The Effects of Revealing Borrowers' Information on Credit Allocation, Defaults and Entrepreneurship

Roberto Hsu Rocha* Javier Feinmann Mariana Mercucci Josival Leite

UC Berkeley UC Berkeley USP SERASA

October 13, 2024

Click Here for the Most Updated Version

Abstract

This paper investigates the effects of increasing the information that lenders have about borrowers on their credit access, default rates, and entrepreneurial activity. Partnering with the largest credit bureau in Brazil, we explore a unique policy that increased the information available for credit bureaus to construct credit scores. Changes in the credit allocation are consistent with a simple conceptual framework of decision-making under imperfect information that rationalizes how new credit scores built with more information should affect credit allocation. This includes not only increasing credit for those who are suggested to be more creditworthy given the new information but also reflects the increase in the precision of the signal, widening the credit distribution. We show that credit allocated by the policy has substantially lower default rates than the counterfactual credit without the policy, but reallocation increased credit inequality, also causing an increase in the racial gap in credit access. We then explore effects beyond the credit market by investigating if the policy affected entrepreneurial activity. We find no effect on firm creation but evidence of better (worse) outcomes of firms created by positively (negatively) exposed individuals. These effects translate into an improvement in the average firm quality of new cohorts of entrepreneurs, suggesting productivity gains from the reallocation of credit.

*Write Thanks

1 Introduction

The consequences of imperfect information in credit markets have long been of interest to economists (Stiglitz and Weiss, 1981; De Meza and Webb, 1987). The development of credit reporting systems is one of the most common policies to reduce the asymmetries between lenders and borrowers. The rationale behind them is that the previous credit history of a potential borrower serves as a signal to lenders about their ability to repay loans. However, the depth of information available in such systems varies widely worldwide, with potentially important consequences not only on credit markets' efficiency but also on the allocation of resources to productive activities (Djankov et al., 2007). Despite their relevance, not much is known about how revealing information about borrowers through improvements in credit reporting systems affects credit allocation and efficiency, nor if it has consequences beyond credit markets.

In this paper, we study the consequences of revealing information about borrowers to lenders through improvements in credit reporting systems. We explore a unique policy in Brazil called *Cadastro Positivo*, that changed the information available for credit bureaus to construct credit scores from a delinquency registry to a complete registry of credit history. Partnering with the biggest credit bureau in the country, we access individual-level information on credit scores, credit access, and financial delinquency. We outline a conceptual framework that rationalizes changes in credit allocation when creditworthiness signals are built with more information. Our framework guides us through empirical analysis of how credit scores built with more information affect individuals' credit access and default rates. We then go beyond credit markets and study if credit reallocation had effects on the entrepreneurial activity of individuals. We pair the credit bureau data with self-collected firm records that comprise the universe of new firms in the biggest state of the country and investigate if changes in personal credit affect both firm creation and the outcomes of these new firms, discussing potential important implications on average firm quality and the allocation of resources in productive activities.

Our paper starts by rationalizing how the increase in lenders' information about borrowers changes their credit access. In our conceptual framework, a lender faces a decision problem in which they choose their credit supply to potential borrowers who differ in their true creditworthiness. The lender does not observe the borrowers' true type and makes decisions based on beliefs formed by observing an unbiased signal.

We propose a policy that generates a more informative signal. Our framework *rationalizes* changes in credit over the joint distribution of signals with and without new information.

They occur due to two mechanisms: a potential change in the signal value, which occurs if the new signal indicates that a given individual is more (less) creditworthy, and a change in the precision of the signal. The latter establishes that a reduction in the signal's noise increases how much weight the lender puts on the signal versus their original prior. This implies that, for similar changes in the signal's value, changes in the expected value of the individuals' true type are increasing in the value of the signals. Under small assumptions on how lenders choose optimal credit to borrowers given their signal, our framework rationalizes how credit access should change over the joint distributions of less and more informative signals, giving us testable implications for our empirical analysis.

We turn to empirically identifying if the additional information changed individuals' credit access. To do so, we explore credit scores constructed with the information made available by the *Cadastro Positivo* policy, henceforth referred to as the *new system of credit scores*, and credit scores constructed without the new information, which we refer to as the *old system of credit scores*.

With the availability of new information to assess creditworthiness, an individual's credit score can potentially increase or decrease. We test if changes in the value of credit scores affect individuals' credit access. To do so, we estimate difference in differences models comparing individuals who increased or decreased their credit score with those whose credit score did not change with the introduction of the new system. We find that changes in the value of credit scores have significant effects on an individual's credit access. One standard deviation increase in credit scores causes, on average, a 15 thousand Brazilian Reais (BRL) increase in credit 2 years after the policy, representing a 20% increase in their credit. Conversely, one standard deviation decrease causes a decrease of around 8 thousand BRL, also representing a 20% decrease in their credit access relative to the group's counterfactual credit without the policy.

Credit scores built with additional information should also be more precise in their creditworthiness assessment, and a rational lender would consider this in their decision-making. Lenders should weigh more the signal relative to their prior if it is built with more information. This would increase (decrease) the lenders' belief of the true creditworthiness of individuals if they have a signal substantially larger (smaller) than the lenders prior, even if they do not have any change in the value of their signal. To empirically test this, we use a sample of individuals who did not have substantial changes in their credit scores with the new system. We then compare those with credit scores further from the average values with those with average credit scores. We find that individuals with one standard deviation higher than the population average observe an increase of 5 thousand BRL in their

credit access two years after the policy. In contrast, those with credit scores one standard deviation below the population mean show a decrease of around 12 thousand BRL.

The two analyses are useful for identifying the drivers of changes in credit, but they are limited to specific parts of the joint distribution of credit scores. We build upon them by mapping how credit changes over the full joint distribution of credit scores. This allows us to test the predictions of our conceptual framework and approximate numerical values for the effects of increasing precision and changes in signal value.

We first employ a semi-parametric approach, which consists of dividing the joint distribution of credit scores into a grid of equally sized bins. We then estimate a difference in differences model that gives us the change in credit from a given bin before and after the policy, relative to the group with no change in credit scores and was at the population average. According to our framework, this latter group has zero effects on both the effects of signal's value and the effects of signal's precision. We complement our semi-parametric approach, estimating changes in credit over the joint distribution parametrically, assuming a linear relationship between credit scores and changes in credit. This allows us to estimate the average effects of increasing precision and changes in signal value.

Credit allocation changes are consistent with the signal value change and the increase in signal precision over the joint distribution of credit scores in the new and old systems. For fixed values of the old system credit scores, we observe that the effects of the policy on credit increase monotonically in the new credit score values. We also observe that when comparing changes in credit between groups with the same difference in credit scores between the new and the old systems, the value of credit changes increases in the value of credit scores, consistent with the effects of the precision of the signal. A linear approximation to changes in credit over the joint distribution estimates the effects of the signal value at 11 thousand BRL and the effects of the change in the signal's precision at 5 thousand BRL.

Revealing information increases the inequality in credit access. Using our estimates of how credit changes over the joint distribution of credit scores, we show that changes in credit access are positively correlated with existing credit. We estimate that the variance of credit distribution increases by 2.5% because of the policy. We also show that information revelation increases racial differences in credit access in our context. This is explained largely because of compositional effects, i.e., nonwhite individuals are disproportionately represented in parts of the distribution of credit scores that were negatively affected in terms of credit access.

After characterizing how credit allocation changes with the revelation of information, we investigate whether revealing information reallocated credit to more or less risky credit.

Since our data does not allow us to observe loan-specific defaults, we approximate default rates by calculating the ratio between an individual's total financial delinquency and their total amount of credit.

We replicate our empirical strategy described above, looking at financial delinquency as the outcome. Our estimates show that the total value of financial delinquency, on average, increases for the groups that had credit increases and decreases for those with credit decreases. This is somewhat expected, as credit increases the potential "risk" of financial delinquency.

Credit was reallocated from more to less risky loans, increasing the efficiency of the credit market. Combining financial delinquency and credit changes, we estimate the default rates on the marginal credit reallocated due to Cadastro Positivo. Under the assumption that the policy did not affect default rates of credit that would have been observed in the absence of the policy, we can calculate the default rate of the marginal credit as the ratio between the change in financial delinquency and changes in credit. We find that, on average, credit given because of the policy had a default rate of around 3%. On the other hand, credit that would have been given in the absence of the policy but was not given because of it would have had a default rate of around 15%. The comparison between both values highlights that credit was reallocated from more to less risky credit because of the revelation of information.

