firms that predicted as constrained (=1) or not (=0) by the ML algorithm (random forest).

firms that are eligible according to the GF assignment mechanism (see Section 2.2). On the other hand, the ML rule would select about 75 per cent of the firms that are not eligible according to the GF rule. This evidence is in line with the rationale of the ML algorithm, which grounds eligibility on both creditworthiness and the actual need for external funds. As a result, GF eligible firms which have fair access to credit are not targeted by ML; on the other hand, firms that have low capacity to access credit, while still being creditworthy, are targeted by ML. Table 2.4 shows in detail the characteristics of the 16.860 firms that are eligible according to the GF but not selected by the ML algorithm. About 70 per cent of these firms are creditworthy but not constrained; about 25 per cent are constrained but not creditworthy, while only 5 per cent are neither creditworthy nor constrained.

In order to shed more light on the differences between the GF eligibility mechanism and our ML targeting rule, we consider the full set of about 90.000 firms in our dataset and estimate a simple linear model where the dependent variable y is a dummy taking value 1 if the firm is eligible according to the GF scoring mechanism and 0 otherwise. Our indepen-

Table 2.5: GF eligibility and ML predicted firm characteristics

| Dependent variable: Eligibility for the Fund | Coef. |
|---|---|
| ML predicted probability of being creditworthy | $0.2506003***$<br>(0.0044215) |
| ML predicted probability of being credit constrained | $-0.1186898***$<br>(0.0044105) |
| Manufacturing, construction, fishing and tradable sector | $0.0166327***$<br>(0.0015128) |
| Constant | $0.8238488***$<br>(0.0045158) |
| Observations<br>Adj R-squared | 88,502<br>0.0407 |

*Notes.* $***$ p-val $\leq 0.01$ . 2012 sample, only firms belonging to the sectors that are currently eligible for the GF scheme. Linear probability model. The dependent variable is binary, taking value=1 if firms are eligible for the guarantee according to the Fund's rules, and zero otherwise. Standard errors in parentheses. The predicted probabilities refer to the random forest model.

dent variables are: the ML predicted probability of being creditworthy; the ML predicted probability of being credit constrained; a dummy equal to 1 if the firm belongs to the sectors of manufacturing, construction and fishing and 0 if the firm belongs to the tradable sector, hospitality industry, transportation and other private service sectors. The results in Table 2.5 show a positive and statistically significant correlation between the probability of being eligible for the GF and that of being creditworthy. On the other hand, being credit constrained is negatively correlated with the probability of being eligible. These results strengthen our claim that the GF eligibility rule places too much weight on firms' creditworthiness, while neglecting credit rationing.

In Figure 2.6 we compare our ML predictions with the GF eligibility score s (see Section 2.2). The criteria for GF eligibility are met when s crosses the 0 threshold. Given that the GF scoring procedure essentially refers to the financial soundness of the firm, the eligibility score is positively correlated with our ML predicted probability to be creditworthy and negatively with the predicted probability to be credit constrained (panels a and b). Yet, even for high values of the GF score there is a sizable share of firms that are not predicted to be creditworthy according to ML, and vice versa. If we look at the predicted joint status of being worthy and constrained (panel c), we can notice that the association with the eligibility score is quite flat. Hence there are several firms that would meet both requirements of ML targeting but are far from the GF eligibility conditions.

## 2.5 Evidence on ML-targeting effectiveness: ex-post evaluation

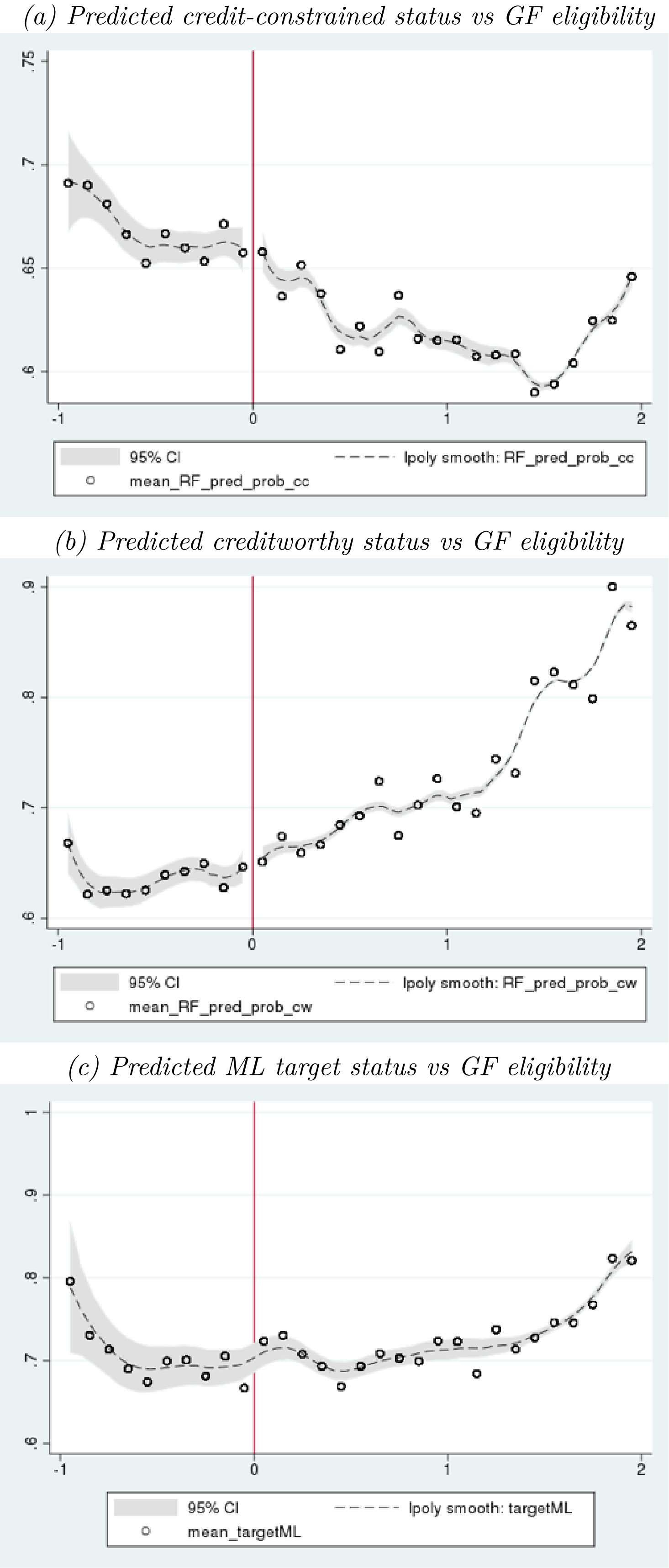
We now assess whether replacing the actual GF eligibility rule with our ML-based assignment mechanism would increase the impact of the policy. This is not warranted: the ML prediction will not automatically ensure higher program effectiveness insofar it fails to target the firms for which the impact is higher. We start (Subsection 5.1) by showing some crude comparisons between ML-targeted and non ML-targeted firms, among the GF beneficiary ones, in terms of financial and real outcomes. Next, in Subsection 5.2, we exploit the threshold for assignment implied under the GF rules and run an RDD experiment (as in de Blasio et al. (2018)), separately for ML-targeted and non ML-targeted groups of firms.

Before analyzing whether the ML-based targeting rule leads to an improvement in the effectiveness of the Fund, we first show how the actual recipients of the public guarantee (as well as the associated funds) are distributed across different ML-targeting groups. In our sample of roughly 90.000 firms, about 7.000 firms received the Fund guarantee in the years 2012-2013. Among them, about 4.000 firms (60%) are also selected as target by the ML algorithm, while 2.869 beneficiary firms are not selected (Table 2.6). Among the latter, about 70% are discarded because they are not predicted by ML as credit-constrained firms.

Table 2.7 shows that the amount of guarantees granted to non ML-targeted firms is 46.5 per cent of the total. On average, these firms are characterized by larger public-guarantee backed loans and larger guarantees (+26 per cent and +22 per cent, respectively). If non ML-targeted treated firms benefit less from the guarantee in terms of credit access, then nearly half of the guarantees have been misallocated.

Guarantee recipients that are not ML targets can be of three types: (1) not constrained

Figure 2.6: Probability of being an ML target vs actual GF eligibility



*(a) Predicted credit-constrained status vs GF eligibility*

*(b) Predicted creditworthy status vs GF eligibility*

*(c) Predicted ML target status vs GF eligibility*

*Notes.* 2012 sample, only firms belonging to the sectors that are currently eligible for the GF scheme (see Subsection 4.4). The x−axis is a continuous measure of GF eligibility (see de Blasio et al., 2018). Eligible firms have $x \geq 0$. The y-axis is the fraction with predicted status equal to 1 (random forest predicted probability $\geq 0.5$ in panels a and b and joint predicted status for panel c).

Table 2.6: ML targeted vs beneficiary firms

| | ML target (B) | | |
|---|---|---|---|
| GF beneficiary (A) | 0 | 1 | Total |
| 0 | 15.165 | 66.426 | 81.591 |
| 1 | 2.869 | 4.042 | 6.911 |
| Total | 18.034 | 70.468 | 88.502 |

*Notes.* 2012 sample, only firms belonging to the sectors that are currently eligible for the GF scheme. (A): firms that obtained the Fund guarantee in the period 2012-13. (B): firms predicted as target (=1) or not (=0) by the ML algorithm (random forest).

Table 2.7: Public resources to GF beneficiary firms by ML targeting status

| GF beneficiary | Amount financed (million Euros) | Guarantees (million Euros) | Firms | Average amount financed (thousand Euros) | Average guarantee (thousand Euros) |
|---|---|---|---|---|---|
| Non-ML target | 1.200,3 | 718,3 | 2.869 | 418,4 | 250,4 |
| of which: | | | | | |
| ► non CC, CW | 836,5 | 510,0 | 1.722 | 485,7 | 296,2 |
| ► CC, non CW | 258,7 | 147,0 | 852 | 303,7 | 172,5 |
| ► non CC, non CW | 105,1 | 61,3 | 295 | 356,4 | 207,8 |
| ML target | 1.335,6 | 828,0 | 4.042 | 330,4 | 204,9 |

*Notes.* 2012 sample, only firms that received the guarantee. CC=credit constrained; CW=credit worthy.

and creditworthy (60 per cent of 2,869 firms); (2) constrained and not creditworthy (30 per cent); (3) not constrained and not creditworthy (10 per cent). Within the non ML-targeted group, the guarantees are mostly channeled to group (1), attracting about 70 per cent of the financed amounts. Hence, the bulk of guarantees are directed towards firms that have, presumably, a good capacity to access credit. Although in this case the risk of not recovering the public guarantee is rather low, the large amount of public collateral assigned to not constrained firms could have been used for firms that face more difficulties in accessing credit. The remaining 30 per cent of guarantees channeled to non ML-targeted firms involves firms that are not creditworthy (either constrained or not constrained).

### Performance of GF beneficiary firms by ML-targeting status

In order to understand whether prioritizing ML-targeted firms could lead to an improvement, we compare the average observed performance of the ML-targeted vs non ML-targeted firms with respect to financial and real outcomes over the period 2009-15. The idea behind this exercise is that, if the group of ML-targeted firms performs better than the other group, the policymaker might increase the average effectiveness of the policy by simply excluding a subset of the Fund's currently eligible firms. This would correspond to the approach referred to as a 'contraction experiment' by Kleinberg et al. (2018).

The firms in our sample received the GF treatment in 2012 or 2013. We look, therefore, at their average performance in two sub-periods: before the GF treatment (2012 for some firms and 2013 for others) and after it. The comparisons can be made only for the subset of about 6,000 treated firms (out of about 7,000) for which we observe both financial and real outcomes in the entire period of interest [18].

As expected, in the period before receiving the GF guarantee, ML targeted firms experienced a lower growth of total bank loans, sales and investments with respect to those that were creditworthy but not credit constrained (see Table 2.8). In the subsequent period, which broadly overlaps with the sovereign debt crisis (leading to a severe credit crunch in Italy; see: Bank of Italy, 2014), both ML-targeted and not ML-targeted firms worsened their average

[18] These firms correspond to about 86 per cent of ML-targeted firms and 90 per cent of the non ML-targeted ones.

condition, but their performance was similar. This suggests that the guarantee had a stronger impact on credit constrained (though creditworthy) firms.

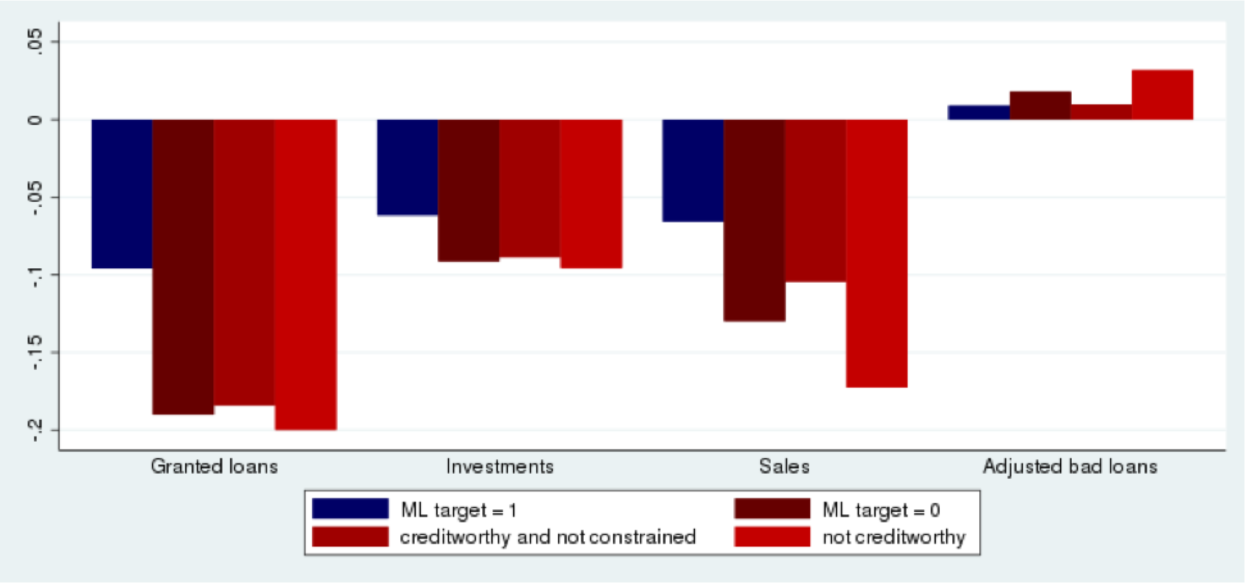
Table 2.8: Firms performance before and after receiving the GF

| | ML target = 1 | ML target = 0 | Difference | p-value | ML target = 0 | |
|---|---|---|---|---|---|---|
| | | | | | Creditworthy & not credit constrained | Not creditworthy |
| Granted loans | | | | | | |
| Before GF | 0.05 | 0.14 | -0.08 | 0.00 | 0.15 | 0.11 |
| After GF | -0.04 | -0.05 | 0.01 | 0.06 | -0.03 | -0.09 |
| Investments | | | | | | |
| Before GF | 0.04 | 0.09 | -0.05 | 0.00 | 0.10 | 0.08 |
| After GF | -0.02 | 0.00 | -0.02 | 0.01 | 0.01 | -0.02 |
| Sales | | | | | | |
| Before GF | 0.05 | 0.10 | -0.05 | 0.00 | 0.10 | 0.10 |
| After GF | -0.01 | -0.03 | 0.02 | 0.01 | -0.01 | -0.07 |
| Adjusted bad loans | | | | | | |
| Before GF | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 |
| After GF | 0.01 | 0.02 | -0.01 | 0.00 | 0.01 | 0.03 |
| Observations | 3,470 | 2,602 | | | 1,631 | 971 |

*Notes.* 2012 sample, only firms that received the guarantee and belong to the sectors that are currently eligible for the GF scheme. For each variable, the rows show the performance in the period before GF (from $t-3$ to $t-1$, where $t \in \{2012, 2013\}$ is the year in which the firm received the GF guarantee) and after GF (from t to 2015). For the variables Granted loans, Investments and Sales, the performance is measured as average annual growth rate; for the variable Adjusted bad loans is equal to 1 if the firm has bad loans in $t-2$ during the period before GF and in $t+2$ during the period after GF.

Figure 2.7 further highlights this suggestive evidence by plotting, for each kind of firm, the difference between the performance after receiving the guarantee and the one before. The

Figure 2.7: Change in firms' performance before and after receiving the GF



*Notes.* 2012 sample, only firms that received the guarantee and belong to the sectors that are currently eligible for the GF scheme. For each variable indicated on the x-axis, the bars plot the performance in the period from t to 2015 minus the performance in the period from $t-3$ to $t-1$, where $t \in \{2012, 2013\}$ is the year in which the firm received the GF guarantee. For the variables Granted loans, Investments and Sales, the performance is measured as average annual growth rate; for the variable Adjusted bad loans, the performance is measured as the difference between the probability that the firm has bad loans in $t+2$ minus the same probability in $t-3$. ML target $=1$ are firms targeted by our ML algorithm (combined prediction from random forest); ML target $=0$ are firms not targeted by the ML algorithm; creditworthy and not constrained are firms not targeted by ML that are creditworthy but not credit constrained; not creditworthy are firms not targeted by ML that are not credit constrained.

negative change in total bank loans is significantly smaller for ML targeted firms. Among non-targeted firms, it was similar for creditworthy but not constrained firms and for not creditworthy firms. However, while for the first group of firms the fall of total bank loans growth rate arguably reflects demand-side factors mostly, the negative change of bank loans characterizing the second group likely reflects supply-side factors, as such firms also exhibited a substantial increase in the probability of default, which implies a significant cost for the GF.

Although Table 2.8 and Figure 2.7 provide suggestive evidence of a greater GF impact on ML-targeted firms, we exploit a more credible identification strategy to estimate the GF treatment effect in the next Subsection.

### Evidence from RDD

In this Subsection we further investigate whether using the ML-based assignment leads to an improvement in policy effectiveness, by performing a RDD exercise. We follow de Blasio et al. (2018) and exploit the GF eligibility mechanism to estimate the impact of the guarantee using a fuzzy-RDD strategy, which allows for imperfect take-up of the treatment. As in their case, the compliance is imperfect both below and above the eligibility threshold. Above the threshold, we have eligible firms that have not applied to the GF, and eligible applicant firms eventually rejected by the GF . Below the cutoff, noncompliance is associated with the fact that we fail to successfully predict the eligibility status for firms using balance sheet data. The fuzzy-RDD identification critically rests on a discontinuity of the probability of treatment at the threshold, as well as on the absence of manipulation of the assignment variable (see Lee and Lemieux (2010)). We assess the the effectiveness of the GF separately for firms that are targeted or not by ML. A greater impact of the GF for the subgroup of ML-targeted firms would strengthen the previous results.

This analysis is conducted over a sample of about 59,000 firms (out of 88,502) for which we are able to observe a set of outcomes over the post-treatment years, up to 2015 (Table 2.9). This sample includes the above referred 6,000 firms which have benefited from the GF between 2012 and 2013. Figure 2.8 displays the density function of the continuous forcing variable for the full sample and the two subsamples (firms targeted by ML and firms not targeted by ML, respectively). The continuous forcing variable is based on the GF eligibility score as in de Blasio et al. (2018). The eligibility cutoff is set at zero: firms to the right of the cutoff are eligible, while firms to the left are not. In order to check whether possible manipulation of the assignment variable is at work, we test the continuity of the density functions at the cutoff for the full sample and the two sub-samples. We employ the test recently proposed by Cattaneo et al. (2018, 2019). As reported in Figure 2.8, we do not reject the null hypothesis of continuity in all cases. As expected (and necessary for identification), the probability of treatment jumps at the cutoff (Figure 2.9).

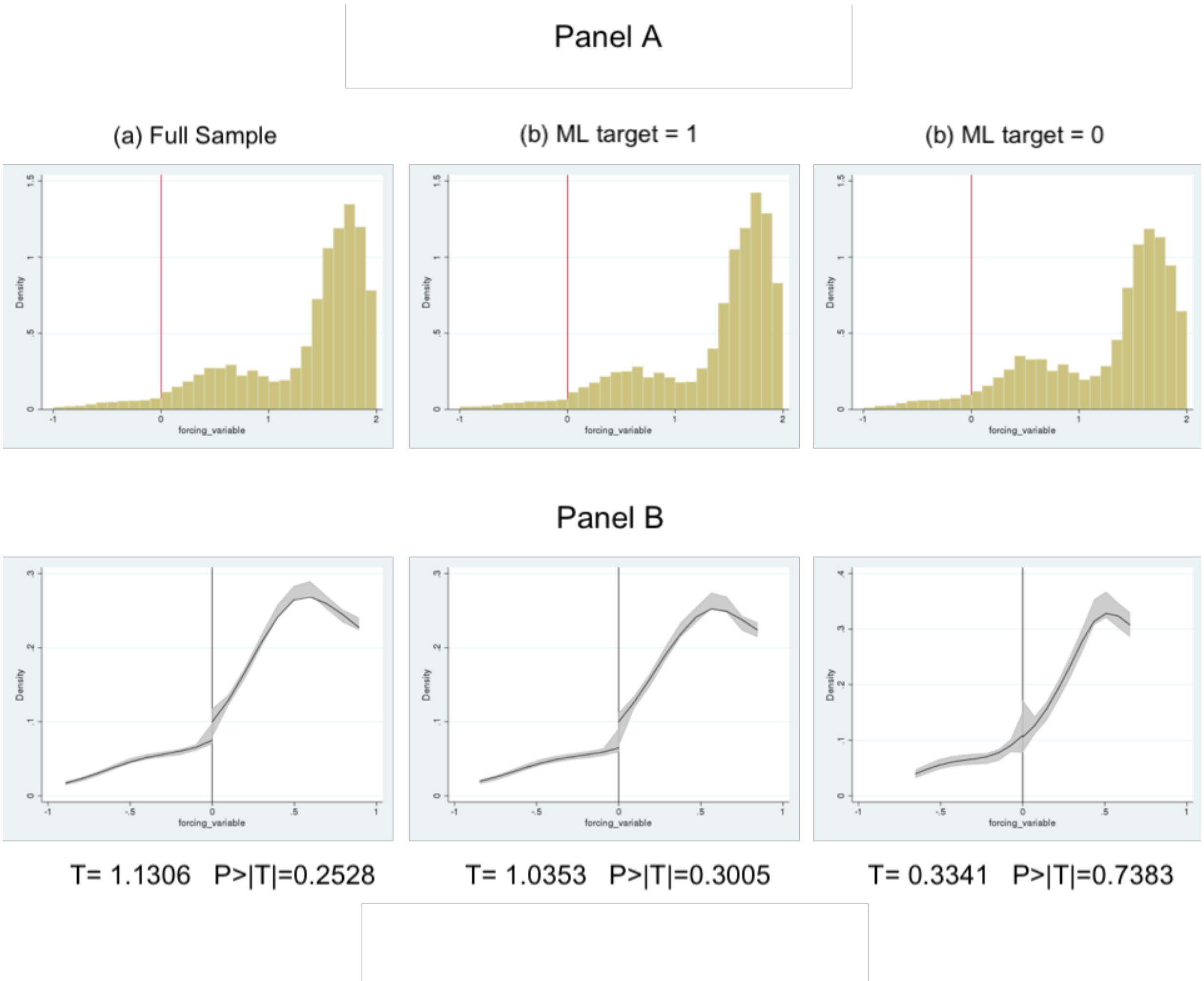
The non-parametric version of the fuzzy RDD corresponds to an instrumental variable estimation in a small neighborhood around the the discontinuity, where the eligibility for the

Table 2.9: Fuzzy-RDD analysis, sample

| | 1. Full sample | 1. ML target = 1 | 1. ML target = 0 |
|---|---|---|---|
| Treated | 6.072 | 3.470 | 2.602 |
| Not treated | 52.992 | 40.060 | 12.932 |
| All firms | 59.064 | 43.530 | 15.534 |

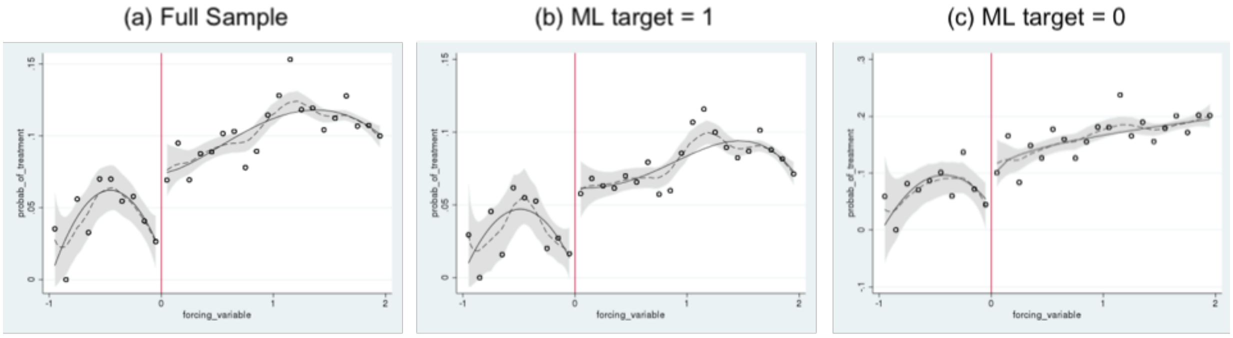
*Notes.* Selected sample of 59.064 firms.

Figure 2.8: Density function of the forcing variable



*Notes.* Selected sample of 59,064 firms. Panel A: density function of the forcing variable (a continuous measure of GF eligibility; eligible firms have $x \geq 0$). Panel B: manipulation tests using local polynomial density estimation (Cattaneo et al., 2018 and 2019). $H_0$ : $\lim_{x\uparrow\bar{x}} f(x) = \lim_{x\uparrow\bar{x}}(\bar{x})$. Under the appropriate assumptions, the test statistic T is distributed as a N(0,1). For each indicator, plots of the manipulation test (above) and test statistics (below) are provided.

Figure 2.9: Probability of treatment at the cutoff



*Notes.* Selected sample of 59,064 firms. The x-axis is the forcing variable (a continuous measure of GF eligibility; eligible firms have $x \geq 0$). The y-axis is the fraction of firms that are treated (i.e. GF beneficiary).

treatment, dichotomously defined over the discontinuity in the forcing variable, serves as an instrument for the actual treatment status Angrist and Pischke (2008). The Wald estimator, which is equal to the ratio of the intention-to-treat (henceforth ITT, numerator) over the first stage associated to compliance (henceforth FS, denominator), captures the causal effect of the treatment on compliers, defined as those whose treatment status changes as we move from a value of the forcing variabe just below the cutoff to a value just above (Local Average Treatment Effect for compliers, henceforth LATE).

To substantiate the assumption of randomization in a neighborhood around the eligibility cutoff, on which the fuzzy-RDD strategy is grounded, we perform a series of balancing tests using a set of covariates measured in pre-treatment years, which include the probability to be credit rationed and to be creditworthy. The results of the non parametric fuzzy-RDD estimates on the baseline covariates are reported in Table 2.10: overall, we find good balancing properties for the baseline covariates for the full sample and, separately, for both the subsamples defined according to the value of the ML targeting rule. In most of the cases in which we get significant ITT effects, these significant ITT effects do not translate into significant LATEs despite moderately powerful FS; the only exception concerns the variable "Pre-treatment bank granted credit" in the ML targeting = 0 subsample, where a strongly significant ITT does not translate into a significant LATE, possibly due to the weak FS. In a few other cases, instead, we do observe some significant ITT effects translating into significant LATEs: this happens for the covariates “Pre-treatment sales (growth rates)” and “Pre-treatment sales (levels)” in the estimation on whole sample, and the covariate "Pre-treatment sales (growth rates)” in the ML targeting = 1 subsample.

In order to get a clearer picture on the balancing properties of our baseline covariates we also run parametric RDD estimates on the three samples[19]. The results, reported in Table 2.11, confirm we have in general good balancing properties, showing an overall picture that is similar to what emerged from non-parametric estimates: only in a few cases, mostly related to the covariates “Pre-treatment sales (growth rate)” and “Pre-treatment bank granted credit (growth rate)”, we get significant ITTs that do translate into significant LATEs both in the whole sample and the two MLtarget subsamples.

To check for the possibility that some of these significant LATE results are due to random chance (Lee and Lemieux (2010)) we combine the multiple discontinuity tests into a single test statistic that measures whether our data support the random treatment hypothesis around the cutoff, testing the joint hypothesis that all discontinuity gaps in all the equations are equal to zero. According to the joint test, executed on the system that includes all base-line covariates' equations, we only marginally fail to reject the null hypothesis for the whole sample and the two subsamples [20], which suggests some of these covariates might not actually be perfectly balanced.

To control for the possible unbalancedness of some of the baseline covariates, on top of the non-parametric Fuzzy-RDD estimates, we also provide the parametric Fuzzy-RDD estimates for our outcome variables of interest: parametric estimates allow us to control, within each sample, for the covariates for which we obtained a significant LATE and ITT in Table 2.11, namely: [21]:

[19] The choice of the polynomial degree to be used in the global parametric estimates relies on a comparative procedure carried out separately for the whole sample and the two subsamples MLtarget=0 and MLtarget=1, where the best model specification is selected based on model fit metrics (mainly the AIC). The analysis of model fit accuracy is based on Second-Stage equations: based on these results we select the polynomial degree to be applied to each covariate in each sample and for consistency the same specification is applied also to Intention-To-Treat and First-Stage equations estimated on the same sample.

[20] The F-statistic associated to the joint test is equal to 2.23 (p-value=0.0226) in the whole sample and to 2.53 and 2.44 (p-values: 0.009 and 0.0124) in the two MLtarget=1 and MLtarget=0 subsamples, respectively.

[21] We largely fail to reject the null hypothesis that all discontinuity gaps are jointly on the remaining baseline covariates: The F-stats associated to the joint test is equal to 0.85 (p-value=0.5285) in the whole sample and to 1.59 and 1.70 (p-values: 0.1463 and 0.1160) in the two MLtarget=1 and MLtarget=0 subsamples, respectively.

- Pre-treatment sales (growth rate) and Prob. of adjusted non-performing loans from the whole sample;
- Pre-treatment sales (growth rate) and Pre-treatment bank granted credit (growth rate) from the MLtarget=1 subsample;
- Pre-treatment bank granted credit (growth rate) and Prob. of being credit constrained from the MLtarget=0 subsample;

Non parametric estimates of the impact of the GF are reported in Table 2.12. Parametric estimates of the impact of the GF are reported in Table 2.13 [22]. The results we obtain from the two estimation procedures are qualitatively the same.

We consider several outcome variables: granted bank loans, sales, investments, probability of adjusted bad loans. Since our sample includes firms that received the GF in 2012 and 2013, outcome variables such as granted bank loans, investments and sales are expressed in terms of their average growth rate in the period 2012-15, or 2013-15 according to the year of treatment. The same averages are computed for non-treated firms, with the initial year being randomly assigned and the proportion of 2012s being the same as that of the treated firms. The variable named "Adjusted bad loans" is, instead, a dummy equal to 1 if the firm has adjusted bad loans in 2015 and 0 otherwise.

The first column displays the results for the full sample, while the second and the third report the estimates related to the sample of ML-targeted firms and ML-non-targeted firms, respectively. As this exercise aims to test the potential heterogeneous effects of the ML targeting rule that we propose, we do not elaborate on the full sample estimates. We, therefore, focus on the results of columns 2 and 3, which display the Fuzzy-RDD estimates carried out separately in the two samples of firms, split according to our ML targeting rule. In line with our previous descriptive findings, the impact of GF on (the growth rate of) granted bank loans is positive and statistically significant for the sample of ML-targeted firms, while no effect at all is detected in the sample of non ML-targeted firms. The fuzzy-RDD estimates in both the subsamples show no significant impact of the GF on adjusted bad loans. This result is reassuring, as it shows that the greater growth of bank loans, reached via ML-targeting, does

The results are qualitatively the same if we run the test after winsorizing outliers below the 5th or above the 95th percentile in each covariate.

[22] For comparison, we also report in the Appendix (Additional Tables section) the results of the Fuzzy-RDD parametric estimation obtained without controlling for covariates.

not undermine the financial sustainability of the GF. Turning to real outcomes, no impact is detected on investments and sales over the first two years following the issuance of the guarantee. The latter result was largely expected, as the years considered in our sample are characterized by a persistent depressed growth of investments in Italy.

## 2.6 Pitfalls and implementation issues

### 2.6.1 Prediction bias when the policy is already in place

One issue pertaining to the comparison with the actual GF rule is that we estimated and validated the ML models on years during which the Fund was already operational. For this reason, the dataset is "contaminated" as it also contains firms that already received the Fund guarantee. Our actual aim is to predict the credit constrained and creditworthy conditions in the counterfactual scenario without the guarantee (we define both of them as a binary variable $S_0$), but for some of the firms we actually observe these conditions in the scenario with the guarantee ($S_1$). If the guarantee has an impact on constraints and default rates, then $S_1$ and $S_0$ are different. Our algorithm has been trained and evaluated to predict the observed status, which is a combination of the two counterfactuals, because what we observe is $S_0 \cdot (1 - GF) + S_1 \cdot GF$, where $GF = 1[guarantee]$.

In general, this implies that the predictive power with respect to $S_0$ is lower, and therefore the true misclassification error is larger. This should translate in lower gains from ML targeting. The contamination may even lead to exclude groups for which the guarantee is actually extremely successful. For instance, if the policy fully removes the credit constraints of a specific group with initial high $\Pr(S_0)$ (for instance, very small firms), then the ML algorithm may end up predicting that that group actually has very low credit constraints, and therefore should not be targeted.

This issue boils down to the question of gains from ML targeting, as discussed in Subsections 5.1 and 5.2. If the contamination problems annihilate the predictive power of the algorithm, or even make it worse (with respect to $S_0$), than a random classifier by excluding relevant groups, then we should find that the estimated impact of the guarantee is not larger (or even smaller) in the ML targeted group. Our results from Subsections 5.1 and 5.2 show the

Table 2.10: Non-Parametric Fuzzy-RDD analysis: Balancing properties (Part 1 of 2)

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel A. Pre-treatment bank granted credit (growth rate) | | | |
| ITT | 0.011<br>(0.009) | -0.008<br>(0.010) | 0.057***<br>(0.019) |
| FS | 0.033***<br>(0.011) | 0.031***<br>(0.010) | 0.042*<br>(0.025) |
| LATE | 0.336<br>(0.305) | -0.250<br>(0.349) | 1356<br>(0.899) |
| Panel B. Pre-treatment investments (growth rate of fixed assets) | | | |
| ITT | -0.010<br>(0.009) | -0.013<br>(0.011) | 0.001<br>(0.019) |
| FS | 0.029***<br>(0.009) | 0.025***<br>(0.009) | 0.038*<br>(0.023) |
| LATE | -0.360<br>(0.348) | -0.499<br>(0.479) | 0.029<br>(0.516) |
| Panel C. Pre-treatment sales (growth rate) | | | |
| ITT | 0.046***<br>(0.007) | 0.055***<br>(0.008) | 0.023*<br>(0.012) |
| FS | 0.032***<br>(0.010) | 0.032***<br>(0.011) | 0.040*<br>(0.021) |
| LATE | 1.451***<br>(0.553) | 1.734**<br>(0.688) | 0.570<br>(0.436) |
| Panel D. Prob. of adjusted non-performing loans | | | |
| ITT | 0.013<br>(0.007) | 0.013*<br>(0.007) | 0.005<br>(0.015) |
| FS | 0.034***<br>(0.010) | 0.030***<br>(0.009) | 0.044*<br>(0.022) |
| LATE | 0.372<br>(0.255) | 0.442<br>(0.274) | 0.106<br>(0.346) |
| Panel E. Pre-treatment bank granted credit (level) | | | |
| ITT | -0.049<br>(0.074) | 0.076<br>(0.086) | -0.202*<br>(0.120) |
| FS | 0.040***<br>(0.011) | 0.037***<br>(0.012) | 0.052**<br>(0.025) |
| LATE | -1.234<br>(-1.920) | 2.059<br>(-2.438) | -3.905<br>(-2.966) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD non parametric estimates. The optimal bandwidth was retrieved by Imbens and Kalyanaraman (2012) procedure. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets.

Table 10: Non-Parametric Fuzzy-RDD analysis: Balancing properties (Part 2 of 2)

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel F. Pre-treatment sales (level) | | | |
| ITT | -0.171***<br>(0.061) | -0.122<br>(0.075) | -0.199*<br>(0.102) |
| FS | 0.040***<br>(0.012) | 0.042***<br>(0.013) | 0.046*<br>(0.024) |
| LATE | -4.234**<br>(-1.981) | -2.941<br>(-2.044) | -4.352<br>(-3.237) |
| Panel G. Prob. of being credit constrained | | | |
| ITT | -0.002<br>(0.005) | -0.001<br>(0.005) | -0.023*<br>(0.012) |
| FS | 0.034***<br>(0.010) | 0.033***<br>(0.010) | 0.050**<br>(0.024) |
| LATE | -0.064<br>(0.152) | -0.033<br>(0.157) | -0.455<br>(0.315) |
| Panel H. Prob. of being creditworthy | | | |
| ITT | -0.003<br>(0.008) | -0.010<br>(0.009) | 0.003<br>(0.016) |
| FS | 0.041***<br>(0.012) | 0.040***<br>(0.013) | 0.057**<br>(0.024) |
| LATE | -0.066<br>(0.213) | -0.239<br>(0.254) | 0.055<br>(0.287) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD non parametric estimates. The optimal bandwidth was retrieved by Imbens and Kalyanaraman (2012) procedure. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets.

Table 2.11: Parametric Fuzzy-RDD analysis: Balancing properties (Part 1 of 2)

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel A. Pre-treatment bank granted credit (growth rate) | | | |
| ITT | -0.0029<br>(0.008) | 0.0204***<br>(0.006) | 0.0355***<br>(0.010) |
| FS | 0.0153<br>(0.013) | 0.0288***<br>(0.009) | 0.039**<br>(0.018) |
| LATE | -0.1864<br>(0.562) | 0.709**<br>(0.298) | 0.9036*<br>(0.468) |
| Panel B. Pre-treatment investments (growth rate of fixed assets) | | | |
| ITT | -0.0076<br>(0.006) | -0.0128**<br>(0.007) | -0.0005<br>(0.010) |
| FS | 0.0381***<br>(0.008) | 0.0282***<br>(0.008) | 0.0413**<br>(0.018) |
| LATE | -0.1999<br>(0.152) | -0.4535*<br>(0.269) | -0.0118<br>(0.252) |
| Panel C. Pre-treatment sales (growth rate) | | | |
| ITT | 0.0109**<br>(0.004) | 0.0134***<br>(0.004) | 0.0018<br>(0.007) |
| FS | 0.0355***<br>(0.008) | 0.0266***<br>(0.009) | 0.0396**<br>(0.019) |
| LATE | 0.3064**<br>(0.125) | 0.5045**<br>(0.236) | 0.0451<br>(0.174) |
| Panel D. Prob. of adjusted non-performing loans | | | |
| ITT | -0.0059**<br>(0.002) | -0.0016<br>(0.003) | -0.0106<br>(0.008) |
| FS | 0.0380***<br>(0.008) | 0.0291***<br>(0.008) | 0.0329<br>(0.027) |
| LATE | -0.1541**<br>(0.069) | -0.0559<br>(0.093) | -0.3232<br>(0.348) |
| Panel E. Pre-treatment bank granted credit (level) | | | |
| ITT | 0.0207<br>(0.057) | -0.0556<br>(0.043) | -0.0372<br>(0.059) |
| FS | 0.0178<br>(0.013) | 0.0307***<br>(0.009) | 0.0350**<br>(0.018) |
| LATE | 1.1605<br>(-3.247) | -1.8105<br>(-1.531) | -1.0638<br>(-1.790) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD parametric estimates. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets. The selection of the best polynomial degree for the RDD global parametric estimate is based on the Akaike Information Criterion (AIC) and the same polynomial degree specification is applied to 1st stage, 2nd stage and ITT regressions.

Table 11: Parametric Fuzzy-RDD analysis: Balancing properties (Part 2 of 2)

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel F. Pre-treatment sales (level) | | | |
| ITT | -0.0670<br>(0.054) | -0.017<br>(0.067) | 0.0615<br>(0.055) |
| FS | 0.0247*<br>(0.014) | 0.0244<br>(0.015) | 0.0364**<br>(0.018) |
| LATE | -2.7141<br>(-2.708) | -0.6975<br>(-2.782) | 1.6914<br>(-1.727) |
| Panel G. Prob. of being credit constrained | | | |
| ITT | 0.0053<br>(0.004) | 0.0050<br>(0.004) | -0.0089*<br>(0.005) |
| FS | 0.0164<br>(0.013) | 0.0160<br>(0.014) | 0.0393**<br>(0.019) |
| LATE | 0.3228<br>(0.392) | 0.3117<br>(0.400) | -0.2275<br>(0.166) |
| Panel H. Prob. of being creditworthy | | | |
| ITT | 0.0043<br>(0.006) | 0.0018<br>(0.006) | 0.0067<br>(0.014) |
| FS | 0.0198<br>(0.013) | 0.0168<br>(0.013) | 0.0403<br>(0.034) |
| LATE | 0.2147<br>0.334 | 0.1090<br>(0.377) | 0.1665<br>(0.399) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD parametric estimates. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets. The selection of the best polynomial degree for the RDD global parametric estimate is based on the Akaike Information Criterion (AIC) and the same polynomial degree specification is applied to 1st stage, 2nd stage and ITT regressions.

Table 2.12: Non-Parametric Fuzzy-RDD analysis: Outcome variables

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel A. Bank granted credit (growth rate) | | | |
| ITT | 0.014**<br>(0.006) | 0.016**<br>(0.008) | 0.009<br>(0.012) |
| FS | 0.037***<br>(0.009) | 0.037***<br>(0.009) | 0.044**<br>(0.021) |
| LATE | 0.365*<br>(0.205) | 0.435*<br>(0.241) | 0.208<br>(0.310) |
| Panel B. Investments (growth rate of fixed assets) | | | |
| ITT | -0.004<br>(0.008) | -0.007<br>(0.009) | 0.006<br>(0.013) |
| FS | 0.035***<br>(0.010) | 0.031***<br>(0.010) | 0.047**<br>(0.019) |
| LATE | -0.108<br>(0.245) | -0.238<br>(0.322) | 0.116<br>(0.292) |
| Panel C. Sales (growth rate) | | | |
| ITT | 0.018***<br>(0.006) | 0.015**<br>(0.007) | 0.023<br>(0.014) |
| FS | 0.039***<br>(0.010) | 0.025**<br>(0.010) | 0.061**<br>(0.025) |
| LATE | 0.465**<br>(0.216) | 0.613<br>(0.391) | 0.377<br>(0.281) |
| Panel D. Prob. of adjusted bad loans | | | |
| ITT | -0.001<br>(0.007) | 0.010<br>(0.006) | -0.023<br>(0.016) |
| FS | 0.032***<br>(0.009) | 0.032***<br>(0.010) | 0.044**<br>(0.019) |
| LATE | -0.024<br>(0.237) | 0.317<br>(0.239) | -0.516<br>(0.462) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD non parametric estimates. The optimal bandwidth has been retrieved by Imbens and Kalyanaraman (2012) procedure. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets.

Table 2.13: Parametric Fuzzy-RDD analysis: Outcome variables

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel A. Bank granted credit (growth rate) | | | |
| ITT | 0,012***<br>(0,004) | 0,0154***<br>(0,005) | -0,002<br>(0,011) |
| FS | 0,038***<br>(0,008) | 0,0301***<br>(0,009) | 0,029<br>(0,028) |
| LATE | 0,321***<br>(0,124) | 0,512***<br>(0,216) | -0,056<br>(0,380) |
| Panel B. Investments (growth rate of fixed assets) | | | |
| ITT | -0,008*<br>(0,005) | -0,007<br>(0,006) | -0,001<br>0,014 |
| FS | 0,036***<br>(0,008) | 0,029***<br>(0,009) | 0,038<br>(0,028) |
| LATE | -0,216<br>(0,139) | -0,248<br>(0,206) | -0,030<br>(0,358) |
| Panel C. Sales (growth rate) | | | |
| ITT | -0,002<br>(0,003) | -0,004<br>(0,004) | 0.004<br>(.009) |
| FS | 0,036***<br>(0,008) | 0,026***<br>(0,009) | 0,045<br>(0,027) |
| LATE | -0,042<br>(0,090) | -0,139<br>(0,158) | .091<br>(0,210) |
| Panel D. Prob. of adjusted bad loans | | | |
| ITT | -0,002<br>(0,003) | 0,002<br>(0,003) | -0,010<br>(0,007) |
| FS | 0,037***<br>(0,008) | 0,028***<br>(0,008) | 0,040**<br>(0,017) |
| LATE | -0,045<br>(0,072) | 0,085<br>(0,094) | -0,258<br>(0,194) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD parametric estimates. The polynomial degree is the same chosen for the main parametric estimation where we do not control for covariates, reported in the Appendix. The same polynomial degree specification is applied to 1st stage, 2nd stage and ITT regressions. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets.
Covariates: In the parametric estimations on the whole sample we control for the growth rate of pre-treatment sales. In the parametric estimations on the MLtarget=1 subsample we control for the growth rates of pre-treatment sales and granted bank credit. In the parametric estimations on the MLtarget=0 subsample we control for the growth rate of pre-treatment granted bank credit and the probability of being credit-constrained.

contrary and, therefore, ease the concerns about the impact of contamination.