Having established how more informative credit scores affected credit access and default rates, we investigate if it had consequences beyond credit markets. We focus on entrepreneurial activity, exploring the longstanding link between credit constraints and the decision to become an entrepreneur (Evans and Jovanovic, 1989), which has been posited as an important barrier to economic development (Banerjee and Newman, 1993).

To understand how the policy affected entrepreneurial activity, we match the credit bureau data with self-collected firm records that comprise the universe of formal firms in the state of São Paulo, the most populous of the country. We then pair this data with firm-level characteristics collected by our partner institution. This allows us to have a data set of individuals, with information on whether they own a firm or not and the characteristics of firms for those who own one.

We find no effect of changes in credit access on firm creation. To estimate this, we divide our sample into three groups according to their values of old and new credit scores. A group of positively exposed individuals is defined as those who are in parts of the joint distribution of credit scores that had gains in credit. The opposite holds for defining negatively exposed individuals. The third group consists of individuals in parts of the distribution for which we estimate small or no change in their credit access. With our treatment groups defined, we

estimate hazard models with time-dependent covariates to investigate the probability that an individual opens their first firm. These work in the spirit of a difference in differences, where we compare the hazard rates of differently exposed groups before and after the policy. Our point estimates for both *treatment* groups are close to and not statistically different than zero. For positively (negatively) exposed individuals, we can reject effects as big as 3% (5%) in the probability of creating a new firm because of the policy.

However, the reallocation of credit affected new firms' outcomes. Comparing firms created by individuals in the three groups, we find that new firms created after the policy by positively (negatively) exposed individuals have a higher (lower) likelihood of surviving two years after their creation. We show that firms from positively exposed individuals are also, on average, in more productive industries, employ more, and have higher average wages.

Our findings indicate that the reallocation of credit caused by revealing information positively affected the average quality of new firms. Extrapolating average differences to the marginal firms' survival changes, our results show that new cohorts of firms are substituting less productive firms with more productive ones, indicating a positive effect on the average quality of new firms' cohorts.

This paper contributes to different threads of literature. First, we contribute to the literature on the effects of borrowers' information on credit access. Recent empirical work has extensively examined how adding/reducing information affects borrowers' access to credit (Musto, 2004; Bos and Nakamura, 2014; Dobbie et al., 2020; DeFusco et al., 2022; Jansen et al., 2022; Bos et al., 2018; Liberman et al., 2018; Herkenhoff et al., 2021; Gross et al., 2020). However, the focus of these papers have been in "first moment" changes in information[1]. Our paper adds to this literature by estimating not only first-moment changes in the creditworthiness assessment but also the effects of changes in the precision of the signal on credit access and default rates. We show that the overall effects of information combine both mechanisms. Furthermore, we estimate the effects on different parts of the credit score distribution, which is different from most papers in this literature, which are focused on local effects on the population of previously bankrupt individuals.

We also add to the literature that studies information-sharing institutions. Several empirical studies using cross-country regressions show a positive relation between information-sharing institutions such as credit registries and private credit bureaus and financial and

[1]That is, a shock occurs that changes the availability of information for a group (namely bankruptcy flags). This increases or decreases the creditworthiness assessment of that given group, and then credit outcomes are compared to a control group.

economic development(Pagano and Jappelli, 1993; Jappelli and Pagano, 2002; Djankov et al., 2007). In turn, within-country studies that leverage changes in these institutions are rarer and focused on the effects of information on firms, comparing firms' financing with more/less available information(Hertzberg et al., 2011; Behr and Sonnekalb, 2012). We contribute to this literature by characterizing who and how much individuals are affected by information-sharing institutions. Furthermore, with important policy implications, we can quantify how default rates change in such an institutional change.

This paper also contributes to the literature that studies the role of credit constraints on entrepreneurial activity. This has loon been studied in developed countries (Evans and Jovanovic, 1989; Black and Strahan, 2002; Hurst and Lusardi, 2004; Robb and Robinson, 2014; Schmalz et al., 2017; Herkenhoff et al., 2021; Cahn et al., 2021; Dobbie et al., 2020) but a debate still withstands with conflicting estimates[2]. In developing countries, there is substantial evidence that existing businesses are credit constrained (De Mel et al., 2008, 2009; McKenzie, 2017), but also evidence that credit access has no substantial effects in spurring new business (Karlan and Zinman, 2010, 2011; Banerjee et al., 2015), despite the influential early theories highlighting the importance of this link (Banerjee and Newman, 1993). Our setting allows us a rare empirical context where we can estimate both "extensive" (firm creation) and "intensive" (firm quality) margins of credit on entrepreneurship in developing countries. Our findings go in line with what has been found in the literature, suggesting no effects on the extensive margin but a positive relation between credit and firm outcomes for firms that would have existed anyway. Furthermore, we can characterize the effects of one of the most advocated policies in credit markets in developing countries (World Bank, 2012), going beyond specific credit products explored in other studies.

2 Institutional Background

Brazil is the largest country in Latin America and the 7th largest in the world, with 212 million inhabitants. Despite having the 8th largest GDP in the world, it is still considered a developing country, ranking 89th in HDI and around 80th in GDP per capita.

The country is characterized by a high level of banking penetration relative to their income level. Over 80% of the population had a bank account, and on average an individual had accounts in 5.4 different banks[3]. There are over 150 operating banks in the country and

[2]For example, in two recent papers Dobbie et al. (2020) and Herkenhoff et al. (2021) study the effects of removing bankruptcy flags on self-employment and entrepreneurship and find conflicting estimates on the effects of credit on this type of occupational choice.

[3]We show average account ownership and share of individuals borrowing from financial institutions by

more than 600 other financial institutions, such as fintechs and credit cooperatives, that also provide loans.

Before *Cadastro Positivo*, credit bureaus were already relevant sources of assessment of creditworthiness. Their main source of information was a delinquency registry often referred to as *Cadastro Negativo de Devedores*. It tracked defaults, protests, bounced checks, overdue debts, and other financial irregularities. These complaints could be made by creditor companies such as banks, financial institutions, stores, service providers, among others. Since 1990[4], the inclusion of a consumer's name in defaulters' lists must follow the guidelines established by federal law, ensuring the right to information and the right to challenge by the consumer. On top of delinquency records, credit bureaus also used publicly available information to calculate credit scores. These included prosecution records, firm ownership, bankruptcy records, and demographic characteristics.

Cadastro Positivo - increase in information available to Credit Bureaus

Cadastro Positivo is the name given to the legislation that changed the information available for credit bureaus to create credit scores. With this change, on top of the previously available sources, credit bureaus gained access to *positive financial data*, including records of on-time payments, responsible credit usage, good financial behavior, the opening of new credit accounts, successful loan repayments, the length of an individual's credit history.

Lenders do not access the registry information directly. The data is shared only with authorized credit bureaus responsible for processing the information and generating credit scores for individuals and businesses. These credit scores are subsequently made available to potential lenders and creditors[5].

Timeline: The Cadastro Positivo was initially established through the Law No. 12.414 on June 9, 2011. After the law, a long implementation period started where regulations were being established by government agencies. The first information originated from Cadastro Positivo became available to credit bureaus in the first semester of 2014.

Until 2019, individuals and businesses had to opt in to have their information included in the Cadastro Positivo. However, the opt-in system faced challenges in garnering a sufficient number of registrants as by 2019, less than 5% of the population had opted-in to Cadastro

country relative to their GDP per capita in Figure A1. We observe that Brazil has higher levels on both statistics relative to their prediction based on GDP per capita.

[4]Credit Bureaus in Brazil existed since at least the 1950s, but their operation was only regulated by Law 8.078/1990 also known as *Código de Defesa do Consumidor*

[5]Brazil has authorized four credit bureaus to access and employ Cadastro Positivo data for credit assessment purposes. These official credit bureaus include Serasa Experian, Boa Vista SCPC (Serviço Central de Proteção ao Crédito), Quod, and SPC Brasil (Serviço de Proteção ao Crédito).

Positivo. BACEN (2021) documents that due to the low take-up of the policy, none of the major banks used Cadastro Positivo information in their lending decision during the opt-in phase[6].

In April 2019 Congress approved Law 166/2019, which changed the default status in Cadastro Positivo to include all individuals and businesses with financial records in the Cadastro Positivo unless they expressly opted out. This change generated an increase of more than 15 times in the number of individuals with active information in the registry. This represented over 100 million individuals with active information in Cadastro Positivo. Even though individuals still could opt out of the system, by the end of 2020 less than 350 thousand people in the country had done so.

The execution of the new law took place over the following 2 years. Between April and December of 2019 year, regulating agencies set up a new set of conditions for credit bureaus to gain access to the data. In December 2019, the change in the system took place, and individuals who were previously outside of Cadastro Positivo were now included in the system. During the first semester of 2020, information from institutions registered with the Central Bank started to be shared with credit bureaus. These data included consumers' payment records from the previous 2 years. Given the complexity of analyzing and adapting credit assessment to the new set of information available, credit scores under the new phase of Cadastro Positivo only began to be commercialized by May of 2021.

Aggregate Credit Patterns Around Policy Implementation: Before *Cadastro Positivo*, Brazil already had a robust and established credit market. Therefore large changes in aggregate credit patterns would not be expected to occur due to the policy. We show some aggregate statistics of household credit in Figure A2. We observe no clear break in the trends of total household credit, nor in their composition across different sources (Panels (a) and (b)). At the same time, there is a slight increase in average cost of credit, mostly accompanying the spike in interest rates that happened around the time of the policy. Lastly, we also see that there are no major changes in the level of credit concentration in the market. The aggregate pattern suggest a continuation of an existing pattern of decrease in concentration, although the magnitude of these changes is relatively small[7].

[6] On top of low take-up, BACEN (2021) documents that banks also reported large selection patterns in individuals registered in the new system. Lenders report that most registered individuals were formal credit delinquents who opted-in the system when they were renegotiating their debts.

[7] The normalized HHI decreases from around 0.125 to 0.10 between 2016 and 2023. This slight decrease in concentration is often attributed to the entry of digital credit products, that increased their participation in the Brazilian market in the end of the 2010's decade

3 Conceptual Framework

In this section, we outline a conceptual framework to understand the effects of additional information in the construction of credit scores. We consider a lender's decision problem, in which they face borrowers who vary in their creditworthiness. The lender does not observe the true types of potential borrowers. Instead, they receive an unbiased signal and update their beliefs accordingly. Signals in our framework are not endogenously determined by borrowers in the spirit of Spence (1973). Instead, we think of them as information structures as in Green and Stokey (1978) or more recently Brooks et al. (2022a,b)

Our framework allows us to define two different effects of revealing information about potential borrowers on their credit access. The first one, which we refer to as the *effect of the signal's value*, is determined by changes in the assessment of creditworthiness given the new information. The second one, which we call *effect of signal's precision*, is driven by the change in the precision of the signal observed by the lender.

3.1 Setup

Borrowers: Potential borrowers differ in their (true) creditworthiness, denoted by $\theta_i \in \mathbb{R}$, with $\theta \sim G()$, where G is a well-defined density function continuous on $\mathbb{R}$, One could conceptualize this parameter as one governing the profitability of loans and the probability of defaults.

Information: ADJUST HERE A BIT - MAKE THE SIGNAL REALIZATION AND INFOMRATION STRUCTURE CONSISTENT WITH LATER

A lender does not directly observe the true creditworthiness of the borrower but receives informative signals about it. The lender then forms expectations of the creditworthiness of the potential borrowers and offers them loans accordingly.

More precisely, we define a signal S, ρ as a set of Signal realizations $S = \Theta$ and a joint distribution ρ over $\Theta \times S$. The marginal distribution of ρ over Θ is $G()$, and that ρ marginal distribution over S is a governed by a distribution $F()$.

We assume that a signal realization $s_i \in S$ is an accurate estimate of the true creditworthiness θ_i. That is: $s_i = \mathbb{E}_\rho[\theta_i | s_i]$.

Decision Problem: The lender faces a decision problem. They have an objective function $\pi(\theta_i, C_i)$ where they decide how much credit they supply to an individual (C_i), based on i's creditworthiness.

We assume that for every θ, there exists a credit supply C^* that maximizes the objective function of the lender. Or formally, we define a function $C_i^*(\theta)$:

$$C_i^*(\theta) = \underset{C}{\text{argmax}}\, \mathbb{E}[\pi(\theta_i, C_i)|s_i]$$

3.2 Gaussian Parametrization

To gain intuition on the effects of the policy, it is useful to include a common parametrization of the signals as a function of the true value of individuals' types. We assume that an individual i's type θ_i is drawn from an underlying distribution $\theta_i \sim N(\mu, \sigma^2)$. Their signal is a noisy measure of their creditworthiness:

$$s_i = \theta_i + u_i, \quad u_i = \epsilon_i - \Delta_i$$

where Δ_i and ϵ_i are independent and distributed according to $N(0, \sigma_\Delta)$ and $N(0, \sigma_\epsilon)$ respectively. Thus $u_i \sim N(0, \sigma_\Delta + \sigma_\epsilon)$. This is an adaptation of the common formulation of signals in the statistical discrimination literature in labor markets (Phelps, 1972; Aigner and Cain, 1977).

The lender's expected value about a given borrower's type is then given by:

$$\mathbb{E}[\theta_i|s_i] = \mu + \underbrace{\frac{\sigma}{\sigma + \sigma_\Delta + \sigma_\epsilon}}_{\text{Shrinkage Factor}} (s_i - \mu)$$

where the expected value is the average of the population updated towards the signal s_i. The shrinkage factor can be interpreted as how valuable the signal is. If it was fully informative ($\sigma_\epsilon + \sigma_p = 0$) the expected value of the individual's type would be equal to the signal. The only difference between the expression above and standard frameworks is that the variance of the *noise* is determined by $\sigma_u = \sigma_\epsilon + \sigma_\Delta$.

3.3 Revealing Information

Policy: We now consider a policy that allows allows Δ_i to be observed. An interpretation of it is that signals are constructed with additional information that recovers Δ_i. Thus, lenders observe a new signal $s_i' = s_i + \Delta_i$[8], which in turn reflects to a new expected value of the borrowers' creditworthiness given the signal:

$$\mathbb{E}[\theta_i|s_i'] = \left(\mu + \frac{\sigma}{\sigma + \sigma_\epsilon}(s_i + \Delta_i - \mu)\right)$$

[8]In other words, s_i is a specific case of a mean preserving spread of s_i' with a difference between variances of σ_Δ

To understand the changes in credit given the new signal, we compare the two expected values of creditworthiness, $E[\theta_i|s_i^{'}]$ and $E[\theta_i|s_i]$. We can write the difference between them as:

$$\mathbb{E}[\theta_i|s_i^{'}] - \mathbb{E}[\theta_i|s_i] = \underbrace{\left(\mu + \frac{\sigma}{\sigma + \sigma_\epsilon}\big(s_i + \Delta i - \mu\big)\right)}_{\mathbb{E}[\theta_i|s_i^{'}]} - \underbrace{\left(\mu + \frac{\sigma}{\sigma + \sigma_\Delta + \sigma_\epsilon}\big(s_i - \mu\big)\right)}_{\mathbb{E}[\theta_i|s_i]}$$

which, by rearranging the terms, can be rewritten as:

$$\mathbb{E}[\theta_i|s_i'] - \mathbb{E}[\theta_i|s_i] = \underbrace{\frac{\sigma\Delta_i}{\sigma + \sigma_\epsilon}}_{\text{Signal Value Change}} + \underbrace{\frac{\sigma \cdot \sigma_\Delta(s_i - \mu)}{\big(\sigma + \sigma_\Delta + \sigma_\epsilon\big) \cdot \big(\sigma + \sigma_\epsilon\big)}}_{\text{Change in Precision}}$$

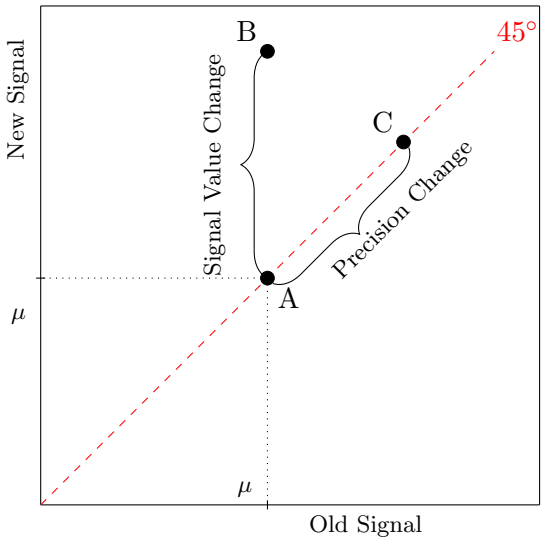
The expression above provides us with important intuition on how an increase in the information changes the expected values of creditworthiness. A more intuitive implication of our framework is that the changes in the expected value of creditworthiness is increasing in the value of Δ_i. This is straightforward from the *signal value change* part in the expression above. In other words, if the information revealed indicates that the individual is more creditworthy, lenders will update their expectations accordingly.