### Transparency and manipulability

In principle, each prediction model can be used to assess whether a single firm is ML target or not, on the basis of the characteristics that are observable at the time of the application for the GF. However, the models differ in terms of transparency. Our favorite ML prediction model (the random forest) is more of a black box, as it does not provide an easily interpretable decision rule. Being an average across a large set of estimated decision trees, the prediction cannot be interpreted by simply looking at thresholds across different variables. Although some measures of variable importance are available (see Appendices A.4 and A.5), there is not a simple rule linking each observable characteristic to the final prediction. This might be a concern for a policymaker who favors transparency. Furthermore, it might lead to raise issues of discrimination, because firms cannot easily understand why they have been excluded (Athey (2017)). The prediction provided by the decision tree is, instead, the most transparent one, as the final selection rule involves looking at relatively few variables and comparing them to specific thresholds. This could be more easily communicated as it resembles most of the ordinary policy allocation rules. The LASSO model should be more interpretable, but the presence of interactions makes it less so. Furthermore, the final prediction for LASSO depends on a linear index of a large set of covariates (see the Appendix), and therefore it is not simple to evaluate which characteristics determine whether a firm is eligible or not.

The main trade-off in choosing a simpler algorithm is in terms of accuracy. As already argued in Subsection 4.3, random forest performs better than the other methods, decision tree included, in out-of-sample prediction. This is particularly true for the creditworthy status. Furthermore, we have mentioned that random forest predicts a higher number of observations that are correctly predicted to be credit constrained (or creditworthy) among those with a higher predicted probability to be in that status. This is even more evident if one looks at shares rather than absolute numbers. Figures 2.10( panel a and b) show the fraction of applications correctly classified if we were to pick only the top x per cent in terms of predicted probability. Indeed, the policy maker may want to include as eligible only the top x per cent of firms with the highest probability of belonging to each of the two status. It can be noticed that the decision tree, being overly simple, is not able to discriminate among the highest predicted probability. To decide which methods to implement, a decision maker should trade-off transparency with the amount of misclassification that arises with the chosen rule.

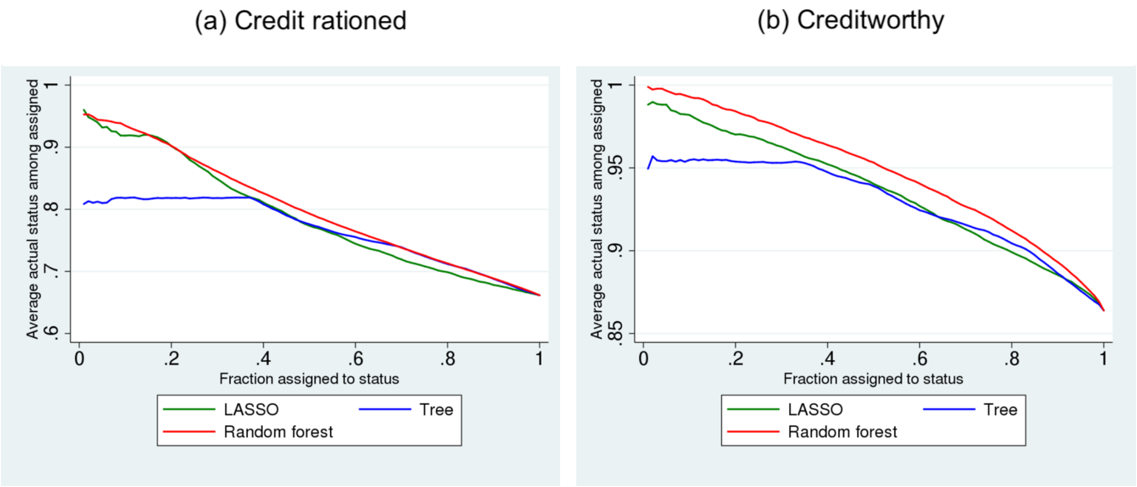
Another possible advantage of a simpler method is the amount of information required. The decision tree, as highlighted in Figures A6.1 and A6.2 in the Appendix, would require relatively few variables. On the other hand, the random forest model requires a large set of covariates. This is potentially costly, although it should not be forgotten that all the information we used is digitalized in administrative databases and that the GF administrators, and, more in general, policymakers, have access to all the information that we use for our predictions.

The trade-off between accuracy on the one hand, and the transparency and information burden on the other, might therefore lead a policymaker to choose a simpler model. Nevertheless, for our application we aim at improving the existing eligibility criteria. With respect to the random forest model, the GF scoring is possibly easier to assess, as it is based on few budget indices and thresholds that can be summarized in an Excel spreadsheet. In terms of formal transparency, therefore, the GF rule is more interpretable. However, there is another dimension of transparency, which we can call substantive transparency. This dimension concerns the accountability of the policy maker for accomplishing her mission, which is using public money in an effective way. In this respect, our random forest algorithm might be preferable, because the gains in terms of effectiveness associated with ML targeting are substantial.

Finally, if we look at the actual implementation, we should not neglect that a simple rule allows each firm and bank to assess eligibility before formally applying for the guarantee. This facilitates the process but hinders the ability of the GF board to assess how many firms would be interested but are not eligible. This information might be particularly useful for ex-post evaluation of the GF performance. A more complicated rule, which requires an assessment made through an online platform (after registration) or by the GF itself (after application), may allow a perhaps independent evaluator to focus only on interested firms and use as a control group those that were excluded because they were not eligible.

Another important implementation issue concerns manipulability. Ex-post, when the rule has been defined and made public, applicant firms might alter their variables in order to access the guarantee. This can be done at different levels. The first is to misreport information in

Figure 2.10: Fraction in the actual status in the x-fraction with highest predicted probability



*Notes.* Testing sample (2011). On the horizontal axis the percentage of observations classified as positive, choosing first those with the highest predicted probability (and, therefore, assigned to status). In cases in which multiple observations have the same predicted probability, we chose among them randomly. On the vertical axis the fraction of true positive cases over those classified as positive.

the application, but as we use data recorded in digitalized archives, we believe this is a minor risk. The second is to alter the variables reported in the archive. This, however, involves fraud and implies a strong legal risk for the applicant. The third is to make some (possibly costly) financial adjustments aimed at meeting the eligibility criteria. We believe this is possible, but this risk is equally shared by the GF rule, over which we aim to improve. As the random forest eligibility rule is even more of a black box, we find it hard for an applicant to carry out this operation. Manipulability can also be an issue ex-ante, where firms behave strategically to alter the variables that we use as proxies for the credit-constrained and creditworthy status. This seems more relevant with respect to the preliminary information system, where requests for access to the system (that we use as a proxy for credit requests) may be performed to alter the dataset and therefore the estimated algorithms. However, individual firms have hardly any chance of influencing the estimates by filling out a loan application (which, in turn, might lead to a request for preliminary information by the bank) when the loan is not needed. Each request counts as one (in a very large dataset) and we also aggregate multiple requests in the same quarter.

### Additional policy objectives

Our ML targeting rule is trained with the aim of increasing the GF effectiveness in raising bank loan availability and reducing the share of loans that go into default. However, any targeting rule, including the current one, might end up having other effects ('omitted payoffs', see Kleinberg et al. (2018)), which might or might not be desirable.

Given that the GF fund was strongly advocated as a counter-measure for the recession, we examine two important issues. The first is whether the rule tends to favor or not firms in disadvantaged territories that have been strongly hit by the crisis, mostly in the Southern regions. The second is whether the fund tends to flow to banks with certain characteristics, such as being part of a group and having a variety of funding sources. Table 12 shows the correlation between a set of pre-treatment firm characteristics (main bank belonging to a group, number of lending banks, funding gap of the main bank, firm headquarters in Southern Italy) and, in turn, the GF eligibility rule (dummy equal to 1 if the firm is eligible) and the ML targeting rule (dummy equal to 1 if the firm is targeted by ML).

GF eligibility tends to have a bias in favor of firms whose main reference bank (in terms of granted credit) belongs to a group or in favor of firms that have already taken out loans with several banks. Conversely, our favorite ML eligibility is negatively correlated with the firm being more indebted towards a bank that belongs to a group and tends to prioritize those with few lending relationships. Moreover, ML targeting seems to favor firms whose main bank has a lower funding gap, i.e. its funding source mainly consists of households' deposits. In terms of regional differences, GF eligibility is negatively correlated with the firm being located in the South of Italy, where firms generally face more difficulties in accessing credit, while the opposite holds for ML eligibility. The former, therefore, seems to favor more developed areas.

These correlations illustrate that, despite being focused on specific issues, each targeting rule might end up prioritizing firms with certain characteristics. This might satisfy additional (omitted) payoffs that the policymaker has in mind, or even work against them. Nevertheless, Kleinberg et al. (2018b) argue that the presence of other policy objectives should not change the way the algorithm is designed, but rather the way in which predictions are employed in

Table 2.14: GF and ML eligibility and omitted payoffs

| | Y = main bank belongs to a group (pre-treatment) | | Y = number of funding banks (pre-treatment) | | Y = funding gap of the main bank (pre-treatment) | | Y = firm head-quartered in Southern Italy | |
|---|---|---|---|---|---|---|---|---|
| | Eligible | ML target | Eligible | ML target | Eligible | ML target | Eligible | ML target |
| Eligible | 0.0319$^{***}$ | | 0.486$^{***}$ | | 0.906 | | $-0.0288^{***}$ | |
| | (0.0124) | | (0.0427) | | (0.9930) | | (0.0084) | |
| ML Target | | 0.000773 | | $-2.057^{***}$ | | $-1.571^{***}$ | | 0.0194$^{***}$ |
| | | (0.0061) | | (0.0422) | | (0.6070) | | (0.00511) |
| Manuf. & constr. sectors | 0.0137$^{*}$ | 0.0143$^{*}$ | 0.967$^{***}$ | 0.853$^{***}$ | 0.532 | 0.454 | $-0.0206^{***}$ | $-0.0200^{***}$ |
| | (0.0081) | (0.0081) | (0.0451) | (0.0422) | (0.6810) | (0.6770) | (0.0043) | (0.0044) |
| Constant | 0.745$^{***}$ | 0.774$^{***}$ | 2.567$^{***}$ | 4.636$^{***}$ | 14.13$^{***}$ | 16.22$^{***}$ | 0.213$^{***}$ | 0.171$^{***}$ |
| | (0.0427) | (0.0425) | (0.0484) | (0.0610) | (3.1710) | (3.2050) | (0.0210) | (0.0166) |
| Observations | 72.300 | 72.300 | 72.470 | 72.470 | 63.299 | 63.299 | 72.470 | 72.470 |

*Notes.* $***$ p-val $\leq$ 0.01) , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1. Robust standard errors in brackets. 'Eligible': binary variable identifying firms that are eligible to the GF; 'ML Target': binary variable identifying firms that are targeted by ML.

the final decision rule. For instance, even if the trained model favors a specific geographical area, the policy maker may desire a more uniform distribution across regions. To achieve this, she should select firms by setting region-specific thresholds of the predicted probability to be target, in order to re-balance the composition of eligible firms across areas. In this way fairness and efficiency are addressed separately. If instead we force the algorithm to give predictions that are orthogonal to geography (or other variables) we may hurt (prediction) efficiency without necessarily improving on fairness.

## 2.7 Conclusion

Gains from ML targeting seem to be relevant. Using the current GF selection mechanism, around 47 per cent of the guarantees (approximately 1,2 billion Euros) went to firms that are not ML targets and showed smaller benefits in terms of access to credit.

We have shown that ML algorithms also come with downsides in terms of transparency and administrative burden. The GF rule might seem formally less opaque, but it fails to be accountable with regard to explaining how it was designed and whether it meets the policy goal of facilitating access to credit for firms that are financially sound, but credit constrained. Hence, it is not clear whether we would lose transparency by using, instead, an ML algorithm trained on data and fully evaluated. ML-based rules also come with a stronger informational requirement, but the development of administrative archives has greatly reduced the cost of recovering this information. In fact, the variables necessary to calculate the predicted ML-status for the single firm are available upon request to the GF management.

While our prediction exercise was framed within the GF operations, it has a more general relevance. The prediction of creditworthy firms is also important for private banks. Credit scoring models are already often based on ML algorithms but, since these models are proprietary, we are not in a position to compare our predictions to those. Our ML algorithms might also be useful for supervisory purposes, to double check the accuracy of private forecasts. The prediction of credit-constrained firms is probably even more important from the point of view of aggregate welfare. Knowing who the creditworthy but constrained firms are is important for designing the public interventions justified by credit-market failures. For instance, an important share of the European Union public funds (structural funds) is channeled to lagging regions on the assumption that firms located there have limited access to credit facilities. Our ML targeting might be useful to substantiate this assumption.

# Bibliography

Albertazzi, U., Bottero, M., and Sene, G. (2017). Information externalities in the credit market and the spell of credit rationing. *Journal of Financial Intermediation*, 30(-):61–70.

Andini, M., Ciani, E., de Blasio, G., D'Ignazio, A., and Salvestrini, V. (2018). Targeting with machine learning: An application to a tax rebate program in italy. *Journal of Economic Behavior and Organization*, 156(-):86–102.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Ascarza, E. (2018). Retention futility: Targeting high risk customers might be ineffective. *Journal of Marketing Research*, 55(1):80–98.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 255(6324):483–485.

Athey, S. (2019). The impact of machine learning on economics. *A chapter in The Economics of Artificial Intelligence: An Agenda from National Bureau of Economic Research, Inc*, pages 507–547.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *PNAS*, 113(27):7353–7360.

Barone, G., Mirenda, L., and Mocetti, S. (2016). Losing my connection: The role of interlocking directorates. *Rimini Centre for Economic Analysis Working Papers*, 16(09).

Beck, T., Klapper, L. F., and Mendoza, J. C. (2008). The typology of partial credit guarantee funds around the world. *Journal of Financial Stability*, 6(1):10–25.

Carmignani, A., de Blasio, G., Demma, C., and D'Ignazio, A. (2019). Urban agglomeration and firms. *Bank of Italy Working Papers n. 1222.*

CF (Various years). Annual reports. Technical report, Comitato di gestione del Fondo di garanzia.

de Blasio, G., Mitri, S. D., D'Ignazio, A., Russo, P. F., and Stoppani, L. (2018). Public guarantees to sme borrowing. a rdd evaluation. *Journal of Banking and Finance*, 96(-):73–86.

Deelen, L. and Molenaar, K. (2004). Guarantee funds for small enterprises. a manual for guarantee fund managers.

ECB (2015). Survey on the access to finance of enterprises in the euro area. Technical report, European Central Bank.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B*, 70(5):849–911.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Galardo, M., Lozzi, M., and Mistrulli, P. E. (2017). Social capital, uncertainty and credit supply: Evidence from the global crisis. *Bank of Italy, mimeo.*

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *New York: Springer.*

Honohan, P. (2010). Partial credit guarantees: Principles and practice. *Journal of Financial Stability*, 6(1):1–9.

Jimenez, G., Ongena, S., Peydro, J. L., and Saurina, J. (2012). Credit supply and monetary policy: Identifying the bank balance-sheet channel with loan applications. *American Economic Review*, 102(5):2301–2326.

Jimenez, G., Ongena, S., Peydro, J. L., and Saurina, J. (2014). Hazardous times for monetary policy: What do twenty three million bank loans say about the effects of monetary policy on credit risk taking? *Econometrica*, 82(2):463–505.

Kleinberg, J., Lakkaraju, H., Leskovec, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.

McBride, L. and Nichols, A. (2015). Improved poverty targeting through machine learning: An application to the usaid poverty assessment tools.

MSE (2015). Relazione sugli interventi di sostegno alle attività economiche e produttive. Technical report, Ministero and dello and Sviluppo and Economico.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

OECD (2013). Sme and entrepreneurship financing: The role of credit guarantee schemes and mutual guarantee societies in supporting finance for small and medium-sized enterprises. Technical report, Organization for Economic Cooperation and Development.

Poolsawad, N., Kambhampati, C., F., J., and Cleland (2014). Balancing class for performance of classification with a clinical dataset. *Proceedings of the World Congress on Engineering*, I.

Rajan, U., Seru, A., and Vig, V. (2015). The failure of models that predict failure: Distance, incentives, and defaults. *Journal of Financial Economics*, 115(2):237–260.

Riding, A., Madill, J., and Haines, G. J. (2007). Incrementality of sme loan guarantees. *Small Business Economics*, 29(1):47–61.

Saadani, Y., Zsofia, A., and de Rezende, R. (2011). A review of credit guarantee schemes in the middle east and north africa region. *World Bank Policy Research Working Paper n. 5612.*

Schivardi, F., Sette, E., and Tabellini, G. (2017). Credit misallocation during the european financial crisis. *Bank of Italy Working Papers n. 1139.*

Uesugi, I., Sakai, K., and Yamashiro, G. (2010). The effectiveness of public credit guarantees in the japanese loan market. *Journal of the Japanese and International Economies*, 24(4):457–480.

Vogel, R. C. and Adams, D. W. (1997). Costs and benefits of loan guarantee programs. *The Financier*, 4(1-2):22–29.

WB (2015). Principles for public credit guarantee schemes for smes. Technical report, The World Bank.

Y. Zhao, Y. and Cen, Y. (2014). Data mining applications with r. *Oxford Academic Press: Elsevier*.

Zia, B. H. (2008). Export incentives, financial constraints, and the (mis)allocation of credit: Micro-level evidence from subsidized export loans. *Journal of Financial Economics*, 87(2):498–527.

## 2.8 Appendix

This Appendix provides additional information on a number of topics: the dataset and its peculiarities (Sections A1-A2); the strategy we follow for model selection and training (Section A3); details on the implementation of the ML algorithms to predict credit-constrained and creditworthy firms (Sections A4-A5, which are meant for readers interested in more technical elements); a comparative description of model prediction results as well as model selection (Section A6); additional figures and tables (Sections A7-8).

### A.1 Covariates description and data cleaning

Our main data sources are: Credit register (CR) data on firms credit history and bad loans; Cerved data on firms' balance sheets. From CR we extract quarterly data covering the two years preceding the quarter when the firm issues the loan request. In particular, we consider: (i) the amount of total bank loans granted to the firms; (ii) the amount of total bank loans granted and actually used by the firm; (iii) the total number of banks lending to the firm; and (iv) a dummy variable indicating whether the firm has been reported as having bad loans. In addition, we include (v) a binary variable to identify firms about which we have no credit history data within the CR dataset (most of these firms likely have lending relationships with some banking institutions, but they do not show up in the data because the total amount of loans granted by each institution does not reach the 30.000 Euros CR threshold). As for balance sheet data, we select from the Cerved database the two most recent annual observations available before the PI request (loan application). In particular, we consider a set of balance-sheet items, taken from both balance sheets and income statements. We also include some indicators such as the return on assets, operating margin on assets and the leverage index. In addition, we generate dummy variables identifying firms with negative or null equity. The list of covariates is reported in Table A1, while descriptive statistics can be found in Table A2, both reported in Appendix A8.

After a data cleaning procedure designed to remove missing data, we try to reduce the information redundancy by analyzing the pairwise correlation among the covariates. Since we are dealing with both categorical and numerical variables, we rely on three different correlation statistics: the Pearson correlation index, the Polyserial correlation index and the Tetrachoric correlation index [23]. Using these statistics, we: (i) select some variables among the ones too highly correlated (more than 95 per cent); (ii) discard variables that are almost not correlated with the dependent variable (a correlation coefficient smaller than 5 per cent). The variables that pass the screening procedure are reported in Tables A3 and A4, Appendix A8.

## A.2 Some peculiarities of the 2011 and 2012 datasets

In the 2011 dataset the unit of observation is the loan application of a given firm in a given quarter of the year. The number of observations in the sample as well as the number of corresponding firms is reported in the main text. The same firm might appear more than once (up to 4 times) within the dataset. As a consequence, the same firm can be observed both as credit constrained and not constrained, depending on the quarter when the PI is issued. This is due to the fact that a firm is defined as credit constrained according to the dynamics of its bank loans in the six months following the PI request. Hence, PI issued in different quarters are associated with different time windows. Concerning the observed status of creditworthy, instead, there is no such variability, as we only consider whether firms have adjusted bad loans or not at the end of 2014.

A similar pattern is observed for ML-based predictions. In particular, the same firm can be ML-predicted as credit constrained or not, or creditworthy or not, depending on the quarter when the PI was issued. In particular, if a firm has a PI issued before June, then the ML algorithms will use firm balance sheet data at t-2, while if the firm has a PI issued after June the ML algorithms will use data at t-1, because balance sheet data at t-1 are usually made available in June. For instance, consider a firm that has two PIs, one in May 2011 and one in July 2011. The firm does not have adjusted bad loans in 2014; hence the observed creditworthy status is 1. When we predict the creditworthy status of this firm, we will be using 2009 balance sheet data in the first case and 2010 balance sheet data in the second. It is possible that in one case the firm will be predicted as not creditworthy and, in the other case, it will be predicted as such.

[23] The Pearson correlation index measures linear correlation between two numerical variables. The Polyserial correlation index is an index of bivariate association among numerical and categorical variables, resulting from an underlying continuous variable. The Tetrachoric correlation index measures the agreement for binary data. It estimates what the correlation would be if measured on a continuous scale.

Unlike the 2011 dataset, the 2012 one is a cross sectional dataset obtained after a random sample selection of one quarter occurrence for each firm. We apply our ML rule to firms for which banks have issued a PI request in 2012. As in the 2011 dataset, it is possible that the same firm has PI requests in different quarters of the year. However, we use this sample to simulate a policy scenario, where each firm is either a beneficiary or not of the GF, and each firm is either a ML target or not. Since the same firm with a PI request issued in different quarters might be associated with different ML predictions (if they rely on balance sheet data in different years), in those cases we randomly selected only one occurrence and discarded the remaining one(s). This leads to a drop in the number of observations, but not of the firms. This also means that, in the resulting 2012 dataset, observations and firms coincide (differently than the 2011 dataset). A further drop depends on the fact that, in order to replicate the GF eligibility mechanism, we need to gather a large set of balance sheet data from 2009, which are not available for all the firms in our sample. Finally, as we want to compare the GF eligibility mechanism with the ML targeting rule, we also restrict our sample of firms to those who belong to the GF eligible sectors. This leaves us with a sample of about 88.000 firms.

### A.3 Strategy for model selection and training

The decision tree is a classification algorithm that provides the researcher with a clear scheme (the tree) to follow for targeting. Intuitively, the decision tree divides the set of possible values of all the variables into J non-overlapping regions $j = 1, ..., J$. At step 1, starting from the whole sample, the algorithm identifies the variable $x_{pi}$ from $X_i$ and the threshold $s_1$ such that, by splitting the sample into two regions $x_{pi} < s_1$ and $x_{pi} \geq s_1$, we obtain the highest reduction in the sum of the Gini impurity index across the two regions [24]. At each subsequent step, the tree continues splitting the sample by finding a variable and a threshold that lead to the highest reduction in the impurity index. The tree can be grown as long as there are at least some observations in each node. However, a high number of levels in a tree (i.e., a very complex tree) is likely to overfit the data, leading to poor out-of-sample

[24] For each region, the Gini impurity index is equal to $2f(1-f)$ where f is the fraction with the outcome equal to 1 (that is the fraction belonging to the status).

predictions. By setting a regularization parameter $c_p$, it is possible to reduce the complexity of the tree (see Hastie et al. (2009)). Formally, the tree choice solves an optimization problem:

$$min_T \sum_{l=1}^{|T|} N_l L_l(T, y_l) + c_p|T| \qquad (a)$$

where T is the tree used to forecast the status y, $|T|$ is the total number of leaves, l is a leaf of tree T, $N_l$ is the number of observations in the leaf, $L_l(T, y_l)$ is a loss function (the Gini impurity index in our case), and $y_l$ is the vector of outcomes for observations in the leaf. Setting a low $c_p$ would lead to a large tree with a good fit in the training sample, but possibly with large out-of-sample error. By setting a higher $c_p$ we reduce its size (we "prune" the tree) and therefore we reduce the risk of overfitting. The complexity parameter ($c_p$) for pruning the tree is chosen using 10-fold cross-validation over the interval $[\alpha, \infty)$. The value $\alpha$ is chosen by considering a not too small $\alpha$ so that we do not deal with splits leading to leaves in which the classes' frequencies are almost equal. Looking at the cross-validated misclassification error for a grid of possible complexity parameters, we choose the smallest $c_p$ whose associated error is larger than the minimum error achieved in the cross-validation plus its standard deviation. This is done because the error usually reaches a plateau around the $c_p$ which gives the minimum error, and therefore by taking a larger (but close enough) $c_p$ we reduce the risk of over-fitting (by reducing complexity) keeping a similar cross-validated error.

The random forest algorithm provides an improved prediction by averaging the classification produced by n decision trees. Each tree is estimated on a new sample bootstrapped from the original training, but allowing only for a (randomly drawn) subset m of the P predictors. Each tree is grown to its maximum extension, without pruning (and therefore without setting an optimal $c_p$). These adjustments are aimed to reduce the correlation between the trees, in order to reduce the variance of the prediction. In order to optimally define the parameters on which the algorithm is based, we look at the *out-of-bag* (OOB) misclassification error [25]. We allow the number of variables m to vary from 1 to $\sqrt{P}$ where P is the total number of covariates in the (post-screening) X matrix. We instead allow the maximum number of trees

[25] The OOB error is computed as follows: for each observation we consider all the trees estimated on bootstrapped training sets where that observation does not appear, and we use their predictions to compute the misclassification error.

to be such that the probability that each variable does not appear in any tree is very low (approximately $10^{-6}$). We then choose the combination $(n, m)$ that has the minimum OOB error.

The logistic LASSO algorithm provides a prediction that is based on a logit model (with a linear index) where the estimated coefficients are penalized according to their magnitude. In this framework, one may account for the potential role played by non-linearities by generating all pairwise interactions between the explanatory variables included in the observables set (say $X_1$ for the rationing exercise and $X_2$ for the creditworthy exercise). Since this procedure leads to a marked increase in the dimension of the covariates matrix in each exercise, we apply an additional screening process to select only those that are more correlated with the respective dependent variable (dropping those with correlation coefficient smaller than 5 per cent) [26]. We use the two matrices thus obtained to estimate our predictive models, including 32 variables in the first exercise and 71 in the second. LASSO solves the following optimization problem (Hastie et al. (2009)):

$$max_{\beta_0,\beta_1} \sum_{i=1}^{N}[y_i(\beta_0 + \beta'\tilde{X}_i) - log(1 + e^{\beta_0+\beta'\tilde{X}_i})] - \sum_{j=1}^{\tilde{P}} |\beta_j| \qquad (b)$$

where $\tilde{X}_i$ of dimension $\tilde{P}$ is the vector of variables including also the (post-screening) pairwise interactions between the variables in $X_i$, $\lambda$ is a penalization parameter, $\beta_0$ is a constant and $\beta'$ is a transpose vector of the $\beta$ coefficients to be estimated (together with the constant). The penalization implies that only a subset of indicators will have coefficients other than zero. In line with Debashis and Chinnaiyan (2005), we choose $\lambda$ by looking at the 10-fold cross-validated misclassification error [27]. The optimal $\lambda$ is selected through cross-validation using the one-standard-error rule described above for the decision tree.

If there are strongly unbalanced classes in the sample, the ML classifier might be biased towards the over-represented class, ending up with a high misclassification error for the under-represented one. In this circumstance, a rebalancing procedure should be applied. This is the case for the "creditworthy status", where the distribution of the creditworthy vs not creditworthy is strongly unbalanced as the not creditworthy observations are about 14 per cent

[26] After the inclusion of all interaction terms, the set of explanatory variables counts, respectively, 152 units in the rationing exercise and 189 in the creditworthy exercise (in both cases, we exclude interactions that generated uninformative and invariant constant terms).

[27] The $\lambda$ parameter is validated over a grid of multiple values within the interval $[\lambda_{min}, \lambda_{max}]$, where $\lambda_{max}$ is defined as discussed in Friedman et al. (2010).

of the total. Following Poolsawad et al. (2014) and Y. Zhao and Cen (2014), we adopt an under-sampling strategy to solve the class imbalance. In particular, we randomly select only a subset of the observations belonging to the over-represented class and discard the remaining ones, so that the number of majority class observations (creditworthy firms, in our case) in the training set equals twice the number of under-represented class (not creditworthy firms) observations [28].

While we estimate two separate models, another approach could be to predict directly the target firms as those that are both constrained and creditworthy without using the balancing procedure. With this new dependent variable, we observe two things: (i) the percentage of observations in the constrained status is about 66 per cent of the training set, only 10 percentage points higher than that of the observations in the final target (56 per cent) meaning that, in this case, the balance feature of the target vs non-target status is informed by the credit-constrained status rather than by the creditworthy one; (ii) no improvement is reached in terms of misclassification error, as when we directly predict the jointly target firms, we obtain roughly the same misclassification error as when we predict the constrained firms and the creditworthy firms separately. We therefore choose to keep the two predictions separate.

## A.4 Details on the forecasting of credit-constrained firms

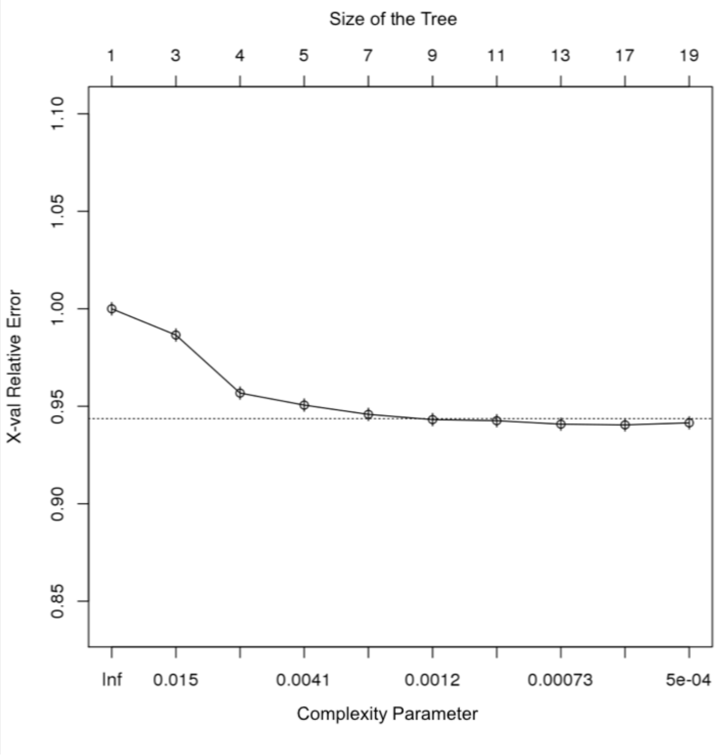
The first exercise is designed to predict the credit-constrained firms by means of the ML algorithms described above. In order to implement the first algorithm (decision tree), one needs to choose the complexity parameter $c_p$. We do this through cross-validation over the interval $[0.0005, \infty)$. As one can see from Figure A4.1, the optimal $c_p$ is 0.00142.

Figure A4.2 shows the variables relative importance, a numeric value ranking the relative importance of variables. This includes not only variables that are primary splits and therefore are relevant for the final prediction (i.e. they appear in Figure A6.1 of Section A6), but also surrogate variables that, in some of the splits, would have done almost as well as the primary ones. In this way we also understand the role played by variables that are

[28] After the under-sampling procedure the training set counts 75,777 observations (initially the training set contained 185,256 observations). The testing set remains the same.

very highly correlated with those that appear in the final decision tree, although they do not actually show up as primary splits. The list and the order by which variables appear in the ranking do not necessarily correspond to that of the pruned tree graph of Figure A6.1. For instance, the variable ranked as first in Figure A4.2 may not be the variable chosen for the first split in Figure A6.1, and some variables not showing up in the pruned tree graph may be present in the relative importance graph. This happens because, given that a variable may appear many times in the tree, either as a primary or a surrogate splitting variable, its overall relative importance value is defined additively, as the sum of goodness of split measures for each split in which it was the primary variable, plus the sum of adjusted goodness measures for splits in which it was a surrogate [29].

Figure A4.1: Complexity parameter validation of the tree for the credit-constrained exercise

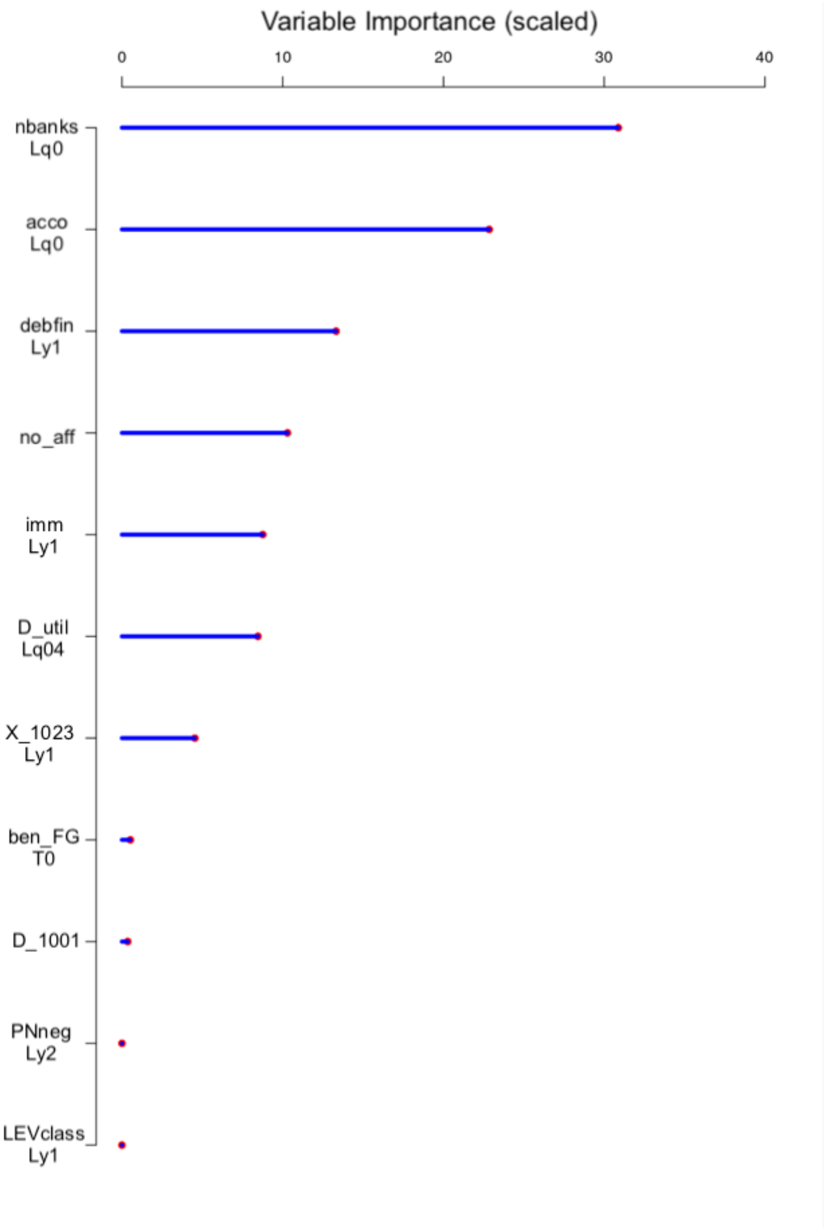


*Notes.* On the vertical axis the cross-validation error of the tree is built with the correspondent complexity parameter on the horizontal axis.

As one can see from Figure A4.2, the list of most important variables for the decision tree

[29] The misclassification is used as a ranking criterion: each observation is classified using the best feasible surrogate rule.

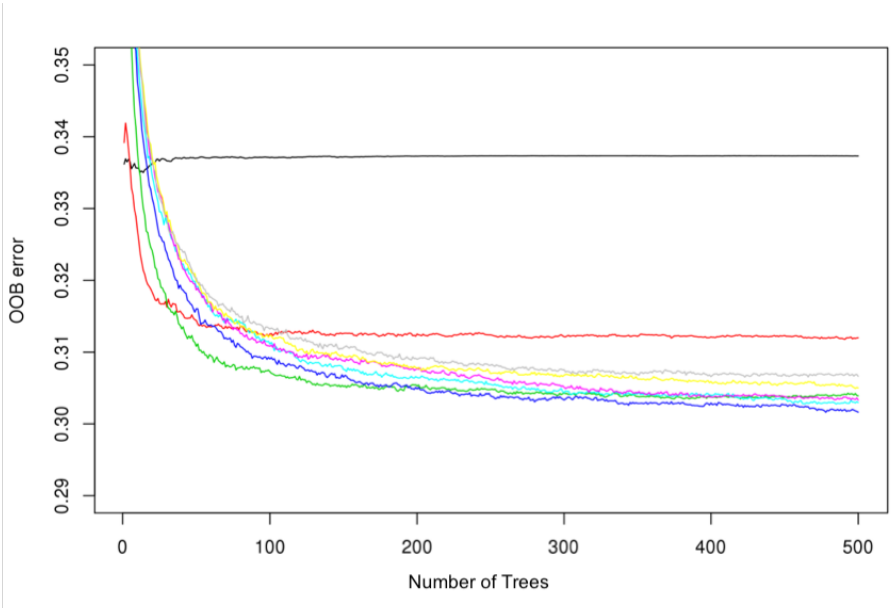
Figure A4.2: Variables importance in the tree for the credit-constrained exercise



*Notes.* The vertical axis shows the scaled importance of variables in the tree. Variable description as follows, in the same order as showed in the figure: $nbanks_{Lq0}$=Number of banks lending money to the firm in the quarter in which the PI request is issued (Lq0); $acco_{Lq0}$=Amount of total bank loans granted to the firm in the quarter in which the PI request is issued (Lq0); $debfin_{Ly1}$=Total amount of short and long term debts, based on the most recent balance-sheet data available when the PI was issued; $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset; $imm_{Ly1}$=Total assets (intangible + tangible assets) based on the most recent balance-sheet data available when the PI was issued (Ly1); $Dutil_{Lq04}$=Change in the total amount of bank loans granted and actually used by the firm, between the quarter when the PI request was issued and the same quarter in the previous year; ; $X1023_{Ly1}$=Long term debts; $benFG_{T0}$=Binary variable identifying whether the firm has already been a beneficiary of the GF-guarantee program (=1) or not (=0) before the PI request was issued; $D1001$=Change in the variable; $X1001_{Ly1}$ (intangible assets) with respect to the previous year; $PNneg_{Ly2}$=Binary variable identifying whether the firm has negative equity (=1) or not (=0) lagged by 1 year with respect to $PNneg_{Ly1}$ (Ly2); $LEVclass_{Ly1}$=Leverage class, based on the most recent balance-sheet data available when the PI was issued (Ly1).

algorithm, in addition to those that already show up in the tree, includes: the total amount of loans granted to the firm at time t (labeled as $acco$), the total amount of financial debts at t-1 (labeled as $debfin_{Ly1}$), the dummy identifying firms that have at least one lending relationship with a total loan amount exceeding the 30.000 Euros CR threshold (labeled as $no-aff$), total assets at t-1 (labeled as $imm_{Ly1}$) and the dummy identifying those firms that have already been a beneficiary of the GF guarantee in the past (labeled as $benFG_{T0}$). As for the second algorithm, the random forest, we need to choose the number of trees of the forest and the number of variables randomly selected for each tree. To do this, we look at the out of bag error (OOB) of the random forest. In Figure A4.3, we can see the OOB errors for the number of trees going from 1 to 500 and the number of variables going from 2 to 8. We choose the parameters with the lowest OOB error, that is: $n$=478 and $m$=5.

Figure A4.3: Out of bag error of the random forest for the credit-constrained exercise

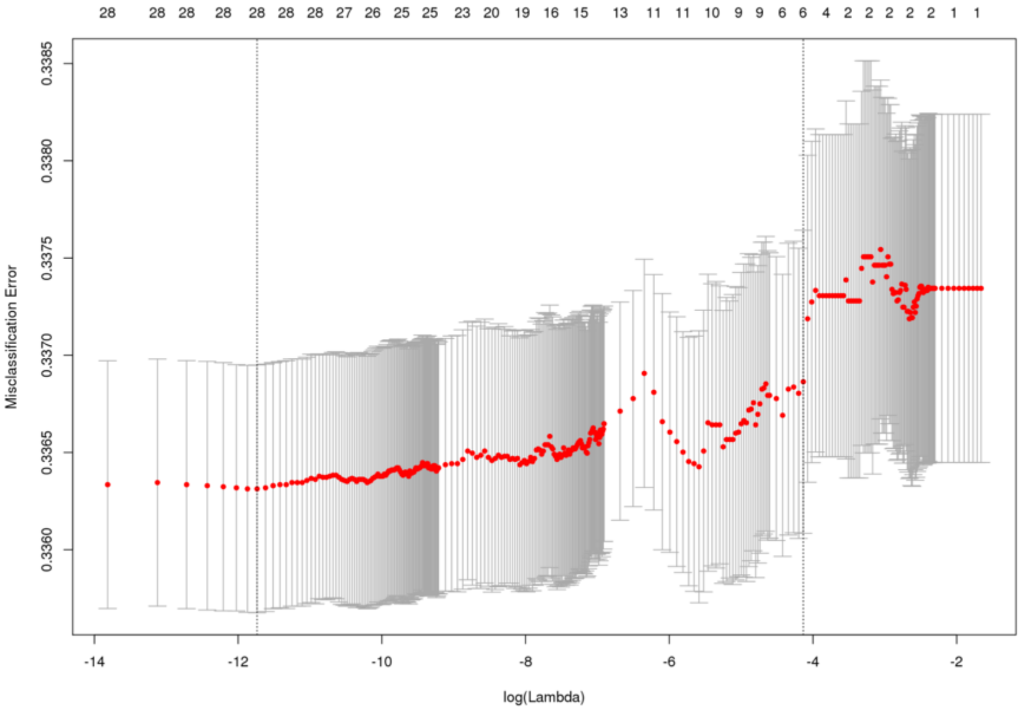


*Notes.* Each line in the graph corresponds to the random forest built with different numbers of variables. colors legend: black stands for 1 variable; red for 2 variables; green for 3 variables; blue for 4 variables; light blue for 5 variables; deep pink for 6 variables; yellow for 7 variables; light grey for 8 variables.

As expected, since the random forest is essentially the average of n not pruned decision trees, the list of important variables selected by the random forest algorithm contains the variables that are important in the decision tree. As we can see from Figure A4.4, in addition to all the variables already selected by the decision tree, other variables such as the age of the firm and Cerved rating of the firm in t-1 (labeled as $rating_{Ly1}$) also appear among the most important predictors.

As for the LASSO regression, we select the regularization parameter $\lambda$ as to minimize the misclassification error according to the one-standard-error rule. Figure A4.5 shows the optimal $\lambda$ chosen is 0.016 (whose logarithm is equal to -4.135), which is associated with the presence of six non-null coefficient in the regression model, shown in Table A4.1 [30].

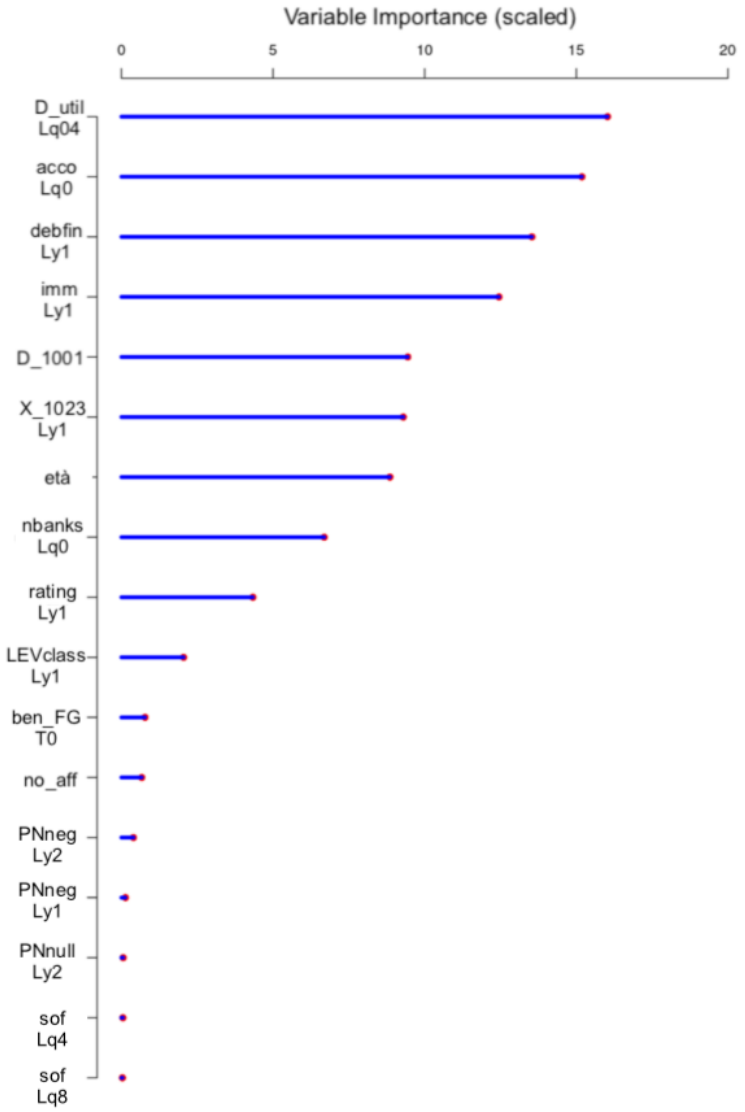
Figure A4.5: Errors of the penalizing parameter for the credit-constrained exercise



*Notes.* The graph shows the misclassification error (computed with cross validation) of regressions calculated using different penalizing parameters (on the bottom horizontal axis) and the number of nonzero coefficients (on the top horizontal axis).