Another implication of our framework is that the revealed information also increases the *precision* of our signal. This implies that expected values of creditworthiness should change even for individuals for whom the information revealed does not change their signal. The intuition behind this is that with a more precise signal, the lenders' shrinkage factor decreases, thus they update the population prior closer to the signal than before.

We illustrate these changes in the diagram below, which represents the joint distribution of signals. The x-axis represents values of the old signal, whereas the y-axis shows the new signal constructed with the revealed information. Points over the 45-degree line represent individuals whose signals did not change with the policy.

We highlight three points in the diagram. Point A represents an individual whose initial signal was equal to the population prior and did not change with the revealed information. Our framework predicts no change in this individual's expected creditworthiness. In turn, point B represents an individual whose signal increases with the revealed information. This is the case for all individuals above the 45-degree line. The opposite holds for those below the 45-degree line. Lastly, we highlight point C, which represents an individual whose signal did not change but was above the population prior. Our framework predicts that the

Diagram of Joint Distribution of Old and New Signals



lenders' expectation of C's creditworthiness increases with the revelation of new information as the shrinkage factor decreases.

Generalizing the Gaussian Parametrization:

The Gaussian parametrization of the distributions is useful for exposition purposes but not necessary for our rationalization of changes in credit with more informative signals. In what follows, we state general assumptions on the expected value of creditworthiness given the signal and how the less and more informative signals relate to each other.

Assumption 1: *The expected value of creditworthiness given the signal can be written as a convex combination between the prior and the signal:*

$$\mathbb{E}[\theta|s_i] = \mathbb{E}[\theta] + \frac{1}{a}s_i$$

where a is increasing in the variance of the signal realization given the true type.

Assumption 2: *A pre-policy signal s_i is a mean-preserving spread of the new more informative signal s_i'. That implies for a given constant t:*

$$\text{(i)} \quad \int_{-\infty}^{\infty} sf(s|\theta)\,ds = \int_{-\infty}^{\infty} s'f(s'|\theta)\,ds'$$

$$\text{(ii)} \quad \int_{-\infty}^{t} F(s|\theta)\,ds > \int_{-\infty}^{t} F(s'|\theta)\,ds'$$

‘ Diaconis and Ylvisaker (1979) proves sufficiency conditions for distributions to follow

the assumption above[9]. In particular, they show that in conjugate priors in the exponential family, it is possible to write the posterior expectation of a random variable with a known distribution given a signal as a convex combination of the expected value and the signal. Changes in the variance of the signal that generate mean-preserving spreads, generate changes in the convex combination parameter that go according to the shrinkage factor described above.[10]

3.4 Implications on Credit Allocation

We now describe how the policy reflects in the lender's solution to their decision problem. In what follows, we assume that the distribution of types and signal realizations follow assumptions 1 and 2.

We outline propositions that characterize how credit should change over the set comprising values of new and old signals $(s_i, s_i') \in \mathbb{R}^2$. We define $h(s_i, s_i') : \mathbb{R}^2 \to \mathbb{R}$, the function that determines how credit decisions should change with the revelation of information that creates signal s_i':

$$h(s_i, s_i') = C^*(E[\theta|s_i']) - C^*(E[\theta|s_i])$$

where $C^*(E[\theta|s_i'])$ is the solution of the lender's decision problem with signal s_i' and $B^*(E[\theta|s_i])$ the solution with signal s_i.

Proposition 1: *For any given pair of old and new signals, s_i, s_i', and a given positive constant c, if C^* is increasing, rationalizable changes in credit should follow*:

i. $h(s_i, s_i' + c) - h(s_i, s_i') > 0$

ii. $h(s_i + c, s_i') - h(s_i, s_i') < 0$

Statement (i) and (ii) are what we define as the *Effects of the signal's value*. (i) states that changes in credit are increasing in the new signal's value. This comes from the fact that, for a fixed value of the old signal s_i, the function that defines changes in credit $h(s_i, s_i')$ is increasing in s_i' which in turn comes from the solution of the lenders problem being increasing in its argument, which is increasing in s_i'. The same argument is valid for

[9] This result is more explicitly explained in Michael Jordan's lecture notes from Stat 260/CS 294 Bayesian Modeling and Inference Spring 2010 at UC Berkeley.

[10] Chambers and Healy (2012) provide more general conditions for what they define as *Update Towards the Signal* types of posteriors. In these cases we still observe a linear relation between the expected value of the random variable and a signal realization. However, they do not explicitly establish how the *shrinkage factor* a relates to the variance of the signal realization.

statement (ii). It states that for a given value of the new signal, the changes in credit should be decreasing in the old signal.

The intuition for statement (i) is reasonably straightforward. A signal that implies that individual i is more creditworthy should imply bigger changes in credit. In turn, statement (ii) is slightly less intuitive. We could explain it as for a given value of the new signal, an old signal that implies i is more creditworthy would represent a bigger pre-policy value of credit, which would, in turn, represent a decrease in the credit change.

We can also think about how these statements would fail empirically. If the lender does not perceive the new information available in the new signal as useful, we would expect equality to hold in statement (i). At the same time, if lenders did not perceive information in the old signal as useful, we would expect equality in statement (ii).

Proposition 2: *If $h(s_i, s_i')$ is increasing in $E[\theta|s_i'] - E[\theta|s_i]$, for any given pair of old and new signals, s_i, s_i', and a given positive constant c, rationalizable changes in credit should follow:*

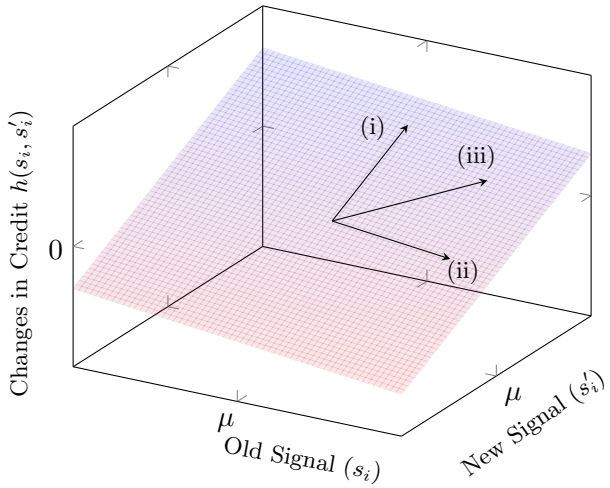
iii. $h(s_i + c, s_i' + c) - h(s_i, s_i') > 0$

Statement (iii) defines what we call the *Effects of the signal's Precision*. It states how, conditional on the same difference between signals, changes in credit should behave. If changes in credit are increasing in changes in the posterior of types given a signal, changes in credit should be increasing as we move towards larger values of the signal conditional on the same difference between new and old signal. It requires that we make assumptions on how function $h(s_i, s_i')$ varies with the differences in $E[\theta|s_i'] - E[\theta|s_i]$ instead of assumptions over $C^*()$. It holds because under our assumption over how posteriors are formed given a signal, $(E[\theta|s_i' + c] - E[\theta|s_i + c]) > (E[\theta|s_i'] - E[\theta|s_i])$. Intuitively, this occurs because, conditional on the same differences in the signal values, the *shrinkage* factor is increasing in the value of the signal. The complete proof is written down in Appendix B.

The diagram illustrates how Propositions 1 and 2 allow us to rationalize changes in credit over the joint distribution of signals. It shows changes in credit in the *z-axis*, values of the old signal (s_i) in the *x-axis* and values of the new signal (s_i') in the *y-axis*. The arrows correspond to each of the statements from both propositions.

We observe that: (i) conditional on the old signal, changes in credit are increasing in the value of the new signal. (ii) conditional on the new signal, changes in credit are decreasing in the value of the new signal. (iii) conditional on the difference between signals, credit changes are increasing in the value of the signal.

Rationalizable Changes in Credit Under a Linear Relationship Between Expected Creditworthiness and Credit



Statement (iii) provides theoretical implications and important consequences of the changes in the credit distribution. As it posits that, conditional on the difference between signals, those with more credit scores should have higher changes, it implies an increase in the inequality in credit access and a widening in the credit distribution, conditional on the changes in credit scores.

When estimating the effects on credit access, we will return to statements (i), (ii), and (iii) to test if our framework can rationalize the changes in credit.

4 Data and Sample Construction

We use two main data sources. First, we use information from SERASA, the largest credit bureau in Brazil. We combine that with firm ownership records from São Paulo's trade board, *Junta Comercial do Estado de Sao Paulo* (JUCESP) an autarchy of the government that is responsible for organizing and keeping firm records[11].

SERASA is the largest credit bureau in Brazil and one of the four authorized by the Central Bank to access the information from Cadastro Positivo. It collects information not only from institutions in the national financial system, but also from retail, utility, and insurance companies.

We use both individual and firm-level data from SERASA. On the individual side, we observe their credit scores built with and without information from Cadastro Positivo,

[11]We translate *Junta Comercial* to Trade Board. Still, it is possible to translate it to the commercial registry or business board.

which allows us to build the counter-factual credit score in the absence of the policy for each individual. On top of that, we observe detailed information on credit access, including loans and purchases made on credit, and loan payments made by those individuals. One caveat of our data, common to most credit bureau information (Liberman et al., 2018), is that we do not observe interest rates of loans.

Firm Ownership Records (JUCESP): According to the Brazilian constitution, every formal business must register in its state's trade boards. In Brazil, firm ownership information is in theory public. All states' Trade Boards are obliged to make all information they collect public but can charge for it[12]. JUCESP is the only Trade Board that allows Brazilian Citizens to access information for free. To recover firm records, it is necessary to have a Brazilian Social Security Number and registration on JUCESP's website. Each citizen has access to 799 firm records per day.

We collect more than 4 million identified firm records, which cover all formal firms created in São Paulo between 2003 and 2023[13]. These records contain all information from the firm ownership history. Importantly, they include the name and social security number of all founders, on top of firm identifiers. This allows us to match firm records from JUCESP with information from the Credit Bureau.

4.1 Sample Construction

Our main sample comprises around 200,000 randomly selected individuals aged between 20 and 60 years old in 2019 living in the state of São Paulo. This represents around 1% of the population of the State. We restrict our empirical focus to São Paulo due to data agreements. It is the most populous state in the Country with over 44 million inhabitants and concentrates around 1/3 of the country's economic activity[14].

We begin with individual-level credit access information from SERASA for all individuals in the sample and match that with firm ownership information from the JUCESP firm records for those who eventually open a firm. For those who are firm owners, we then find information on their firm's outcomes using SERASA's information. In Appendix C we detail the sample construction procedures, indicating all the steps done by the company's

[12] Almost all states in the country charge exorbitant price rates for firm records stating who are the owners or who created each firm.

[13] We would like to thank Marina Dias and Todd Messer for providing substantial help in the code to scrape these files. Furthermore, we would like to thank 39 other Brazilian Citizens who generously helped us in this process. Unfortunately, we cannot name them all here due to space limitations, but our sincerest gratitude.

[14] Compared to other Latin American countries, Sao Paulo has the same population as Argentina, and its economic activity would be on par with Mexico's GDP.

Table 1: Summary Statistics

Demographic Characteristics		Credit Characteristics		
	(1)		(2)	(3)
	Mean		Mean	Std. Dev.
Female	0.45	C. Sc. - Old System	467.15	310.59
Nonwhite	0.24	C. Sc. - New System	552.89	222.37
Less than H.S.	0.21	Total Credit	30325.63	100943.65
High School	0.56	Loans	26870.59	97722.95
Some College	0.22	Credit Purchases	3455.04	9243.21
Age	40.19	Financial System Default	2435.93	95206.02
		Other Default	600.01	8291.51
N.Obs	194247			

This Table presents summary statistics of our sample. Statistics are calculated in the last period before the policy's implementation. Monetary values are presented in Dec. 2023 Brazilian Reais (1 USD = 5.03 BRL).

researchers and de-identification processes that were required in our data agreement.

We show summary statistics of our sample in Table 1. Statistics are calculated in the last period before the implementation of the policy. Monetary values are presented in Dec. 2023 Brazilian Reais. At the time, the corresponding exchange rate was 1 dollar to 5.03 Brazilian Reais. On the left-hand side of the panel, we describe the demographic characteristics of our sample.[15] On the right side of the panel, we show credit characteristics. Given the novelty of these data, next, we discuss the characteristics and definitions of these key variables:

Credit Scores: we access two credit scores for each individual in our panel. The old system credit scores were constructed using only data sources available to the credit bureau before the policy. In contrast, the new system credit scores represent their updated creditworthiness assessment measure, using data made available by Cadastro Positivo and an updated prediction model[16]. Under the new system, the credit bureau defined values above 700 as *excellent* credit scores, those between 501 and 700 as *good*, 301 and 500 as *low*, and below 300 as *very low* credit scores.

We can see in Table 1 that new and old system credit scores differ in both average and variance. This happens because the new credit scores are subject to a different scale than the old ones. To make them comparable measures, we normalize them in our empirical

[15] Nonwhite individuals group those who declare themselves *Pretos, Pardos* and *Indigenas* (Black, Mixed and Indigenous). Less than High School groups all individuals without a High School diploma, whereas Some College groups individuals with at least some college education, thus including college dropouts and post-secondary technical degrees.

[16] Serasa has multiple credit score measures with specific goals, which they commercialize for different purposes. The ones we use are their most standard measures.

analysis by subtracting the population's average and dividing by the standard error.

Total Credit: is our main outcome variable in the empirical analysis. It is defined at a given period as the sum of Loans and Credit Purchases (Mostly Credit Cards) from a given individual. This includes all types of credit, including credit for consumption and other personal use, and real estate related credit equivalent to mortgages in the US. In Brazil, real estate related credit represents a much smaller share of the personal credit market than in the US (around 25% relative to around 70% in the US). Unfortunately, in our data we cannot decompose credit by their type, only by type of institution, which we use for heterogeneity analysis. [17]

Financial Delinquency: Financial institutions report the amount owed to credit bureaus when individuals fail to meet their financial obligations on time. Typically, the lender sends a notice indicating the debt and sets a deadline for regularization. If this deadline is not met, the overdue debt may be marked as delinquent and reported to the credit bureau. In our data, each individual's amount reported as delinquent from financial system obligations is observed at a given period.

Our credit bureau partner receives delinquency measures from other sources besides the financial system. Our data allows us to disentangle between these sources. In our empirical analysis of default rates, we focus on delinquency originating from the financial system. Table 1 shows that delinquency from other sources is, on average, smaller than financial system delinquency.

5 Effects on Individuals' Credit Access

In this section, we empirically analyze the effects of the increase in the information available to construct credit scores on individuals' credit access. We begin by characterizing the marginal distributions of credit scores constructed under the new and old systems and the joint distribution of both measures. We then estimate how changes in credit scores affected the credit access of individuals over the joint distribution of credit scores. We focus on the transition from the opt-in to the opt-out phase as a source of variation on lenders' information about borrowers, as the opt-in period had low take-up, and lenders report not using that information.

[17] See Gonzalez et al. (2023) for a summary of how personal credit is divided across types in Brazil during our analysis period.

5.1 Credit Scores with and without Cadastro Positivo

We begin by characterizing credit scores constructed with and without the information from *Cadastro Positivo*. Let I_{it}^n be a vector that contains credit delinquency information from individual i at period t which was available to credit bureaus before the policy. I_{it}^p is a vector of additional information about individual i at period t that is made available to credit bureaus with Cadastro Positivo.