[30] The sequence of $\lambda$ parameters used in the cross-validation counts 600 values, generated by a sequence ranging within the interval [0.2,0.0005] with a uniform increment of 0.0005.

Figure A4.4: Variables importance in the random forest for the credit-constrained exercise



*Notes.* The vertical axis shows the scaled importance of variables in the random forest. Variable description as follows, in the same order as showed in the figure: $Dutil_{Lq04}$=Change in the total amount of bank loans granted and actually used by the firm, between the quarter when the PI request was issued and the same quarter in the previous year.; $acco_{Lq0}$=Amount of total bank loans granted to the firm in the quarter in which the PI request is issued (Lq0); $debfin_{Ly1}$=Total amount of short and long term debts, based on the most recent balance-sheet data available when the PI was issued; $imm_{Ly1}$=Total assets (intangible + tangible assets) based on the most recent balance-sheet data available when the PI was issued (Ly1); $D1001$=Change in the variable; $X1001_{Ly1}$ (intangible assets) with respect to the previous year; $X1023_{Ly1}$=Long term debts; $eta'$=Firm age (expressed in years); $nbanks_{Lq0}$=Number of banks lending money to the firm in the quarter in which the PI request is issued (Lq0); $rating_{Ly1}$=Rating index produced by Cerved measuring firms' level of riskiness, based on the elaboration of balance-sheet data available when the PI is issued (Ly1); $LEVclass_{Ly1}$=Leverage class, based on the most recent balance-sheet data available when the PI was issued (Ly1); $benFG_{T0}$=Binary variable identifying whether the firm has already been a beneficiary of the GF-guarantee program (=1) or not (=0) before the PI request was issued; $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset; $PNneg_{Ly2}$=the same as before but lagged by 1 year with respect to $PNneg_{Ly1}$ (Ly2); $PNneg_{Ly1}$=Binary variable identifying whether the firm has negative equity (=1) or not (=0), based on the most recent balance-sheet data available when the PI was issued (Ly1); $PNnull_{Ly2}$=Binary variable identifying whether the firm has null equity (=1) or not (=0) lagged by 1 year with respect to $PNnull_{Ly1}$ (Ly2); $sof_{Lq4}$=Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) 4 quarters before the PI request (Lq4); $sof_{Lq8}$=Binary variable defined as before but 8 quarters before (Lq8).

Table A4.1: Non-null coefficients of the LASSO regression

| **Variable** | **Coef.** |
|---|---|
| $nbanks_{Lq0}$ | -0.331841806204264 |
| $imm_{Ly1}$ x $debfin_{Ly1}$ | -0.000038241057647 |
| $acco_{Lq0}$ x $debfin_{Ly1}$ | -0.000034496744523 |
| $nbanks_{Lq0}$ x $imm_{Ly1}$ | 0.001697873471909 |
| $rating_{Ly1}$ | 0.024015592002447 |
| $no-aff$ | 0.436245372314423 |

*Notes.* Variables are ordered based on the magnitude of the associated estimated coefficient. Variable description as follows. $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset; $rating_{Ly1}$=Rating index produced by Cerved measuring firms' level of riskiness, based on the elaboration of balance-sheet data available when the PI is issued (Ly1); $nbanks_{Lq0}$ x $imm_{Ly1}$= interaction term between $nbanks_{Lq0}$ and $imm_{Ly1}$; $acco_{Lq0}$ x $debfin_{Ly1}$= interaction term between $acco_{Lq0}$ and $debfin_{Ly1}$; $imm_{Ly1}$ x $debfin_{Ly1}$= interaction term between $imm_{Ly1}$ and $debfin_{Ly1}$, where: $nbanks_{Lq0}$=Number of banks lending money to the firm in the quarter in which the PI request is issued (Lq0), $imm_{Ly1}$=Total assets (intangible + tangible assets) based on the most recent balance-sheet data available when the PI was issued (Ly1); $acco_{Lq0}$=Amount of total bank loans granted to the firm in the quarter in which the PI request is issued (Lq0); $debfin_{Ly1}$=Total amount of short and long term debts, based on the most recent balance-sheet data available when the PI was issued;

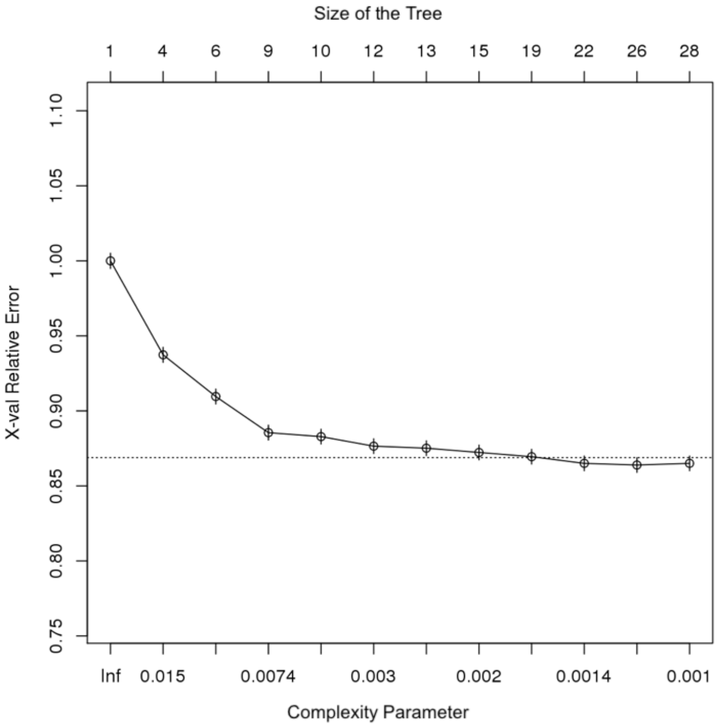
## A.5 Details on forecasting creditworthy firms

As before, the complexity parameter $c_p$ for the decision tree algorithm is chosen through cross-validation. The cross validation interval is $[0.001, \infty]$. as a result of the trade-off between the accuracy and interpretability of the resulting tree, in favor of a more readable tree structure. If we validate the $c_p$ over a larger interval, we obtain a decision tree extremely hard to interpret and nevertheless dominated by other methods such as the random forest in terms of prediction accuracy. As one can see from Figure A5.1, the optimal $c_p$ selected is 0.00141.

Figure A5.2 reports variables relative importance. As one can see, in addition to the splitting variables that already appeared in the tree graph (see Figure A12 in Section A6), the relative importance graph includes: a dummy variable (labeled as $PNnull_{Ly1}$) that describes a firm with null equity or not and a dummy variable (labeled as $PNneg_{Ly1}$) that identifies firms with negative equity or not.

As for the random forest, Figure A5.3 shows the OOB errors graph, which allows us to choose the combination of number of trees and number of variables that minimizes such an error. As happened for the constrained firms forecasting, the important variables of the tree are important also for the random forest (Figure A5.4).

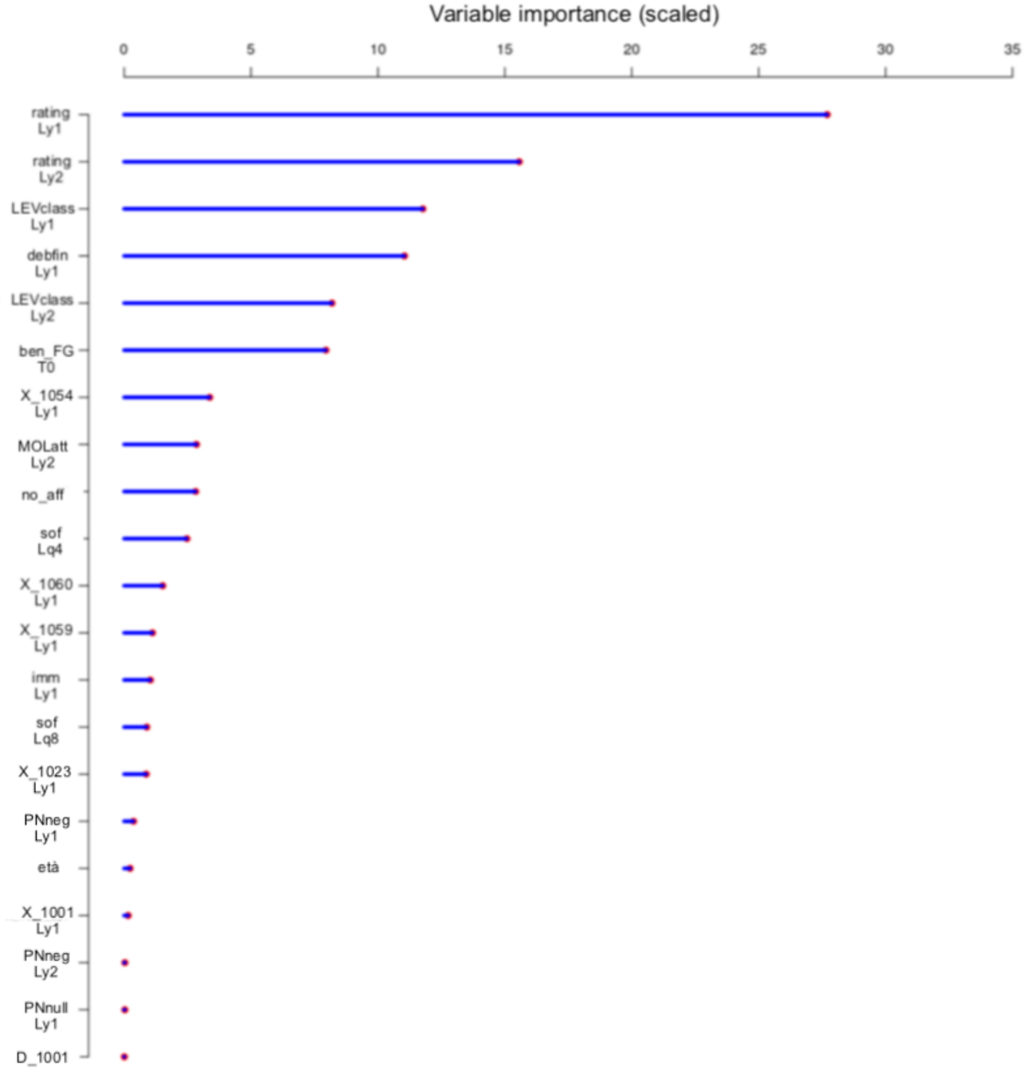
Figure A5.1: Complexity parameter validation of the tree for the creditworthy exercise



*Notes.* On the vertical axis the cross-validation error of the tree is built with the correspondent complexity parameter on the horizontal axis.

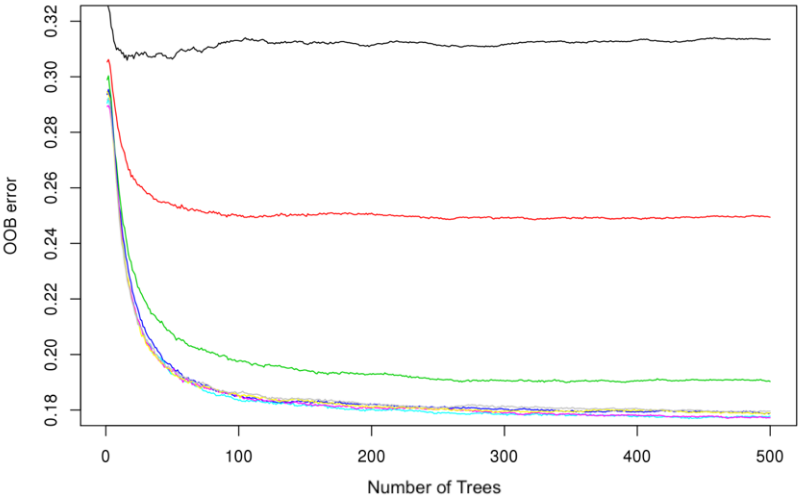
Before fitting the LASSO regression, we validate the penalizing parameter $\lambda$ through 10-folds cross-validation. Figure A5.5 shows the cross validation error graph: the best $\lambda$ selected according to the one-standard-error rule is equal to 0.00039 (whose logarithm is equal to -7.849), which is associated with the presence of 45 non-null coefficients in the estimated model, listed in Table A5.1.

Figure A5.2: Variables importance in the tree for the creditworthy exercise



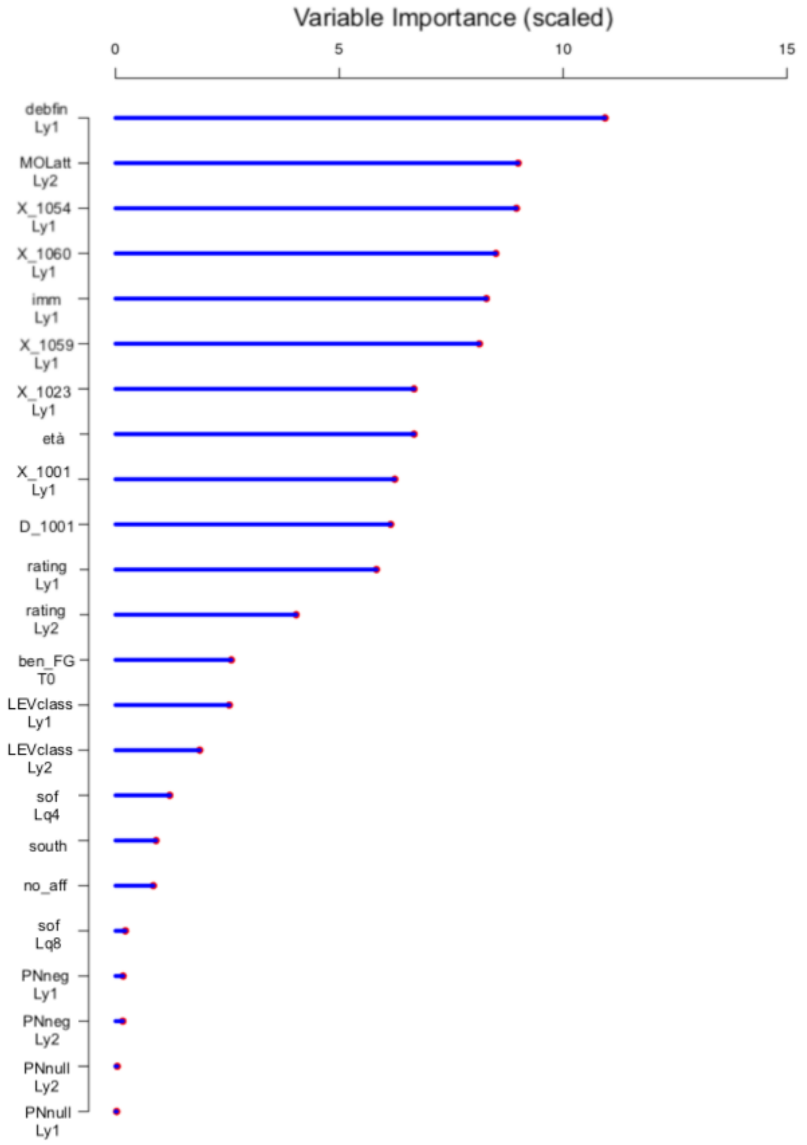
*Notes.* The vertical axis shows the scaled importance of variables in the tree. Variable description as follows, in the same order as showed in the figure: $rating_{Ly1}$=Rating index produced by Cerved measuring firms' level of riskiness, based on the elaboration of balance-sheet data available when the PI is issued (Ly1); $rating_{Ly2}$=the same as before but lagged by 1 year with respect to $rating_{Ly1}$ (Ly2); $LEVclass_{Ly1}$=Leverage class, based on the most recent balance-sheet data available when the PI was issued (Ly1); $debfin_{Ly1}$=Total amount of short and long term debts, based on the most recent balance-sheet data available when the PI was issued; $LEVclass_{Ly2}$=same as before but lagged by one year (Ly2); $benFG_{T0}$=Binary variable identifying whether the firm has already been a beneficiary of the GF-guarantee program (=1) or not (=0) before the PI request was issued; $X1054_{Ly1}$=Production value; $MOLatt_{Ly2}$=Operating margin on assets index, based on the most recent balance-sheet data available when the PI was issued (Ly1) lagged by 1 year with respect to $MOLatt_{Ly1}$ (Ly2); $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset; $sof_{Lq4}$=Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) in the CR 4 quarters before the PI request (Lq4); $X1060_{Ly1}$=Gross operating margin; $X1059_{Ly1}$=Labor cost; $imm_{Ly1}$=Total assets (intangible + tangible assets) based on the most recent balance-sheet data available when the PI was issued (Ly1); $sof_{Lq8}$=Binary variable defined as before but 8 quarters before the PI request (Lq8); $X1023_{Ly1}$=Long term debts; $PNneg_{Ly1}$=Binary variable identifying whether the firm has negative equity (=1) or not (=0) based on the most recent balance-sheet data available when the PI was issued (Ly1); età=Firm age (expressed in years); $X1001_{Ly1}$=Intangible assets; $PNneg_{Ly2}$=the same as before but lagged by 1 year with respect to $PNneg_{Ly1}$ (Ly2); $PNnull_{Ly1}$=Binary variable identifying whether the firm has null equity (=1) or not (=0) based on the most recent balance-sheet data available when the PI was issued (Ly1); $D1001$=Change in the variable $X1001_{Ly1}$ with respect to the previous year.

Figure A5.3: Out of bag error of the random forest for the creditworthy exercise



*Notes.* Each line in the graph corresponds to the random forest built with different numbers of variables. Colors legend: black stands for 1 variable; red for 2 variables; green for 3 variables; blue for 4 variables; light blue for 5 variables; deep pink for 6 variables; yellow for 7 variables; light grey for 8 variables.

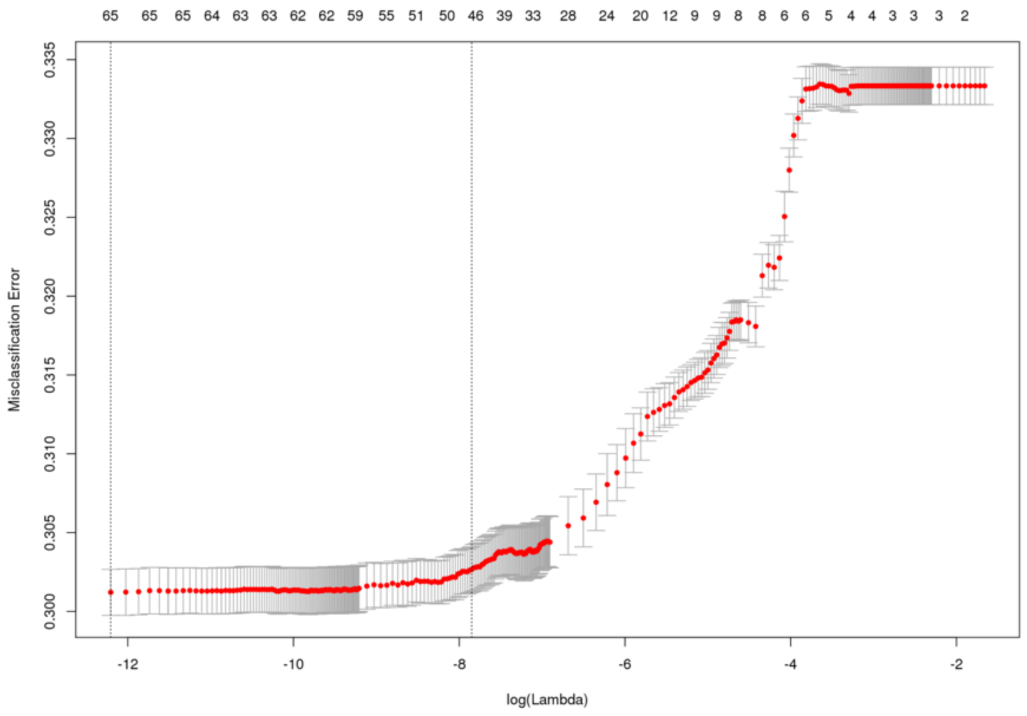
Figure A5.4: Variables importance in the random forest for the creditworthyexercise



*Notes.* The vertical axis shows reported the scaled importance of variables in the random forest. Variable description as follows, in the same order as showed in the figure: $debfin_{Ly1}$=Total amount of short and long term debts, based on the most recent balance-sheet data available when the PI was issued; $MOLatt_{Ly2}$= operating margin on assets index based on the most recent balance-sheet data available when the PI was issued (Ly1) lagged by 1 year with respect to $MOLatt_{Ly1}$ (Ly2); $X1054_{Ly1}$=Production value; $X1060_{Ly1}$=Gross operating margin; $imm_{Ly1}$=Total assets (intangible + tangible assets) based on the most recent balance-sheet data available when the PI was issued (Ly1); $X1059_{Ly1}$=Labor cost; $X1023_{Ly1}$=Long term debts; $eta'$=Firm age (expressed in years); $X1001_{Ly1}$=Intangible assets; $D1001$=Change in the variable $X1001_{Ly1}$ with respect to the previous year; $rating_{Ly1}$=Rating index produced by Cerved measuring firms level of riskiness, based on the elaboration of balance-sheet data available when the PI is issued (Ly1); $rating_{Ly2}$=same as before but lagged by 1 year with respect to $rating_{Ly1}$ (Ly2); $benFG_{T0}$=Binary variable identifying whether the firm has already been a beneficiary of the GF-guarantee program (=1) or not (=0) before the PI request was issued; $LEVclass_{Ly1}$=Leverage class, based on the most recent balance-sheet data available when the PI was issued (Ly1); $LEVclass_{Ly2}$=same as before but lagged by 1 year with respect to $LEVclass_{Ly1}$ (Ly2); $sof_{Lq4}$= Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) in the CR 4 quarters before the PI request (Lq4); $south$=Binary variable identifying whether the firm is located in the South of Italy (=1) or not (=0); $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset; $sof_{Lq8}$=Binary variable defined as before but 8 quarters before (Lq8);

$PNneg_{Ly1}$=Binary variable identifying whether the firm has negative equity (=1) or not (=0), based on the most recent balance-sheet data available when the PI was issued (Ly1); $PNneg_{Ly2}$=same as before but lagged by 1 year with respect to $PNneg_{Ly1}$ (Ly2); $PNnull_{Ly2}$=Binary variable identifying whether the firm has null equity (=1) or not (=0), based on the second-to-most recent balance-sheet data available when the PI was issued (Ly2); $PNnull_{Ly1}$=Binary variable identifying whether the firm has null equity (=1) or not (=0), based on the most recent balance-sheet data available when the PI was issued (Ly1).

Figure A5.5: Errors of the penalizing parameter for the creditworthy exercise



*Notes.* The graph shows the misclassification error (computed with cross validation) of regressions calculated using different penalizing parameters (on the bottom horizontal axis) and the number of nonzero coefficients (on the top horizontal axis).

Table A5.1: Non-null coefficients of the LASSO regression

| **Variable** | **Coef.** |
|---|---|
| $sof_{Lq8}$ | -1.491700 |
| $sof_{Lq4}$ x $PNneg_{Ly1}$ | -1.220300 |
| $benFG_{T0}$ | -1.083400 |
| $sof_{Lq4}$ x $south$ | -0.864931 |
| $rating_{Ly2}$ | -0.625037 |
| $PNneg_{Ly1}$ | -0.423501 |
| $rating_{Ly1}$ x $south$ | -0.308869 |
| $LEVclass_{Ly2}$ | -0.241184 |
| $rating_{Ly1}$ x $benFG_{T0}$ | -0.225360 |
| $rating_{Ly2}$ x $eta'$ | -0.204530 |
| $south$ | -0.124396 |
| $sof_{Lq4}$ x $LEVclass_{Ly2}$ | -0.111410 |
| $rating_{Ly1}$ x $LEVclass_{Ly2}$ | -0.066829 |
| $no-aff$ x $LEVclass_{Ly1}$ | -0.065553 |
| $rating_{Ly1}$ x $eta'$ | -0.062732 |
| $rating_{Ly2}$ x $LEVclass_{Ly2}$ | -0.052011 |
| $rating_{Ly1}$ x $debfin_{Ly1}$ | -0.051245 |
| $rating_{Ly2}$ x $debfin_{Ly1}$ | -0.045650 |
| $PNneg_{Ly1}$ x $debfin_{Ly1}$ | -0.043521 |
| $no-aff$ x $LEVclass_{Ly2}$ | -0.041634 |
| $debfin_{Ly1}$ | -0.021730 |
| $eta'$ x $south$ | -0.017883 |
| $rating_{Ly2}$ x $PNneg_{Ly2}$ | 0.001894 |
| $rating_{Ly2}$ x $imm_{Ly1}$ | 0.009092 |
| $imm_{Ly1}$ | 0.011241 |
| $X1059_{Ly1}$ x $imm_{Ly1}$ | 0.011459 |
| $X1060_{Ly1}$ x $debfin_{Ly1}$ | 0.011492 |
| $PNneg_{Ly1}$ x $south$ | 0.018185 |

| | |
|---|---|
| $X1001_{Ly1}$ | 0.162070 |
| $PNnull_{Ly2}$ | 0.234353 |
| $rating_{Ly2}$ x $no-aff$ | 0.253314 |
| $rating_{L}y2$ x $south$ | 0.347448 |
| $rating_{Ly2}$ x $benFG_{T0}$ | 0.374981 |
| $PNneg_{Ly2}$ x $LEVclass_{Ly2}$ | 0.650047 |
| $no-aff$ | 0.899711 |

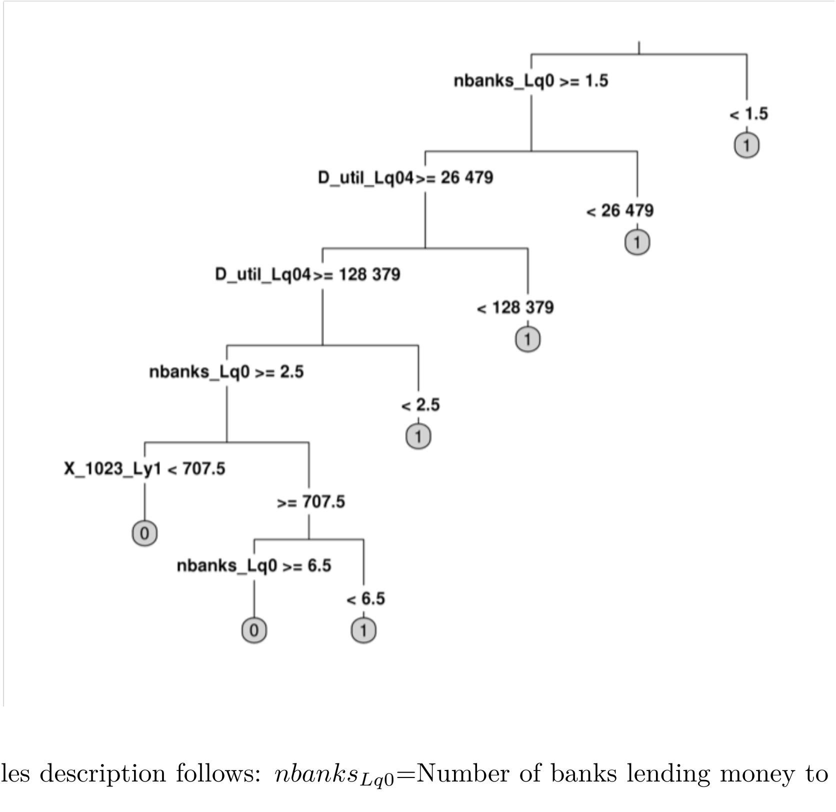
*Notes.* Variables are ordered based on the magnitude of the associated estimated coefficient. Variable description, for the variables whose associated coefficient in terms of magnitude is among the top-three most relevant features for the prediction follows. For the description of the other variables please refer to the Table A1 reported in Appendix A8.

Top-3 features include: [positive coeff.] $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset, and the interaction terms between $PNneg_{Ly2}$ and $LEVclass_{Ly2}$, $rating_{Ly2}$ and $benFG_{T0}$, where $PNneg_{Ly2}$= Binary variable identifying whether the firm has negative equity (=1) or not (=0), based on the second-to-most recent balance-sheet data available when the PI was issued (Ly2), $LEVclass_{Ly2}$=Leverage class, based on the second-to-most recent balance-sheet data available when the PI was issued (Ly2), $rating_{Ly2}$= Rating index produced by Cerved measuring firms level of riskiness, based on the elaboration of balance-sheet data available two years before the PI request was issued , $benFG_{T0}$=Binary variable identifying whether the firm has already been a beneficiary of the GF-guarantee program (=1) or not (=0) before the PI request was issued // [negative coeff.] $sof_{Lq8}$= Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) in the CR 8 quarters before the PI request (Lq8); the interaction between $sof_{Lq4}$ and $PNneg_{Ly1}$, where $sof_{Lq4}$= Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) in the CR 4 quarters before the PI request (Lq4) and $PNneg_{Ly1}$= Binary variable identifying whether the firm has negative equity (=1) or not (=0), based on the most recent balance-sheet data available when the PI was issued (Ly1), and $benFG_{T0}$.

## A.6 Models prediction results and model selection

An initial understanding of the characteristics of targeted firms can be provided by the decision tree, which is a good compromise between flexibility and interpretability. For this tool, the estimated (trained) algorithm essentially resembles a decision rule, in which each step (node) discriminates firms according to the value of a specific variable. Figure A6.1 shows the decision rule to predict credit-constrained firms, which tend to be those with few lending relationships with banks and a small variation in used credit, or those that have a larger number of lending relationships and greater exposure to total medium-long term debts. Figure A6.2 shows a more complicated prediction for creditworthy firms, which essentially depends on the Cerved-rating score, which is a balance-sheet summary of the firms' financial soundness, the presence of past defaults and exposure to the bank. In this case, also the past presence of a GF guarantee plays a role. The prediction from the random forest is less interpretable, as it combines many different trees. One can construct measures of variable importance, but we do not get a neat decision rule. The LASSO predictions are in principle more interpretable, but the presence of interactions and powers makes it less so. The difficulty in interpreting the forecasting rules raises some transparency concerns that we discuss in Subsection 6.2.

Figure A6.1: Classification tree for the credit-constrained exercise



*Notes.* Variables description follows: $nbanks_{Lq0}$=Number of banks lending money to the firm in the quarter in which the PI request is issued (Lq0); $Dutil_{Lq04}$=Change in the total amount of bank loans granted and actually used by the firm, between the quarter when the PI request was issued and the same quarter in the previous year; $X1023_{Ly1}$=Long term debts.

Our main aim is to have a forecasting rule that performs well out of sample. We therefore compare the models by looking at misclassification in the testing sample. The misclassification tables focus on the *false positive rate (FP)*, which is the fraction of actually negative observations that are predicted as positive, and at the *false negative rate (FN)*, which is the fraction of actually positive observations that are predicted as negative. Positive means that they are in the target status, and vice versa for negative. We define the predicted status as positive when the forecast probability of being so at least equal to 0.5 [31]. For the credit-constrained exercise (Table A6.1), the decision tree and random forest performances are similar overall. The decision tree tends to do worse in classifying the actually non-constrained firms (as the FP rate is higher), while the random forest does worse in classifying the actually constrained. LASSO has a higher misclassification rate. For the creditworthy prediction (Table A6.2), the lowest misclassification rate is reached by the random forest. In this case the decision tree has the worst performance and LASSO is in between.

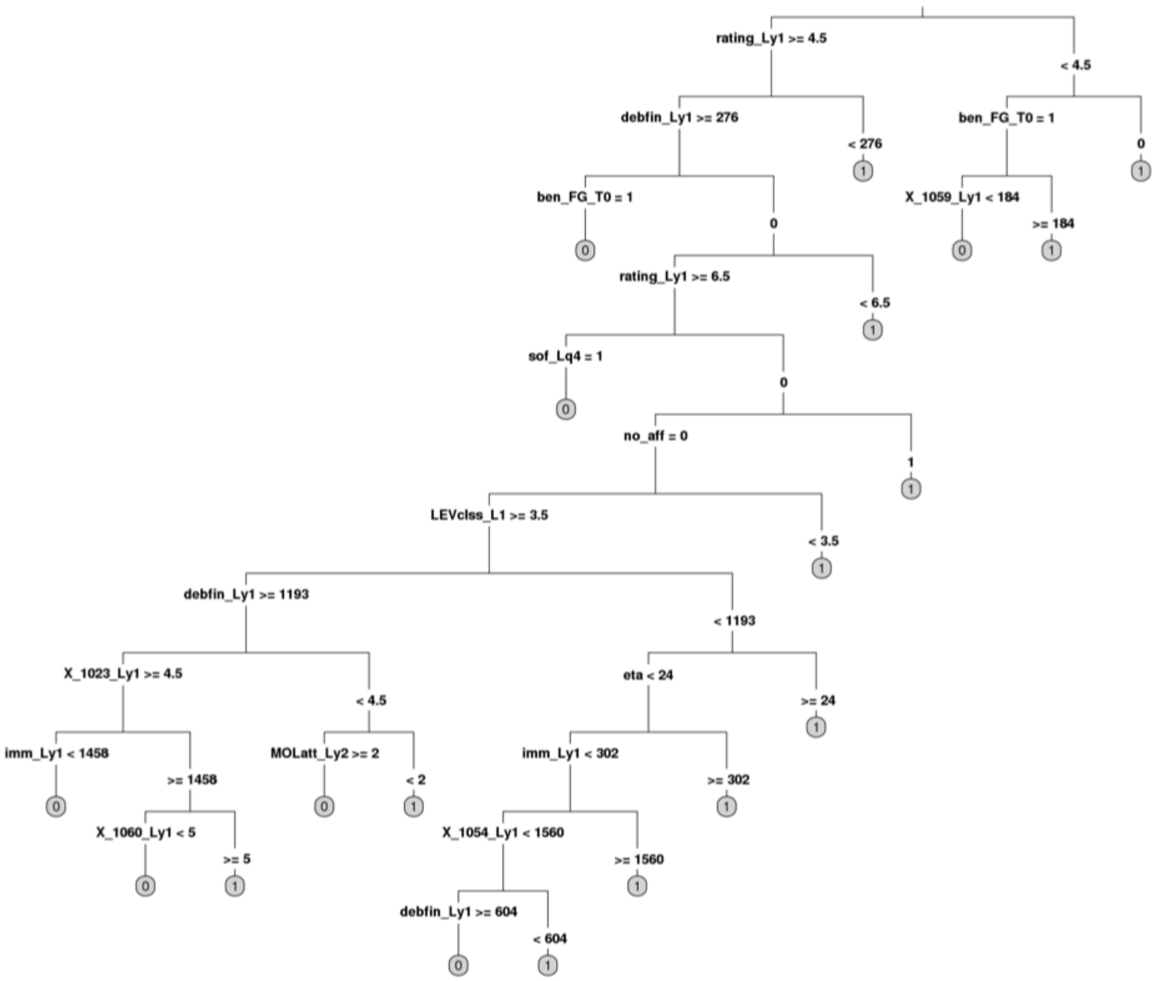
Comparing different models might be misleading if the total fraction of predicted positive cases is different across different algorithms [32].Instead of classifying firms as target if the predicted probability is at least equal to 0.5, we can follow two approaches. The first approach orders all the observations according to the predicted probability and assigns to the target group the fraction x with the highest forecasted probability. In this way we can compare the algorithms performance keeping fixed the fraction of predicted positive cases. The Lift curve (Figure A6.3) looks at the how the *true positive rate (TP)*, the fraction of actually positive observations that are forecasted as positive) changes with x (Hastie et al. (2009)). For example, the point x=0.20 means that the 20 per cent with the highest predicted probability of being in the target status is classified as 1 and all the rest as 0. The diagonal line is a random classifier (gives equal probability 0.5 to each observation): with this kind of classifier, at x=0.2 one should predict correctly the 20 per cent of positive observations. If one uses a better classifier, she should expect to have more than 20 per cent of correctly predicted observations in the top 20 per cent of predicted probability. Again, random forest

[31] If the predicted probabilities of a given status is equal to 0.5, the status assignment is random (this happens only in the credit constrained prediction exercise and characterizes very few cases).

[32] Furthermore, accuracy rates can be unreliable metrics of performance for unbalanced data sets: for example, if we imagine that we have an extremely unbalanced set with 95 per cent of red balls and 5 per cent of blue balls a totally red-classifier (predicting all balls red) will have high accuracy in terms of misclassification error (only 0.05) but it will nevertheless be completely useless.

does slightly better in both cases.

Figure A6.2: Classification tree for the creditworthy exercise



*Notes.* Variables description follows: $rating_{Ly1}$=Rating index produced by Cerved measuring firms' level of riskiness, based on the elaboration of balance-sheet data available when the PI is issued (Ly1); $debfin_{Ly1}$=Total amount of short and long term debts based on the most recent balance-sheet data available when the PI was issued; $benFG_{T0}$=Binary variable identifying whether the firm has already been a beneficiary of the GF-guarantee program (=1) or not (=0) before the PI request was issued; $X1059_{Ly1}$=Labor cost; $sof_{Lq4}$=Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) to the CR 4 quarters before the PI request; $no-aff$=Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset; $LEVclass_{Ly1}$=Leverage class, based on the most recent balance-sheet data available when the PI was issued (Ly1); $X1023_{Ly1}$=Long term debts; $eta'$=Firm age (expressed in years); $imm_{Ly1}$=Total assets (intagible + tangible assets) based on the most recent balance-sheet data available when the PI was issued (Ly1); $MOLatt_{Ly2}$=Operating margin on assets index lagged by 1 year with respect to $MOLatt_{Ly1}$ (Ly2); $X1060_{Ly1}$=Gross operating margin (most recent balance-sheet data available when the PI was issued; $X1054_{Ly1}$=Production value.

Table A6.1: Confusion matrices for each ML algorithm in the credit-constrained exercise

| | | | | |
|---|---|---|---|---|
| Panel A. Decision tree | | | | |
| | $Y_{\text{pred}} = 0$ | $Y_{\text{pred}} = 1$ | Misclassification rate: 31.85% | |
| $Y_{\text{actual}} = 0$ | 8,408 | 23,100 | TN: 26.68% | FN:10.64% |
| $Y_{\text{actual}} = 1$ | 6,558 | 55,033 | FP: 73.3% | TP: 89.35% |
| Panel B. Random forest | | | | |
| | $Y_{\text{pred}} = 0$ | $Y_{\text{pred}} = 1$ | Misclassification rate: 32.09% | |
| $Y_{\text{actual}} = 0$ | 10,243 | 21,265 | TN: 32.5% | FN: 13.98% |
| $Y_{\text{actual}} = 1$ | 8,616 | 52,975 | FP: 67.49% | TP: 86% |
| Panel C. LASSO regression | | | | |
| | $Y_{\text{pred}} = 0$ | $Y_{\text{pred}} = 1$ | Misclassification rate: 33.83% | |
| $Y_{\text{actual}} = 0$ | 2,362 | 29,146 | TN: 7.49% | FN: 3.82% |
| $Y_{\text{actual}} = 1$ | 2,354 | 59,237 | FP: 92.5% | TP: 96.17% |

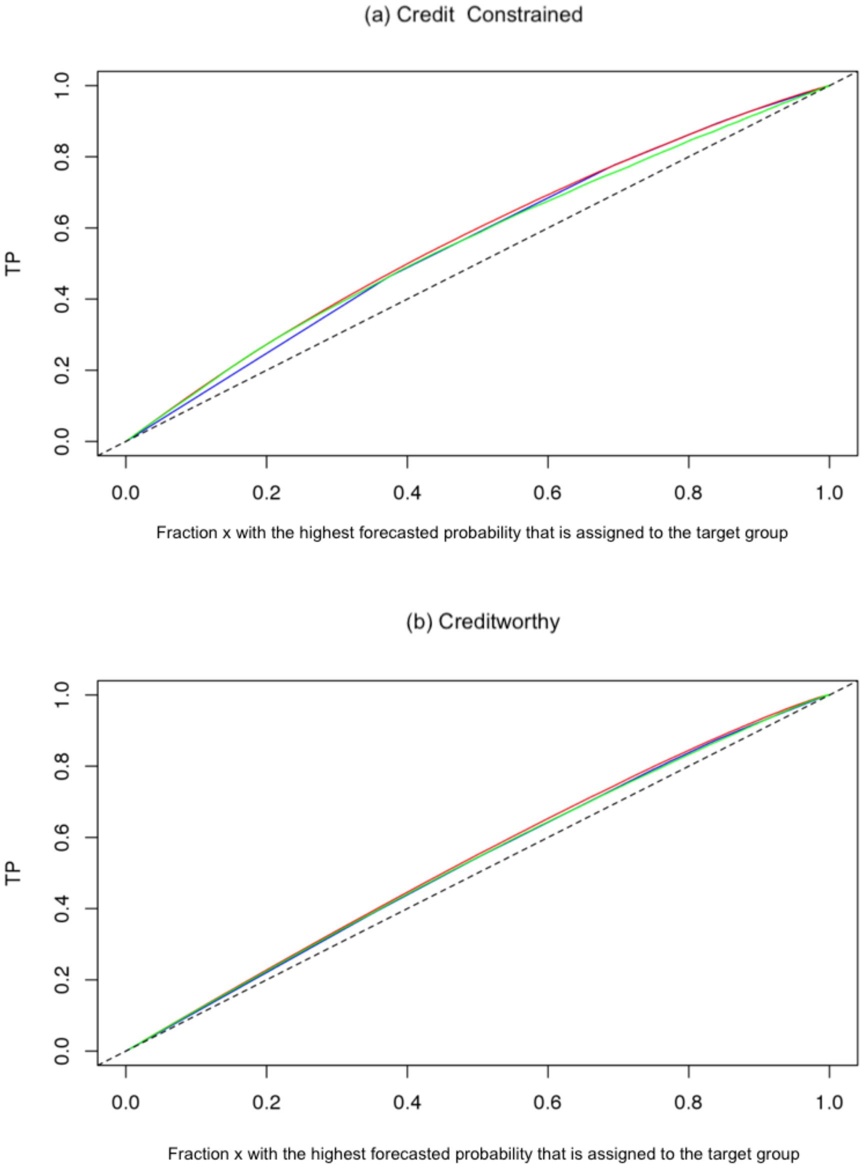
*Notes.* Testing sample (2011). $Y_{\text{actual}}$ is 1 if the actual status is to be credit constrained, 0 otherwise; $Y_{\text{pred}}$ is 1 if a credit-constrained observation is predicted (predicted probability $\geq 0.5$ ), 0 otherwise. FP is the *false positive rate* computed as the percentage of observations predicted positive, but that are actually negative, over the total number of actually negative observations; TP is the *true positive rate* computed as the percentage of observations predicted positive, that are actually positive, over the total number of actually positive observations; FN is the *false negative rate* computed as the percentage of observations predicted negative, but that are actually true, over the total number of actually positive observations; TN is the *true negative rate* computed as the percentage of observations predicted negative, but that are actually negative, over the total number of actually negative observations.

Table A6.2: Confusion matrices for each ML algorithm in the creditworthy exercise

| | | | | |
|---|---|---|---|---|
| Panel A. Decision tree | | | | |
| | $Y_{pred} = 0$ | $Y_{pred} = 1$ | Misclassification rate: 20.02% | |
| $Y_{actual} = 0$ | 5,097 | 7,574 | TN: 40.22% | FN: 13.75% |
| $Y_{actual} = 1$ | 11,066 | 69,362 | FP: 59.77% | TP: 86.24% |
| Panel B. Random forest | | | | |
| | $Y_{pred} = 0$ | $Y_{pred} = 1$ | Misclassification rate: 17.66% | |
| $Y_{actual} = 0$ | 4,948 | 7,723 | TN:39.05% | FN: 10.84% |
| $Y_{actual} = 1$ | 8,726 | 71,702 | FP: 60.95% | TP: 89.15% |
| Panel C. LASSO regression | | | | |
| | $Y_{pred} = 0$ | $Y_{pred} = 1$ | Misclassification rate: 18.55% | |
| $Y_{actual} = 0$ | 3,661 | 9,010 | TN: 28.89% | FN: 10.27% |
| $Y_{actual} = 1$ | 8,264 | 72,164 | FP: 71.1% | TP: 89.72% |

*Notes.* Testing sample (2011). $Y_{actual}$ is 1 if the actual status is to be creditworthy, 0 otherwise; $Y_{pred}$ is 1 if a creditworthy observation is predicted (predicted probability $\geq 0.5$), 0 otherwise. FP is the *false positive rate* computed as the percentage of observations predicted positive, but that are actually negative, over the total number of actually negative observations; TP is the *true positive rate* computed as the percentage of observations predicted positive, that are actually positive, over the total number of actually positive observations; FN is the *false negative rate* computed as the percentage of observations predicted negative, but that are actually true, over the total number of actually positive observations; TN is the *true negative rate* computed as the percentage of observations predicted negative, but that are actually negative, over the total number of actually negative observations.

Figure A6.3: Lift curves



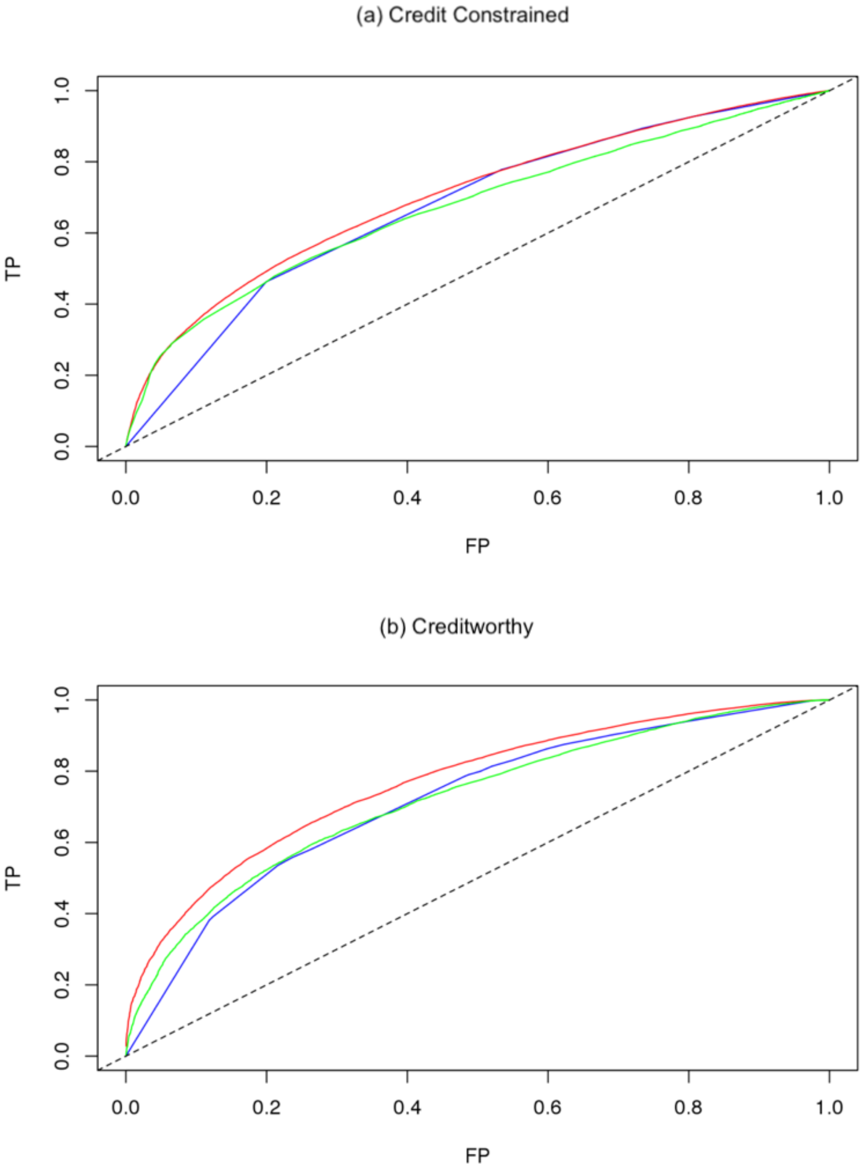
(a) Credit Constrained

(b) Creditworthy

*Notes.* Testing sample (2011). The vertical axis shows the true positive ratio. On the horizontal axis the percentage of observations classified as positive, choosing first those with the highest predicted probability. Color legend: red is the random forest Lift curve; blue is the decision tree Lift curve; green is the LASSO Lift curve.

The second approach considers the entire set of possible thresholds that can be used to classify each observation as target or not. By changing the threshold, we obtain, for each algorithm, different combinations of the false positive rate (FP) and true positive rate (TP). The Receiver Operating Characteristic (ROC) curve shows all possible combinations for each algorithm (Hastie et al., 2009). Again, the diagonal line is a random classifier leading to equality between FP and TP rates. If one uses a better classifier, she should expect to have a TP rate higher than that obtained from the random classifier for each FP rate. This provides a graphical representation of the trade-off between the benefits of good positive classification and the costs implied by prediction errors. Looking at the ROC, the best classifier in both exercises is random forest (Figure A6.4).

Figure A6.4: ROC curves



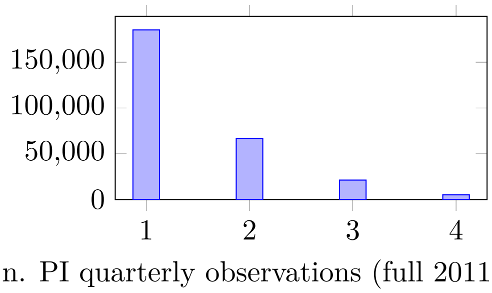
(a) Credit Constrained

(b) Creditworthy

*Notes.* Testing sample (2011). The vertical axis shows the true positive ratio. The horizontal axis shows the false positive ratio. Color legend: red is the random forest ROC curve; blue is the decision tree ROC curve; green is the LASSO ROC curve.

### A.7 Additional Figures

Figure A1: Frequency of the numbers of quarterly loan applications by the same firm



## A.8 Additional Tables

Table A1: Complete list of variables and a brief description

| Variable | Source | Description |
|---|---|---|
| draz | CR (elaboration) | Binary response variable identifying whether a firm is constrained (=1) or not (=0) |
| credit_worthy | CR (elaboration) | Binary response variable identifying whether a firm is creditworthy (=1) or not (=0) |
| util_Lq0 | CR | Amount of bank loans granted and actually used by the firm in the quarter in which the PI request is issued (Lq0) |
| util_Lq1 | CR | Variable util_Lq0 lagged by 1 quarter (Lq1) |
| util_Lq2 | CR | Variable util_Lq0 lagged by 2 quarters (Lq2) |
| util_Lq3 | CR | Variable util_Lq0 lagged by 3 quarters (Lq3) |
| util_Lq5 | CR | Variable util_Lq0 lagged by 5 quarters (Lq5) |
| util_Lq6 | CR | Variable util_Lq0 lagged by 6 quarters (Lq6) |
| util_Lq7 | CR | Variable util_Lq0 lagged by 7 quarters (Lq7) |
| Dutil_Lq04 | CR | Change in the total amount of bank loans granted and actually used by the firm, between the quarter when the PI request was issued and the same quarter in the previous year |
| Dutil_Lq08 | CR | Change in the total amount of bank loans granted and actually used by the firm, between the quarter when the PI request was issued and the same quarter two years earlier |
| acco_Lq0 | CR | Amount of total bank loans granted to the firm in the quarter in which the PI request is issued (Lq0) |
| acco_Lq1 | CR | Variable acco_Lq0 lagged by 1 quarter (Lq1) |
| acco_Lq2 | CR | Variable acco_Lq0 lagged by 2 quarters (Lq2) |
| acco_Lq3 | CR | Variable acco_Lq0 lagged by 3 quarters (Lq3) |
| acco_Lq5 | CR | Variable acco_Lq0 lagged by 5 quarters (Lq5) |
| acco_Lq6 | CR | Variable acco_Lq0 lagged by 6 quarters (Lq6) |
| acco_Lq7 | CR | Variable acco_Lq0 lagged by 7 quarters (Lq7) |

| | | |
|---|---|---|
| Dacco_Lq04 | CR | Change in the total amount of loans granted to the firm, between the quarter when the PI request was issued and the same quarter in the previous year |
| Dacco_Lq08 | CR | Change in the total amount of loans granted to the firm, between the quarter when the PI request was issued and the same quarter two years earlier |
| nbanks_Lq0 | CR | Number of banks lending money to the firm in the quarter in which the PI request is issued (Lq0) |
| nbanks_Lq1 | CR | Variable nbanks_Lq0 lagged by 1 quarter (Lq1) |
| nbanks_Lq2 | CR | Variable nbanks_Lq0 lagged by 2 quarters (Lq2) |
| nbanks_Lq3 | CR | Variable nbanks_Lq0 lagged by 3 quarters (Lq3) |
| D_nbanksLq04 | CR | Change in the total number of banks lending money to the firm, between the quarter when the PI request was issued and the same quarter in the previous year |
| sof_Lq0 | CR | Binary variable identifying whether a firm has been reported to have bad loans (=1) or not (=0) in the CR in the quarter in which the PI request is issued (Lq0). A firm has bad loans if she is reported as insolvent by any bank, regardless of the amount of loans borrowed from that bank |
| sof_Lq1 | CR | Variable sof_Lq0 lagged by 1 quarter (Lq1) |
| sof_Lq2 | CR | Variable sof_Lq0 lagged by 2 quarters (Lq2) |
| sof_Lq3 | CR | Variable sof_Lq0 lagged by 3 quarters (Lq3) |
| sof_Lq4 | CR | Variable sof_Lq0 lagged by 4 quarters (Lq4) |
| sof_Lq5 | CR | Variable sof_Lq0 lagged by 5 quarters (Lq4) |
| sof_Lq6 | CR | Variable sof_Lq0 lagged by 5 quarters (Lq5) |
| sof_Lq7 | CR | Variable sof_Lq0 lagged by 6 quarters (Lq6) |
| sof_Lq8 | CR | Variable sof_Lq0 lagged by 7 quarters (Lq7) |
| no_aff | CR | Binary variable identifying whether data on firm's credit history is available (=1) or not (=0) in the CR dataset |
| X_1001_Ly1 | Cerved | Intangible assets |
| D_1001_ | Cerved | Change in the variable X_1001_Ly1 with respect to the previous year |
| X_1002_Ly1 | Cerved | Tangible fixed assets |

| | | |
|---|---|---|
| X_1023_Ly1 | Cerved | Long term debts |
| D_1023_ | Cerved | Change in the variable X_1023_Ly1 with respect to the previous year |
| X_1024_Ly1 | Cerved | Long term debts towards banks |
| D_1024_ | Cerved | Change in the variable X_1024_Ly1 with respect to the previous year |
| X_1047_Ly1 | Cerved | Long term debts: other financial liabilities |
| D_1047_ | Cerved | Change in the variable X_1047_Ly1 with respect to the previous year |
| X_1027_Ly1 | Cerved | Short term debts towards banks |
| D_1027_ | Cerved | Change in the variable X_1027_Ly1 with respect to the previous year |
| X_1048_Ly1 | Cerved | Short term debts: other financial liabilities |
| D_1048_ | Cerved | Change in the variable X_1048_Ly1 with respect to the previous year |
| X_1033_Ly1 | Cerved | Short-term total liabilities |
| D_1033_ | Cerved | Change in the variable X_1033_Ly1 with respect to the previous year |
| X_1034_Ly1 | Cerved | Liabilities, net of advances received |
| D_1034_ | Cerved | Change in the variable X_1034_Ly1 with respect to the previous year |
| X_1051_Ly1 | Cerved | Net revenues |
| D_1051_ | Cerved | Change in the variable X_1051_Ly1 with respect to the previous year |
| X_1054_Ly1 | Cerved | Production value |
| D_1054_ | Cerved | Change in the variable X_1054_Ly1 with respect to the previous year |
| X_1058_Ly1 | Cerved | Operating value added |
| D_1058_ | Cerved | Change in the variable X_1058_Ly1 with respect to the previous year |
| X_1059_Ly1 | Cerved | Labor cost |
| D_1059_ | Cerved | Change in the variable X_1059_Ly1 with respect to the previous year |
| X_1060_Ly1 | Cerved | Gross operating margin |
| D_1060_ | Cerved | Change in the variable X_1060_Ly1 with respect to the previous year |
| X_1067_Ly1 | Cerved | Net financial income |
| D_1067_ | Cerved | Change in the variable X_1067_Ly1 with respect to the previous year |
| X_1068_Ly1 | Cerved | Current profit before financial charges in the current year |
| D_1068_ | Cerved | Change in the variable X_1068_Ly1 with respect to the previous year |
| X_1069_Ly1 | Cerved | Financial charges |
| D_1069_ | Cerved | Change in the variable X_1069_Ly1 with respect to the previous year |

| | | |
|---|---|---|
| rating_Ly1 | Cerved | Rating index produced by Cerved measuring firms' level of riskiness, based on the elaboration of balance-sheet data: the index ranges from 1 to 9, higher values are associated to higher risk. The index refers to the most recent balance-sheet data available when the PI is issued (Ly1) |
| rating_Ly2 | Cerved | Variable rating_Ly1 lagged by 1 year (Ly2) |
| imm_Ly1 | Cerved (elaboration) | Total assets (intangible + tangible assets); it is based on the most recent balance-sheet data available when the PI was issued (Ly1) |
| imm_Ly2 | Cerved (elaboration) | Variable imm_Ly1 lagged by 1 year (Ly2) |
| roa_Ly1 | Cerved (elaboration) | Return on assets index; it is based on the most recent balance-sheet data available when the PI was issued (Ly1) |
| roa_Ly2 | Cerved (elaboration) | Variable roa_Ly1 lagged by 1 year (Ly2) |
| MOLatt_Ly1 | Cerved (elaboration) | Operating margin on assets index; it is based on the most recent balance-sheet data available when the PI was issued (Ly1) |
| MOLatt_Ly2 | Cerved (elaboration) | Variable MOLatt_Ly1 lagged by 1 year (Ly2) |
| PNnull_Ly1 | Cerved (elaboration) | Binary variable identifying whether the firm has null equity (=1) or not (=0); it is based on the most recent balance-sheet data available when the PI was issued (Ly1) |
| PNnull_Ly2 | Cerved (elaboration) | Variable PNnull_ly1 lagged by 1 year (Ly2) |
| PNneg_Ly1 | Cerved (elaboration) | Binary variable identifying whether the firm has negative equity (=1) or not (=0); it is based on the most recent balance-sheet data available when the PI was issued (Ly1) |
| PNneg_Ly2 | Cerved | Variable PNneg_ly1 lagged by 1 year (Ly2) |

| | | |
|---|---|---|
| LEVclass_Ly1 | Cerved (elaboration) | Leverage class, based on the most recent balance-sheet data available when the PI was issued (ly1). Leverage classes are defined as: Class 1: 25 < LEV_Ly1 ≤ 50; Class 2: 0 ≤ LEV_Ly1 ≤ 25; Class 3: 50<LEV_Ly1 ≤ 75; Class 4: 75 < LEV_Ly1 ≤ 100; Class 5: LEV_Ly1 < 0 or LEV_Ly1 > 100. The Leverage index is obtained as the ratio between total debts and the sum of total debts and equity, i.e. LEV_Ly1=debfin_Ly1/(debfin_Ly1+_1020_Ly1) |
| LEVclass_Ly2 | Cerved (elaboration) | Variable LEVclass_Ly1 lagged by 1 year (Ly2) |
| South | Cerved (elaboration) | Binary variable identifying whether the firm is located in the South of Italy (=1) or not (=0) |
| Industria | Cerved (elaboration) | Binary variable identifying whether the firm works in the industrial cluster (=1) or not (=0) according to the ATECO07 classification rules |

Table A2: Summary statistics

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| draz | 278,355 | 0.662291 | 0.4729296 | 0 | 1 |
| credit__worthy | 278,355 | 0.8637352 | 0.3430702 | 0 | 1 |
| nbanks__Lq0 | 278,355 | 3.500742 | 3.925006 | 0 | 65 |
| nbanks__Lq1 | 278,355 | 3.381549 | 3.886809 | 0 | 63 |
| nbanks__Lq2 | 278,355 | 3.379946 | 3.868736 | 0 | 65 |
| nbanks__Lq3 | 278,355 | 3.387832 | 3.862846 | 0 | 68 |
| rating__Ly1 | 278,355 | 5.139926 | 1.950705 | 1 | 9 |
| rating__Ly2 | 278,355 | 5.16766 | 1.946496 | 1 | 9 |
| X__1001__Ly1 | 278,355 | 759.3384 | 33890.19 | 0 | 8006477 |
| X__1016__Ly1 | 278,355 | 1382.7 | 26958.24 | 0 | 8515841 |
| X__1020__Ly1 | 278,355 | 3057.165 | 37774.63 | -805292 | 7137686 |
| X__1058__Ly1 | 278,355 | 1972.476 | 30238.87 | -248055 | 7929319 |
| X__1059__Ly1 | 278,355 | 1303.892 | 19034.31 | 0 | 5689109 |
| X__1069__Ly1 | 278,355 | 149.6218 | 3088.111 | 0 | 764986 |
| X__1076__Ly1 | 278,355 | 95.99631 | 7723.9 | -1243793 | 1765924 |
| X__1021__Ly1 | 278,355 | 327.778 | 9691.388 | 0 | 2481209 |
| X__1067__Ly1 | 278,355 | 81.77763 | 3248.203 | -616209 | 1180472 |
| X__1073__Ly1 | 278,355 | 150.8776 | 3197.765 | -161024 | 944818 |
| X__1068__Ly1 | 278,355 | 410.9461 | 10553.39 | -1201972 | 2854104 |
| X__1074__Ly1 | 278,355 | 95.7519 | 7717.045 | -1243793 | 1759069 |
| X__1060__Ly1 | 278,355 | 668.5834 | 14991.26 | -440237 | 3862118 |
| X__1024__Ly1 | 278,355 | 1039.929 | 18104.29 | 0 | 4145568 |
| X__1047__Ly1 | 278,355 | 57.51599 | 2282.624 | 0 | 759100 |

| | | | | | |
|---|---|---|---|---|---|
| X_1048_Ly1 | 278,355 | 19.51724 | 25.58316 | 0 | 3694.11 |
| sof_Lq0 | 278,355 | 0.0152791 | 0.1226607 | 0 | 1 |
| sof_Lq1 | 278,355 | 0.011956 | 0.1086879 | 0 | 1 |
| sof_Lq2 | 278,355 | 0.0101489 | 0.1002295 | 0 | 1 |
| sof_Lq3 | 278,355 | 0.0087335 | 0.0930441 | 0 | 1 |
| sof_Lq4 | 278,355 | 0.007609 | 0.0868972 | 0 | 1 |
| sof_Lq5 | 278,355 | 0.0065492 | 0.0806618 | 0 | 1 |
| sof_Lq6 | 278,355 | 0.0059025 | 0.076601 | 0 | 1 |
| sof_Lq7 | 278,355 | 0.0052271 | 0.0721099 | 0 | 1 |
| sof_Lq8 | 278,355 | 0.0046919 | 0.0683363 | 0 | 1 |
| eta | 278,355 | 15.72364 | 12.54063 | 1 | 158 |
| no_aff | 278,355 | 0.130427 | 0.3367732 | 0 | 1 |
| ben_FG_T0 | 278,355 | 0.1046901 | 0.3061542 | 0 | 1 |
| roa_Ly1 | 278,355 | 2.168137 | 166.7608 | -78600 | 2644 |
| roa_Ly2 | 278,355 | 3.911367 | 32.66447 | -2500 | 13400 |
| MOLatt_Ly1 | 278,355 | 6.63502 | 132.5956 | -59000 | 2650 |
| MOLatt_Ly2 | 278,355 | 8.254592 | 34.88289 | -2480 | 13400 |
| PNnull_Ly1 | 278,355 | 0.0022741 | 0.0476331 | 0 | 1 |
| PNnull_Ly2 | 278,355 | 0.0027339 | 0.0522155 | 0 | 1 |
| PNneg_Ly1 | 278,355 | 0.0610156 | 0.2393594 | 0 | 1 |
| PNneg_Ly2 | 278,355 | 0.0575811 | 0.2329501 | 0 | 1 |
| defret_t | 278,355 | 0.0573548 | 0.2325198 | 0 | 1 |
| LEVclass_Ly1 | 278,355 | 3.434046 | 1.043287 | 1 | 5 |
| LEVclass_Ly2 | 278,355 | 3.437941 | 1.034682 | 1 | 5 |
| South | 278,355 | 0.2081766 | 0.4060046 | 0 | 1 |
| Industria | 278,355 | 0.2875608 | 0.4526261 | 0 | 1 |
| D_nbanksLq04 | 278,355 | 0.1170053 | 1.21296 | -29 | 24 |
| D_1001_ | 278,355 | 25.71761 | 7763.913 | -664320 | 3417254 |
| D_1002_ | 278,355 | 37.75182 | 8490.817 | -3238995 | 846731 |
| D_1005_ | 278,355 | 123.612 | 10507.52 | -2390097 | 1117130 |

| | | | | | |
|---|---|---|---|---|---|
| D__1069__ | 278,355 | -41.947 | 1238.145 | -367852 | 147223 |
| D__1076__ | 278,355 | -15.43817 | 7101.941 | -1170543 | 1619477 |
| D__1021__ | 278,355 | 18.61368 | 2263.007 | -389231 | 828918 |
| D__1067__ | 278,355 | -23.37891 | 3321.527 | -1029975 | 304159 |
| D__1073__ | 278,355 | -2.192438 | 1625.085 | -252359 | 500797 |
| D__1074__ | 278,355 | -14.7069 | 7093.923 | -1170543 | 1619477 |
| D__1060__ | 278,355 | -14.83233 | 4283.792 | -701717 | 610715 |
| D__1024__ | 278,355 | 33.70418 | 7625.363 | -1701778 | 1173540 |
| D__1047__ | 278,355 | 6.703725 | 3721.354 | -903000 | 757749.8 |
| D__1027__ | 278,355 | 94.8289 | 14571.34 | -3383334 | 1403034 |
| D__1048__ | 278,355 | .6057571 | 19.13432 | -1062.5 | 3133.4 |
| D__1023__ | 278,355 | 88.75335 | 10911.99 | -1737727 | 2680583 |
| D__1068__ | 278,355 | -60.80531 | 5795.715 | -1323831 | 364816 |
| util__Lq0 | 278,355 | 2903870.75 | 20827374 | 0 | 2654373632 |
| util__Lq1 | 278,355 | 2873342.25 | 20780410 | 0 | 3019080448 |
| util__Lq2 | 278,355 | 2835801.25 | 21281142 | 0 | 3801057536 |
| util__Lq3 | 278,355 | 2796632.25 | 21530432 | 0 | 3942093056 |
| util__Lq5 | 278,355 | 2734884.75 | 22740976 | 0 | 4270024192 |
| util__Lq6 | 278,355 | 2720034.5 | 23230754 | 0 | 4263706368 |
| util__Lq7 | 278,355 | 2704553.5 | 23754744 | 0 | 4270024192 |
| acco__Lq0 | 278,355 | 4490085.5 | 31087684 | 0 | 3998619648 |
| acco__Lq1 | 278,355 | 4467964.5 | 30979026 | 0 | 4017632512 |
| acco__Lq3 | 278,355 | 4462579.5 | 31500472 | 0 | 4546283520 |
| acco__Lq2 | 278,355 | 4462436 | 31043964 | 0 | 4522943488 |
| acco__Lq5 | 278,355 | 4468116 | 33159358 | 0 | 4900827648 |
| acco__Lq6 | 278,355 | 4475285.5 | 33659944 | 0 | 4897276416 |
| acco__Lq7 | 278,355 | 4480276 | 34099640 | 0 | 4945260032 |
| X__1002__Ly1 | 278,355 | 2672.047363168 | 48989.71094 | 0 | 14006136 |
| X__1005__Ly1 | 278,355 | 4727.964355 | 83007.60156 | 0 | 15737555 |
| X__1014__Ly1 | 278,355 | 6312.870605 | 130269.8359 | 0 | 45944344 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| debfin_Ly2 | 278,355 | 6472.198242 | 130350.25 | 0 | 43878748 |
| D_util_Lq04 | 278,355 | 147899.1406 | 7653159 | -1778668928 | 977245184 |
| D_util_Lq08 | 278,355 | 197887.8594 | 12428891 | -2410919680 | 1261780608 |
| D_acco_Lq04 | 278,355 | 22902.0957 | 9866105 | -2004762112 | 1467343744 |
| D_acco_Lq08 | 278,355 | -4529.17041 | 14285915 | -2588611328 | 1484282240 |

Table A3: List of variables after the screening for the credit-constrained exercise

| **Variable name** |
|---|
| acco_Lq0 |
| ben_FG_T0 |
| D_1001 |
| debfin_Ly1 |
| D_util_Lq04 |
| eta |
| imm_Ly1 |
| LEVclass_Ly1 |
| nbanks_Lq0 |
| no_aff |
| PNneg_Ly1 |
| PNneg_Ly2 |
| PNnull_Ly2 |
| rating_Ly1 |
| sof_Lq4 |
| sof_Lq8 |
| X_1023_Ly1 |

Table A4: List of variables after the screening for the creditworthy exercise

| **Variable** |
|---|
| rating_Ly1 |
| rating_Ly2 |
| X_1001_Ly1 |
| X_1054_Ly1 |
| X_1059_Ly1 |
| X_1060_Ly1 |
| X_1023_Ly1 |
| sof_Lq4 |
| sof_Lq8 |
| eta |
| no_aff |
| ben_FG_T0 |
| imm_Ly1 |
| MOLatt_Ly2 |
| PNnull_Ly1 |
| PNnull_Ly2 |
| PNneg_Ly1 |
| PNneg_Ly2 |
| debfin_Ly1 |
| LEVclass_Ly1 |
| LEVclass_Ly2 |
| South |
| D_1001_ |

Table A5: Parametric Fuzzy-RDD analysis: Outcome variables

| | (a) Full sample | (b) ML target = 1 | (c) ML target = 0 |
|---|---|---|---|
| Panel A. Bank granted credit (growth rate) | | | |
| ITT | 0,013***<br>(0,004) | 0,016***<br>(0,005) | 0,002<br>(0,011) |
| FS | 0,039***<br>(0,008) | 0,030***<br>(0,009) | 0,029<br>(0.028) |
| LATE | 0,337***<br>(0,124) | 0,535***<br>(0,218) | 0,078<br>(0,385) |
| Panel B. Investments (growth rate of fixed assets) | | | |
| ITT | -0,007<br>(0,005) | -0,006<br>(0,006) | 0,002<br>(0,014) |
| FS | 0,037***<br>(0,008) | 0,029***<br>(0,009) | 0,038<br>(0,029) |
| LATE | -0,186<br>(0,133) | -0,212<br>(0,198) | 0,043<br>(0,361) |
| Panel C. Sales (growth rate) | | | |
| ITT | -0,001<br>(0,003) | -0,003<br>(0,004) | 0,003<br>(0,009) |
| FS | 0,037***<br>(0,008) | 0,026***<br>(0,009) | 0,046*<br>(0,027) |
| LATE | -0,022<br>(0,087) | -0,107<br>(0,152) | 0,075<br>(0,205) |
| Panel D. Prob. of adjusted bad loans | | | |
| ITT | -0,003<br>(0,003) | 0,002<br>(0,003) | -0,013**<br>(0,007) |
| FS | 0,038***<br>(0,008) | 0,029***<br>(0,008) | 0,040**<br>(0,017) |
| LATE | -0,072<br>(0,071) | 0,064<br>(0,091) | -0,325<br>(0,212) |

*Notes.* $***$ p-val $\leq$ 0.01 , $**$ p-val $\leq$ 0.05 , $*$ p-val $\leq$ 0.1 . Selected sample of 59,064 firms (see Subsection 5.2). Fuzzy-RDD parametric estimates. Outliers below the 5th or above the 95th percentile were dropped. Standard errors in brackets.

The selection of the best polynomial degree for the RDD global parametric estimate is based on the Akaike Information Criterion (AIC): in the whole sample and in the subsample MLtarget=1 we always rely on a 1st order polynomial degree specification; in the subsample MLtarget=0, we rely on a 1st order polynomial degree specification for the outcome “Prob. of adjusted bad loans”, on a 2nd order polynomial degree specification for the outcome “Sales (growth rate)” and a 3rd order polynomial degree specification for the outcomes Bank granted credit (growth rate) and Investments (growth rate of fixed assets). The same polynomial degree specification is applied to 1st stage, 2nd stage and ITT regressions.

## Chapter 3

# Social preferences and strategic incentives for cooperation in infinitely repeated Prisoner Dilemmas [1]

[1] We thank all the authors who made the data we use in this article available.

## 3.1 Introduction

In this work we want to investigate the determinants of individuals' cooperative behavior in infinitely repeated Prisoner Dilemmas (henceforth PDs).

The main contribution of this work is to shed light on how the strategic incentives of the game, set by structural game parameters, and individual social preferences affect cooperation and interact with each other across strategically different contexts. The ultimate objective is bridge the two strands of the experimental literature that, separately, looked at how strategic incentives, on one side, and individual characteristics, on the other side, can account for subjects' cooperativeness in infinitely repeated PDs (see Roth and Murnighan (1978),Murnighan and Roth (1983), Dal Bó (2005), Blonski et al. (2011),Dal Bó and Fréchette (2011), Dal Bó and Fréchette (2018) for the former strand of the literature and Sabater-Grande and Georgantzis. (2002), Dreber et al. (2014), Davis et al. (2016), Proto et al. (2019) for the latter) .

First we present a meta-analysis run on an extended version of the dataset collected by Dal Bó and Fréchette (2018), where we rely on simple supervised-learning algorithms to test the ability of structural game parameters to predict sunjects' cooperative behavior in infinitely repeated PDs. Our findings, in line with the evidence brought by previous literature, confirm that structural game parameters have some predictive power, which increases as subjects gain experience, that the two composite indicators derived by Dal Bó and Fréchette (2011) ($sizeBAD$) and Blonski et al. (2011) ($\delta - \delta_{RD}$) to explain cooperation levels seem to, indeed, capture a great share of the information relevant to predict cooperation. However, models that use only structural game characteristics as predictors exhibit a poorer prediction performance when it comes to the ability to predict cooperative choices in contexts that are not 'strategically' conducive to cooperation - namely where cooperation is not sustainable as a long-run equilibrium - compared to cases where cooperation can be sustained in equilibrium.

To further explore the role of strategic incentives and social preferences on cooperation, we propose a novel experimental design to collect data that would allow us to answer our main research questions:

- What is the role of social preferences as cooperation predictors? Do social preferences perform better as cooperation predictors when cooperation is not sustainable as an equilibrium?

- How do strategic and non-strategic motives for cooperation interact? Are subjects who exhibit non-strategic taste for cooperation less sensitive to changes in strategic incentives to cooperation?

- What is the role of subjects with non-strategic taste for cooperation in shaping overall cooperativeness? Is the share of non-strategic cooperators a relevant factor in shaping the level of cooperation attainable when cooperation is and is not sustainable as an equilibrium?

## 3.2 Motivation & Theoretical Framework

The study of social dilemmas, such as PDs, where individual and collective interests are in conflict, has long attracted the interest of economists, who contributed to this field of research both theoretically and experimentally. In particular, studying how individuals behave in contexts where they face the same social dilemma an indefinite number of times, is extremely relevant from a policy perspective since these contexts more closely mirror real-world situations where subjects are not informed ex-ante of the duration of their future interactions with others.

Contrary to the case of one-shot or finitely-repeated interactions, however, when we introduce infinitely repeated interactions, standard economic theory fails to postulate univocal predictions on subjects' behavior, opening for the possibility that even among purely self-interested individuals cooperation could be sustained in equilibrium, if players are sufficiently patient [2].

Likewise, the empirical evidence collected so far did not succeed in isolating what factors can best predict the emergence of cooperative long-run equilibria and in explaining the heterogeneity in cooperativeness observed in contexts that should be equivalent to subjects from a theoretical point of view.

We focus our attention on infinitely repeated PDs because they represent the simplest form of infinitely repeated game where the essence of the tension between personal interest and social optimum, which is at the heart of every social dilemma, is well captured.

[2] Folk Theorem, see Fudenberg and Maskin (1986)

In a canonical 2x2 PD, see Table 3.1, subjects face a binary choice on whether to cooperate or defect, given the following payoffs:

T: Temptation's payoff from defecting when the other cooperates

R: Reward from mutual cooperation

P: Punishment from mutual defection

S: Sucker's payoff from cooperating when the other defects

where $T > R > P > S$ and, typically, $2R > (T + S)$, which makes joint cooperation more profitable than alternating between cooperation and defection.

Table 3.1: Prisoners' Dilemma Row Player's Payoffs

| | Original | | | Normalized | |
|---|---|---|---|---|---|
| | C | D | | C | D |
| C | R | S | C | $\frac{R-P}{R-P} = 1$ | $\frac{S-P}{P-P} = -l$ |
| D | T | P | D | $\frac{T-P}{P-P} = 1 + g$ | $\frac{P-P}{P-P} = 0$ |

If we consider the normalized version [3] of the payoffs' matrix, the number of relevant parameters in the stage game reduces to two: the gain from unilateral defection ($g$) and the loss from unilateral cooperation ($l$). When they are implemented in the laboratory, following the pioneering contribution of Roth and Murnighan (1978) and Murnighan and Roth (1983), infinitely repeated PDs essentially transform into 'indefinitely repeated' games where subjects play the stage PD game an indefinite number of times and new relevant parameters emerge: in 'supergame', subjects are matched to a partner and play the stage PD game with the same partner a number of rounds that depends on a pre-set 'continuation probability' ($\delta$); when the supergame is over, subjects are re-matched to new partners and play the same repeated PD game again; this procedure is iterated for each supergame, with the total number of supergames to be played being pre-determined.

[3] The normalized version of the payoffs' matrix is obtained by applying a monotonic linear transformation to the original matrix.

Two strands of the experimental literature on infinitely repeated PDs, so far largely unrelated, focused on the role of structural game parameters (like the continuation probability $\delta$, the gain from unilateral defection $g$, etc.) on cooperation, on one side, and on the role of personal characteristics, including preferences, on the other side.

The strand of the economic literature focusing on the role of structural game parameters leveraged on the most recent advances in the theory of infinitely repeated games to study whether and to what extent structural game parameters can impact cooperation levels in PDs. A recent work by Dal Bó and Fréchette (2018) offers a comprehensive review of the main contributions on this topic (Roth and Murnighan (1978),Murnighan and Roth (1983), Dal Bó (2005), Blonski et al. (2011),Dal Bó and Fréchette (2011)) while providing some empirical evidence on how structural game parameters affect cooperation by relying on meta-data that bring together more than 150.000 observations collected from 15 different experimental papers. Cooperation is generally found to be increasing in the probability of future interactions and, on average, greater when cooperation can be supported as a Subgame Perfect Nash Equilibrium (SGPE) or as a Risk Dominant Equilibrium (RD) [4] , although a large amount of variation is left unexplained. In the attempt to dig deeper in the unexplained variation in cooperation levels observed, different approaches have been followed, which tried to best combine the information contained by the structural parameters of the game into composite indicators: the two most prominent examples are provided by Blonski et al. (2011), who build a continuous measure of 'how risk-dominant' is cooperation [5] based on the distance between $\delta$ and $\delta_{RD}$ (where $\delta_{RD} = \frac{g+l}{1+g+l}$), and by Dal Bó and Fréchette (2011), who build a continuous measure of how resistant is cooperation to strategic uncertainty based on the basin of attraction of the Always Defect (AD) strategy ($sizeBAD$) [6]. Dal Bó and Fréchette (2018) separately test the ability of these two composite indicators to predict cooperation in their

[4] The concept of Risk Dominance is borrowed from the literature on coordination games, where Harsanyi et al. (1988) define an equilibrium to be risk-dominant to another if the opportunity costs of unilaterally deviating from that equilibrium is higher. Blonski et al. (2011) and Blonski and Spagnolo (2015) develop an equilibrium selection theory for infinitely repeated PDs, moving from the concept of 'strategic risk' by Harsanyi et al. (1988).

[5] Assuming subjects are uncertain about their opponent's moves, we consider a strategy to be *risk-dominant* if it is a best-response to the other player randomizing with a 50-50% probability between a cooperative strategy (Grim) and a non-cooperative strategy (Always Defect).

[6] The basin of attraction of AD against a cooperative strategy like Grim corresponds to the maximum probability of the other player playing Grim that makes playing AD optimal. When cooperation can be supported in equilibrium it is equal to $\frac{(1-\delta)l}{(1-(1-\delta)(1+g-l))}$ . When cooperation is not supported in equilibrium this maximum probability is equal to 1.

meta-data: they show that cooperation is positively correlated with the distance $(\delta - \delta_{RD})$, especially when cooperation is Risk Dominant (treatments where $\delta > \delta_{RD}$), and negatively correlated to the size of the basin of attraction of AD ($sizeBAD$) when cooperation is Risk Dominant. The question on which of the two indices predicts best cooperation is left unanswered, given the high correlation between the two indices in the meta-data.

Another growing strand of the literature, instead, recently focused on the role of personal characteristics - including preferences - on cooperation in infinitely repeated games. The role of individual preferences on cooperation has already been analyzed in the context of finitely repeated games, where free-riding and defection are the only possible outcomes for rational and self interested individuals, and thus preferences - in particular social preferences - are deemed necessary to justify the emergence of cooperation (Fehr and Fischbacher (2003)). Some studies focused on the role of social preferences on cooperation in one-shot PDs and found some evidence of a positive correlation between cooperation and other-regarding attitudes, which were typically measured through other games: Blanco et al. (2011) report the presence of a positive correlation between cooperative behavior in a sequential one-shot PD and other-regarding behavior measured in terms of giving in a modified dictator game and in an ultimatum game, and likewise, Capraro et al. (2014) find a positive correlation between cooperative behavior in a one-shot continuous-choice PD and giving in a dictator game. These results seem to support the hypothesis by Peysakhovich et al. (2014) that individuals display a "cooperative phenotype", which is not correlated with norm-enforcing punishment or non-competitiveness but is valid in a general domain and substantiates in a temporally stable inclination towards paying costs to benefit others that makes subjects' behavior consistent across different decision scenarios.

The picture appears, however, different when we introduce infinite repetition. Dal Bó and Fréchette (2018) offer a comprehensive review of the main findings from this strand of the literature, where no conclusive evidence has yet been found in favor of the presence of a systematic effect of individual characteristics such as risk aversion (Sabater-Grande and Georgantzis. (2002), Dreber et al. (2014), Davis et al. (2016), Proto et al. (2019)), social preferences (Dreber et al. (2014), Davis et al. (2016)), intelligence (Proto et al. (2019)), patience (Davis et al. (2016), Kim (2019)) etc. on cooperativeness. Interestingly, the effect of some of these characteristics seems to be sensitive to the strategic environment defined by the structural parameters of the infinitely repeated game, as it is the case for social preferences, which are found to be predictive of cooperation only when cooperation is not sustainable as an equilibrium: Dreber et al. (2014), by letting subjects play a standard dictator game (DG) after a series of indefinitely repeated PDs where cooperation either is or is not an equilibrium (between-subjects design), find that cooperation is related to giving in the DG only when the structural game parameters of the PD make cooperation not sustainable as a long-run equilibrium. Similarly, Arechar et al. (2018), who let their subjects play a standard DG before and after an infinitely repeated PD with varying continuation probability (within-subjects design), find that the giving behavior in the first DG predicts both the level of cooperation and the strategies played by subjects in the PD: givers are found more likely to cooperate and less likely to choose a non-cooperative strategy like 'Always Defect", but only when cooperation is not sustainable as an equilibrium. Consistently, Davis et al. (2016), who let their subjects play an infinitely repeated PD where cooperation is an equilibrium and later ask the same subjects to perform a series of tasks aimed to measure their personal attitude over a series of dimensions, find that cooperation is not systematically related to social preferences, measured in terms of altruism and behavior in a trust game.

This evidence could be consistent with the idea that individuals could also have a *non-strategic* taste for cooperation, in addition to a *strategic* taste for cooperation. In this framework, individuals having a strategic taste for cooperation would exhibit a cooperative attitude only when the structural characteristics of the game and their expectations are such that cooperation can be a profitable strategy, while individuals having a non-strategic taste for cooperation would exhibit a cooperative attitude even when the structural characteristics of the game and their expectations do not guarantee cooperation to be a profitable strategy. If this was the case, the structural characteristics of the game would be effective predictors of cooperativeness when cooperation is sustainable as an equilibrium but not necessarily otherwise, while, vice-versa, factors explaining the non-strategic taste for cooperation would be relevant predictors of cooperativeness when cooperation is not an equilibrium and not necessarily otherwise.

The behavior of individuals who exhibit a non-strategic taste for cooperation - the *other-*

*regarding types* - could be explained by the presence of some forms of social preferences, which, if strong enough, can alter incentives for cooperation favoring the emergence of cooperative equilibria even in situations where the structural parameters of the game are such that for a purely self-interested and payoff-maximizing individual cooperation would never be sustainable neither as a SGPE nor as a RD equilibrium. These other-regarding types would then be willing to start by playing a cooperative strategy, irrespective of the structural characteristics of the game, thus independently from the presence of strategic incentives to cooperate.

As we will briefly discuss later, different models of social preferences could be used to model how individuals map PD's monetary payoffs into utilities when deciding over their actions and strategies, and for given social preferences parameters, it clearly emerges - irrespective of the model specification chosen - that cooperation can emerge as a possible equilibrium or even become a dominant strategy in a one-shot or infinitely repeated PDs where cooperation would not be in principle sustainable as an equilibrium among self-interested players.

The rest of individuals - the *self-interested types* - might only exhibit a strategic taste for cooperation, according to which they would initiate a cooperative strategy only if they consider cooperation to be profitable from a strategic point of view, based on their expectations on how many individuals in the population would also cooperate, either because they rationally internalize that the structural parameters of the game are such that cooperating would be profitable in the long run, or because of non-strategic reasons. A residual part of the self-interested types, instead, would never exhibit a taste for cooperation, always opting for a defective strategy irrespective of the structural parameters of the game and of their expectations on the fraction of cooperators in the population.

In this framework, when the structural game parameters are such that cooperation is not sustainable as a long-run equilibrium, we would observe cooperation only by other-regarding types and by the fraction of self-interested types who have at least a small positive expectation on the fraction of individuals in the population who would be motivated to cooperate by non-strategic reasons. This prediction would be consistent with what was originally postulated by Kreps et al. (1982), who focused on the case of finitely repeated PDs and rationalized the emergence of cooperation in a context where cooperation is never sustainable as an equilibrium for rational self-interested individuals: according to Kreps et al. (1982) it would be sufficient to assume that players have incomplete information concerning preferences for cooperation in the population - so that players assign a small positive probability to the possibility that their opponent might choose to cooperate because he/she 'enjoys cooperation' - in order to produce a sequential cooperative equilibrium.

To elaborate on what type of social preferences could shape a *non-strategic taste* for cooperation, we refer to a modified version of the canonical 2x2 PD matrix, where utilities associated to each action are displayed instead of crude payoffs, see Table 3.2.

If we assume players are self-interested and rational, the mapping of payoffs through the utility function does not affect the structure of the matrix, and, under the most simple and standard assumption of linearity $U_i(A_i, A_j) = \pi_i(A_i, A_j)$ [7], the utility that subjects assign to each of the possible actions will exactly correspond to their own monetary payoff associated to that action, see Table 3.2 (panel a). If both players are self-interested and rational, (Defect, Defect) will be the unique Nash Equilibrium (NE) of the stage game and the same will hold when the game is repeated an infinite time of times but the structural parameters of the game are such that cooperation would not be sustainable as a SGPE or a RD equilibrium. Instead, if players exhibit some forms of social preferences, this would affect the mapping of game's monetary payoffs into utilities (as shown also by Duffy and Muñoz-García (2012)) , possibly allowing for other equilibria even in the stage game.