Credit bureaus use these data to assess the creditworthiness of individual i. Let $f_{it}^n(I_{it}^n) : \mathbb{R}^n \to \mathbb{R}$ be a function that transforms the information into a creditworthiness measure before Cadastro Positivo and $f_{it}^p(I_{it}^p, I_{it}^n) : \mathbb{R}^m \to \mathbb{R}$ be its counterpart under the new system. Credit scores under each system are defined by:

$$s_{it}^{'} = f_{i\tau}^p(I_{i\tau}^p, I_{i\tau}^n) \quad s_{it} = f_{i\tau}^n(I_{i\tau}^n)$$

Marginal Distribution of Credit Scores: Despite only accessing delinquency records under the old system, credit scores varied substantially over the population. This is maintained under the new system. In Figure A4, we show the distribution of both credit scores. To make them comparable, we use the Z-score[18] of both measures. Next, we show how both measures correlate with observable characteristics of individuals.

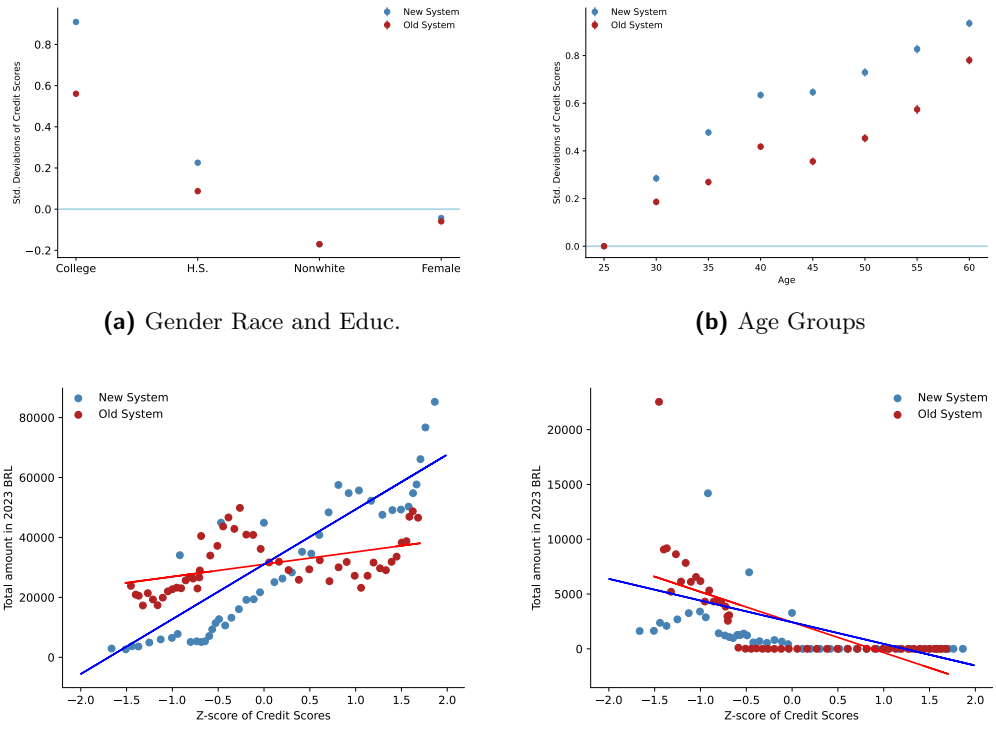
In Panels A and B of Figure 1, we show the correlation between the Z-score of credit scores and the demographic characteristics of individuals. The plotted coefficients Γ are the OLS estimates of a linear model $Z_i = \Gamma X_i + \epsilon_i$. We include age, gender, race, and education in the vector of observable characteristics.

We observe in Panel A that women and nonwhite individuals have, on average, lower credit scores than men and white individuals. There are no clear differences in this estimate between the new and old systems of credit scores. When we look at education levels, we observe that both H.S. and college-educated individuals have a higher credit score than those with less than H.S. education. This difference is amplified when information from the positive credit score system is included. In Panel B, we show that credit scores are positively correlated with age in both the new and old systems.

In Panels C and D of Figure 1, we show the correlation of the Z-score of credit scores with the amount of credit individuals take. We observe that credit is increasing in both old and new system credit scores. However, the correlation between both measures is substantially stronger in the new system. The opposite holds when we observe default measures in Panel C. Default is decreasing on credit scores, but the correlation is stronger in the old system

[18] $Z_i = \frac{s_i - \bar{s}}{sd(s)}$ where $\bar{s} = \sum_i \frac{s_i}{N}$

Figure 1: Credit Scores Correlation with Observable Characteristics



(a) Gender Race and Educ.

(b) Age Groups

(c) Loans

(d) Financial Delinquency

Panels (a) and (b) plot the coefficients of a regression of $\frac{cs_i - \bar{cs}}{sd(cs_i)}$ on observable characteristics. The sample is restricted to the last period before the implementation of the policy. Coefficients in panels (a) and (b) are estimated in the same regression that includes dummies for gender, race, education groups, and age groups. We omit from the regression white men with less than high school education in the youngest age group. In panels (c) and (d), we show binscatters of Credit and Default with the Z-scores of credit scores in both positive and negative systems. The sample is restricted to the last period before the implementation of the policy.

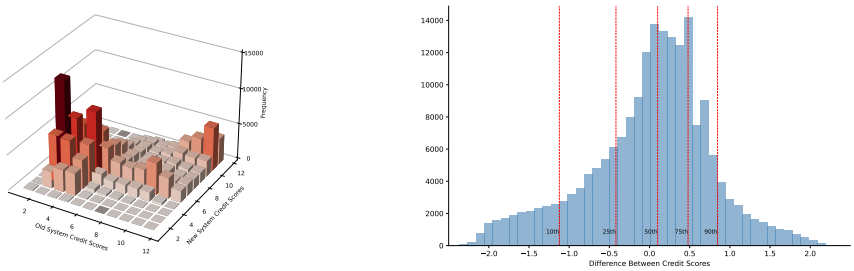
measures.

Joint Distribution of Credit Scores:

Next, we describe the joint distribution of credit scores in both systems. In Panel A of Figure 2, we show the histogram of our sample over both measures of credit scores. We restrict the sample to observations in the last period before the implementation of our policy and divide our sample into 11 equally sized bins.

The policy generates substantial variation in credit scores. For each individual, we measure the difference in their credit scores in the positive and negative systems. We refer to this measure below when describing our empirical strategies to estimate the effects of the signal's value and the effects of the signal's precision:

Figure 2: Joint Distribution of Credit Scores in Positive and Negative System



(a) Joint Distribution

(b) Distribution of Difference in Credit Scores

In Panel (a), we show a histogram of the joint distribution of credit scores in the new and old systems. Individuals are divided in a grid of 11 equally sized bins of values in each system. In Panel (b), we show a histogram of the difference between the z-scores of credit scores in the new system and credit scores in the old systems. Vertical red lines represent the 10th, 25th, 50th, 75th and 90th percentile of the distribution.

$$\Delta_i = s'_{i\tau} - s_{i\tau}$$

where we fix τ as the period before the policy implementation.

In Panel B of Figure 2, we plot the distribution of Δ_i. It has a shape similar to a normal distribution, with a median centered at 0. It has a slightly higher mass in the left tail than in the distribution's right tail. In Figure A5 we show the correlation of Δ_i with the demographic and credit characteristics of individuals. Δ_i has virtually no correlation with gender of race. It is positively correlated with education levels and age. Furthermore, Δ_i positively correlates with individuals' credit before the policy.

5.2 Effects on Credit Access

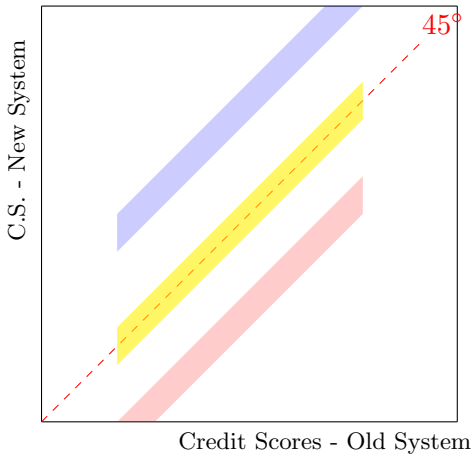
Next, we estimate how individuals' credit access changed according to the values of credit scores in both negative and positive systems. We start by testing both signal changes and precision mechanisms outlined above. We then estimate how credit access changes according to the joint distribution of credit scores in both systems.