When accounting for social preferences, we assume players exhibit social preferences à la Charness and Rabin (2002). In the two-players case, in absence of reciprocity concerns, the utility of the individual $i$ in the pair would be given by:

$$U_i(A_i, A_j) = f(\pi_i(A_i, A_j), \pi_j(A_i, A_j)) = (\beta_i r + \alpha_i s)\pi_j + (1 - \beta_i r - \alpha_i s)\pi_i$$

where

$r = 1$ if $\pi_i > \pi_j$, and $r = 0$ otherwise;

$s = 1$ if $\pi_i < \pi_j$, and $s = 0$ otherwise;

[7] $U_i(A_i, A_j)$ is the utility subject $i$ gets based on his own action $A_i$ and the action of the other person in the pair $A_j$ and $\pi_i(A_i, A_j)$ is the payoff subject $i$ realizes based on his own action $A_i$ and the action of the other person in the pair $A_j$.

so that the utility of individual $i$ can also be expressed as:

$$\begin{cases} if\ \pi_i = \pi_j \longrightarrow U_i = \pi_i \\ if\ \pi_i > \pi_j \longrightarrow U_i = \beta_i \pi_j + (1 - \beta_i)\pi_i \\ if\ \pi_i < \pi_j \longrightarrow U_i = \alpha_i \pi_j + (1 - \alpha_i)\pi_i \end{cases}$$

This framework mirrors the behavioral model adopted by Bruhin et al. (2018) to shape social preferences in absence of reciprocity concerns, which is itself inspired by the behavioral models developed by Charness and Rabin (2002) and Fehr and Schmidt (1999). Parameters $\beta_i$ and $\alpha_i$ measure the weights assigned by player $i$ to the payoff of the other player both in a situation of advantageous and disadvantageous inequality, and based on the values of these two paramters it is possible to classify individuals with different types os social preferences.

When both $\beta_i = 0$ and $\alpha_i = 0$ invividuals are purely selfish and do not show any forms of social preferences, caring exclusively about their own payoff, irrespective of their relative position in the pair.

When both $\beta_i > 0$ and $\alpha_i > 0$, instead, individuals are altruistic and always care about the payoff of the other player no matter what is their relative position in the pair, showing concerns both for the maximization of the payoff of the worst-off player and for efficiency. An increase in both $\beta_i$ and $\alpha_i$ signals an increase in the weight player $i$ attaches to the social good, as compared to this own material payoff. When, instead, $\beta_i$ increases and $\alpha_i$ decreases, or more in general the ratio $\frac{\beta_i}{\alpha_i}$ increases, this signals that player $i$ puts relatively more weight to the maximization of the payoff of the worst-off player, and less to the maximization of the total surplus.

When $\beta_i > 0$ but $\alpha_i < 0$, individuals are inequality averse, which implies they are behindness averse but do care about the payoff of the other player when they are better off. Based on the relative size of parameters $\mid \alpha_i \mid$ and $\mid \beta_i \mid$, individuals would either care more about advantageous inequality than disadvantageous inequality ($\mid \alpha_i \mid < \mid \beta_i \mid$) or viceversa ($\mid \alpha_i \mid > \mid \beta_i \mid$), where the latter represents the case originally studied by Fehr and Schmidt (1999).

When both $\beta_i < 0$ and $\alpha_i < 0$ individuals are spiteful and always attach a negative weight to the payoff of the other player, irrespective of their relative position in the pair.

If we assume players have some forms of social preferences ($\beta_i \neq 0$ and $\alpha_i \neq 0$), these preferences will have an impact on how players map payoffs into utilities when playing a PD, see Table 3.2 (panel b).

Table 3.2: Prisoners' Dilemma Row Player's Utilities - C&R

| General | | | Social Preferences à la Charness and Rabin (2002) | | |
|---|---|---|---|---|---|
| | C | D | | C | D |
| C | $U_i(C,C)$ | $U_i(C,D)$ | C | $U_i(C,C) = R$ | $U_i(C,D) = \alpha_i T + (1-\alpha_i)S$ |
| D | $U_i(D,C)$ | $U_i(D,D)$ | D | $U_i(D,C) = \beta_i S + (1-\beta_i)T$ | $U_i(C,C) = P$ |

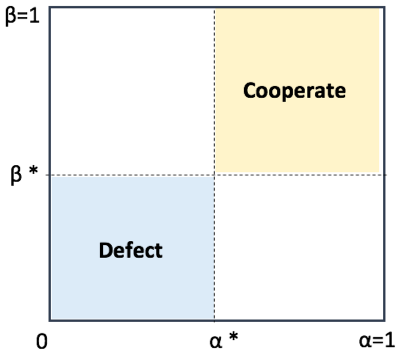
If we assume players have perfect information about social preferences, we have that under some circumstances a cooperative equilibrium can arise even in the one-shot stage game interaction. Indeed, under some circumstances, Cooperation will be a dominant strategy for both players, which will result in an efficient (Cooperate, Cooperate) Equilibrium.

$$\begin{cases} Cooperate\ is\ BR\ to\ Cooperate \longrightarrow when\ \beta_i > \beta_i^* = \frac{(T-R)}{(T-S)} \\ Cooperate\ is\ BR\ to\ Defection \longrightarrow when\ \alpha_i > \alpha_i^* = \frac{(P-S)}{(T-S)} \end{cases}$$

In order to have Cooperation as a best response to Cooperation, intuitively, we need player $i$ to have a high enough $\beta_i$, which implies player $i$ cares enough about the other player's payoff when $\pi_i > \pi_j$, so to compensate the loss in terms of higher material payoff he could have obtained by means of a unilateral defection. When $(T-R) < (T-S)$, which is always the case in a canonical PD where $T > R > P > S$, the threshold value $\beta_i^*$ will be bounded between 0 and 1, which implies the condition will be met only when players have an high enough degree of concern for others ($\beta_i > \beta_i^*$).

Similarly, in order to have Cooperation as a best response to Defection, we need player $i$ to have a high enough $\alpha_i$, which implies player $i$ cares enough about the other player's payoff even when $\pi_i < \pi_j$, so to compensate the loss in terms of material payoff he incurs as a consequence of his opponent's unilateral defection. When $(P-S) < (T-S)$, which is always the case in a canonical PD where $T > R > P > S$, the threshold value $\alpha_i^*$ will be bounded between 0 and 1, which implies the condition will be met only when players have an high enough degree of concern for others ($\alpha_i > \alpha_i^*$).

Figure 3.1: Players Pure Dominant strategies



Therefore, if both players are selfish and self-interested ($\beta_i = \alpha_i = 0$) or exhibit low concerns the social welfare ($\beta_i < \beta_i*$ and $\alpha_i < \alpha_i^*$), the unique NE of the stage game will be (Defect, Defect). If, instead, both players have strong enough concerns for the social welfare, the unique NE of the stage game will be (Cooperate, Cooperate). It is therefore possible to observe cooperative outcomes even in absence of any scope for future interactions.

When we move to the context of infinitely repeated PDs, players are called to play the same stage game for an indefinite number of times with the same partner and every player $i$ discounts the flow of his future payoffs according to a discount factor $0 < \delta_i < 1$. In this context, even in absence of social preferences, the outcome (Cooperate, Cooperate) can be sustained as an equilibrium if players are sufficiently patient, as predicted by the Folk theorem (Fudenberg and Maskin (1986)). This prediction holds whenever:

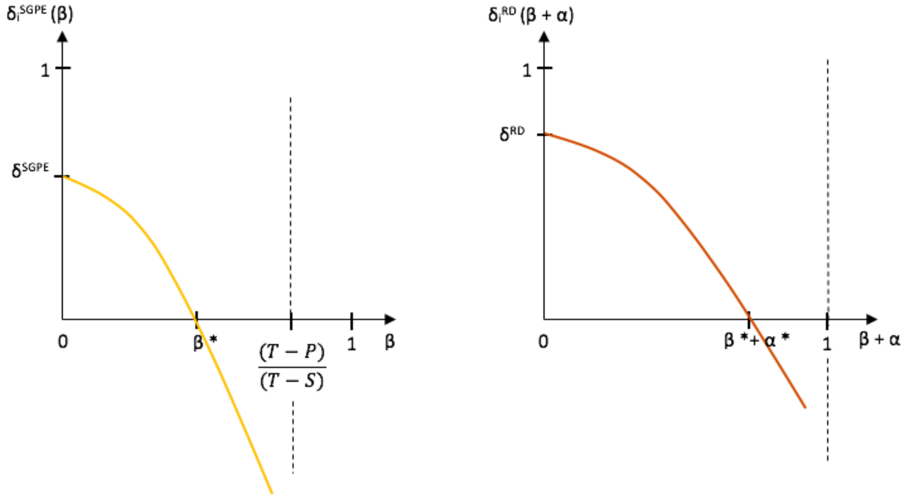
$$\delta^{SGPE} : \sum_{t=0}^{\infty} \delta^t R > T + \sum_{t=1}^{\infty} \delta^t P$$

$$\delta_i > \delta^{SGPE} = \frac{(T-R)}{(T-P)}$$

where $\delta^{SGPE}$ is the threshold value that makes a player indifferent between playing a grim strategy - where the player starts by cooperating in the first round and then keeps cooperating until a defection is observed, switching to defection ever after - and an always-defect strategy, under the assumption that the opponent plays grim.

Under the assumption of both players having strong enough social preferences and perfect

Figure 3.2: $\delta_i^{SGPE}$ and $\delta_i^{RD}$ as a function of $\beta_i$ and $\beta_i + \alpha_i$



information, mutual cooperation can be feasible as a Subgame Perfect NE (SGPE) of the infinitely repeated game under a wider set of parameters. In particular, (Cooperate, Cooperate) will be sustainable as an equilibrium outcome whenever:

$$\delta_i^{SGPE} : \sum_{t=0}^{\infty} \delta^t R > \beta_i S + (1 - \beta_i) T + \sum_{t=1}^{\infty} \delta^t P$$

$$\delta_i > \delta_i^{SGPE} = \frac{(T - R) - \beta_i (T - S)}{(T - P) - \beta_i (T - S)}$$

where $\delta_i^{SGPE}$ is the threshold value that makes a player $i$ indifferent between playing a grim strategy and an always-defect strategy.

When social preferences are absent and $\beta_i = 0$, $\delta_i^{SGPE}$ and $\delta^{SGPE}$ coincide. For positive values of $\beta_i$, below the threshold $\beta_i^*$, $\delta_i^{SGPE} < \delta^{SGPE}$, given that $\delta_i^{SGPE}$ is decreasing in $\beta_i$. For high values of $\beta_i$, above the threshold $\beta_i^*$, $\delta_i^{SGPE} < 0$, which implies the condition $\delta_i > \delta_i^{SGPE}$ is always met and player $i$ would always prefer to play a grim strategy.

The threshold value $\delta^{SGPE}$ is obtained assuming the player $i$ is choosing what strategy would be best to play, assuming the opponent is playing the cooperative grim strategy. Accounting for the strategic uncertainty arising from not knowing what strategy the opponent will be playing, we can still observe the outcome (Cooperate, Cooperate) being sustainable as an equilibrium among self-interested players if:

$$\delta_i > \delta^{RD} = \frac{T - S - R + P}{T - S}$$

where $\delta^{RD}$ is the value that makes playing grim a risk dominant strategy to the other player randomizing 50-50 between the two grim and always defect strategies.

$$\delta_i^{RD} : \frac{1}{2}\left[\sum_{t=0}^{\infty} \delta^t R\right] + \frac{1}{2}\left[S + \sum_{t=1}^{\infty} \delta^t P\right] > \frac{1}{2}\left[T + \sum_{t=1}^{\infty} \delta^t P\right] + \frac{1}{2}\left[\sum_{t=0}^{\infty} \delta^t P\right]$$

Under the assumption of both players having strong enough social preferences and perfect information, mutual cooperation can be feasible as a Risk Dominant (RD) equilibrium of the infinitely repeated game under a wider set of parameters. In particular, (Cooperate, Cooperate) will be sustainable as an equilibrium outcome whenever:

$$\delta_i > \delta_i^{RD} = \frac{T - S - R + P - (\beta_i + \alpha_i)(T - S)}{T - S - (\beta_i + \alpha_i)(T - S)}$$

When social preferences are absent, so that $\beta_i = \alpha_i = 0$, $\delta_i^{RD} = \delta^{RD}$. For positive values of $\beta_i, \alpha_i$ such that $\beta_i + \alpha_i < \beta_i^* + \alpha_i^*$, $\delta_i^{RD} < \delta^{RD}$ since $\delta_i^{RD}$ is decreasing in $\beta_i + \alpha_i$. For positive values of $\beta_i, \alpha_i$ such that $\beta_i + \alpha_i > \beta_i^* + \alpha_i^*$, $\delta_i^{RD} < 0$, which implies the condition $\delta_i > \delta_i^{RD}$ always holds and player $i$ would always find profitable to play a grim strategy

These conclusions are not specific to the choice of modeling social preferences using a model à la Charness and Rabin (2002). Indeed, we could also model social preferences relying on the original model by Fehr and Schmidt (1999) and we would obtain qualitatively equivalent results [8].

In this framework, we would identify the *other-regarding types* as the individuals showing strong enough altruistic preferences (with positive and large $\alpha_i$ and $\beta_i$).

In general, we expect the behavior of the other-regarding types and of the self-interested types to differ when cooperation is not an equilibrium, especially if self-interested types have low expectations on the fraction of cooperators in the society, and to be more comparable

[8] For a discussion, see the Appendix A.1.

when cooperation is sustainable in equilibrium. In particular, we expect:

- Other-regarding types to cooperate, on average, more than the self-interested types when cooperation is not sustainable in equilibrium

- Sessions with a higher density of other-regarding types to sustain higher levels of cooperation over time when cooperation is not an equilibrium, through a beliefs updating mechanism; instead, we expect no striking difference across sessions with a different concentration of other-regarding types when cooperation is sustainable as an equilibrium.

- Other-regarding types to be less sensitive to changes in the strategic incentives to cooperation.

- Individuals' social preference type to be a good predictor of cooperation at the individual level in treatments where cooperation is not an equilibrium, while not necessarily when cooperation is an equilibrium.

Except for a few works, the evidence on motives behind cooperation in infinitely repeated PDs across strategically different scenarios is scarce. Reuben and Suetens (2012) study an indefinitely repeated PD where cooperation is not sustainable in equilibrium employing the strategy-method to disentangle strategically and non-strategically motivated behavior and to identify to what extent strategically motivated individuals are responsible for observed cooperative patterns. By adopting a sequential design where both players can submit their actions conditional on whether or not the round they are playing is the last one, Reuben and Suetens (2012) are able to study the end-game effect and to distinguish strategically from non-strategically motivated second movers: they find that cooperation is greater when the round played is not the last one, which suggests a prevalent end-game effect and a large scope for strategically motivated cooperation, although a role for non-strategically motivated cooperation driven by individual preferences also emerges. They further document that individuals' motivation to cooperate over time is stable, which suggests individuals choose to cooperate either for strategic or non-strategic considerations and behave consistently over time.

Our objective with this paper is to bridge the two strands of the literature that separately studied the role of strategic incentives to cooperate and of individual characteristics on cooperativeness in infinitely repeated PDs, looking at how these two dimensions interact in shaping individuals' cooperativeness.

First, we rely on a meta-analysis to test the ability of structural game parameters, which determine strategic incentives to cooperate, to predict cooperativeness across a wide range of treatments. In particular, we are interested in analyzing and comparing the predictive power of structural game parameters in contexts in which cooperation can and cannot be supported in equilibrium under standard assumptions, and in testing the predictive power of the two composite indicators developed by the literature ($\delta - \delta_{RD}$ and $sizeBAD$).

Second, we propose a novel experimental design, which would allow us to collect new data on both individuals' preferences and individuals' behavior across strategically-different scenarios - which are currently not available from previous studies - in order to directly test our hypotheses on the role of individual preferences and strategic incentives.

## 3.3 Meta-Analysis

Our meta-analysis relies on the application of some simple off-the-shelf (Athey (2017)) supervised learning algorithms on a wide set of experimental data collected from previous experimental works. We rely on machine learning (ML) techniques because they allow for a high degree of flexibility in the model structure, which is derived through a purely a-theoretical and data-driven procedure aimed to maximize the out-of-sample prediction performance of the model.

In our case, we are interested in testing the ability of structural game parameters to predict cooperation and we aim to obtain results that could be generalized also to PDs that are not included in our sample: in this respect, a ML approach that exploits the regularities hidden in the data to estimate models that are designed as to maximize the prediction performance on out-of-sample observations is particularly useful (Mullainathan and Spiess (2017)). In addition, ML routines exploit all the information available in the data to identify what are the most relevant predictors of the behavior of interest even when there's information redundancy and the range of candidate predictors is wide: this would allow us to 'let the data speak' in terms of the relative predictive performace of the different structural game parameters considered by the theory [9].

Although ML applications on experimental data have only recently started to grow (Fudenberg and Peysakhovich (2016)),Naecker (2015),Nay and Vorobeychik (2016), Naecker and Peysakhovich (2017), Fudenberg and Liang (2019)) it emerged that ML algorithms can serve as useful instruments for experimental economists, allowing them to uncover unexplored regularities in the data that can help in better explaining the mechanisms behind subjects' behavior, complementing the information provided by theoretical models.

Through this analysis we are interested in understanding: (i) how well the main structural parameters of the game can predict first-round cooperation choices; (ii) whether the composite indicators derived from the theory - $(\delta - \delta_{RD})$ and $sizeBAD$ - would be selected among the most relevant predictors of cooperation by ML completely a-theoretical routines solely based on parameters' ability to describe the patterns observed in the data; (iii) whether one of the two composite indicators outperforms the other in terms of prediction accuracy or 'model parsimony', which we interpret in terms of proximity between the best model structure selected by the algorithm and the model structure suggested by the theory, typically based on the indicator only; and (iv) whether the prediction power of structural game characteristics over actual choices differs, depending on whether the decision environment is or is not conducive to cooperation.

Our analysis is conducted over an extended version of the dataset collected by Dal Bó and Fréchette (2018), which brings together data from 15 different randomly-terminated 'standard' PD experiments with perfect monitoring and fixed matching across supergames [10].

[9] ML techniques rely on highly flexible functional forms, where greater complexity improves the in-sample fit but increases the out-of-sample error of the selected model: the level of complexity of the model is then set through a regularization parameter that is chosen by cross validation in order to minimize the out-of-sample error (Hastie et al. (2009) ).

[10] Dal Bó and Fréchette (2018) main inclusion criteria are: (1) the stage game is a fixed 2x2 PD game; (2) there is perfect monitoring; (3) there's one-shot interaction or repeated interaction through a random continuation rule (and this does not change inside a supergame); (4) pairs are fixed inside a supergame. They further condition the inclusion on the data availability and on a publication date no before 2014, if the paper is not their own.

We follow the same inclusion criteria extending the publication date up until 2019 and conditioning on the availability of at least 7 supergames per treatment. We restrict our attention to papers where the authors report detailed information on subjects' payments. We use Internet searches to find articles satisfying these conditions. The papers included in our meta-analysis are: Andreoni and Miller. (1993); Cooper et al. (1996), Dal Bó (2005), Dreber et al. (2008), Aoyagi and Freéchette. (2009), Duffy and Ochs (2009), Dal Bó et al. (2010), Dal Bó and Fréchette (2011), Blonski et al. (2011), Fudenberg et al. (2012), Bruttel and Kamecke (2012), Table 3.3 displays the variety of treatments encompassed in the dataset, including those that are already included in the metadata by Dal Bó and Fréchette (2018).

We have data from 18 papers, involving 33 different treatments - identified as non-overlapping combinations of structural parameters $\delta, l, g$ - with a total of 3267 subjects and observations on almost 270.000 choices.

We focus on first-round (henceforth 1R) cooperation choices, although it may be argued that they only provide an imcomplete picture of individuals' cooperative attitude, because it simplifies our analysis over a series of dimensions: first, different treatments and different supergames within the same treatment may have a different number of rounds, which would complicate the analysis; second, 1R choices can be though as solely reflecting subjects' own individual strategies, net of the impact of other players' strategies that are likely to impact subjects' subsequent choices in the supergame; third the binary choice of cooperating/defecting in the first round can be univocally mapped into subjects' willingness to engage in a cooperative (Always Cooperate, Grim, etc.) or non-cooperative (Always Defect) strategy.

We analyze the predictive power of structural game parameters on subjects' choices in the first round across different supergames: our main focus will be on supergames 1 to 7 as the 7th supergame is the highest supergame we can study without losing any treatment. As regressors, we include a series of treatment-specific characteristics of the game ($S_t$) that include:

- $\delta$ the continuation probability
- $g$ the 'normalized' gain from unilateral defection
- $l$ the 'normalized' loss from unilateral cooperation
- $I_{RD}$ a dummy identifying treatments where cooperation is a Risk Dominant Eq. (RD)
- $I_{SGPE}$ a dummy identifying treatments where cooperation is a Subgame Perfect Nash Eq. (SGPE)
- *matching* an ordered discrete variable that describes the between supergames matching procedures [11]

KaterinaSherstyuk et al. (2013), Fréchette and Yuksel (2017), Dal Bó and Fréchette (2019), Peysakhovich and Rand (2016), Romero and Rosokha (2018), Ghidoni et al. (2019) and Proto et al. (2019)
In the Appendix A.2 we replicate one of the main analysis proposed by Dal Bó and Fréchette (2018) to study the effect of $\delta - \delta_{RD}$ and $sizeBAD$ on cooperation on our dataset, in order to show that it is not systematically different from the dataset originally collected by Dal Bó and Fréchette (2018)

[11] *matching* takes value '1' if subjects were matched according to complete stranger protocol, aka turnpike

- $m_{(R-P)}$ the mark-up from mutual cooperation with respect to mutual defection that is equal to the distance between R and P payoffs from the original matrix [12]
- $P_{DDpayoff}$ the payoff from mutual defection from the original matrix [13]
- *totsup* the total number of supergames
- *showup* the amount of the show-up fee [14]

Since we are also interested in testing the predictive accuracy of the two indices $(\delta - \delta_{RD})$ and *sizeBAD* proposed by the literature, we separately estimate the predictive model for each supergame three times, as summarized in Table 3.4: first, we estimate the model including all treatment characteristics and the $(\delta - \delta_{RD})$ indicator only ($S_t^1 = S_t + (\delta - \delta_{RD})$); second, we estimate the model including all treatment characteristics and the *sizeBAD* indicator only ($S_t^2 = S_t + sizeBAD$); third, we estimate the model including all treatment characteristics and both the $(\delta - \delta_{RD})$ and *sizeBAD* indicators ($S_t^3 = S_t + (\delta - \delta_{RD}) + sizeBAD$).

We employ three different supervised learning algorithms to deal with our classification problem, where $Y = 1[1Rchoice = cooperation]$: the Decision Tree, the Random Forest and the Logistic LASSO. Before fitting our models we randomly split our sample in two subsamples, a training set and a testing set, following the 2/3 and 1/3 division rule (as suggested by Y. Zhao and Cen (2014)). We then fit our models on the training set and test the predictive accuracy of the models over the testing set. We chose to report only the results obtained by using the Logistic LASSO algorithm as it shows a good predictive performance with respect to the other algorithms in most of the cases and, at the same time, it performs well on the ground of interpretability, providing easy-to-interpret outputs. The logistic LASSO algorithm provides a prediction that is based on a logit model (with a linear index) where

or zipper; takes value '2' if subjects were matched according to perfect stranger protocol, aka round robin; takes value '3' if subjects were matched according to perfect stranger protocol until the pool is exhausted; takes value '4' if subjects were matched according to a random matching protocol.

[12] The amount is measured in UD dollars: if the experiments were originally paid in dollars, the amounts are obtained by multiplying experimental currency units (ECU) by the exchange rate to dollars; if the experiments were originally paid in other currencies, the amounts are obtained by (1) multiplying experimental currency units (ECU) by the exchange rate to the relevant currency (2) converting the amounts obtained to american dollars. The same procedure was adopted by Dal Bó and Fréchette (2018).
Blonski et al. (2011) was conducted between May and November of 2006; accordingly the exchange rate is set at 0.785 Euro = 1 dollar. Bruttel and Kamecke (2012) was conducted between January and February of 2007; accordingly the exchange rate is set at 1 Euro = 1.307 dollars. Ghidoni et al. (2019) was conducted between June 2016 and November 2017; accordingly the exchange rate is set at 1 Euro = 1.1156 dollars. Proto et al. (2019) was conducted between June 2013 and June 2016; accordingly the exchange rate is set at 1 GBP = 1.5718 dollars.

[13] see footnote 12.

[14] see footnote 12

the estimated coefficients are penalized according to their magnitude:

$$max_{\beta_0,\beta} \sum_{i=1}^{N} [y_{i,t}(\beta_0 + \beta' \tilde{X_{i,t}}) - log(1 + e^{\beta_0 + \beta' \tilde{X_{i,t}}})] - \sum_{j=1}^{\tilde{P}} |\beta_j|$$

where $\tilde{X_{i,t}}$ (of dimension $\tilde{P}$) is the vector of all predictors including also the pairwise interactions between the variables in $S_t$, $\lambda$ is a penalization parameter [15], $\beta_0$ is a constant and $\beta'$ is a transpose vector of the $\beta$ coefficients to be estimated (together with the constant) [16]. The LASSO penalization implies that only a subset of indicators will have estimated coefficients other than zero and the non-null coefficients of the sparse model will be the only ones relevant for prediction.

If we look at the list of non-null coefficients selected by the LASSO algorithm for the three models throughout supergames (see Table 3.5), we observe that the composite indicators $(\delta - \delta_{RD})$ and $sizeBAD$ are always selected among the most relevant cooperation predictors, either alone or in combination with other structural game parameters.

[15] The optimal $\lambda$ is selected by looking at the 10-fold cross-validated misclassification error, using the one-standard-error rule. The cross-validation process aims at identifying the value of $\lambda$ that minimizes the misclassification error, but acknowledging that the estimation of misclassification rates also comes with an error, we adopt the "one-standard error" rule: according to this rule, we choose the most parsimonious model whose error is no more than one standard error above the error of the best model, which represents a more conservative approach (Hastie et al. (2009)).

[16] The index $i = 1, .., N$ identifies each single individual in the sample, who is exposed to one treatment $t = 1, .., T$ only. We observe cooperation realizations at the individual level ($y_{i,t}$, such that $y_{i,t} \neq y_{l,t}$), but we feed our algorithm with treatment-specific variables only ($\tilde{X_t} = \tilde{X_{i,t}} = \tilde{X_{l,t}}$). The set of regressors in $\tilde{X_t}$ changes depending of which of the three models, see Table 3.4, we are estimating $\tilde{X_t^k}$, where $k = \{1, 2, 3\}$.

Table 3.3: General information on the meta-data

| | Session | Subjects | $\delta$ | $g$ | $l$ | Supergames |
|---|---|---|---|---|---|---|
| **Andreoni and Miller (1993)** | **1** | **14** | 0 | 1.67 | 1.33 | 200 |
| **Cooper et al. (1996)** | **3** | **33** | 0 | 0.44 | 0.78 | 10 |
| **Dal Bó (2005)** | **6** | **276** | | | | |
| | 2 | 72 | 0 | 1.17 | 0.83 | 7 |
| | 2 | 102 | 0 | 0.83 | 1.17 | 9 |
| | 1 | 42 | 0.75 | 1.17 | 0.83 | 7 |
| | 1 | 60 | 0.75 | 0.83 | 1.17 | 10 |
| **Dreber et al. (2008)** | **2** | **50** | | | | |
| | 1 | 28 | 0.75 | 2 | 2 | 21 |
| | 1 | 22 | 0.75 | 1 | 1 | 27 |
| **Aoyagi and Fréchette (2009)** | **4** | **74** | | | | |
| | 2 | 36 | 0 | 0.33 | 0.11 | 75 |
| | 2 | 38 | 0.9 | 0.33 | 0.11 | 10 |
| **Duffy and Ochs (2009)** | **9** | **102** | 0.9 | 1 | 1 | 13 |
| **Dal Bó, Foster and Putterman (2010)** | **28** | **424** | 0 | 1 | 3 | 10 |
| **Dal Bó and Fréchette (2011)** | **18** | **266** | | | | |
| | 3 | 44 | 0.5 | 2.57 | 1.86 | 71 |
| | 3 | 50 | 0.5 | 0.67 | 0.87 | 72 |
| | 3 | 46 | 0.5 | 0.09 | 0.57 | 77 |
| | 3 | 44 | 0.75 | 2.57 | 1.86 | 33 |
| | 3 | 38 | 0.75 | 0.67 | 0.86 | 47 |
| | 3 | 44 | 0.75 | 0.09 | 0.57 | 35 |
| **Blonski, Ockenfels and Spagnolo (2011)** | **10** | **200** | | | | |
| | 1 | 20 | 0.5 | 2 | 2 | 11 |
| | 2 | 40 | 0.75 | 2 | 2 | 11 |
| | 1 | 20 | 0.75 | 1 | 8 | 11 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Bruttel and Kamecke (2012)** | **3** | **36** | 0.8 | 1.17 | 0.83 | 2 |
| **Sherstyuk, Tarui, and Saijo (2013)** | **4** | **56** | 0.75 | 1 | 0.25 | 29 |
| **Frechette and Yuksel (2017)** | **3** | **60** | 0.75 | 0.4 | 0.4 | 12 |
| **Dal Bó and Fréchette (2019)** | **41** | **672** | | | | |
| | **3** | **50** | 0.5 | 2.57 | 1.86 | 37 |
| | **8** | **140** | 0.5 | 0.09 | 0.57 | 46 |
| | **8** | **114** | 0.75 | 2.57 | 1.86 | 25 |
| | **10** | **164** | 0.75 | 0.09 | 0.57 | 24 |
| | **10** | **168** | 0.9 | 2.57 | 1.86 | 21 |
| | **2** | **36** | 0.95 | 2.57 | 1.86 | 7 |
| **Peysakhovich and Rand (2016)** | **6** | **96** | | | | |
| | **3** | **52** | 0.125 | 0.66 | 0.33 | 45 |
| | **3** | **44** | 0.875 | 0.33 | 0.33 | 10 |
| **Romero and Rosokha (2018)** | **6** | **82** | | | | |
| | **3** | **44** | 0.95 | 2.57 | 1.86 | 10 |
| | **3** | **38** | 0.95 | 2.57 | 1.86 | 20 |
| **Ghidoni, Cleave and Suetens (2019)** | **4** | **80** | 0 | 0.73 | 0.46 | 10 |
| **Proto, Rustichini and Sofianos (2019)** | **40** | **586** | | | | |
| | **32** | **476** | 0.75 | 0.09 | 0.57 | 12 |
| | **8** | **110** | 0.5 | 0.09 | 0.57 | 13 |
| | **201** | **3267** | Choices: 269.832 | | | |

*Notes.* The column ‘Supergame’ reports the minimum number of supergames observed across similar sessions.

Table 3.4: Logistic LASSO: set of predictors

| **(1)** $\delta - \delta_{RD}$ **model** | **(2)** $sizeBAD$ **model** | **(3) Unconstrained model** |
|---|---|---|
| $S_t^1$: main features of the game<br>$\delta$, $g$, $l$, $I_{RD}$, $I_{SGPE}$,<br>$matching$, $m_{(R-P)}$,<br>$P_{DDpayoff}$, $showup$,<br>$totsup$, $\delta - \delta_{RD}$<br>$I_t^1$: all pairwise interactions<br>between variables in $S_t^1$ | $S_t^2$: main features of the game<br>$\delta$, $g$, $l$, $I_{RD}$, $I_{SGPE}$,<br>$matching$, $m_{(R-P)}$,<br>$P_{DDpayoff}$, $showup$,<br>$totsup$, $sizeBAD$<br>$I_t^2$: all pairwise interactions<br>between variables in $S_t^2$ | $S_t^3$: main features of the game<br>$\delta$, $g$, $l$, $I_{RD}$, $I_{SGPE}$,<br>$matching$, $m_{(R-P)}$,<br>$P_{DDpayoff}$, $showup$,<br>$totsup$, $\delta - \delta_{RD}$ , $sizeBAD$<br>$I_t^3$: all pairwise interactions<br>between variables in $S_t^3$ |
| $\tilde{X}_t^1$ contains 66 predictors | $\tilde{X}_t^2$ contains 66 predictors | $\tilde{X}_t^3$ contains 78 predictors |

Table 3.5: Logistic LASSO: non-null estimated coefficients - Supergame 1 to 7

| **(1)** $\delta - \delta_{RD}$ **model** | **(2)** $sizeBAD$ **model** | **(3) Unconstrained model** |
|---|---|---|
| Supergame 1 | | |
| (+) $I_{RD}\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$<br>$m\cdot totsup$, $I_{SGPE}\cdot totsup$<br>(-) $l, g\cdot l$ | (+) $I_{RD}\cdot totsup$<br>(-) $sizeBAD$, $l, g\cdot$ $sizeBAD$<br>$g, g\cdot l$ | (+) $I_{RD}\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$<br>$I_{SGPE}\cdot totsup$<br>(-) $l, g\cdot$$sizeBAD$,$g\cdot l$ |
| Supergame 2 | | |
| (+) $I_{RD}\cdot$ $(\delta-\delta_{RD})$,<br>$I_{SGPE}\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$,<br>$totsup\cdot I_{RD}$<br>(-) $g\cdot P, l$ | (+) $I_{RD}\cdot totsup, \delta\cdot showup$<br>(-) $sizeBAD$, $g\cdot P$,<br>$sizeBAD\cdot g, l$, | (+) $I_{RD}\cdot$ $(\delta-\delta_{RD})$,<br>$I_{SGPE}\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$,<br>$totsup\cdot I_{RD}$<br>(-) $g\cdot P, g\cdot$ $sizeBAD$, $l$ |
| Supergame 3 | | |
| (+) $I_{SGPE}\cdot$ $(\delta-\delta_{RD})$,<br>$g\cdot$ $(\delta-\delta_{RD})$, $m\cdot I_{RD}$<br>$(\delta-\delta_{RD})$, ...<br>(-) $\delta\cdot p, l, P\cdot showup, ...$ | (+) $matching, P\cdot totsup$, ...<br>(-) $sizeBAD$, $I_{SGPE}\cdot P$<br>$sizeBAD\cdot I_{RD}, l$, ... | (+) $I_{SGPE}\cdot$ $(\delta-\delta_{RD})$,<br>$(\delta-\delta_{RD})$, $P\cdot m$,<br>$m\cdot$ $(\delta-\delta_{RD})$, $m\cdot I_{RD}$, ...<br>(-) $\delta\cdot P, l\cdot$$sizeBAD$,<br>$P\cdot showup, m\cdot showup, ...$ |
| Supergame 4 | | |
| (+) $\delta\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$,<br>$I_{SGPE}\cdot$ $(\delta-\delta_{RD})$,<br>$I_{RD}$ $(\delta-\delta_{RD})$<br>(-) $l$ | (+)<br>(-) $sizeBAD$, 1 | (+) $\delta\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$,<br>$I_{SGPE}\cdot$ $(\delta-\delta_{RD})$<br>(-) $l$ |
| Supergame 5 | | |
| (+) $I_{SGPE}\cdot$ $(\delta-\delta_{RD})$,<br>$\delta\cdot(\delta-\delta_{RD})$, $g\cdot$ $(\delta-\delta_{RD})$,<br>$m\cdot$ $(\delta-\delta_{RD})$, $(\delta-\delta_{RD})$, ...<br>(-) $\delta\cdot P, g\cdot P, l$<br>$P\cdot showup, ...$ | (+)<br>(-) $sizeBAD$, 1, $sizeBAD$ $\cdot l$ | (+) $P\cdot m, I_{SGPE}\cdot$ $(\delta-\delta_{RD})$,<br>$\delta\cdot(\delta-\delta_{RD})$, $l\cdot m$,<br>$m\cdot(\delta-\delta_{RD})$<br>(-) $P\cdot I_{RD}, g\cdot P, m, g\cdot m$,<br>$\delta\cdot l$, ... |
| Supergame 6 | | |
| (+) $m\cdot$ $(\delta-\delta_{RD})$, $\delta\cdot$ $(\delta-\delta_{RD})$, | (+) | (+) $m\cdot$ $(\delta-\delta_{RD})$, $\delta\cdot$ $(\delta-\delta_{RD})$, |

The logistic LASSO outputs a predicted probability to cooperate for each individual in the sample. However, since we feed the algorithm with treatment-specific variables only, without including any individual-specific information, we obtain the same estimated probability to cooperate for all individuals exposed to the same treatment.

Standard metrics used to evaluate predition accuracy of ML classification algorithms typically convert predicted probabilities into predicted categories by setting a predicted probability cutoff $c$ such that those observations whose predicted probability is above $c$ are classified as belonging to the $Y = 1$ category, while the others are classified as belonging to the $Y = 0$ category. In this context, given that the predicted probability is the same for all individuals exposed to the same treatment, we would have that all individuals exposed to the same treatment would either be classified as 1R cooperators or defectors with no within-treatment variability: misclassified individuals will then be cooperators exposed to treatments where the predicted probability to cooperate is below the cutoff or defectors exposed to treatments where the predicted probability to cooperate is above the cutoff.

Table 3.6: Classification accuracy: 1st Round Cooperation – Supergame 1

| | Area under ROC Curve | Misclassification (c=0.5) | FP (c=0.5) | FN (c=0.5) | Precision (c=0.5) | Recall (c=0.5) |
|---|---|---|---|---|---|---|
| **(1)** $(\delta - \delta_{RD})$ **model** | 0.64 | 39% | 20% | 19% | 0.60 | 0.62 |
| **(2)** *sizeBAD* **model** | 0.64 | 39% | 20% | 19% | 0.60 | 0.62 |
| **(3) Unconstrained model** | 0.64 | 39% | 20% | 19% | 0.60 | 0.62 |

*Notes.* FP = False Positives, FN = False Negatives.

Table 3.7: Classification accuracy: 1st Round Cooperation – Supergame 7

| | Area under | Misclassification | FP | FN | Precision | Recall |
|---|---|---|---|---|---|---|
| | ROC Curve | (c=0.5) | (c=0.5) | (c=0.5) | (c=0.5) | (c=0.5) |
| **(1)** $(\delta - \delta_{RD})$ **model** | 0.77 | 27% | 12% | 15% | 0.69 | 0.61 |
| **(2)** *sizeBAD* **model** | 0.78 | 27% | 11% | 16% | 0.69 | 0.62 |
| **(3) Unconstrained model** | 0.77 | 27% | 12% | 15% | 0.68 | 0.62 |

*Notes.* FP = False Positives, FN = False Negatives.

Table 3.8: Classification accuracy: SGPE vs. non-SGPE treatments

| | **SGPE** | | | | | **Non-SGPE** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ROC** | **Misscl.** | **Corr.** | **FP** | **FN** | **ROC** | **Misscl.** | **Corr.** | **FP** | **FN** |
| **Supergame 1** | | | | | | | | | | |
| **(1)** $\delta - \delta_{RD}$ **model** | 0.63 | 39% | 0.56 | 29% | 10% | 0.55 | 40% | 0.29 | 2% | 38% |
| **(2)** *sizeBAD* **model** | 0.64 | 39% | 0.54 | 29% | 10% | 0.56 | 40% | 0.14 | 2% | 38% |
| **(3) Unconstr. model** | 0.63 | 39% | 0.56 | 29% | 10% | 0.55 | 40% | 0.23 | 2% | 38% |
| **Supergame 7** | | | | | | | | | | |
| **(1)** $\delta - \delta_{RD}$ **model** | 0.69 | 33% | 0.73 | 17% | 16% | 0.58 | 13% | 0.05 | 0% | 13% |
| **(2)** *sizeBAD* **model** | 0.72 | 33% | 0.71 | 16% | 17% | 0.58 | 13% | . | 0% | 13% |
| **(3) Unconstr. model** | 0.70 | 33% | 0.74 | 16% | 17% | 0.60 | 13% | 0.21 | 0% | 13% |

*Notes.* ROC = Area under the ROC curve, Misscl. = Misclassification, Corr. = Correlation between predicted and observed cooperation rate, FP = False Positives, FN = False Negatives.

Table 3.6 reports some classification accuracy metrics for 1R cooperative choices in Supergame 1: while the Area Under the ROC (Receiver Operating Characteristics) Curve provides a measure of classification accuracy considering the entire set of all possible thresholds $c \in [0, 1]$, the other metrics assume the predicted probability cutoff for classification is set at $c = 0.5$. The area under the ROC curve - which varies within the interval [0,1], where 0 is the performance of a random classifier and 1 of a perfect classifier - measures the ability of the classifier to distinguish between cooperators and defectors [17]. The misclassification rate quantifies the intensity of misclassification irrespective of the misclassification error type (false positives or false negatives), while the 'precision' and the 'recall' indicators are more focused on the ability of the classifier to identify cooperators, our target of interest: the 'precision' indicator measures how accurate positive predictions ($\hat{Y} = 1$) are, that is the probability to correctly detect positive values, while the 'recall' indicator measures the coverage of the actual positive sample ($Y = 1$) achieved by the classifier [18].

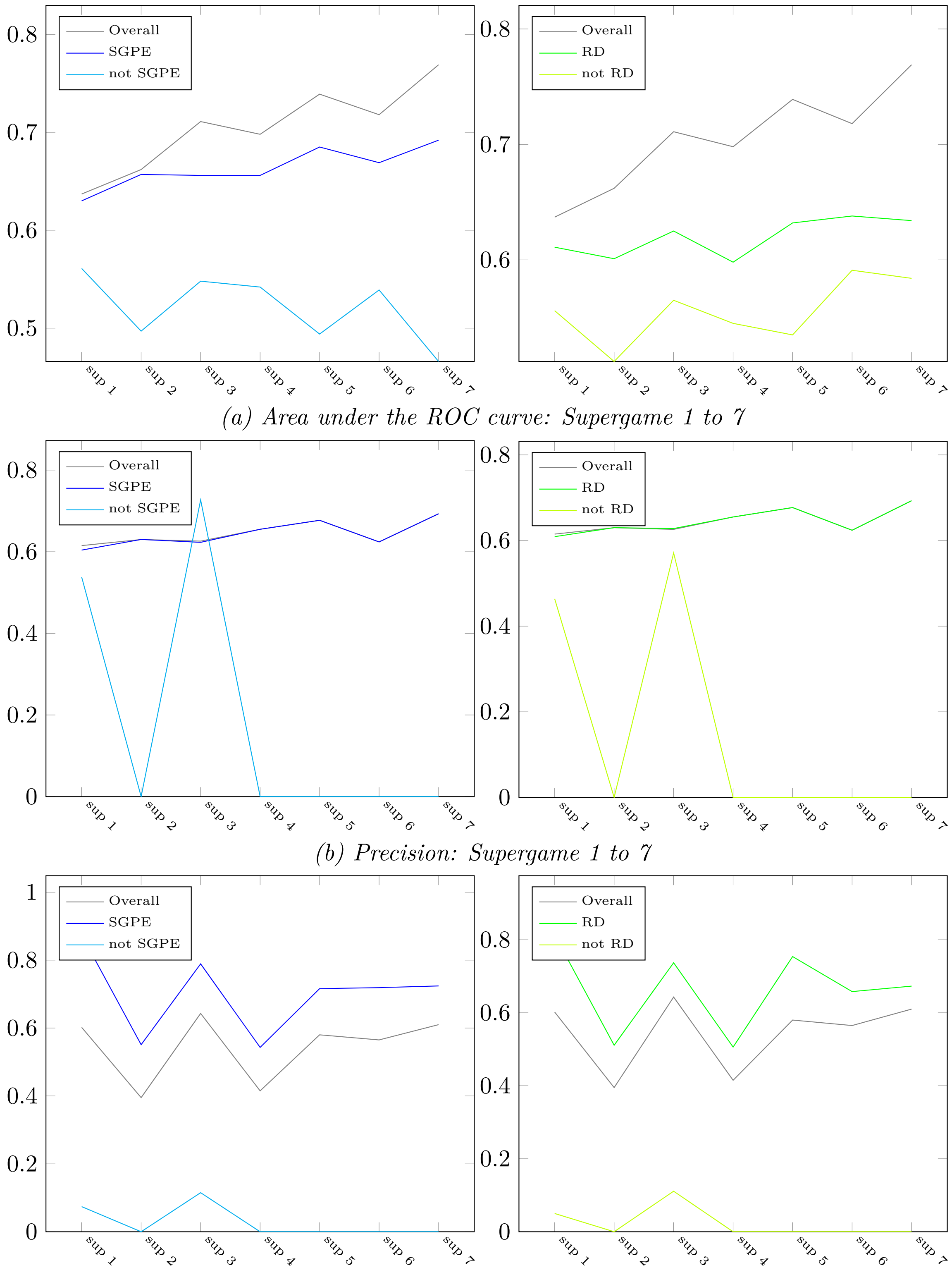
Looking at Tables 3.6 and 3.7 and Figures 3.3[19], we can see that all the three models reach almost the same level of prediction accuracy across different supergames and that the prediction accuracy of the model increases as subjects move to later supergames and gain some experience. It further emerges from Table 3.8 and Figures 3.3 an asymmetry in terms of prediction accuracy along the SGPE or RD equilibrium dimensions: ML models seem to produce classifiers that are more accurate - especially in terms of models' ability to detect cooperators - in treatments where structural game parameters are such that cooperation is sustainable in equilibrium.

[17] The ROC curve summarizes the trade-off between the True Positive rate (TPR) and the False Positive rate (FPR), where $TPR = TP/(TP + FN)$, $FPR = FP/(FP + TN)$, TP=positives correclty predicted as positives, FN=positives incorrectly predicted as negatives, FP=negatives incorrectly predicted as positives and TN=negatives correctly predicted as negatives.