5.2.A Effects of the Signal's Value

We begin testing whether changes in the signal value affect how much credit individuals access. To do so, we compare individuals whose credit scores increased or decreased with the additional information with those who did not have any change.

The diagram below provides a visual description of our empirical exercise. It represents points in the joint distribution of credit scores. Our exercise compares individuals at the joint distribution's blue, yellow, and red parts before and after the new credit scores were made available. Being in the blue part of the diagram implies that the additional information improved the creditworthiness assessment of that given individual. The opposite holds for those in the part of the distribution highlighted in red. Those in the yellow part have similar creditworthiness assessments with and without information made available by the policy.

Diagram of the Empirical Strategy to Estimate the Effects of Changes in Signals' Value



More formally, we define the three groups as:

$$D_i^+ = \mathcal{I}\Big[\Delta_i \in [0.75, 1.25]\Big] \qquad D_i^- = \mathcal{I}\Big[\Delta_i \in [-1.25, -0.75]\Big] \quad D_i^0 = \mathcal{I}\Big[\Delta_i \in [-0.25, 0.25]\Big]$$

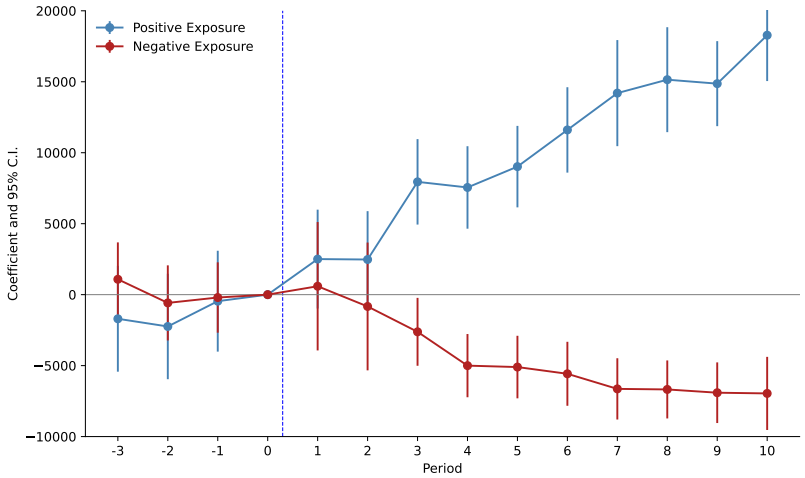
We restrict the sample for those individuals lying in one of the three groups (i.e., $D_i^+ = 1|D_i^- = 1|D_i^0 = 1$) and estimate the following difference in differences model.

$$Y_{it} = \alpha_i + \delta_t + \sum_{t \in T} \beta_t^+ \cdot D_i^+ \cdot \delta_t + \sum_{t \in T} \beta_t^- \cdot D_i^- \cdot \delta_t + \varepsilon_{it} \tag{1}$$

where α_i are individual fixed effects and δ_t are time dummies. Our identifying assumption is that, on average, individuals in different groups would follow parallel trends in the absence of the policy.

We show our estimates of $\{\beta_t^+, \beta_t^-\}$ in Figure 3. Vertical bars represent 95% confident intervals, centered around our coefficient estimates.

Figure 3: Effects of Signal Value Change on Credit Access



This Figure shows how credit changes for individuals with increases and decreases in their credit scores due to the policy. The connected dots plot the $\{\beta_t^+, \beta_t^-\}$ estimates from equation 1 with their respective 95% confidence intervals. Positive change is defined as individuals with $\Delta_i \in [0.75, 1.25]$ and negative change as $\Delta_i \in [-1.25, -0.75]$.

We find significant increases in credit access for those who had a positive change in their credit scores, shown in the blue connected dots. Two years after the policy, this group observed an increase of 15 thousand BRL in their credit access. On the other hand, we estimate a decrease in credit access for those with negative changes in their credit scores. We estimate a reduction of 8 thousand reais 2 years after the policy. This represents a change of 20% in credit access for both groups relative to their counterfactual credit access without the policy[19].

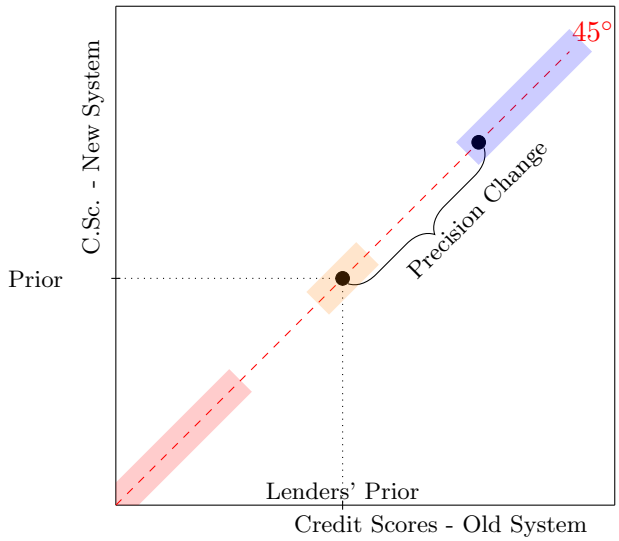
5.2.B Effects of Signal's Precision

We next test if the increase in the precision of the signal had effects on individuals' credit access. With a reduction in the noise of the signal, our conceptual framework predicts that lenders expected values of the individuals' true creditworthiness should be closer to the received signal. This generates changes in credit access even for those without changes in their credit scores.

[19] We arrive at percentage effects by dividing our coefficient estimates by the baseline level of the respective group in period=0 summed with the time fixed effects (for positive exposure $\frac{\hat{\beta_t^+}}{E[Y_{it}|t=0,D_i^+]+\hat{\delta_t}}$, and for negative exposure $\frac{\hat{\beta_t^-}}{E[Y_{it}|t=0,D_i^-]+\hat{\delta_t}}$) . In Figure A6, we show our estimates normalized.

To test this hypothesis, we make comparisons among individuals who did not have changes in their credit scores between positive and negative systems. We illustrate this in the diagram below. Individuals in the blue area have high credit scores, and the lender's posterior about their true creditworthiness should increase with a more precise measure. The opposite holds for those in the red area. Those in the orange area should not observe changes in their creditworthiness assessment.

Diagram of the Empirical Strategy to Estimate the Effects of Changes in Signals' Precision



We formally describe our exercise. First, we restrict our sample to individuals with $\Delta_i \in (-0.25, 0.25)$. We then define the following three groups

$$D_i^+ = \mathcal{I}\Big[s_i > 1\Big] \qquad D_i^- = \mathcal{I}\Big[s_i < -1\Big] \quad D_i^0 = \mathcal{I}\Big[s_i \in [-0.25, 0.25]\Big]$$

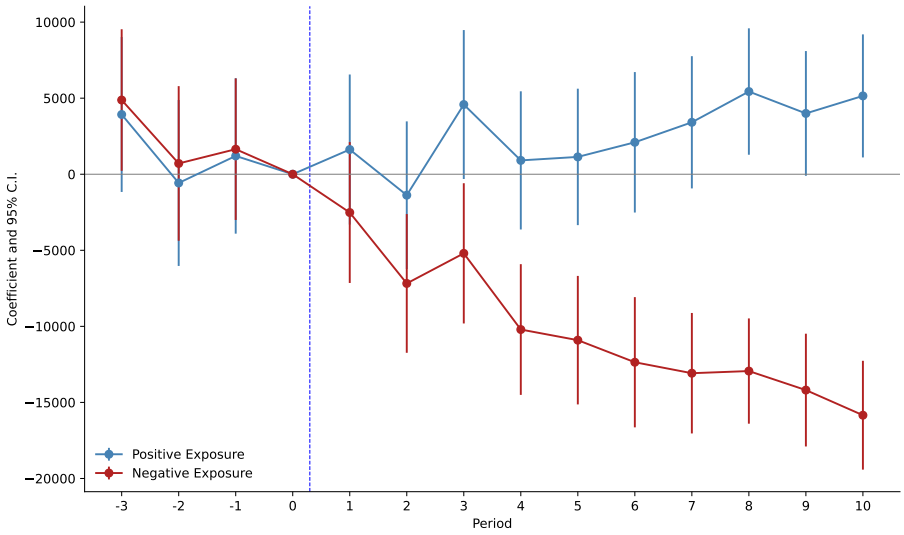
We restrict the sample for those individuals lying in one of the three groups (i.e., $D_i^+ = 1 | D_i^- = 1 | D_i^0 = 1$) and estimate the following difference in differences model.

$$Y_{it} = \alpha_i + \delta_t + \sum_{t \in T} \beta_t^+ \cdot D_i^+ \cdot \delta_t + \sum_{t \in T} \beta_t^- \cdot D_i^- \cdot \delta_t + \varepsilon_{it} \tag{2}$$

where again α_i are individual fixed effects and δ_t are time dummies.