[18] Precision = $\frac{TP}{(TP+FP)}$; Recall = $TPR = \frac{TP}{(TP+FN)}$

[19] Figures 3.3 are based on prediction data obtained from model (1) where $\delta - \delta_{RD}$ is included among regressors. The other prediction models yield qualitatively similar results, which are reported in the Appendix A.3

Figure 3.3: Classification Accuracy metrics over Supergames 1 to 7 ($\delta - \delta_{RD}$) model



*(a) Area under the ROC curve: Supergame 1 to 7*

*(b) Precision: Supergame 1 to 7*

*(c) Recall: Supergame 1 to 7*

Figure 3.3 (panel a) shows that the overall ability of the ML model to discriminate between cooperators and defectors increases over supergames, with a steadily higher disciminatory ability over observations from treatments where cooperation is a SGPE or RD equilibrium. Figure 3.3 (panel b) shows a modest increase in model precision over supergames, which appears to be mostly driven by an increasing ability in accurately detecting cooperators over SGPE or RD equilibrium treatments: the model ability to predict positive occurrences over treatments where cooperation is not an equilibrium, instead, soon decays to zero. A similar dynamics is shown by Figure 3.3 (panel c), where we observe that the coverage of the actual cooperators' sample soon decays to zero in treatments where cooperation is not an equilibrium, while remains somewhat stable, fluctuatiing around higher values, in treatments where cooperation is an equilibrium.

The same evidence emerges if we collapse our observations at the treatment level and we look at how predicted probabilities estimated at the treatment level, which we interpret as the treatment-specific 'predicted cooperation rate', compare to actual observed rates of cooperation (see Tables 3.9 and 3.10 and Figure 3.4). The correlation increases over supergames and is steadily higher for treatments where cooperation is an equilbrium, compared to those where it is not.

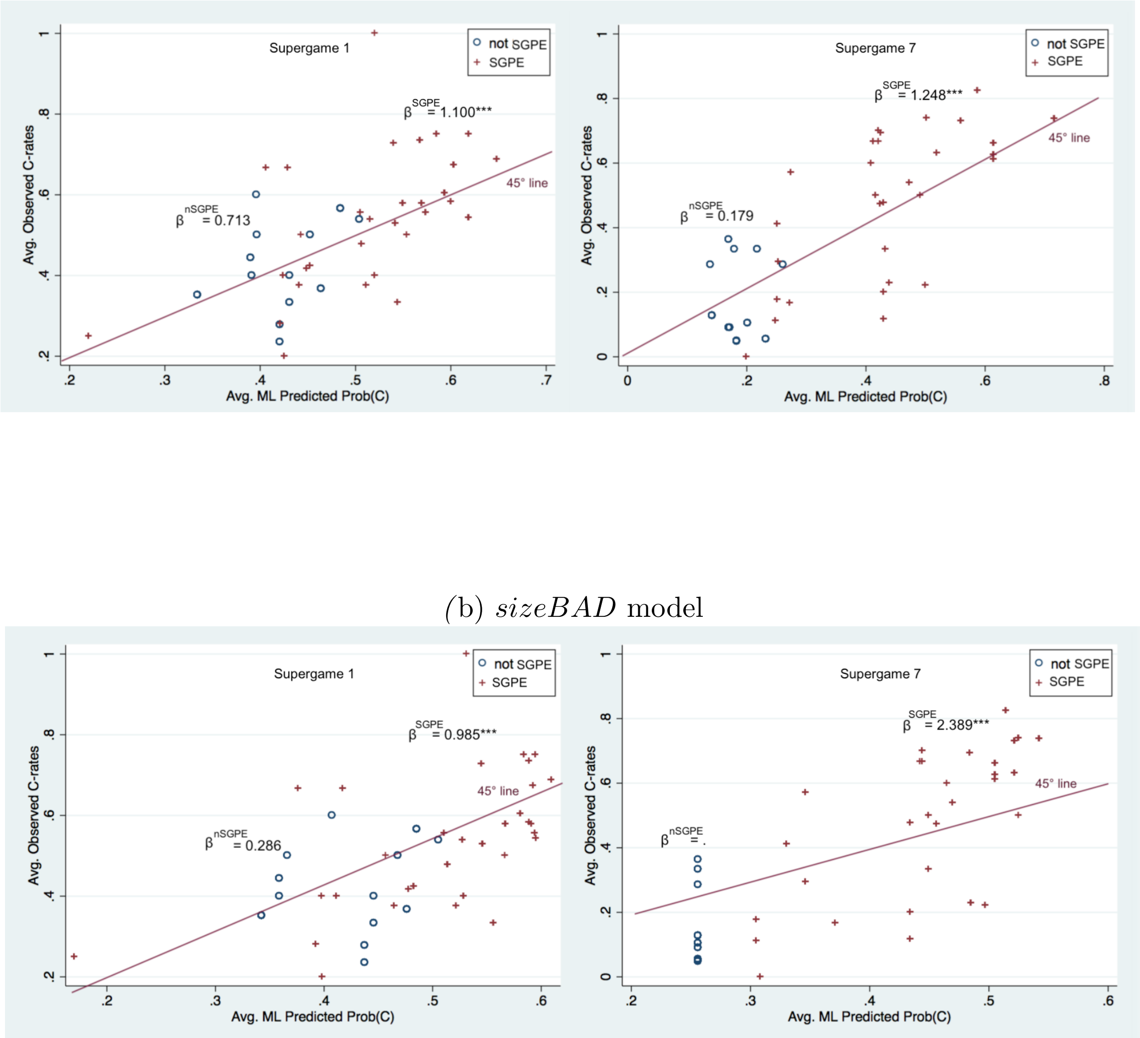
Table 3.9: Prediction accuracy: 1st Round Cooperation – Supergame 1

| | Correlation Predicted vs. Actual Cooperation Rate | | | | |
|---|---|---|---|---|---|
| | Overall | SGPE | Not SGPE | RD | Not RD |
| **(1)** $\delta - \delta_{RD}$ **model** | 59% | 56% | 29% | 52% | 28% |
| **(2)** *sizeBAD* **model** | 55% | 54% | 14% | 50% | 24% |
| **(3) Unconstrained model** | 58% | 56% | 23% | 50% | 28% |

Table 3.10: Prediction accuracy: 1st Round Cooperation – Supergame 7

| | Correlation Predicted vs. Actual Cooperation Rate | | | | |
|---|---|---|---|---|---|
| | Overall | SGPE | Not SGPE | RD | Not RD |
| **(1) $\delta - \delta_{RD}$ model** | 80% | 73% | 5% | 60% | 33% |
| **(2) *sizeBAD* model** | 78% | 71% | . % | 60% | 32% |
| **(3) Unconstrained model** | 81% | 74% | 21% | 61% | 44% |

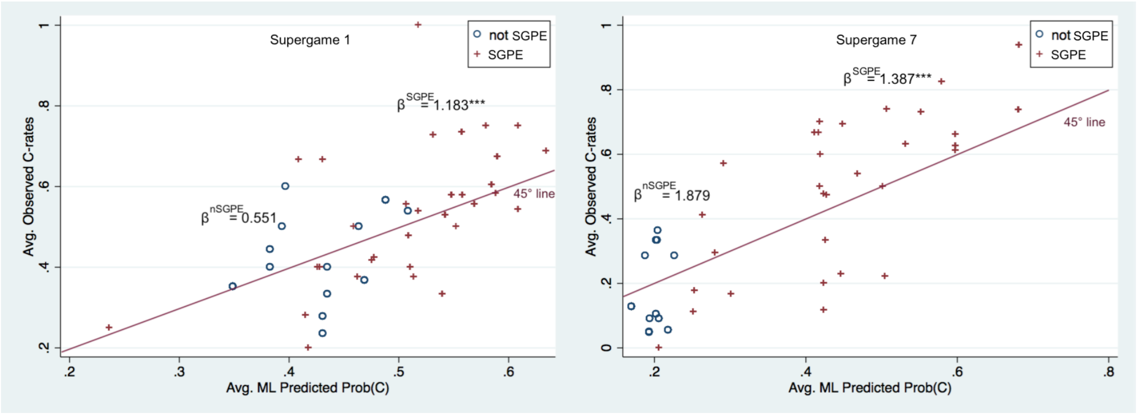
Figure 3.4: Average ML Predicted vs. Observed Cooperation Rates: Supergame 1 and 7



(a) $\delta - \delta_{RD}$ model

(b) *sizeBAD* model

(c) Unconstrained model



*Notes.* The unit of observation is the treatment. $\beta$ coefficients reported are obtained by separately estimating - for SGPE and non SGPE treatments - the following linear regression model $Y[Observed\ Crate] = \beta_0 + \beta_1 \cdot X[Predicted\ Crate]$. We report estimated coefficients and statistical significance of coefficients $\beta_1$.

Judging which of the two indicators $(\delta - \delta_{RD})$ and $sizeBAD$ performs best in the prediction is not straightforward. If we look at model's 'parsimony', that is the proximity between the model structure selected by the algorithm and the structure implied by the theory - typically based solely on the indicator itself - models trained including $sizeBAD$ among the regressors seem to perform better: more parsimonious models achieve the same level of prediction accuracy reached by the other models, which are obtained through more complex model structures, suggesting the $sizeBAD$ indicator alone is able to capture a great share of the variability needed to produce accurate predictions. However, when we estimate the unconstrained model, where the ML algorithm is fed with both the indicators $(\delta - \delta_{RD})$ and $sizeBAD$ and let free to select which one is the most useful for prediction, we observe that the ML algorithm is more likely to select the $(\delta - \delta_{RD})$ among the most relevant predictors (see Table 3.5). Another argument in favor of $(\delta - \delta_{RD})$ comes from the analysis on the correlation between predicted and observed cooperation rates, where models trained including $(\delta - \delta_{RD})$ slightly outperform model trained including $sizeBAD$, especially in treatments where cooperation is not a SGPE or RD equilibrium.

## 3.4 Experimental Design

The evidence brought by the meta-analysis could be consistent with the idea that the main drivers for cooperation in contexts where cooperation is not sustainable as an equilibrium are essentially *non-strategic*. For this reason, models trained using only the information on structural game parameters, which capture solely the intensity of *strategic* incentives for cooperation, perform worse on the ground of the ability to detect cooperators when applied to contexts in which cooperation cannot be supported in equilibrium under standard assumptions.

This narrative would also provide a comprehensive explanation to the (relatively scarce) evidence coming from the experimental literature studying the predictive power of individual characteristics, in particular social preferences, in infinitely repeated PDs, which we discussed in Section 2.

We propose to test the research questions behind our narrative through a novel experimental design, through which we aim to be able to observe both:

- a measure of players' social preferences, as to distinguish the *other-regarding* from the *self-interested* types
- players actual behavior in infinitely repeated PDs

The proposed experimental design is divided into two parts:

PART 1: Measurement of social preferences
& Questionnaire

PART 2: Infinitely Repeated PDs
& Elicitation of subjects' beliefs on the share of cooperators across supergames

At the beginning of the experiment, subjects learn about the two-parts structure of the experiment and that they will be informed of their earnings and paid only at the end of the last part of the experiment. Parts 1 is the same across treatments, while the design of Part 2 changes across treatments. Each subject is exposed to one treatment only.

The experiment has been programed using *Otree* and run entirely online. Subjects were recruited from the local pool of students of the University of Bologna using ORSEE (Greiner

(2004)). The first wave of data collection has been run between May 16 and May 26, 2020 [20] a total of 192 subjects completed both parts of the experiment [21]. Subjects spent on average 20 minutes to take part in Part 1 of the experiment and sessions of Part 2 of the experiment lasted on average one hour and five minutes. Subjects earned on average, 14.60 Euros (payments ranged within 8.8 and 27.8 Euros), including the 4 Euros show-up fee.

Subjects never interacted with the same counterparts in Part 1 and Part 2. The structure of the experimental design is tailored to minimize the risk of spillover effects between Part 1 and Part 2. Indeed, there is some evidence that having subjects playing both the infinitely repeated PDs and other games aimed to measure their social preferences could lead to contamination and spillover issues: Peysakhovich and Rand (2016), for example, by letting their subjects play a series of cooperation games, including a DG after an infinitely repeated PD with varying continuation probabilities, document that subject cooperativeness is significantly impacted by how conducive to cooperation was the environment they faced in the PD, with subjects exposed to the treatment where cooperation was an equilibrium exhibiting higher cooperativeness. This suggests that, even in absence of pure hedging or income effects that may be triggered by the multi-games structure of the experiments, contamination and across-games spillovers might still be an issue, possibly leading to attenuated or exacerbated results.

Parts 1 and 2 took place about four days apart.

Upon registration, subjects have been invited to participate in Part 1 of the experiment. Since the decision environment faced by subjects in Part 1, which will be further described in this section, is essentially non-strategic, subjects were invited to complete Part 1 whenever they wanted over a pre-defined time frame that expired some days prior to the moment when Part 2 of the experiment actually took place. Only subjects who completed all tasks from Part 1 within the due date were invited to participate in Part 2, where subjects actually interacted in real time with their counterparts. Subjects were paid only at the end of Part 2, provided that they completed Part 1 within the due date and logged-in on time for Part 2, to prevent attrition.

[20] The experiment was pre-registered on the Open Science Framework (OSF) Registry: https://osf.io/mzt74

[21] Attrition from Part 1 to Part 2 of the experiment was on average equal to 6.75%. Additional 16 participants took part in the pilot session of the experiment, run in early May 2020: data from this session are discarded from the analysis due to a slight difference in experimental procedures, namely a shorter time window between Part 1 and Part 2.

### 3.4.1 Part 1

In Part 1 we estimate subjects' social preferences relying on the procedure recently proposed by Bruhin et al. (2018). Different alternative procedures have been proposed over the past years to estimate social preferences in the literature, which mainly differ in terms of: (i) the degree flexibility in terms of the number of social preferences categories considered, which are either determined ex-ante or determined endogenously; and (ii) the choice to adopt a fully non-parametric or parametric approach [22]. Bruhin et al. (2018) propose a parametric approach based on a behavioral model inspired by the work by Fehr and Schmidt (1999) and Charness and Rabin (2002), which is extended to account also for positive and negative reciprocity concerns:

$$U_i(\pi_i, \pi_j) = (\beta_i r + \alpha_i s + \gamma q + \eta v)\pi_j + (1 - \beta_i r - \alpha_i s - \gamma q - \eta v)\pi_i$$

where

$r = 1$ if $\pi_i > \pi_j$, and $r = 0$ otherwise;

$s = 1$ if $\pi_i < \pi_j$, and $s = 0$ otherwise;

$q = 1$ if player $j$ behaved kindly toward $i$, and $q = 0$ otherwise (positive reciprocity);

$v = 1$ if player $j$ behaved unkindly toward $i$, and $v = 0$ otherwise (negative reciprocity);

The resulting behavioral model provides a parsimonious characterization of subjects' social preferences through a vector of 4 parameters $\theta = \{\alpha, \beta, \gamma, \eta\}$, which are estimated from experimental choice. This approach does not impose any ex-ante constraints on the number or the characteristics of social preferences' types, which are endogenously determined through a Finite Mixture model.

In our case, we rely on a simplified version of their behavioral model, where only distributional preferences are considered and reciprocity concerns are not accounted for ($\gamma = 0$ and $\delta = 0$

[22] For a discussion, see Bruhin et al. (2018).

so that $\theta = \{\alpha, \beta\}$):

$$U_i(\pi_i, \pi_j) = (\beta_i r + \alpha_i s)\pi_j + (1 - \beta_i - \alpha_i s)\pi_i$$

Indeed, Bruhin et al. (2018) claim that distributional preferences turn out to be considerably more important than reciprocity preferences, and in the context of our analysis, which is focused on $1^{st}$ Round choices in infinitely repeated PDs, reciprocity preferences are likely to play a minor role. We choose to rule out reciprocity concerns for a matter of simplicity but accounting for reciprocity concerns will surely provide a more complete and comprehensive picture, especially in the analysis of cooperative outcomes' survival over rounds and across supergames, and we aim to extend our analysis in this direction in a future work.
In the orginal work by Bruhin et al. (2018), social preferences parameters $\theta = \{\alpha, \beta, \gamma, \eta\}$ are estimated relying on a set of 117 binary decisions data. In each binary decision situation, subjects have to choose between two possible payoff allocations between themselves and an anonymous player $j$. These binary decision situations are represented by: (i) a series of 39 dictator games for identifying the parameters $\alpha$ and $\beta$, and (ii) a series of 78 reciprocity games for identifying $\gamma$ and $\eta$. Since we are interested in estimating only distributional preferences parameters $\theta = \{\alpha, \beta\}$, we rely only on a set of 39 dictator games to obtain the binary decision choice data necessary for the estimation. This procedural difference does not deeply affect neither the results of the estimation of parameters $\alpha$ and $\beta$ and the characterization of preferences types, nor the categorization of subjects into preference types and the quality of out-of-sample predictions [23].

In Part 1, subjects play the same set of 39 Dictator Games designed by Bruhin et al. (2018), which are shown in Table 3.11, where amounts are expressed in terms of Experimental Currency Units (ECUs). In each dictator game, subjects play in the role of the dictator (player $i$), who can either increase or decrease the payoff of player $j$ by choosing one of two possible payoff allocations, $X = (\Pi^i_X; \Pi^j_X)$ or $Y = (\Pi^i_Y; \Pi^j_Y)$. In order to identify subject's distributional preferences, governed by $\alpha$ and $\beta$, the cost of changing the other player's payoff systematically varies across the dictator games. The dictator games are presented in blocks

[23] see the Appendix A.4 for a discussion based on the data from Bruhin et al. (2018)

and appear in random order across subjects.

Subjects are paid only for one binary decision situation in which they played as player $i$, that is randomly selected for payment and paid at the end of the experiment [24]. The distribution of payoffs in this decision situation is paid out both to subjects playing as Person $i$ and to the randomly matched partners, selected to play the role of the receiver (Person $j$). In Part 1, ECUs are converted Euros at a conversion rate such that 100 ECUs = 0.4 Euros.

After playing the Dictator games, subjects are asked to answer a Questionnaire, soliciting personal and demographic data (gender, age, major), numeracy ability, based on the 8-items measure developed by Weller et al. (2013), and personality traits, assessed through the big five personality dimensions measured using the 44-items Big Five Inventory developed by John and Srivastava (1999).

### 3.4.2 Part 2

In Part 2 each subject plays two series of infinitely repeated PDs: the PD matrix will remain fixed but the continuation probability $\delta$ will change across series, where each series is composed by 10 supergames. The matrix of the PD monetary payoffs, in terms of Experimental Currency Units (ECUs), is shown in Figure 3.5.
The set of continuation probabilities faced by subjects changes across treatments. Irrespective of what the treatment they are exposed to, all subjects play two blocks of 10 supergames each. In one of these two blocks, they play a series of 10 one-shot games, where the continuation probability is set equal to zero ($\delta^{'} = 0$); in the other block, they play a series of 10 supergames with a continuation probability ($\delta^{''}$) that varies across treatments.
Therefore, $\delta^{''}$ is the principal treatment variable:

- in T1 $\delta^{''} = 0.35$, so that cooperation is not sustainable among self-interested players

[24] The choice of paying subjects only for one randomly selected choice, out of the many made throughout the experiment ("pay one" approach) can be seen as an alternative to the more traditional choice of paying subjects for all choices made ("pay all" approach). The "pay one" approach, which helps in avoiding wealth effects and issues related to hedging and bankruptcy in experiments involving multiple decisions, is increasingly gaining momentum in the experimental literature, with recent theoretical (Azrieli et al. (2018)) and empirical (Charness et al. (2016)) contributions suggesting the "pay one" approach can prove to be as effective as the "pay all" approach, or eventually better in some cases.

Table 3.11: The Dictator Games from Bruhin et al. (2018), expressed in ECUs

| **DG** | $\Pi_X^i$ | $\Pi_X^j$ | $\Pi_Y^i$ | $\Pi_Y^j$ |
|---|---|---|---|---|
| 1 | 940 | 150 | 800 | 510 |
| 2 | 970 | 490 | 770 | 170 |
| 3 | 1060 | 330 | 680 | 330 |
| 4 | 990 | 480 | 750 | 180 |
| 5 | 930 | 510 | 810 | 150 |
| 6 | 430 | 1030 | 230 | 710 |
| 7 | 370 | 1060 | 290 | 680 |
| 8 | 350 | 1060 | 310 | 680 |
| 9 | 1010 | 190 | 730 | 470 |
| 10 | 420 | 1040 | 240 | 700 |
| 11 | 450 | 1020 | 210 | 720 |
| 12 | 470 | 730 | 190 | 1010 |
| 13 | 870 | 140 | 870 | 520 |
| 14 | 400 | 690 | 260 | 1050 |
| 15 | 350 | 680 | 310 | 1060 |
| 16 | 950 | 510 | 790 | 150 |
| 17 | 910 | 520 | 830 | 140 |
| 18 | 390 | 1050 | 270 | 690 |
| 19 | 330 | 680 | 330 | 1060 |
| 20 | 890 | 140 | 850 | 520 |
| 21 | 410 | 1050 | 250 | 690 |
| 22 | 1050 | 270 | 690 | 390 |
| 23 | 520 | 870 | 140 | 870 |
| 24 | 890 | 520 | 850 | 140 |
| 25 | 510 | 810 | 150 | 930 |
| 26 | 960 | 500 | 780 | 160 |
| 27 | 620 | 790 | 580 | 410 |
| 28 | 670 | 420 | 530 | 780 |
| 29 | 720 | 750 | 480 | 450 |
| 30 | 700 | 760 | 500 | 440 |
| 31 | 680 | 780 | 520 | 420 |
| 32 | 740 | 460 | 460 | 740 |
| 33 | 620 | 410 | 580 | 790 |
| 34 | 790 | 600 | 410 | 600 |
| 35 | 660 | 780 | 540 | 420 |
| 36 | 690 | 770 | 510 | 430 |
| 37 | 600 | 410 | 600 | 790 |
| 38 | 640 | 790 | 560 | 410 |
| 39 | 780 | 540 | 420 | 660 |

neither as a SGPE or a RD equilibrium;

- in T2 $\delta^{''} = 0.55$, so that cooperation is sustainable among self-interested players as a SGPE but not as a RD equilibrium;

- in T3 $\delta^{''} = 0.75$, so that cooperation is sustainable among self-interested players both as a SGPE and a RD equilibrium.

The order of $\delta^{'}$ and $\delta^{''}$ is randomized across sessions to control for order effects.

We adopt a perfect-stranger matching procedure across blocks and a random-matching procedure within blocks. Specifically, to reduce contagion among subjects and increase the speed of convergence towards an equilibrium, we used matching groups of 4 subjects.

Subjects are informed of their opponent's choices and their round-payoff at the end of each round, and of their overall supergame-payoff in ECUs at the end of each supergame. Subjects are paid only for the payoff they realize in one supergame per series, which will be randomly selected at the end of the experiment. In Part 2, ECUs are converted Euros at a conversion rate such that 100 ECUs = 2 Euros.

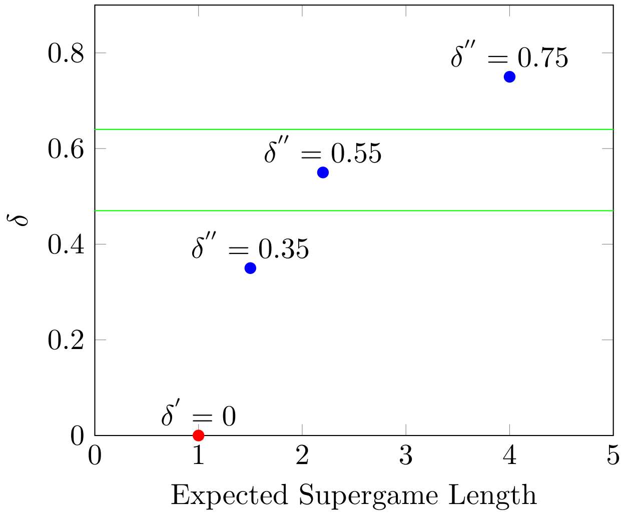
We further elicit subjects' perceptions on the share of cooperators in their session throughout the supergames of both $\delta$-series. In the $1^{st}, 5^{5th}$ and $10^{th}$ supergame of each series, subjects are asked to guess, before actually playing, what is the number of players in their group who would start by cooperating. If one of these supergame is selected to be the one relevant for subjects' payment and subjects' guess matches the actual fraction of individuals who cooperated in the first round of the selected supergame, subjects receive an additional fixed payment of 2 Euros.

Figure 3.5: PD Monetary Payoffs matrix

| | C | D |
|---|---|---|
| C | 73, 73 | 10, 100 |
| D | 100, 10 | 43, 43 |

$$g = l = 0.9$$

Figure 3.6: Part 2 - Differences across Treatments

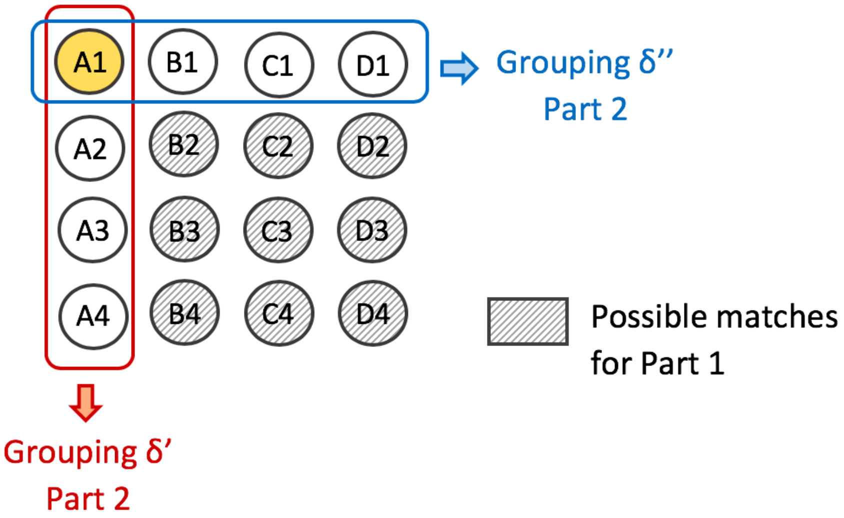


| | $\delta - \delta_{SGPE}$ | $\delta - \delta_{RD}$ | $sizeBAD$ |
|---|---|---|---|
| $\delta^{'} = 0$ | -0.47 | -0.64 | 1 |
| $\delta^{''} = 0.35$ | -0.12 | -0.29 | 1 |
| $\delta^{''} = 0.55$ | +0.08 | -0.09 | 0.736 |
| $\delta^{''} = 0.75$ | +0.28 | +0.11 | 0.3 |

Table 3.12: Experimental Design - Differences across Treatments

| | Treatment 1 | Treatment 2 | Treatment 3 |
|---|---|---|---|
| PART 1 | Pref. Elicitation<br>Questionnaire | Pref. Elicitation<br>Questionnaire | Pref. Elicitation<br>Questionnaire |
| PART 1 | Inf. Rep. PDs<br>$\delta^{'} = 0$; $\delta^{''} = 0.35$<br>Belief Elicitation | Inf. Rep. PDs<br>$\delta^{'} = 0$; $\delta^{''} = 0.55$<br>Belief Elicitation | Inf. Rep. PDs<br>$\delta^{'} = 0$; $\delta^{''} = 0.75$<br>Belief Elicitation |

Figure 3.7: Matching procedure for subject $A1$ in Part 1 and Part 2



### Recruitment & Matching Procedures

We recruit participants in order to have 16 subjects per session in Part 2 of the experiment [25].

When subjects take part in Part 1, playing the DGs in the role of the dictator, they are informed their actions will have monetary consequences both on their payoff and on the payoff of their matched partner, who is a randomly selected individual from their same session, with whom they will never interact again in Part 2 of the Experiment.

When subjects take part in Part 2, and interact in real time to play the infinitely repeated PDs, they are grouped in 4 groups of 4 people ($N_G = 4$) before each of the two $\delta$-series starts. Subjects are informed that they will never interact with the same counterpart across the two different $\delta$-series, and that not even their counterparts from the two different $\delta$-series will ever interact with each other.

The matching structure, as shown in Figure 3.7, ensures that subject $i$:

- will never meet again his/her counterparts from the $\delta'$-series in the $\delta''$-series;
- within each $\delta$-series, will be randomly paired with any of his/her $N_G - 1 = 3$ group mates, with a probability to be rematched to the same partner from the previous round equal to 1/3;

[25] To this aim, we recruit a higher number of participants in order to hit the target of least 20-22 subjects completing Part 1 within the due date and being entitled to participate in Part 2. This allows us to manage attritition issues between Part 1 and Part 2. Redundant participants who complete Part 1 within the due date and log-in on time for Part 2, but do not actually play, are paid a show-up fee of 5 Euros.

- in Part 2, will never interact again with the partners he/she was matched to in Part 1 (subject *i*' counterpart in Part 1 is randomly selected from the pool of the $N - 1 - ((N_G - 1) * 2) = 15 - 6 = 9$ subjects not grouped with subject $i$ in Part 2).

## 3.5 Empirical strategy

### 3.5.1 Hypotheses

The experiment is designed to study whether individuals with and without a *non-strategic* taste for cooperation, which is measured by the type and the intensity of social preferences, behave differently in terms of cooperative attitudes when playing the infinitely repeated PDs. Our conjecture is that individuals without a strategic taste for cooperation would exhibit a cooperative attitude only when the structural characteristics of the game and their expectations are such that cooperation can be a profitable strategy, while individuals having a non-strategic taste for cooperation would exhibit a cooperative attitude even when the structural characteristics of the game and their expectations do not guarantee cooperation would be a profitable strategy.

***Hypothesis 1.*** *Other-regarding (OR) types are more cooperative, on average, than the self-interested (SI) types when cooperation is not sustainable in equilibrium.*

***Hypothesis 2.*** *Groups with a higher density of OR types manage to reach and sustain higher levels of cooperation over time when cooperation is not an equilibrium.*

***Hypothesis 3.*** *The evolution of cooperation levels over supergames does not depend on the fraction of OR types in the group when cooperation is an equilibrium.*

***Hypothesis 4.*** *OR types are less sensitive to changes in the strategic incentives to cooperation with respect to SI types.*

***Hypothesis 5.*** *Structural parameters of the game are effective predictors of cooperativeness when cooperation is sustainable as an equilibrium but not necessarily otherwise, while,* *vice-versa, social preferences are relevant predictors of cooperativeness when cooperation is not an equilibrium and not necessarily otherwise.*

## 3.6 Results

The analysis of the experimental data collected through Part 1 and Part 2 of the experiment allows us to test whether the hypothesis that individuals with and without a strong *non-strategic* taste for cooperation behave differently in terms of cooperative attitudes in infinitely repeated PDs, is supported by the empirical evidence.

Relying on the data collected in Part 1 of the experiment (DGs choices), we are able to retrieve:

- individual-level estimates of social preferences parameters ($\alpha_i$ and $\beta_i$)
- type-specific estimates of social preferences parameters ($\alpha_t$ and $\beta_t$), which allow us to categorize subjects into one of the three social preference types: Behindness-Averse (BA), Moderately Altruistic (MA), and Strongly Altruistic (SA) types.

Table 3.13 shows some summary statistics on the estimates of social preference parameters obtained at the type and at the individual level.

Provided that, as discussed in Section 2, we are interested in testing qualitative predictions based on the distance between the estimated social preference parameters and the threshold values $\alpha^*$ and $\beta^*$ [26], we pool together Behindness-Averse (BA) and Moderately Altruistic (MA) types and rely on a binary classification: we idetify as Other-Regarding (OR) types the individuals classified as Strongly Altruistic (SA) and as Self-Interested (SI) all the others. Throughout the analysis we focus on Round1-Cooperation choices only.

[26] $\alpha^* = \frac{(P-S)}{(T-S)} = \frac{(43-10)}{(100-10)} \simeq 0.37$ is the threshold value that makes Cooperation become a best response to Defection in a one-shot PD game. $\beta^* = \frac{(T-R)}{(T-S)} = \frac{(100-73)}{(100-10)} \simeq 0.30$ is the threshold value that makes Cooperation become a best response to Cooperation in a one-shot PD game, and Grim a preferred strategy to Always Defect in an infinitely repeated PD irrespective of the actual value of $\delta$.

Table 3.13: Summary statistics of social preference parameters estimates

| | Type-specific estimates | | |
|---|---|---|---|
| . | Behindness-Averse | Moderately Altruistic | Strongly Altruistic |
| $\alpha$ | -0.254 | 0.029 | 0.18 |
| $\beta$ | 0.015 | 0.046 | 0.418 |

| | Summary of individual-specific estimates | |
|---|---|---|
| | Strongly Altruistic | Not Strongly Altruistic |
| $\alpha$ | 0.164 [0.188] | -0.157 [0.510] |
| $\beta$ | 0.414 [0.205] | 0.007 [0.303] |

Notes: Type-specific estimates are obtained from the Finite Mixture Model with k=3. Individual-specific estimates are obtained through a procedure that estimates the parameters separately for each subject. For details on the estimation procedure see Bruhin et al. (2018). Standard deviation in squared brackets.

### 3.6.1 Evidence from One-Shot play: Hypothesis 1 & 2

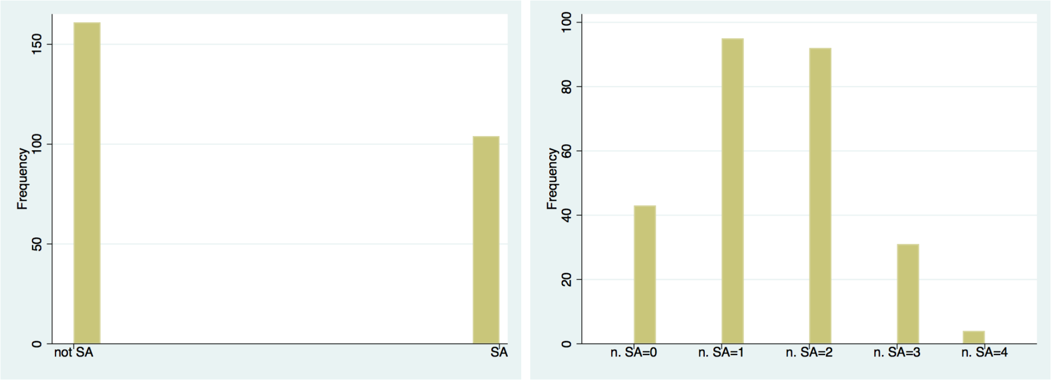
We rely on data on subjects' choices in one-shot PDs ($\delta^{'}$ series, where $\delta = 0$), pooling data from all treatments and sessions (N=265) to test whether Hypothesis 1 and 2 are supported by the data [27].

Figure 3.8 shows the distribution of social preference types within the sample (Left panel) and the, endogenously determined, distribution based on the number of Strongly Altruistic types within the matching group (Right panel).

Figure 3.9 shows average cooperativeness across supergames and over supergames by social preference types (Upper panel) and conditional on the number of Strongly Altruistic types within the matching group (Lower panel).

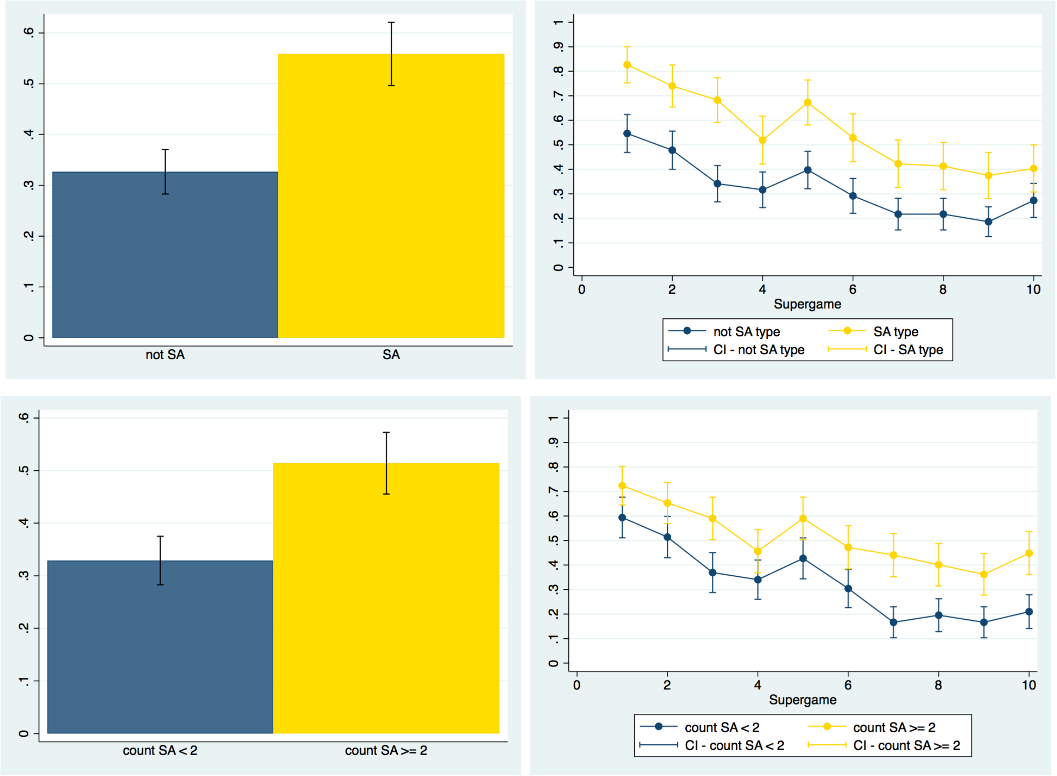
[27] We exclude 23 subjects from the original 288 subjects sample because they behaved very inconsistently in the series of Dictator Games, therefore we were not able to identify their social preference parameters.

Figure 3.8: Distribution by social preference type and SA count (N=265)



*Notes.* Data from all sessions with $\delta = 0$ (N=265); Legend. SA: 'Strongly Altruistic'; not SA: 'not Strongly Altruistic', which combines 'Behindness-Averse' and 'Moderately Altruistic' individuals.

Figure 3.9: Average cooperation by types and SA count



*Notes.* Data from all sessions with $\delta = 0$ (N=265); Legend. SA: 'Strongly Altruistic'; not SA: not 'Strongly Altruistic' combines 'Behindness-Averse' and MA: 'Moderately Altruistic' individuals.

Table 3.14: Cooperation choices across and over supergames by social preference types

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| SA | 0.232*** | 0.141*** | 0.950*** | 1.315*** | 1.324*** | 0.980*** |
| | (0.0432) | (0.0387) | (0.188) | (0.248) | (0.279) | (0.281) |
| Guess | | 0.189*** | | | | |
| | | (0.0179) | | | | |
| Supergame | | | -0.160*** | | | |
| | | | (0.0179) | | | |
| Supergame · (SA=0) | | | | -0.135*** | -0.112*** | -0.0467** |
| | | | | (0.0221) | (0.0206) | (0.0225) |
| Supergame · (SA=1) | | | | -0.201*** | -0.190*** | -0.106*** |
| | | | | (0.0240) | (0.0275) | (0.0295) |
| Belief | | | | | | 0.761*** |
| | | | | | | (0.0828) |
| *Constant* | 0.327*** | -0.0158 | 0.195 | 0.0670 | 0.240 | -1.550*** |
| | (0.0265) | (0.0311) | (0.135) | (0.151) | (0.160) | (0.237) |
| R2 (or Pseudo R2) | 0.128 | 0.408 | 0.088 | 0.091 | 0.114 | 0.274 |
| LogLik | -53.23 | -1.919 | -1365.0 | -1360.7 | -465.5 | -381.5 |
| N | 265 | 265 | 2650 | 2650 | 795 | 795 |

Models (1)-(2): Estimated coefficients of OLS models of average cooperation across rounds on social preference types.

Models (3)-(7): Estimated coefficients of a correlated random effects probit of $1^{st}$ Round cooperation choices on social preference types ("SA" is a dummy variable which is valued 1 if the subject is classified as a Strongly Altruistic type), accounting for within-session trends over supergames with and w/o controlling for beliefs on the share of cooperators in the group.

Models (3)-(4) are estimated on the whole sample of 2650 (N=265 x T=10) observations. Models (5)-(7) are estimated on a reduced sample of 795 (N=265 x T=3) observations, for which we can also observe subjects' beliefs. Model (4)-(6): The difference between "Supergame · (SA=0)" and "Supergame · (SA=1)" is always statistically significant: $\chi^2$=5.34, p-value= 0.0209 in model (4); $\chi^2$=8.79, p-value= 0.0030 in model (5); $\chi^2$=4.09, p-value= 0.0430 in model (6).

Model (5): The impact of "Belief" is not statistically different for the two types.

Note: Clustering at matching group level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

Table 3.15: Cooperation choices across and over supergames by group composition

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Count SA $\geq$ 2 | 0.185*** | 0.0751** | 0.746*** | 0.489* | 0.443 | 0.135 |
| | (0.0504) | (0.0355) | (0.212) | (0.261) | (0.289) | (0.281) |
| Belief (mean) | | 0.196*** | | | | |
| | | (0.0194) | | | | |
| Supergame | | | -0.160*** | | | |
| | | | (0.0179) | | | |
| Supergame $\cdot$ (SA $<$ 2) | | | | -0.184*** | -0.162*** | -0.0855*** |
| | | | | (0.0228) | (0.0217) | (0.0219) |
| Supergame $\cdot$ (SA $\geq$ 2) | | | | -0.135*** | -0.116*** | -0.0467 |
| | | | | (0.0264) | (0.0278) | (0.0325) |
| Belief | | | | | | 0.768*** |
| | | | | | | (0.0870) |
| *Constant* | 0.329*** | -0.00995 | 0.212 | 0.331* | 0.507*** | -1.279*** |
| | (0.0311) | (0.0310) | (0.157) | (0.169) | (0.187) | (0.245) |
| R2 | 0.085 | 0.377 | 0.083 | 0.085 | 0.101 | 0.262 |
| LogLik | -59.54 | -8.731 | -1371.7 | -1369.3 | -472.4 | -388.2 |
| N | 265 | 265 | 2650 | 2650 | 795 | 795 |

Models (1)-(2): Estimated coefficients of OLS models of average cooperation across rounds on social preference types.

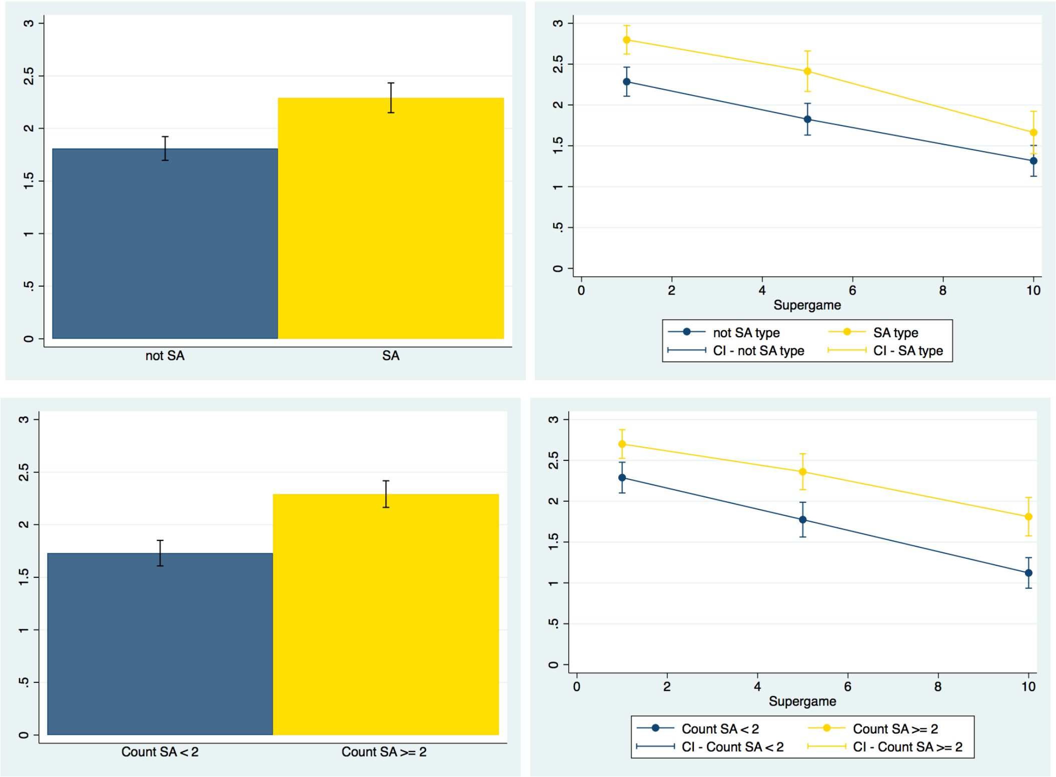
Models (3)-(6): Estimated coefficients of a correlated random effects probit of $1^{st}$ Round cooperation choices on the count of social preference types ("Count SA" measures the number of subjects classified as Strongly Altruistic types within the matching group), accounting for within-session trends over supergames with and w/o controlling for beliefs on the share of cooperators in the group.

Models (3)-(4) are estimated on the whole sample of 2650 (N=265 x T=10) observations. Models (5)-(6) are estimated on a reduced sample of 795 (N=265 x T=3) observations, for which we can also observe subjects' beliefs. Model (4)-(6): The difference between "Supergame $\cdot$ (Count SA $<$ 2)" and "Supergame $\cdot$ (Count SA $\geq$ 2)" is never statistically significant: $\chi^2$=2.03, p-value= 0.1538 in model (4) $\chi^2$=1.91, p-value=0.1669 in model (5); $\chi^2$=1.06, p-value= 0.3039 in model (6).

Model (5): The impact of "Belief" is not statistically different for the two types.

Note: Clustering at matching group level. Sign. levels: * $p$ < 0.10, ** $p$ < 0.05, *** $p$ < 0.01. Clustered standard errors in parentheses.

Figure 3.10: Average beliefs across and over rounds by types and SA count



*Notes.* Data from all sessions with $\delta = 0$ (N=265); Legend. SA: ’Strongly Altruistic’; not SA: not ’Strongly Altruistic’ combines ’Behindness-Averse’ and MA: ’Moderately Altruistic’ individuals.

As shown by Figure 3.9 and confirmed by the econometric analysis reported in Table 3.14, Strongly Altruistic individuals do exhibit a higher cooperative tendency than others, even after controlling for beliefs on the share of cooperators in the group, which have a strong and positive per-se effect on cooperation. Indeed, beliefs explain a large fraction of the variation in observed cooperation levels and - despite a large within-group heterogeneity - Strongly Altruistic types show on average significantly higher beliefs, as shown by Figure 3.10 (Upper panel).

Moreover, as shown by Figure 3.9 and Table 3.15, groups with a higher count of Strongly Altruistic types do reach and sustain a higher level of cooperation. However, most of the difference in cooperation trends and levels between the two groups is explained by between-groups differences in beliefs, provided that groups with a higher count of Strongly Altruistic types show on average significantly higher beliefs, see Figure 3.10 (Lower panel).

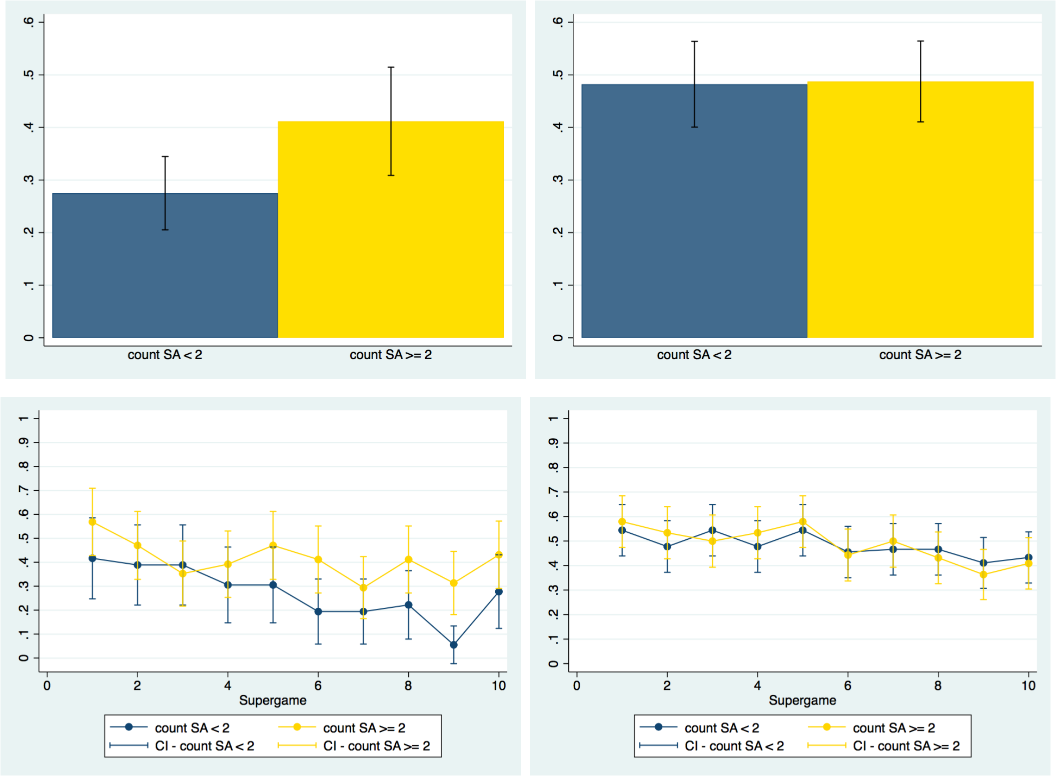
***Result 1.*** *Data on one-shot PD series strongly support Hypotheses 1 and 2: Other-Regarding (OR) types are more cooperative than the Self-Interested (SI) types and groups with a higher count of OR types reach higher levels of cooperation.*

### 3.6.2 Evidence from Infinitely-Repeated play: Hypothesis 2 & 3

To further test Hypothesis 2 and 3, we rely on data on subjects' choices in infinitely repeated PDs ($\delta''$ series, where $\delta > 0$), looking seprarately at data from Treatment 1 ($\delta = 0.35 < \delta^{SPE}$, N=87) and Treatments 2-3 ($\delta = 0.55 > \delta^{SPE}$, N=91 and $\delta = 0.75 > \delta^{SPE}$, N=87).

Figure 3.11 shows average cooperativeness across and over supergames conditional on the number of Strongly Altruistic types within the matching group across treatments where cooperation is not sustainable as an equilibrium (Left panel, Treatment 1 - $\delta = 0.35$) and where cooperation is sustainable as an equilibrium (Right panel, Treatment 2 and 3 - $\delta = 0.55 \& \delta = 0.75$).

Figure 3.11: Average cooperation by SA count: not SPE and SPE Treatments



*Notes.*Data from all sessions with $\delta > 0$ (N=265); Left panel: not SPE treatment; Right panel: SPE treatments. Legend. SA: 'Strongly Altruistic'; not SA: not 'Strongly Altruistic' combines 'Behindness-Averse' and MA: 'Moderately Altruistic' individuals.

Table 3.16: Cooperation choices in IR-PDs by group composition: SPE Treatments

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Count SA $\geq$ 2 | 0.00528 | 0.00319 | 0.109 | 0.263 | 0.212 | 0.202 | -0.492 |
| | (0.0718) | (0.0484) | (0.345) | (0.330) | (0.275) | (0.262) | (0.454) |
| Belief (mean) | | 0.210*** | | | | | |
| | | (0.0176) | | | | | |
| Supergame | | | -0.0753*** | | | | |
| | | | (0.0217) | | | | |
| Supergame $\cdot$ (SA $<$ 2) | | | | -0.0606** | -0.0522* | -0.0287 | -0.0310 |
| | | | | (0.0280) | (0.0274) | (0.0279) | (0.0264) |
| Supergame $\cdot$ (SA $\geq$ 2) | | | | -0.0893*** | -0.0775** | -0.0543* | -0.0543* |
| | | | | (0.0323) | (0.0306) | (0.0287) | (0.0308) |
| Belief | | | | | | 0.699*** | |
| | | | | | | (0.0829) | |
| Belief $\cdot$ (SA $<$ 2) | | | | | | | 0.575*** |
| | | | | | | | (0.0956) |
| Belief $\cdot$ (SA $\geq$ 2) | | | | | | | 0.867*** |
| | | | | | | | (0.137) |
| *Constant* | 0.482*** | 0.00524 | 0.339 | 0.260 | 0.305 | -1.444*** | -1.141*** |
| | (0.0548) | (0.0483) | (0.253) | (0.236) | (0.189) | (0.278) | (0.298) |
| R2 (or PseudoR2) | 0.0000497 | 0.350 | 0.018 | 0.018 | 0.020 | 0.163 | 0.169 |
| LogLik | -77.69 | -39.37 | -847.0 | -846.5 | -326.9 | -279.2 | -277.3 |
| N | 178 | 178 | 1780 | 1780 | 534 | 534 | 534 |

Models (1)-(2): Estimated coefficients of OLS models of average cooperation across rounds on social preference types.

Models (3)-(7): Estimated coefficients of a correlated random effects probit of $1^{st}$ Round cooperation choices on the count of social preference types ("Count SA" measures the number of subjects classified as Strongly Altruistic types within the matching group), accounting for within-session trends over supergames with and w/o controlling for beliefs on the share of cooperators in the group.

Models (3)-(4) are estimated on the whole SPE treatments sample of 1780 (N=178 x T=10) observations. Models (5)-(7) are estimated on a reduced sample of 534 (N=178 x T=3) observations, for which we can also observe subjects' beliefs. Model (4)-(6): The difference between "Supergame $\cdot$ (Count SA $<$ 2)" and "Supergame $\cdot$ (Count SA $\geq$ 2)" is never statistically significant: $\chi^2$=0.45, p-value= 0.5036 in model (4) $\chi^2$=0.39, p-value=0.534 in model (5); $\chi^2$=0.41, p-value= 0.5208 in model (6); $\chi^2$=0.33, p-value=0.5629 in model (7).

Model (7): The impact of "Belief" is marginally (at the 10% level) different for the two types, with a larger estimated effect for the SA type: $\chi^2$=3.12, p-value=0.0773

Table 3.17: Cooperation choices in IR-PDs by group composition: not SPE Treatment

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Count SA $\geq$ 2 | 0.137* | 0.136** | 0.412 | 0.173 | 0.553* | 0.612** | -0.408 |
| | (0.0714) | (0.0517) | (0.287) | (0.304) | (0.309) | (0.255) | (0.392) |
| Belief (mean) | | 0.155*** | | | | | |
| | | (0.0312) | | | | | |
| Supergame | | | -0.0839*** | | | | |
| | | | (0.0260) | | | | |
| Supergame $\cdot$ (SA $<$ 2) | | | | -0.109*** | -0.0547 | -0.00667 | -0.0266 |
| | | | | (0.0287) | (0.0458) | (0.0443) | (0.0406) |
| Supergame $\cdot$ (SA $\geq$ 2) | | | | -0.0633 | -0.0567* | -0.0194 | -0.0105 |
| | | | | (0.0394) | (0.0338) | (0.0301) | (0.0320) |
| Belief | | | | | 0.565*** | | |
| | | | | | | (0.106) | |
| Belief $\cdot$ (SA $<$ 2) | | | | | | | 0.309*** |
| | | | | | | | (0.0995) |
| Belief $\cdot$ (SA $\geq$ 2) | | | | | | | 0.731*** |
| | | | | | | | (0.136) |
| *Constant* | 0.275*** | -0.0285 | -0.348* | -0.217 | -0.295 | -1.674*** | -1.033*** |
| | (0.0391) | (0.0644) | (0.204) | (0.204) | (0.212) | (0.300) | (0.238) |
| R2 (or PseudoR2) | 0.0460 | 0.278 | 0.025 | 0.027 | 0.025 | 0.120 | 0.132 |
| LogLik | -20.61 | -8.499 | -448.7 | -447.9 | -160.8 | -145.2 | -143.3 |
| N | 87 | 87 | 870 | 870 | 261 | 261 | 261 |

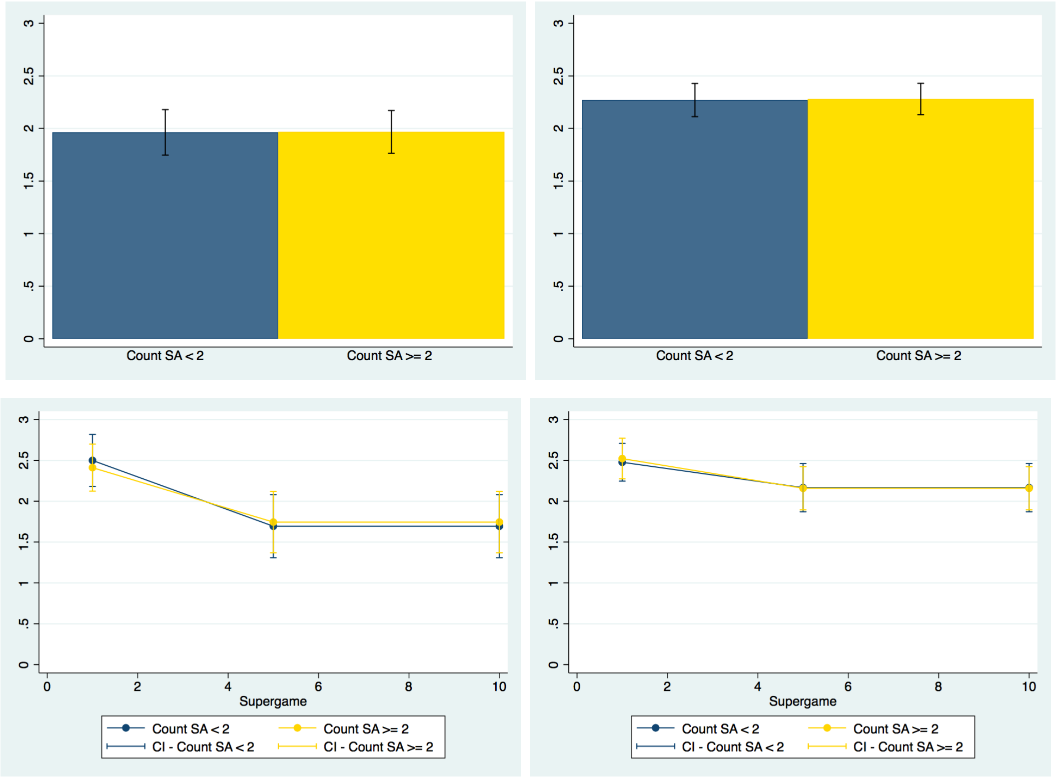
Models (1)-(2): Estimated coefficients of OLS models of average cooperation across rounds on social preference types.

Models (3)-(7): Estimated coefficients of a correlated random effects probit of $1^{st}$ Round cooperation choices on the count of social preference types ("Count SA" measures the number of subjects classified as Strongly Altruistic types within the matching group), accounting for within-session trends over supergames with and w/o controlling for beliefs on the share of cooperators in the group.

Models (3)-(4) are estimated on the not SPE treatment sample of 870 (N=87 x T=10) observations. Models (5)-(7) are estimated on a reduced sample of 261 (N=87 x T=3) observations, for which we can also observe subjects' beliefs. Model (4)-(6): The difference between "Supergame $\cdot$ (Count SA $<$ 2)" and "Supergame $\cdot$ (Count SA $\geq$ 2)" is never statistically significant: $\chi^2$=0.89, p-value=0.3459 in model (4) $\chi^2$=0.00, p-value=0.9723 in model (5); $\chi^2$=0.06, p-value=0.8125 in model (6); $\chi^2$=0.10, p-value=0.7571 in model (7).

Model (7): The impact of "Belief" is statistically different for the two types, with a larger estimated effect for

Figure 3.12: Average beliefs across and over rounds by SA count: not SPE and SPE Treatments



*Notes.* Data from all sessions with $\delta > 0$ (N=265); Left panel: not SPE treatment; Right panel: SPE treatments. Legend. SA: 'Strongly Altruistic'; not SA: not 'Strongly Altruistic' combines 'Behindness-Averse' and MA: 'Moderately Altruistic' individuals.

As shown by Figure 3.11 and confirmed by the econometric analysis reported in Tables 3.16 and 3.17, groups with a higher count of Strongly Altruistic types tend to reach and sustain a higher level of cooperation compared to other groups in treatments where cooperation is not sustainable as a SPE, but not in treatments where cooperation is actually sustainable as a SPE. As in the case of one-shot interactions, in treatments where cooperation is not an equilibrium, most of the difference in cooperation trends and levels between the two groups is explained by how individuals react to their beliefs, provided that groups with different counts of Strongly Altruistic types do not show significantly different beliefs neither in the SPE treatments, see Figure 3.12 (Left panel), nor in the not SPE treatment.

***Result 2.*** *Data on infinitely repeated PD series strongly support Hypotheses 2 and 3: groups with a higher count of OR types reach substantially higher levels of cooperation when cooperation is not sustainable as an equilibrium but not otherwise.*

### 3.6.3 Combining evidence from One-Shot and Infinitely-Repeated play: Hypothesis 4

To test Hypothesis 4 we pool data from all treatments and series. Given the characteristics of our experimental setup, where the payoff matrix is constant across one-shot and infinitely-repeated series, in order to test the effects of changes in strategic incentives for cooperation we look at changes in subjects' behavior between the $\delta^{''} > 0$ and $\delta^{'} = 0$ series.

In order to estimate and test the relevance of the effect of changes in strategic incentives for cooperation on behavior, we adopt a Difference-in-Difference approach, where we identify individuals assigned to Treatment 1 - where $\delta = 0.35$ in the IR series - as the Control group and individuals assigned to Treatment 2 and 3 - where $\delta = 0.55$ and $\delta = 0.75$ in the IR series - as the Treated group. Since all individuals are first exposed to a series of one-shot PDs (which identifies the pre-treatment period in our D-i-D framework, see Table 3.18), we want to test whether the change in cooperativeness moving from one-shot to infinitely repeated series is positive and significant for individuals in the Treated group, who are exposed to a significant change in strategic incentives for cooperation:

$$Y_{igt} = Y_0 + \alpha_g \cdot Treat + \gamma_t \cdot IR + \theta_* \cdot (Treat \cdot IR) + \epsilon_{igt}$$

where $Y_{igt}$ is the binary cooperation choice by individual "i" belonging to group "g" (either Treated or Control) at time "t" (either belonging to the one-shot OS or infinitely repeated IR series) and the treatment effect is captured by parameter $\theta_*$.

Table 3.18: Difference-in-Difference Framework

| | **Treated Group (T2 and T3)** | **Control Group (T1)** |
|---|---|---|
| **Pre-Treatment Period (One-Shot play)** | One-Shot play<br>T2 and T3 | One-Shot play<br>T1 |
| **Post-Treatment Period (Infinitely Repeated play)** | Inf. Rep. play<br>T2 and T3 | Inf. Rep. play<br>T1 |

Table 3.19 reports the results of this estimation, proving there's a positive and significant treatment effect, as suggested by the graphical evidence brought by Figure 3.13 (Upper panel). In order to test whether individuals belonging to different social preference types react differently to the treatment, we estimate a fully interacted model, in order to obtain an estimation of all paramenters of interest, and mainly of $\theta_*^{SA}$ and $\theta_*^{nSA}$:

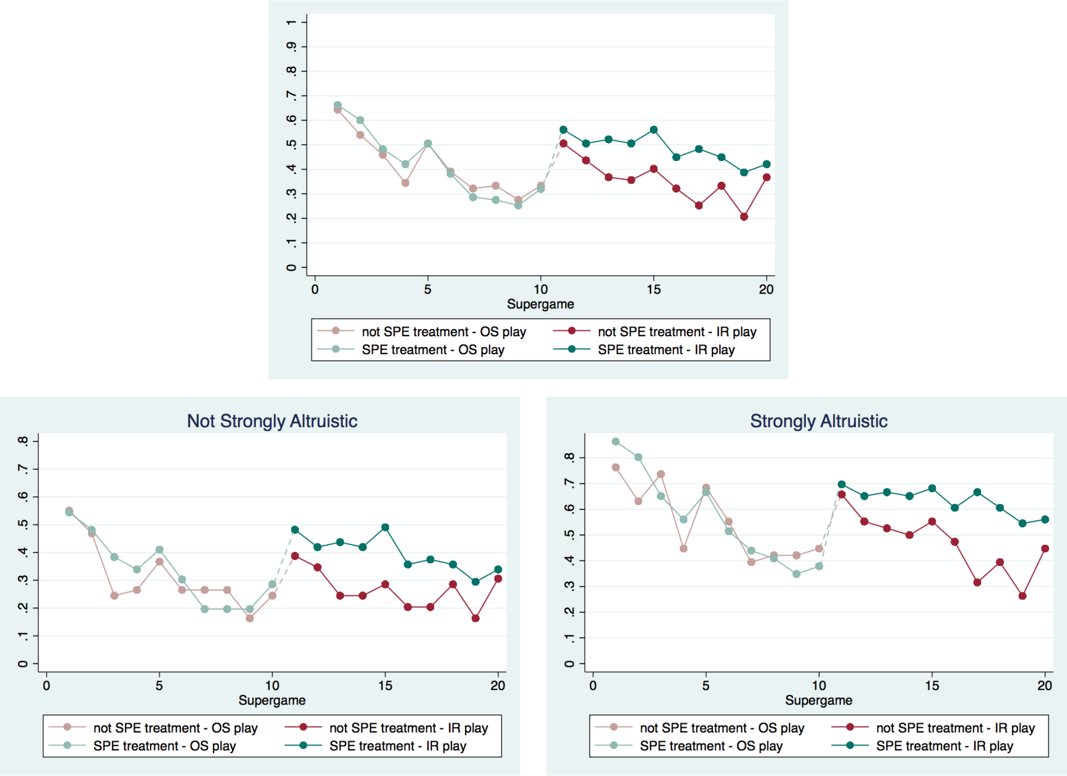
$$\begin{cases} Y_{igt}^{SA} = Y_0^{SA} + \alpha^{SA} \cdot Treat_g + \gamma^{SA} \cdot IR_t + \theta_*^{SA} \cdot (Treat_g \cdot IR_t) + \epsilon_{igt}^{SA} \\ Y_{igt}^{nSA} = Y_0^{nSA} + \alpha^{nSA} \cdot Treat_g + \gamma^{nSA} \cdot IR_t + \theta_*^{nSA} \cdot (Treat_g \cdot IR_t) + \epsilon_{igt}^{nSA} \end{cases}$$

Table 3.20 reports the results of the estimation of the fully interacted model, which confirm what emerges from the graphs reported in Figure 3.13 (Lower panel). Strongly Altruistic types have a significantly higher intercept (the hypothesis that $Y_0^{SA} = Y_0^{nSA}$ is strongly rejected in all specifications) and, as expected, there's no significant "group-assignment effect" for either social preference type (the hypothesis that $\alpha^{SA} = \alpha^{nSA}$ is accepted in all specifications). Similarly to what emerged from the pooled D-i-D model (Table 3.18) there's a negative and significant "time-effect" when moving from one-shot to infinitely-repeated series: the effect is not statistically different between the two social preference types (the hypothesis that $\gamma^{SA} = \gamma^{nSA}$ is accepted in all specifications) and could be explained by the fact that individuals belonging to the Control group de-facto experience a prolonged series of interactions in an environemnt where cooperation is not sustainable as an equilibrium and further learn not to cooperate.

The "treatment effect" is positive and significant in all specifications in which we do not control for individuals' beliefs and the effect is not statistically different between the two social preference types (the hypothesis that $\theta_*^{SA} = \theta_*^{nSA}$ is accepted in all specifications). Once we control for individual beliefs on the share of cooperators in the group, however, the effect vanishes. A cadidate explanation for this evidence, combined with what emerges from Figure 3.14 on the dynamics of beliefs across treatments and social preference types, is that not Strongly Altrustic types do react more strongly than Strongly Altruistic types to changes in strategic incentives for cooperation but only in terms of beliefs. However, the translation of this effect into behavior is weakened by the fact that Strongly Altruistic types have substantially higher perceptions on the share of cooperators in the group, regardless of the strategic environment, and tend to be less reactive to beliefs.

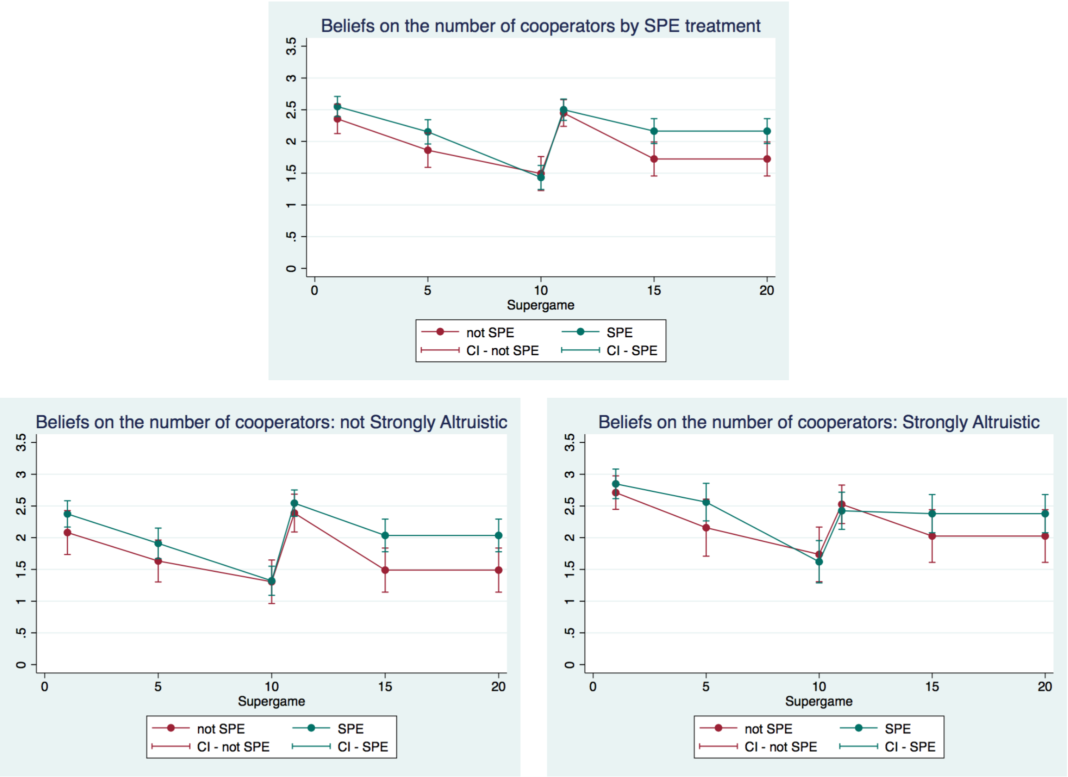
***Result 3.*** *Data on one-shot and infinitely repeated PD series do not support Hypotheses 4: OR types do not react less strongly than SI types to changes in strategic incentives for co-operation in terms of cooperative behavior, although they seem to react less strongly in terms of beliefs.*

Figure 3.13: Average Cooperativeness across supergames between $\delta$-series



*Notes.* Average cooperativeness across supergames: "OS play" identifies observations from One-Shot games from the $\delta$-series (labeled as 'Supergames 1 to 10'), while "IR" identifies observations from Infinitely Repaeated games from the $\delta > 0$ series (labeled as 'Supergames 11 to 20'). Upper panel: All types pooled; Lower panel: separately by type.

Figure 3.14: Average Beliefs over supergames between $\delta$-series



*Notes.* Average Beliefs on the share of cooperators within the matching group: "OS play" identifies observations from One-Shot games from the $\delta$-series (labeled as 'Supergames 1 to 10'), while "IR" identifies observations from Infinitely Repaeated games from the $\delta > 0$ series (labeled as 'Supergames 11 to 20'). Upper panel: All types pooled; Lower panel: separately by type.

Table 3.19: Difference-in-Difference

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| SPE | -0.00265 | 0.000235 | 0.00319 | 0.00967 | -0.0807 |
| | (0.150) | (0.158) | (0.161) | (0.173) | (0.158) |
| IR | -0.218* | -0.226* | -0.672*** | -0.637*** | -0.511*** |
| | (0.115) | (0.120) | (0.161) | (0.193) | (0.188) |
| SPE · IR | 0.452*** | 0.466*** | 0.461*** | 0.299* | 0.186 |
| | (0.144) | (0.151) | (0.151) | (0.181) | (0.176) |
| Supergame (1-10) | | -0.107*** | | | |
| | | (0.00914) | | | |
| Supergame (1-10) · IR=0 | | | -0.149*** | -0.130*** | -0.0692*** |
| | | | (0.0138) | (0.0147) | (0.0159) |
| Supergame (1-10) · IR=1 | | | -0.0657*** | -0.0548*** | -0.0279* |
| | | | (0.0125) | (0.0139) | (0.0152) |
| Belief | | | | | 0.641*** |
| | | | | | (0.0442) |
| *Constant* | -0.266** | 0.306** | 0.531*** | 0.672*** | -0.883*** |
| | (0.122) | (0.135) | (0.148) | (0.165) | (0.181) |
| Pseudo R2 | 0.006 | 0.045 | 0.050 | 0.052 | 0.197 |
| LogLik | -2886.8 | -2775.0 | -2758.1 | -950.0 | -805.2 |
| N | 5300 | 5300 | 5300 | 1590 | 1590 |

Models (1)-(5): Estimated coefficients of a correlated random effects probit of $1^{st}$ Round cooperation choices on: SPE (=1 the subject is assigned to an SPE-treatment, which means is part of the Treatment group), IR (=1 if the observations refer to the Infinitely Repeated series where $\delta > 0$), accounting for within-series trends over supergames with and w/o controlling for beliefs on the share of cooperators in the group.

Models (1)-(3) are estimated on the whole sample of 5300 (N=265 x T=20) observations. Models (4)-(5) are estimated on a reduced sample of 1590 (N=265 x T=6) observations, for which we can also observe subjects' beliefs. Model (3)-(5): The difference between "Supergame · (IR = 0)" and "Supergame · (IR = 1)" is always statistically significant when we do not account for beliefs: $\chi^2$=19.93, p-value= 0.000 in model (3) $\chi^2$=14.10, p-value=0.000 in model (4). The difference is only marginally significant (at 10%) when we account for beliefs: $\chi^2$=3.65, p-value=0.056 in model (5).

Note: Clustering at the individual level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

Table 3.20: Difference-in-Difference by social preference types

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| SA | 0.867*** | 0.931*** | 1.273*** | 1.304*** | 0.997*** |
| | (0.235) | (0.251) | (0.292) | (0.343) | (0.331) |
| SPE · (SA=0) | 0.0690 | 0.0774 | 0.0758 | 0.0956 | -0.0106 |
| | (0.178) | (0.190) | (0.188) | (0.206) | (0.191) |
| SPE · (SA=1) | 0.0308 | 0.0435 | 0.0485 | 0.0179 | -0.0887 |
| | (0.229) | (0.246) | (0.253) | (0.281) | (0.261) |
| IR · (SA=0) | -0.159 | -0.597*** | -0.508** | -0.486* | -0.453* |
| | (0.176) | (0.214) | (0.234) | (0.263) | (0.258) |
| IR · (SA=1) | -0.292** | -0.771*** | -0.932*** | -0.936*** | -0.627** |
| | (0.134) | (0.173) | (0.202) | (0.291) | (0.276) |
| SPE · IR · (SA=0) | 0.377* | 0.378* | 0.378* | 0.275 | 0.111 |
| | (0.206) | (0.216) | (0.214) | (0.245) | (0.240) |
| SPE · IR · (SA=1) | 0.552*** | 0.566*** | 0.569*** | 0.320 | 0.299 |
| | (0.199) | (0.210) | (0.214) | (0.275) | (0.262) |
| Supergame (1-10) · IR=0 | | -0.150*** | | | |
| | | (0.0138) | | | |
| Supergame (1-10) · IR=1 | | -0.0657*** | | | |
| | | (0.0125) | | | |
| Supergame (1-10) · IR=0 (SA=0) | | | -0.126*** | -0.105*** | -0.0485** |
| | | | (0.0181) | (0.0184) | (0.0191) |
| Supergame (1-10) · IR=1 (SA=0) | | | -0.0592*** | -0.0487*** | -0.0121 |
| | | | (0.0159) | (0.0179) | (0.0200) |
| Supergame (1-10) · IR=0 (SA=1) | | | -0.189*** | -0.178*** | -0.108*** |
| | | | (0.0202) | (0.0240) | (0.0264) |
| Supergame (1-10) · IR=1 (SA=1) | | | -0.0764*** | -0.0649*** | -0.0560** |
| | | | (0.0204) | (0.0218) | (0.0232) |
| Belief | | | | | 0.634*** |
| | | | | | (0.0439) |
| *Constant* | -0.644*** | 0.129 | 0.00667 | 0.159 | -1.259*** |
| | (0.148) | (0.169) | (0.178) | (0.195) | (0.214) |
| Peusdo R2 | 0.013 | 0.058 | 0.059 | 0.073 | 0.213 |
| LogLik | -2866.0 | -2736.9 | -2732.3 | -929.8 | -788.6 |
| N | 5300 | 5300 | 5300 | 1590 | 1590 |

### 3.6.4 Evidence from Infinitely-Repeated Games: Hypothesis 5

To test Hypothesis 5, we rely on data on subjects' choices in infinitely repeated PDs ($\delta''$ series, where $\delta > 0$), pooling data from Treatment 1 ($\delta = 0.35 < \delta^{SPE}$, N=87) and Treatments 2-3 ($\delta = 0.55 > \delta^{SPE}$, N=91 and $\delta = 0.75 > \delta^{SPE}$, N=87). Pooling observations from all treatments, we are interested in testing whether accounting for social preference types leads to a higher increase in explained variation in treatments where cooperation is not sustainable as a SPE. As reported in Table 3.21, in contrast to our initial conjecture but in line with the evidence reported on the other hypotheses, the social preference type categorization has a remarkable impact on explained variation regardless of the strategic environment and most of the effect seems to be transmitted through the beliefs channel.

***Result 4.*** *Data on infinitely repeated PD series do not support Hypotheses 5, coherently with previous results: the social preference type categorization is a strong predictor of cooperativeness regardless of the strategic environment.*

Table 3.21: Cooperation choices over supergames across treatments and social preference types

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|---|
| Supergame | -0.0647*** | -0.0734*** | -0.0738*** | -0.0739*** | -0.0637*** | -0.0436** | -0.0444** |
| · (SPE=1) | (0.0206) | (0.0211) | (0.0211) | (0.0211) | (0.0201) | (0.0197) | (0.0197) |
| Supergame | -0.104*** | -0.0879*** | -0.0878*** | -0.0877*** | -0.0590** | -0.00985 | -0.0107 |
| · (SPE=0) | (0.0263) | (0.0269) | (0.0269) | (0.0269) | (0.0286) | (0.0267) | (0.0266) |
| | | | | | | | |
| SPE | | 0.519** | 0.597** | | | | |
| | | (0.240) | (0.234) | | | | |
| SA | | | 1.004*** | 0.898*** | 0.841*** | 0.594** | 0.398 |
| | | | (0.198) | (0.234) | (0.280) | (0.255) | (0.378) |
| SPE · (SA=0) | | | | 0.529** | 0.475** | 0.319 | 0.330 |
| | | | | (0.264) | (0.240) | (0.240) | (0.239) |
| SPE · (SA=1) | | | | 1.591*** | 1.268*** | 1.013*** | 0.813** |
| | | | | (0.308) | (0.281) | (0.265) | (0.374) |
| Belief | | | | | | 0.652*** | |
| | | | | | | (0.0652) | |
| Belief · (SA=0) | | | | | | | 0.618*** |
| | | | | | | | (0.0732) |
| Belief · (SA=1) | | | | | | | 0.711*** |
| | | | | | | | (0.105) |
| *Constant* | 2.927 | 2.300 | 1.404 | 1.388 | 1.326 | -0.978 | -0.933 |
| | (1.787) | (1.834) | (1.765) | (1.770) | (1.706) | (1.449) | (1.438) |
| Pseudo R2 | 0.024 | 0.025 | 0.035 | 0.035 | 0.172 | 0.172 | 0.173 |
| LogLik | -1297.9 | -1295.6 | -1282.5 | -1282.4 | -477.0 | -414.4 | -414.1 |
| N | 2650 | 2650 | 2650 | 2650 | 795 | 795 | 795 |

Models (1)-(7): Estimated coefficients of a correlated random effects probit of 1st Round cooperation choices on: SA (=1 if the subject is classified as a Strongly Altruistic type), SPE (=1 if the subject is assigned to an SPE-treatment, which means is part of the Treatment group), accounting for within-series trends over supergames with and w/o controlling for beliefs on the share of cooperators in the group.

Model (1)-(7): The difference between "Supergame · (IR = 0)" and "Supergame · (IR = 1)" is never statistically significant in model (2) (3) and (4).

Model (5): The impact of "Belief" is not statistically different for the two types.

Note: Clustering at the individual level. Sign. levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Clustered standard errors in parentheses.

## 3.7 Conclusion

This study aims to investigate what shapes individuals' cooperative attitudes in infinitely repeated Prisoner Dilemmas.

From the meta-analysis it emerges that structural incentives for cooperation, measured using compact indicators developed by the literature like $(\delta - \delta_{RD})$ or $sizeBAD$, do have some predictive power: their ability to predict cooperation in terms of Round-1 Cooperation increases over supergames, suggesting subjects actually learn about structural incentives for cooperation while playing and respond to them. It further emerges an asymmetry in terms of prediction accuracy along the SGPE or RD equilibrium dimensions: prediction models seem to produce classifiers that are more accurate - especially in terms of models' ability to detect cooperators - in treatments where structural game parameters are such that cooperation is sustainable in equilibrium.

From the experimental data, collected through an experimental procedure that allows to observe, within subjects, both a measure of individual social preferences parameters and the actual play in infinitely repeated PDs, we find a strong evidence in favor of the role of social preferences in shaping cooperation when cooperation would not be sustainable as an equilibrium: subjects classified as 'Strongly Altrusitic' types, based on their estimated social preference parameters, show a significantly stronger cooperative attitude, and groups with a higher concentration of 'Strongly Altrusitic' types manage to reach and mantain higher levels of cooperation.
However, contrary to what we expected, subjects classified as 'Strongly Altrusitic' types are not less sensitive than others to changes in strategic incentives for cooperation. Although not 'Strongly Altruistic' types seem to react more markedly to changes in strategic incentives to cooperation in terms of updating beliefs on the share of cooperators in the group, 'Strongly Altrusitic' types tend exhibit on average higher levels of cooperativeness, irrespective of the strategic environment. As a result, the social preference caracterization proves to be effective in explaining observed variation in cooperation levels attained both when cooperation is and is not sustainable as an equilibrium. However, in order to fully uncover the mechanisms

through which subjects react to the environment they face when solving social dilemmas would require a further and more specific analysis of the mechanisms that link behavior, beliefs and social preferences' orientation.

# References

# Bibliography

Andreoni, J. and Miller., J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *Economic Journal*, 103(418):570–85.

Aoyagi, M. and Freéchette., G. R. (2009). Collusion as public monitoring becomes noisy: Experimental evidence. *Journal of Economic Theory*, 144(3):1135–65.

Arechar, A., Kouchaki, M., and Rand, D. (2018). Examining spillovers between long and short repeated prisoner's dilemma games played in the laboratory. *Games*, 9(1):5.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 255(6324):483–485.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126(4):1472–1503.

Blanco, M., Engelmann, D., and Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72(2):321–338.

Blonski, M., Ockenfels, P., and Spagnolo., G. (2011). Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence. *American Economic Journal: Microeconomics*, 3(3):164–92.

Blonski, M. and Spagnolo, G. (2015). Prisoners' other dilemma. *International Journal of Game Theory*, 44(1):61–81.

Bruhin, A., Fehr, E., and Schunk, D. (2018). The many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences. *Journal of the European Economic Association*, 17(4):1025–1069.

Bruttel, L. and Kamecke, U. (2012). Infinity in the lab. how do people play repeated games? *Theory and Decision*, 72(2):205–19.

Capraro, V., Jordan, J. J., and Rand, D. G. (2014). Heuristics guide the implementation of social preferences in one-shot prisoner's dilemma experiments. *Scientific reports*, 4:6790.

Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131:141–150.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Cooper, R., DeJong, D. V., Forsythe, R., and Ross, T. W. (1996). Russell cooper and douglas v. dejong and robert forsythe and thomas w. ross. *Games and Economic Behavior*, 12(2):187–218.

Dal Bó, P. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review*, 95(5):1591–604.

Dal Bó, P., Foster, A., and Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, 100(5):2205–29.

Dal Bó, P. and Fréchette, G. (2011). The evolution of cooperation in infinitely repeated games: Experimental evidence. *American Economic Review*, 101(1):411–29.

Dal Bó, P. and Fréchette, G. (2018). On the determinants of cooperation in infinitely repeated games: A survey. *Journal of Economic Literature*, 56(1):60–114.

Dal Bó, P. and Fréchette, G. (2019). Strategy choice in the infinitely repeated prisoners dilemma. *American Economic Review*. (forthcoming).

Davis, D., Ivanov, A., and Korenok, O. (2016). Individual characteristics and behavior in repeated games: An experimental study. *Experimental Economics*, 19(1):67–99.

Dreber, A., Fudenberg, D., and Rand., D. G. (2014). The role of altruism, inequality aversion, and demo- graphics. *Journal of Economic Behavior and Organization*, 98:41–55.

Dreber, A., Fudenberg, D., Rand., D. G., and Nowak, M. A. (2008). Winners don't punish. *Nature*, 452(7185):348–51.

Duffy, J. and Muñoz-García, F. (2012). Patience or fairness? analyzing social preferences in repeated games. *Games*, 3(1):56–77.

Duffy, J. and Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, 66(2):785–812.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fréchette, G. and Yuksel, S. (2017). Infinitely repeated games in the laboratory: Four perspectives on discounting and random termination. *Experimental Economics*, 20(2):279–308.

Fudenberg, D. and Liang, A. (2019). Predicting and understanding initial play. *American Economic Review.* (forthcoming).

Fudenberg, D. and Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, 54(3):533–54.

Fudenberg, D. and Peysakhovich, A. (2016). Recency, records, and recaps: Learning and nonequilibrium behavior in a simple decision problem. *ACM Transactions on Economics and Computation (TEAC)*, 4(4):23.

Fudenberg, D., Rand, D. G., and Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *American Economic Review*, 102(2):720–49.

Ghidoni, R., Cleave, B. L., and Suetens, S. (2019). Perfect and imperfect strangers in social dilemmas. *European Economic Review*, 1(116):148–59.

Greiner, B. (2004). The online recruitment system orsee 2.0-a guide for the organization of experiments in economics. *University of Cologne, Working paper series in economics*, 10(23):63–104.

Harsanyi, J. C., Selten, R., et al. (1988). A general theory of equilibrium selection in games. *MIT Press Books*, 1.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *New York: Springer.*

John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

KaterinaSherstyuk, Tarui, N., and Saijo., T. (2013). Payment schemes in infinite-horizon experimental games. *Experimental Economics*, 16(1):125–53.

Kim, J. (2019). The effects of time preferences on cooperation: Experimental evidence from infinitely repeated games. *Working Paper.*

Kreps, D. M., Milgrom, P., Roberts, J., and Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2):245–252.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Murnighan, J. K. and Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games: A test of several models. *Journal of Conflict Resolution*, 27(2):279–300.

Naecker, J. (2015). The lives of others: Predicting donations with non-choice responses. *Working Paper.*

Naecker, J. and Peysakhovich, A. (2017). Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity. *Journal of Economic Behavior and Organization*, 133:373–384.

Nay, J. J. and Vorobeychik, Y. (2016). Predicting human cooperation. *PLoS ONE*, 11(5).

Peysakhovich, A., Nowak, M. A., and Rand, D. G. (2014). Humans display a 'cooperative phenotype'that is domain general and temporally stable. *Nature communications*, 5:4939.

Peysakhovich, A. and Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647.

Proto, E., Rustichini, A., and Sofianos., A. (2019). Intelligence, personality, and gains from cooperation in repeated interactions. *Journal of Political Economy*, 127(3):1351–1390.

Reuben, E. and Suetens, S. (2012). Revisiting strategic versus non-strategic cooperation. *Experimental Economics*, 15(1):24–43.

Romero, J. and Rosokha, Y. (2018). The evolution of cooperation: The role of costly strategy adjustments. *American Economic Journal: Microeconomics*, 11(1):299–328.

Roth, A. E. and Murnighan, J. K. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, 17(2):189–98.

Sabater-Grande, G. and Georgantzis., N. (2002). Accounting for risk aversion in repeated prisoners' dilemma games: An experimental test. *Journal of Economic Behavior and Organization*, 48(1):37–50.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., and Peters., E. (2013). Development and testing of an abbreviated numeracy scale: A rasch analysis approach. *Journal of Behavioral Decision Making*, 26(2):198–212.

Y. Zhao, Y. and Cen, Y. (2014). Data mining applications with r. *Oxford Academic Press: Elsevier.*

## 3.8 Appendix

### A.1 Theoretical Framework: Modeling social preferences à la Fehr and Schmidt (1999)

When we model social preferences using the original model by Fehr and Schmidt (1999) both concerns for advantageous and disadvantageous inequality are considered. In this case, focusing on a two-players case with complete information, the utility function of individual $i$ in the pair would be given by:

$$U_i(A_i, A_j) = f(\pi_i(A_i, A_j), \pi_j(A_i, A_j)) = \pi_i - a_i max\{\pi_j - \pi_i; 0\} - b_i max\{\pi_i - \pi_j; 0\}$$

such that the utility of individual $i$ can also be expressed as:

$$\begin{cases} if\ \pi_i = \pi_j \longrightarrow U_i = \pi_i \\ if\ \pi_i > \pi_j \longrightarrow U_i = \pi_i - b_i(\pi_i - \pi_j) \\ if\ \pi_i < \pi_j \longrightarrow U_i = \pi_i - a_i(\pi_j - \pi_i) \end{cases}$$

In this framework, $a_i$ and $b_i$ measure weights attached to differences in payoffs within the pair both in situations of advantageous and disadvantageous inequality, where it is usually further assumed that $a_i \geq b_i$ and $1 > b_i \geq 0$, which implies that the weight assigned to the envy-driven dis-utility from having the lowest payoff in the pair is higher than the weight assigned to the guilt-driven dis-utility from having the highest payoff in the pair.

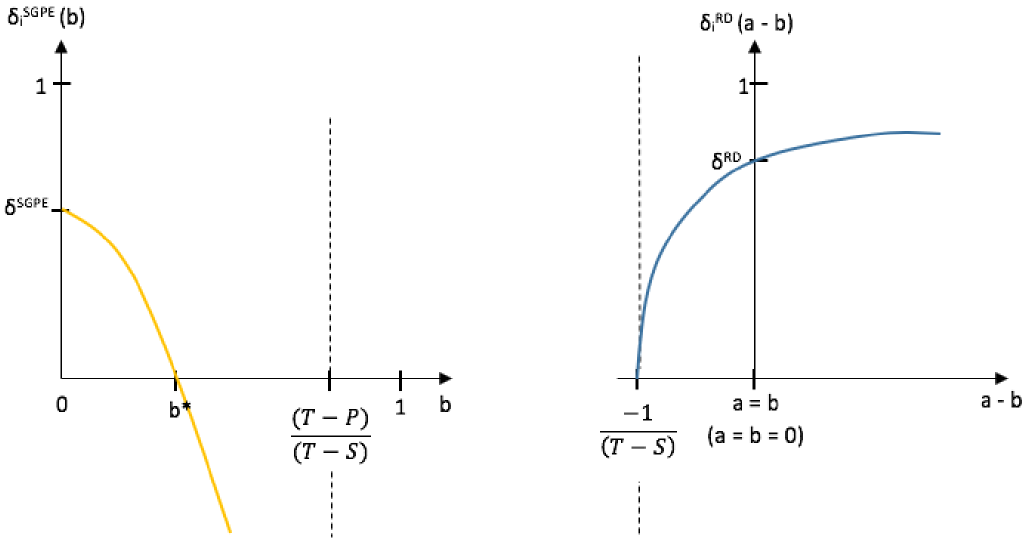
Using a different utility function to map payoffs into utilities alters the reformulation of the stage-game matrix in terms of players' utilities shown in 3.2, as shown in 3.22.

Table 3.22: Prisoners' Dilemma Row Player's Utilities - F&S

| General | | | Social Preferences à la Fehr and Schmidt (1999) | | |
|---|---|---|---|---|---|
| | C | D | | C | D |
| C | $U_i(C,C)$ | $U_i(C,D)$ | C | $U_i(C,C) = R$ | $U_i(C,D) = S - a_i(T-S)$ |
| D | $U_i(D,C)$ | $U_i(D,D)$ | D | $U_i(D,C) = T - b_i(T-S)$ | $U_i(C,C) = P$ |

In this context, if players have perfect information about social preferences, as shown by Duffy

Figure 3.15: $\delta_i^{SGPE}$ and $\delta_i^{RD}$ as a function of $b_i$ and $a_i - b_i$



and Muñoz-García (2012), we have that for no or low levels of social preferences ($b_i = 0$ or $b_i \leq b_i^* = \frac{(T-R)}{(T-S)}$), the only Nash Equilibrium (NE) of the stage game if (Defect, Defect), but if both players have strong enough social preferences ($b_i > b_i^* = \frac{(T-R)}{(T-S)}$), then multiple NE - (Cooperate,Cooperate), (Defect, Defect) and a mixed strategy equilibrium - emerge. Accordingly, once we move to the infinitely repeated version of the game, the presence of social preferences allows cooperation to arise as a:

- Subgame Perfect Nash equilibrium,

  whenever $\delta_i > \delta_i^{SGPE} = \frac{(T-R)-b_i(T-S)}{(T-P)-b_i(T-S)}$

  where $\delta_i^{SGPE} : \sum_{t=0}^{\infty} \delta^t R > T - b_i(T-S) + \sum_{t=1}^{\infty} \delta^t P$

- Risk Dominant equilibrium,

  whenever $\delta_i > \delta_i^{RD} = \frac{(P-R)+(T-S)(1+a_i-b_i)}{(T-S)(1+a_i-b_i)}$

  where, $\delta_i^{RD}$ :

  $\frac{1}{2}\left[\sum_{t=0}^{\infty} \delta^t R\right] + \frac{1}{2}\left[S - a_i(T-S) + \sum_{t=1}^{\infty} \delta^t P\right] > \frac{1}{2}\left[T - b_i(T-S) + \sum_{t=1}^{\infty} \delta^t P\right] + \frac{1}{2}\left[\sum_{t=0}^{\infty} \delta^t P\right]$

In both cases, where $a_i = b_i = 0$, $\delta_i^{SGPE}$ and $\delta_i^{RD}$ coincide with threshold values $\delta^{SGPE}$ and $\delta^{RD}$, while for high enough values of $b_i$, $\delta_i^{SGPE} < \delta^{SGPE}$ and $\delta_i^{RD} < \delta^{RD}$ for $a_i - b_i < 0$ and $\delta_i^{RD} > \delta^{RD}$ for $a_i - b_i > 0$, being both $\delta_i^{SGPE}$ and $\delta_i^{RD}$ decreasing in $b_i$.

## A.2 Meta-Analysis: Dataset

In their paper, Dal Bó and Fréchette (2018) perform a regression analysis to test the ability of the two indices $\delta - \delta_{RD}$ and $sizeBAD$ to describe the levels of cooperativeness observed in their data by relying on a simple Probit model: they study the effect of the two indices on 1st Round Cooperation, separately looking at 1st Round Cooperation in Supergames 7 and 15 (see Table A1.1, a replication of Table 8 in Dal Bó and Fréchette (2018)).

Table A1.1: The Impact of the Indices on 1R Cooperation - Marginal Effects at the average (DF2018)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Sup.7 | Sup.15 | Sup.7 | Sup.15 | Sup.7 | Sup.15 |
| SGPE | -0.0986 | 0.195 | | | | |
| | (0.145) | (0.136) | | | | |
| $(\delta - \delta_{SGPE})$ x SGPE | 0.747*** | 0.979*** | | | | |
| | (0.0780) | (0.0733) | | | | |
| $(\delta - \delta_{SGPE})$ x Not SGPE | 0.566** | -0.349 | | | | |
| | (0.282) | (0.275) | | | | |
| RD | | | 0.113** | 0.121*** | 0.225 | 0.420* |
| | | | (0.0451) | (0.0415) | (0.220) | (0.243) |
| $(\delta - \delta_{RD})$ x RD | | | 1.030*** | 1.677*** | | |
| | | | (0.129) | (0.178) | | |
| $(\delta - \delta_{RD})$ x Not RD | | | 0.238*** | 0.235 | | |
| | | | (0.0574) | (0.273) | | |
| $sizeBAD$ x RD | | | | | -0.902*** | -1.139*** |
| | | | | | (0.326) | (0.372) |
| $sizeBAD$ x Not RD | | | | | -0.429* | -0.342 |
| | | | | | (0.229) | (0.368) |
| Observations | 2,305 | 1,030 | 2,305 | 1,030 | 2,305 | 1,030 |

*Notes:* Clustered standard errors in parentheses.

*** p<0.01, ** p<0.05, * p<0.1

We replicate the same analysis on our metadata, to show that the qualitatively results on the two indices on 1st Round Cooperation are unchanged (see Table A1.2).

Table A1.2: The Impact of the Indices on 1R Cooperation - Marginal Effects at the average (Metadata)

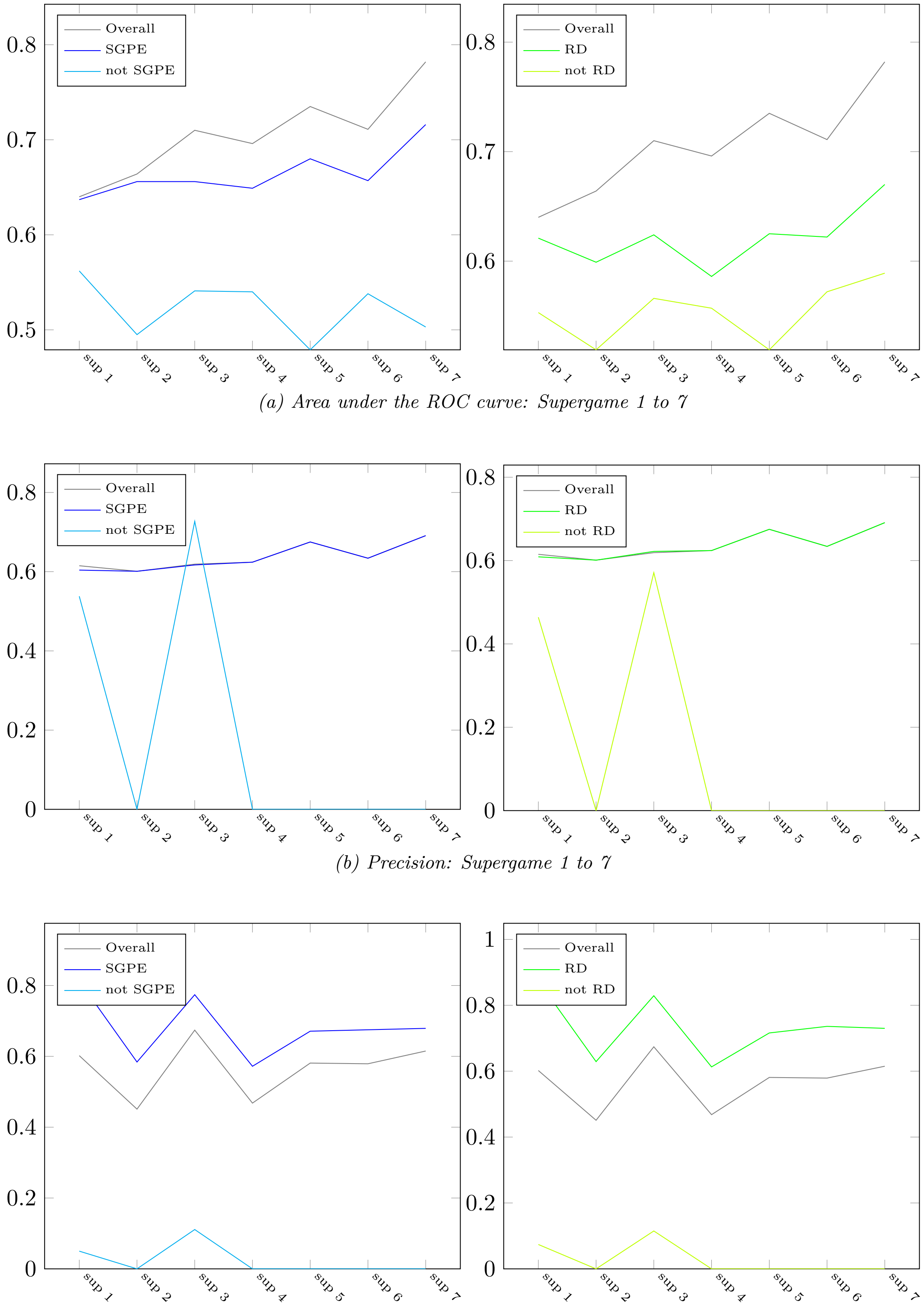
| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Sup.7 | Sup.15 | Sup.7 | Sup.15 | Sup.7 | Sup.15 |
| SGPE | -0.0212 | 0.218** | | | | |
| | (0.116) | (0.103) | | | | |
| $(\delta - \delta_{SGPE})$ x SGPE | 0.591*** | 0.723*** | | | | |
| | (0.121) | (0.156) | | | | |
| $(\delta - \delta_{SGPE})$ x Not SGPE | 0.519** | -0.350 | | | | |
| | (0.235) | (0.240) | | | | |
| RD | | | 0.134** | 0.142*** | 0.240 | 0.325* |
| | | | (0.0536) | (0.0384) | (0.195) | (0.185) |
| $(\delta - \delta_{RD})$ x RD | | | 0.844*** | 1.038*** | | |
| | | | (0.0999) | (0.182) | | |
| $(\delta - \delta_{RD})$ x Not RD | | | 0.249*** | 0.305 | | |
| | | | (0.0589) | (0.197) | | |
| $sizeBAD$ x RD | | | | | -1.085*** | -1.143*** |
| | | | | | (0.253) | (0.173) |
| $sizeBAD$ x Not RD | | | | | -0.463** | -0.428* |
| | | | | | (0.217) | (0.250) |
| Observations | 3,157 | 1,616 | 3,157 | 1,616 | 3,157 | 1,616 |

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

## A.3 Meta-Analysis: Additional Figures

Figure A2.1: Classification Accuracy metrics over Supergames 1 to 7 (*sizeBAD* model)

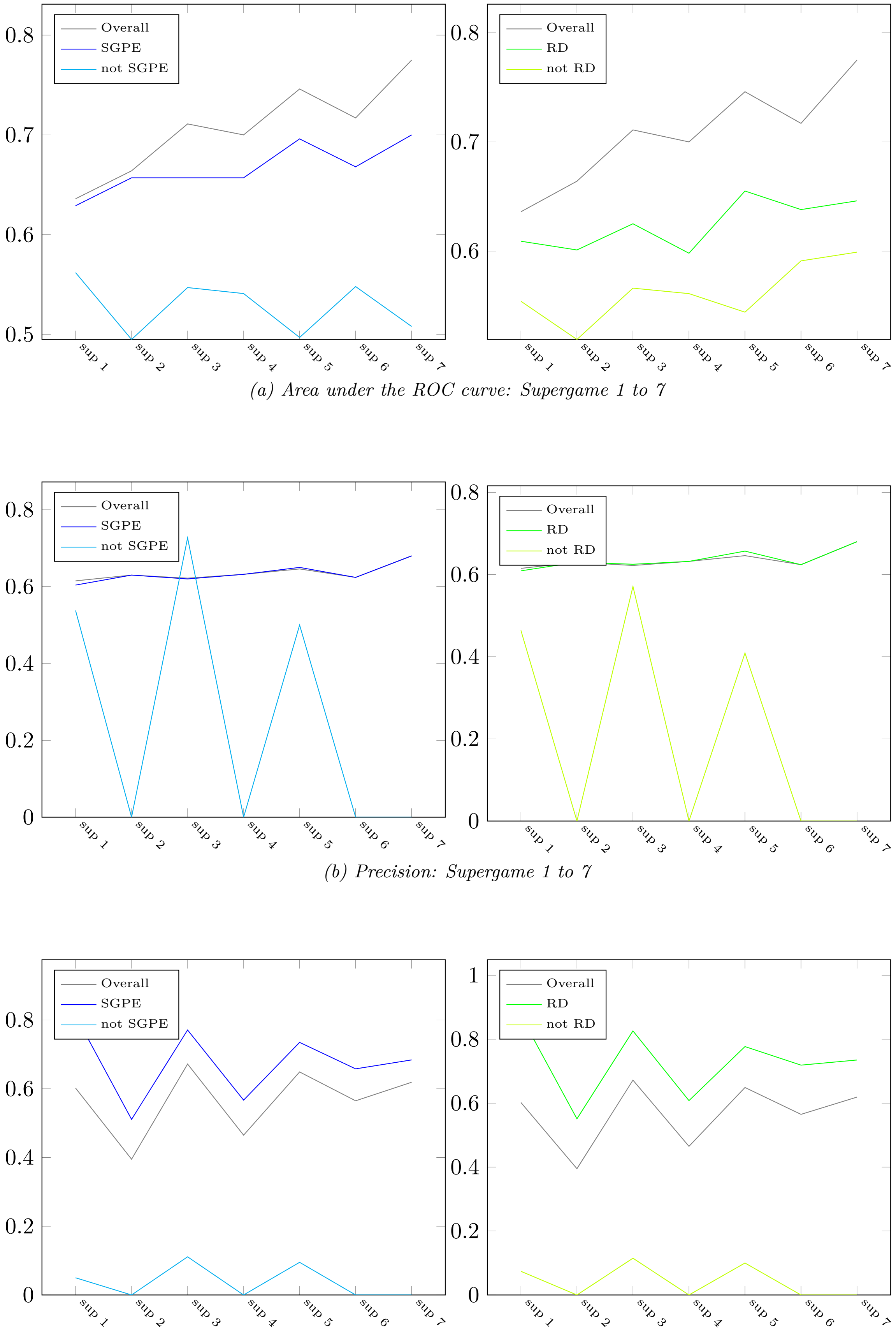


*(a) Area under the ROC curve: Supergame 1 to 7*

*(b) Precision: Supergame 1 to 7*

*(c) Recall: Supergame 1 to 7*

Figure A2.2: Classification Accuracy metrics over Supergames 1 to 7 (Unconstrained model)



*(a) Area under the ROC curve: Supergame 1 to 7*

*(b) Precision: Supergame 1 to 7*

*(c) Recall: Supergame 1 to 7*

## A.4 Experimental Design: Comparing results with and w/o reciprocity in Bruhin et al. (2018)

We evaluate how sensitive the results of the paper are to the inclusion/exclusion of reciprocity concerns in the social preferences model. We compare the estimates reported in the paper, obtained on the whole sample, which includes observations from both the Dictator Games (DG) and the Reciprocity Games (RG) for all individuals, with estimates obtained estimating the constrained model (where $\gamma = 0$ and $\eta = 0$) only on the observations collected from the Dictator Game.

$$U_i = (1 - \alpha s - \beta r - \gamma q - \eta v) \cdot \Pi^i + (\alpha s + \beta r + \gamma q + \eta v) \cdot \Pi^j$$

where:
$s = 1$ if $\Pi^i < \Pi^j$, and $s = 0$ otherwise (disadvantageous inequality);
$r = 1$ if $\Pi^i > \Pi^j$, and $s = 0$ otherwise (advantageous inequality);
$q = 1$ if player $j$ behaved kindly toward $i$, and $q = 0$ otherwise (positive reciprocity);
$v = 1$ if player $j$ behaved unkindly toward $i$, and $v = 0$ otherwise (negative reciprocity);

In the first case, we rely on all the information from the 117 binary decisions taken from the subjects throughout the experiment (117 x 174 = 20358 observations) , in the second case, we restrict our attention to the 39 binary DG decisions (39 x 174 = 6786 observations). We evaluate whether:

A) the estimates of the parameters derived from the aggregate model (n. types K=1) are substantially different

B) the summmary statistics of the individual estimates of the parameters (n. types K=174) are substantially different

C the estimates of the parameters derived from the finite mixture model (n. types K=3) are substantially stable across types

We run this analysis both for observations from Session 1 and Session 2 of the Experiment. We further investigate, for estimates obtained from Session 2 data, whether:

D) the ability to explain the variability in subjects' subsequent choices in the Trust Games (TG) and the Reward and Punishment Games (RPG) using linear models augmented with predictions based on finite-mixture and inidividual model estimates, is substantially different

In the paper, a McFadden's (1981) random utility model for discrete choices is used to estimate the social preference parameters of the behavioral model $\theta = (\alpha,\ \beta,\ \gamma,\ \eta)$. The underlying assumption is that the utility player $i$ gets from choosing the allocation $X_g = (\Pi^i_{X_g},\ \Pi^j_{X_g},\ r_{X_g},\ s_{X_g},\ q_{X_g},\ v_{X_g})$ in game $g = 1, .., G$, within the set of possible allocations $\{X_g,\ Y_g\}$ is given by:

$$U^i(X_g; \theta, \sigma) = U^i(X_g; \theta) + \epsilon_{X_g}$$

where $U^i(X_g; \theta)$ is the deterministic component of the utility deriving from allocation $X_g$ and $\epsilon_{X_g}$ is a random component representing noise in the utility evaluation: the random component $\epsilon_{X_g}$ is assumed to follow a type 1 extreme value distribution with a scale parameter $\frac{1}{\sigma}$. Under this framework, player $i$ would choose allocation $X_g$ over the allocation $Y_g$ whenever $U^i(X_g; \theta, \sigma) \geq U^i(Y_g; \theta, \sigma)$, so that the probability that the choice of player $i$ in game $g$ - $C_g$ - equals $X_g$ is given by:

$$Pr(C_g = X_g;\ \theta,\ \sigma,\ X_g,\ Y_g)$$

$$= Pr(U^i(X_g; \theta) - U^i(Y_g; \theta) \geq \epsilon_{Y_g} - \epsilon_{X_g}$$

$$= \frac{exp(\sigma U^i(X_g; \theta))}{exp(\sigma U^i(X_g; \theta)) + exp(\sigma U^i(Y_g; \theta))}$$

where the parameter $\sigma$ measures choice sensitivity to differences in deterministic utilities, so that when $\sigma = 0$ player $i$ chooses each of the two options with the same 0.5 probability irrespective of the deterministic utility associated to the two options, while when $\sigma$ is arbitrarily large the probability of choosing the most appealing option in terms of deterministic utility approaches 1.

The first approach estimates the random utility model by pooling the data to obtain aggregate estimates of the parameters $(\hat{\theta},\ \hat{\sigma})$. These aggregate estimates represent the most parsimonious characterization of social preferences, where all players are assumed to belong to the same 'representative' type (n. types = 1).

At the opposite extreme, the individual estimates are obtained by separately estimating the parameters of the social preferences model for each individual $(\hat{\theta}_i,\ \hat{\sigma_i})$. This approach is the least parsimonious, and likely to suffer from small sample bias, but is able to fully uncover the behavioral heterogeneity in the data (n. types = N).

The intermediate approach in terms of flexibility and parsimony is represented by the finite mixture model, where the population is assumed to be characterized by a finite number of K distinct preference types, each characterized by a different set of parameters $(\hat{\theta_k},\ \hat{\sigma_k})$. This approach acknowledges latent heterogeneity in the data, although individual type-membership is not directly observable. In this context, the estimation leads to a parsimonious characterization of the K types in the population, providing a set of type-specific preference parameters and types' shares in the population $\hat{\pi_k}$. In the paper, the optimal number of types is fixed to K=3.

Models reported in the following pages are estimated on observations from the full set of 174 players. For individual estimates, summary statistics are reported on the sample of 160 players whose estimated parameters are not classified as erratic, based on the estimates obtained in the paper.

**Data from Session 1**

A) Estimates from the aggregate model

|  | DG Sample |  | Paper Sample |
|---|---|---|---|
| $\hat{\alpha}$ | 0.0628*** | $\hat{\alpha}$ | 0.0835*** |
|  | (0.016) |  | (0.015) |
| $\hat{\beta}$ | 0.279*** | $\hat{\beta}$ | 0.261*** |
|  | (0.021) |  | (0.019) |
| $\hat{\sigma}$ | 0.016*** | $\hat{\sigma}$ | 0.016*** |
|  | (0.001) |  | (0.001) |
|  |  | $\hat{\gamma}$ | 0.072*** |
|  |  |  | (0.014) |
|  |  | $\hat{\eta}$ | -0.042*** |
|  |  |  | (0.011) |

B) Summary of individual estimates

|  | DG Sample |  |  |  |  | Paper Sample |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  | MIN | MAX | MED | MEAN |  | MIN | MAX | MED | MEAN |
| $\hat{\alpha}_i$ | -13.783 | 0.459 | 0.052 | -0.101 | $\hat{\alpha}_i$ | -1.394 | 0.471 | 0.053 | 0.017 |
| $\hat{\beta}_i$ | -11.336 | 1.085 | 0.197 | 0.1669 | $\hat{\beta}_i$ | -1.977 | 0.998 | 0.211 | 0.216 |
| $\hat{\gamma}_i$ | - | - | - | - | $\hat{\gamma}_i$ | -0.366 | 0.783 | 0.042 | 0.0836 |
| $\hat{\eta}_i$ | - | - | - | - | $\hat{\eta}_i$ | -1.106 | 0.598 | -0.008 | 0.055 |
| $\hat{\sigma}_i$ | 0.000 | 0.804 | 0.0599 | 0.2915 | $\hat{\sigma}_i$ | 0.004 | 0.858 | 0.035 | 0.174 |

C) Estimates from the Finite Mixture model

| DG Sample | | | | Paper Sample | | | |
|---|---|---|---|---|---|---|---|
| | BA | MA | SA | | BA | MA | SA |
| $\hat{\pi}_k$ | 0.179*** | 0.352*** | 0.469*** | $\hat{\pi}_k$ | 0.121*** | 0.474*** | 0.405*** |
| $\hat{\alpha}_k$ | -0.38*** | 0.044*** | 0.149*** | $\hat{\alpha}_k$ | -0.435*** | 0.065*** | 0.159*** |
| $\hat{\beta}_k$ | -0.016 | 0.077*** | 0.482*** | $\hat{\beta}_k$ | -0.145 | 0.129*** | 0.463*** |
| $\hat{\gamma}_k$ | - | - | - | $\hat{\gamma}_k$ | 0.17 | -0.001 | 0.151*** |
| $\hat{\eta}_k$ | - | - | - | $\hat{\eta}_k$ | -0.076 | -0.027** | -0.053*** |
| $\hat{\sigma}_k$ | 0.009*** | 0.066*** | 0.020*** | $\hat{\sigma}_k$ | 0.008*** | 0.0316*** | 0.018*** |

Using estimated posterior probabilities to belong to each type, we get the same classification for $15 + 49 + 55 = 119/160 = 74\%$ of the subjects.

| **Paper Sample** | **DG Sample** | | | |
|---|---|---|---|---|
| | BA | MA | SA | Total |
| BA | 15 | 4 | 0 | 19 |
| MA | 10 | 49 | 17 | 76 |
| SA | 4 | 6 | 55 | 65 |
| Total | 29 | 59 | 72 | 160 |

**Data from Session 2**

A) Estimates from the aggregate model

| DG Sample | | Paper Sample | |
|---|---|---|---|
| $\hat{\alpha}$ | 0.079*** | $\hat{\alpha}$ | 0.098*** |
| | (0.013) | | (0.013) |
| $\hat{\beta}$ | 0.255*** | $\hat{\beta}$ | 0.245*** |
| | (0.020) | | (0.019) |
| $\hat{\sigma}$ | 0.020*** | $\hat{\sigma}$ | 0.019*** |
| | (0.001) | | (0.001) |
| | | $\hat{\gamma}$ | 0.029*** |
| | | | (0.010) |
| | | $\hat{\eta}$ | -0.043*** |
| | | | (0.008) |

B) Summary of individual estimates

| DG Sample | | | | | Paper Sample | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MIN | MAX | MED | MEAN | | MIN | MAX | MED | MEAN |
| $\hat{\alpha}_i$ | -1.240 | 0.449 | 0.052 | 0.035 | $\hat{\alpha}_i$ | -1.636 | 0.399 | 0.060 | 0.048 |
| $\hat{\beta}_i$ | -0.362 | 1.012 | 0.160 | 0.232 | $\hat{\beta}_i$ | -0.405 | 0.905 | 0.169 | 0.225 |
| $\hat{\gamma}_i$ | - | - | - | - | $\hat{\gamma}_i$ | -1.087 | 0.679 | 0.005 | 0.032 |
| $\hat{\eta}_i$ | - | - | - | - | $\hat{\eta}_i$ | -0.0553 | 0.229 | -0.009 | 0.045 |
| $\hat{\sigma}_i$ | 0.007 | 0.929 | 0.471 | 0.389 | $\hat{\sigma}_i$ | 0.007 | 0.886 | 0.069 | 0.275 |

C) Estimates from the Finite Mixture model

| DG Sample | | | | Paper Sample | | | |
|---|---|---|---|---|---|---|---|
| | BA | MA | SA | | BA | MA | SA |
| $\hat{\pi_k}$ | 0.102*** | 0.449*** | 0.449*** | $\hat{\pi_k}$ | 0.100*** | 0.544*** | 0.356*** |
| $\hat{\alpha_k}$ | -0.368*** | 0.042*** | 0.160*** | $\hat{\alpha_k}$ | -0.328*** | 0.061*** | 0.193*** |
| $\hat{\beta_k}$ | -0.047 | 0.072*** | 0.469*** | $\hat{\beta_k}$ | -0.048 | 0.095*** | 0.495*** |
| $\hat{\gamma_k}$ | - | - | - | $\hat{\gamma_k}$ | -0.028 | -0.005 | 0.099*** |
| $\hat{\eta_k}$ | - | - | - | $\hat{\eta_k}$ | -0.015 | -0.019*** | -0.082*** |
| $\hat{\sigma_k}$ | 0.018*** | 0.085*** | 0.023*** | $\hat{\sigma_k}$ | 0.015*** | 0.049*** | 0.019*** |

Using estimated posterior probabilities to belong to each type, we get the same classification for $15 + 67 + 56 = 138/160 = 86\%$ of the subjects.

| **Paper Sample** | **DG Sample** | | | |
|---|---|---|---|---|
| | BA | MA | SA | Total |
| BA | 15 | 0 | 1 | 16 |
| MA | 1 | 67 | 19 | 87 |
| SA | 0 | 1 | 56 | 57 |
| Total | 16 | 68 | 76 | 160 |

Following the same approach as in the paper, we further analyze whether linear models augmented with predictions based on finite-mixture and inidividual model estimates can better explain the variability in the choice made by the same set of subjects in a series of Trust Games (TG) and Reward and Punishment Games (RPG). The models are augmented with, respectively:

- Type-specific prediction on the probability to take the choice of interest
- Individual-specific prediction on the probability to take the choice of interest
- Type-specific prediction on the probability to take the choice of interest *and* Difference between the Individual-specific prediction and the Type-specific prediction ($\Delta_{i-p}$).

The augmented models are compared to a baseline model estimated using only individual-specific characteristics, such as Big 5 personality traits, cognitive ability, age, gender, monthly income, and field of study as explanatory variables.

| **DG Sample** | | | | **Paper Sample** | | | |
|---|---|---|---|---|---|---|---|
| Trust Game | | | | | | | |
| Indiv. charact. | → | $R^2 = 0.0589$ | - | Indiv. charact. | → | $R^2 = 0.0589$ | - |
| Indiv. charact.<br>+ Type pred. | → | $R^2 = 0.3655$ | $\hat{\beta}_{tp} = 0.617^{***}$ | Indiv. charact.<br>+ Type pred. | → | $R^2 = 0.3491$ | $\hat{\beta}_{tp} = 0.607^{***}$ |
| Indiv. charact.<br>+ Indiv. pred. | → | $R^2 = 0.3677$ | $\hat{\beta}_{ip} = 0.606^{***}$ | Indiv. charact.<br>+ Indiv. pred. | → | $R^2 = 0.3457$ | $\hat{\beta}_{ip} = 0.580^{***}$ |
| Indiv. charact.<br>+ Indiv. pred.<br>+ $\Delta_{i-t}$ | → | $R^2 = 0.4040$ | $\hat{\beta}_{tp} = 0.686^{***}$<br>$\hat{\beta}_{\Delta} = 0.344^{***}$ | Indiv. charact.<br>+ Indiv. pred.<br>+ $\Delta_{i-t}$ | → | $R^2 = 0.3748$ | $\hat{\beta}_{tp} = 0.650^{***}$<br>$\hat{\beta}_{\Delta} = 0.309^{***}$ |

| **DG Sample** | | | | **Paper Sample** | | | |
|---|---|---|---|---|---|---|---|
| Reward-Punishment Game | | | | | | | |
| Indiv. charact. | → | $R^2 = 0.0354$ | - | Indiv. charact. | → | $R^2 = 0.0354$ | - |
| Indiv. charact.<br>+ Type pred. | → | $R^2 = 0.2127$ | $\hat{\beta}_{tp} = 0.971^{***}$ | Indiv. charact.<br>+ Type pred. | → | $R^2 = 0.2674$ | $\hat{\beta}_{tp} = 1.123^{***}$ |
| Indiv. charact.<br>+ Indiv. pred. | → | $R^2 = 0.194$ | $\hat{\beta}_{ip} = 0.532^{***}$ | Indiv. charact.<br>+ Indiv. pred. | → | $R^2 = 0.253$ | $\hat{\beta}_{ip} = 0.641^{***}$ |
| Indiv. charact.<br>+ Indiv. pred.<br>+ $\Delta_{i-t}$ | → | $R^2 = 0.229$ | $\hat{\beta}_{tp} = 0.898^{***}$<br>$\hat{\beta}_{\Delta} = -.254^{***}$ | Indiv. charact.<br>+ Indiv. pred.<br>+ $\Delta_{i-t}$ | → | $R^2 = 0.302$ | $\hat{\beta}_{tp} = 1.064^{***}$<br>$\hat{\beta}_{\Delta} = -.353^{***}$ |

## A.5 Experimental Design: Instructions

Translation of the instructions from Italian.

### Part 1

Welcome.

Thanks for your participation in this study!

During this study, you will have the opportunity to earn money. The amount of your earnings will depend on the decisions you and the other participants will make. All decisions will remain completely anonymous.

The study is divided into two parts: a preliminary part and the main part of the study.

Before moving to the main part of the study, you are required to take part in the preliminary part of the study, which will last about 30 minutes. You can take part in the preliminary part of the study whenever you prefer before $*date\_time\_ddl\_part1*$. Remember that you will be paid for your participation in the study ONLY IF you complete the preliminary part before the deadline and log in on time for the main part of the study. Payments will only be calculated at the end of the main part of the study.

This preliminary part of the study is divided into two parts: Part A and Part B. At the beginning of each part you will get the corresponding instructions and we will ask you to answer a few comprehension questions.

- In part A we will ask you to decide how certain monetary payments between you (Person “A") and another specific participant in the experiment (Person “B”) should be distributed.
- In Part B we will ask you to complete a questionnaire.

*Part A: Instructions*

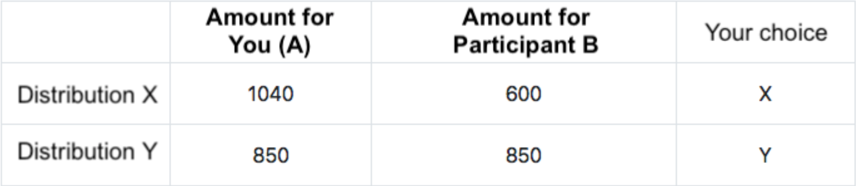
In this part of the study you will have to make 39 decisions that will affect you and another participant, who will be randomly selected among the other participants in this study and will be paired with you in each decision situation. You will never learn who this person is, and the other person will also not learn of your identity. You will no longer interact with this participant for the rest of the study.

In each of the 39 decisions you will have exactly two options: X and Y.

Each option is associated with a monetary amount for you (Player A) and for the other participant paired with you (Player B). With your decision you will determine the distribution of payments between you and the other participant definitively, the other participant has a passive role and cannot change the distribution.

*Please note: We present monetary amounts as points on the computer screen. 100 points are worth 0.4 Euros.*

Payments: You will be paid only for one of the decisions you will make, which will be randomly selected at the end of the entire study. At the end of the study, you will be informed of which decision has been randomly selected for payment and of how much you and the other participant will receive, based on your choices.



| | Amount for You (A) | Amount for Participant B | Your choice |
|---|---|---|---|
| Distribution X | 1040 | 600 | X |
| Distribution Y | 850 | 850 | Y |

The 39 different situations will be presented successively on the screen, like in this example, where the payments associated with each of the two options - X and Y - are shown for you and the other participant: in this case, if you choose X you will receive 1040 points and the other participant 600 points, while if you choose Y you both receive 850 points.
Before we start with Part A, we will ask you to answer some comprehension questions.

*[Part A - Control Questions]*

*Part B: Instructions*

This part consists of a questionnaire. It is important for us that you answer the questions as good as possible.

(1) Demographic data [Year of birth | Gender | Major ]

(2) 44-items Big Five Inventory (John and Srivastava (1999))

(3) 8-items Numeracy test (Weller et al. (2013))

Thank you for your time!
You will receive further instructions and the link to log in for the main part of study in the next days by email. Remember that you will be entitled to receive the show up fee of 5 Euros only if you log in on time to take part in the main part of the study.

**Part 2**

Welcome.

Thanks for your participation in this study.

By logging-in on time, after successfully completing the preliminary part of the study, you have earned the right to receive a payment of minimum 5 Euros.

All the decisions you will make during the study will remain completely anonymous. The final amount of your earnings will depend on the decisions you and the other participants will make during the study.

At the end of today's session, we will inform you of the amount of your earnings based on the decisions that you and the other study participants will make today and made in the preliminary part of the study.

Please shut down all the other programs running on your computer except Zoom, which you should keep open until the very end of the study.

All you will need is a blank sheet of paper. It is important that you do not try to communicate with the other participants during the session.

Today we will ask you to take part in two activities [28].: in both activities you will have to make a series of decisions but at the end of the study you will be actually paid only for one of the decisions you have madein each activity , which will be selected randomly at the end of the study.

*Please note: We present monetary amounts as points on the computer screen. 100 points are worth 1.5 Euros.*

*Activity n.1: Instructions (One-Shot series)*

Task 1 consists of 10 rounds. In each round you will interact with a counterpart and you will be asked to make a decision.

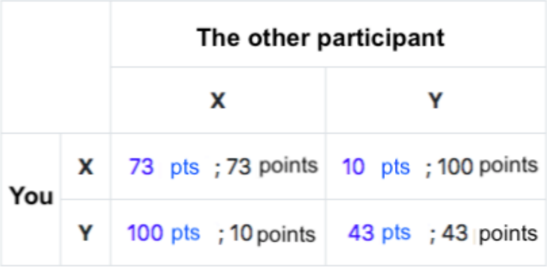
Before the first round starts, you will be paired with three other participants to this study, who will be the members of your group for the entire duration of activity n.1. At the begin-

[28] The order of Activities n. 1 and n. 2 was randomized across sessions.

ning of each interaction round, you will be paired with another participant, randomly selected among your group mates.

You will not learn the identity of the participant paired with you and he will not learn about yours. Over the ten rounds you may be re-paired with a participant with whom you have already interacted in one of the previous rounds but you will not be able to identify when this may happen.

In each round, you and the other participant will have two possible choices: X and Y. Each cell shows the amount of your earnings in points (left, in blue) and that of the other participant (right, in black).



| | | The other participant | |
|---|---|---|---|
| | | X | Y |
| You | X | 73 pts ; 73 points | 10 pts ; 100 points |
| | Y | 100 pts ; 10 points | 43 pts ; 43 points |

Before rounds 1, 5 and 10 begin, we will also ask you to guess what number of participants in your group who will choose the 'X' option. If, at the end of the study, one of these rounds is randomly selected for payment and your conjecture proves correct, you will receive an additional fixed payment of 2 Euros.

Before we start with Activity n.1, we will ask you to answer some comprehension questions.

*[Activity n.1 - Control Questions]*

*Activity n.2: Instructions (Infinitely Repeated series)*

Activity 2 consists of 10 matches. Each match consists of a variable number of rounds.

Before the first match starts, you will be paired with three other participants to this study, who will be the members of your group for the entire duration of activity n.2. None of these participants interacted with you during activity n.1.

Each match can consist of one or more rounds of interaction. After each round of each match, we will randomly draw a number within the interval [1,100].

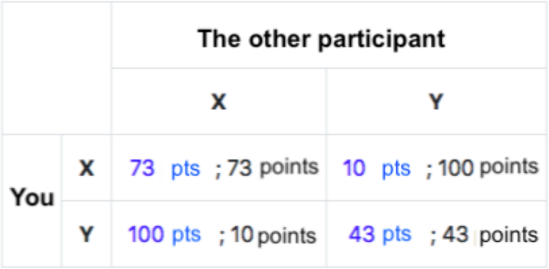
> If the number drawn is <= [$\delta$ : *continuation probability*], the match continues for another round.
>
> If the drawn number is> [$\delta$ : *continuation probability*], the match ends.

The duration of each match is therefore determined randomly and there is a probability of [$\delta$ : *continuation probability*]% that the match continues for another round.

At the beginning of each match, you will be paired with a partner, randomly selected among your group mates. You will interact with the same partner for the entire duration of the match. Over the ten matches you may be re-paired with a participant with whom you have already interacted in one of the previous matches but you will not be able to identify when this may happen.

In each round of each match, you and the other participant will have two possible choices: X and Y.

Each cell shows the amount of your earnings in points (left, in blue) and that of the other participant (right, in black).



| | | The other participant | |
|---|---|---|---|
| | | X | Y |
| You | X | 73 pts ; 73 points | 10 pts ; 100 points |
| | Y | 100 pts ; 10 points | 43 pts ; 43 points |

You will only be paid for the decisions you will make in one of the matches, which will be randomly selected at the end of the study. Your earnings will be equal to your overall earnings, which correspond to the sum of the earnings you have realized through all the rounds of the selected match.

Before matches 1, 5 and 10 begin, we will also ask you to guess what number of participants in your group who will choose the option 'X' in the first round of that match. If, at the end of the study, one of these matches is randomly selected for payment and your conjecture on the first round proves correct, you will receive an additional fixed payment of 2 Euros.

Before we start with Activity n.2, we will ask you to answer some comprehension questions.

*[Activity n.2 - Control Questions]*

## A.6 Results: Additional descriptive statistics

### Mean Round1-Cooperation over supergames by OR types

Table A2.1: Mean of Round1-Cooperation over supergames - Infinitely Repeated series

| | SA = 0 | SA = 0 | SA = 0 | SA = 1 | SA = 1 | SA = 1 |
|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T1 | T2 | T3 |
| | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) |
| Sup: 1 | .3877551 | .5636364 | .4035088 | .6578947 | .6388889 | .7666667 |
| | .4922875 | .5005048 | .4949621 | .4807829 | .4871361 | .4301831 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 2 | .3469388 | .4909091 | .3508772 | .5526316 | .5555556 | .7666667 |
| | .4809288 | .504525 | .4814868 | .5038966 | .5039526 | .4301831 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 3 | .244898 | .4181818 | .4561404 | .5263158 | .5277778 | .8333333 |
| | .434483 | .4978066 | .5025 | .5060094 | .5063094 | .379049 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 4 | .244898 | .4363636 | .4035088 | .5 | .4444444 | .9 |
| | .434483 | .5005048 | .4949621 | .5067117 | .5039526 | .3051286 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 5 | .2857143 | .4727273 | .5087719 | .5526316 | .5 | .9 |
| | .4564355 | .5038572 | .5043669 | .5038966 | .5070926 | .3051286 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 6 | .2040816 | .3272727 | .3859649 | .4736842 | .5 | .7333333 |
| | .4072055 | .4735424 | .4911497 | .5060094 | .5070926 | .4497764 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 7 | .2040816 | .3636364 | .3859649 | .3157895 | .5277778 | .8333333 |
| | .4072055 | .4854794 | .4911497 | .4710691 | .5063094 | .379049 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 8 | .2857143 | .3454545 | .3684211 | .3947368 | .5 | .7333333 |
| | .4564355 | .479899 | .4866643 | .4953554 | .5070926 | .4497764 |

Table A2.2: Mean of Round1-Cooperation over supergames - One Shot series

| | SA = 0 | SA = 0 | SA = 0 | SA = 1 | SA = 1 | SA = 1 |
|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T1 | T2 | T3 |
| | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) | (mean/sd/N) |
| Sup: 1 | .5510204 | .5636364 | .5263158 | .7631579 | .7777778 | .9666667 |
| | .5025445 | .5005048 | .5037454 | .4308515 | .421637 | .1825742 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 2 | .4693878 | .5272727 | .4385965 | .6315789 | .6944444 | .9333333 |
| | .5042338 | .5038572 | .5006262 | .4888515 | .4671766 | .2537081 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 3 | .244898 | .4181818 | .3508772 | .7368421 | .6111111 | .7 |
| | .434483 | .4978066 | .4814868 | .4462583 | .4944132 | .4660916 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 4 | .2653061 | .3454545 | .3333333 | .4473684 | .5 | .6333333 |
| | .4460713 | .479899 | .4755949 | .5038966 | .5070926 | .4901325 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 5 | .3673469 | .5090909 | .3157895 | .6842105 | .5833333 | .7666667 |
| | .4870779 | .504525 | .4689614 | .4710691 | .5 | .4301831 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 6 | .2653061 | .4545455 | .1578947 | .5526316 | .4722222 | .5666667 |
| | .4460713 | .5025189 | .3678836 | .5038966 | .5063094 | .5040069 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 7 | .2653061 | .2545455 | .1403509 | .3947368 | .4444444 | .4333333 |
| | .4460713 | .4396203 | .3504383 | .4953554 | .5039526 | .5040069 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 8 | .2653061 | .2909091 | .1052632 | .4210526 | .3888889 | .4333333 |
| | .4460713 | .4583678 | .3096202 | .5003555 | .4944132 | .5040069 |
| | 49 | 55 | 57 | 38 | 36 | 30 |
| Sup: 9 | .1632653 | .2727273 | .122807 | .4210526 | .3333333 | .3666667 |
| | .3734378 | .4494666 | .3311331 | .5003555 | .4780914 | .4901325 |