Our estimates are of $\{\beta_t^+, \beta_t^-\}$ are shown in Figure 4.

Figure 4: Effects of Signal Value Change on Credit Access



This Figure shows how credit changes for individuals with small changes in credit scores, but who were far from the population average. The connected dots plot the $\{\beta_t^+, \beta_t^-\}$ estimates from equation 2 and their respective 95% confidence intervals. Sample is restricted to individuals with $\Delta_i \in [-0.25, 0.25]$. Positive exposure are individuals with $s_i > 1$ and negative exposure corresponds to individuals with $s_i < -1$. Standard errors are clustered at the individual level.

We observe that individuals with no substantial change in their credit scores are also affected by the policy. Those who, before and after the information revelation had credit scores above the population average, present an increase of around 5 thousand BRL two years after the policy. Conversely, individuals who, before and after the policy had credit scores below the population average, had a decrease of around 12 thousand BRL in their credit two years after the policy.

These findings are consistent with our framework of lenders' decision-making under imperfect information. The lender should positively update their creditworthiness assessment for those with above-average credit scores as the signal becomes more credible. The opposite holds for those with below-average credit scores.

5.2.C Effects over the Joint Distribution of Credit Scores

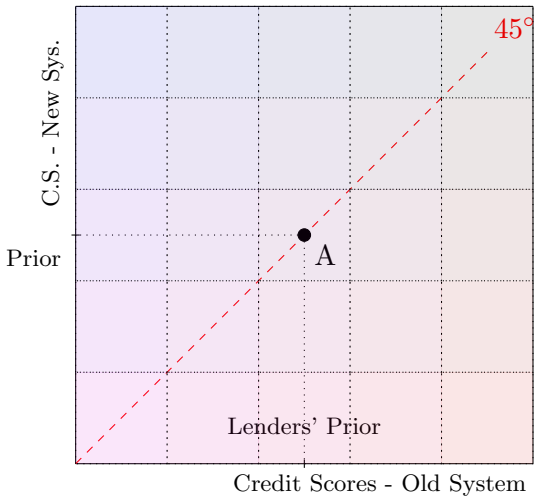
The previous exercises are helpful as they isolate the two channels through which additional information affects individuals' credit access. However, they are restricted to specific parts of the joint distribution of credit scores. In what follows, we show how credit changes over the whole joint distribution of credit scores by estimating the function $h(s_i, s_i')$ over the full set of $s_i \times s_i'$. This allows us to test our conceptual framework's predictions and estimate

and compare the effects of increasing precision and changes in signal value across the joint distribution.

We start with a semi-parametric approach. We divide the joint distribution of credit scores into a grid of equally sized bins, defined by their credit score values in the old and new systems. We then estimate the average credit access change for each group.

The diagram below provides intuition for our bins. Vertical dashed lines represent the division between groups in the old system, and horizontal dashed lines the division into groups of the new credit score system. Each square represents one of the groups used in the estimation.

Diagram of the Empirical Strategy to Estimate the Effects over the Joint Distribution of New and Old Credit Scores



Formally, the sample is divided into five groups of equal size based on the z-score of credit scores in the old system, with groups defined by $k \in \mathcal{K} = \{1, 2, 3, 4, 5\}$. $D_i^k = 1$ if individual i belongs to the group k. Similarly, the sample is divided into equally sized groups based on the z-score of credit scores in the positive system, with groups defined by $j \in \mathcal{J} = \{1, 2, 3, 4, 5\}$ and $D_i^j = 1$ if individual i belongs to the group j. Each group in $\mathcal{K}$ ranges over 0.6 standard deviations of the old system credit score (i.e., $(\max s_i \in k) - (\min s_i \in k) = 0.6$) and Group k=3 is centered around zero. Groups in $\mathcal{J}$ range over 0.8 standard deviations of the new system credit score, and j=3 is also centered around zero. Ranges are slightly different between both systems because the distribution of credit scores is more spread in the new system than in the old one, as we show in Figure A4.[20]

[20] Although rather arbitrary, our choices in the definition of groups are relatively innocuous in terms of our findings, as shown below by our linear and non-parametric assumptions defined below.

The interaction $D_i^k \cdot D_i^j = 1$ defines 25 groups across the joint distribution of credit scores. This is equivalent to approximating the function that determines changes in credit as follows:

$$h(s_i, s_i') \approx h_{sp}(s_i, s_i') = \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \beta^{kj} \cdot D^j(s_i') \cdot D^k(s_i)$$

We estimate $\{\beta^{kj}\}$ with OLS through the following equation:

$$Y_{it} = \alpha_i + \delta_t + \sum_{k \in \mathcal{K}} \sum_{j \in \mathcal{J}} \beta^{kj} \cdot D_i^j \cdot D_i^k \cdot Post_t + \varepsilon_{it} \tag{3}$$

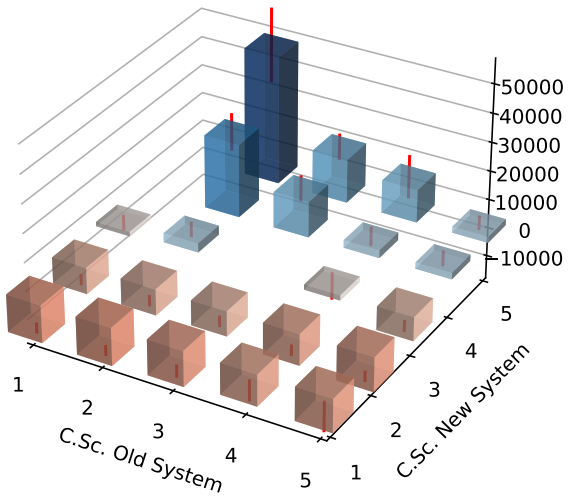
where once more α_i are individual fixed effects, δ_t are time dummies and $Post_t$ is an indicator function that takes value one in periods after the policy ($t > 0$).

In our estimation, we omit $D^3 \cdot D^3$, the group for which both the effects of the signal's value and the effects of the signal's precision are 0 according to our framework. Therefore, coefficients $\{\beta^{kj}\}$ correspond to difference in differences estimates of the change in credit access caused by the policy for individuals in the joint group defined by $D^k \cdot D^j$. Our identifying assumption is that credit in these groups would follow trends parallel to the one of the group $D^3 \cdot D^3$ in the absence of the policy.[21]

We present our estimates in Figure 5. The 3-D visualization is helpful for the interpretation of our findings. Each bar corresponds to a given coefficient, with 95% confidence intervals plotted in the red lines. Bars are organized such that the x-axis (labeled C. Sc. old system) indexes coefficients for a given group k, and the y-axis (labeled C.Sc. new system) indexes coefficients for a given group j. Thus, for example, the bar plotted in C. Sc. old system k= 2, and C. Sc. new system j= 5, corresponds to $\beta^{2,5}$, the difference in differences estimate of the change in credit for individuals who were in group 2 of credit scores in the old system, and 5 with credit scores under the new system. Figure A7 shows the same plot from different angles. Figure A8 shows a 2-dimensional visualization of the estimates in a heatmap.

[21] This implies as well that we restrict the function $h_{sp}(s_i, s_i')$, imposing $h_{sp}(s_i, s_{i'}) = 0$ if $s_i \in 3$ and $s_i' \in 3$

Figure 5: Estimates of Change in Credit over the Joint Distribution of Credit Scores



This Figure shows our estimates of changes in credit over the joint distribution of credit scores, i.e., the estimates of coefficients $\{\beta^{kj}\}$ from equation 3. Each bar corresponds to a given coefficient, with 95% confidence intervals plotted in the red lines. Standard Errors are clustered at the individual level. Bars are organized such that the x-axis (labeled C. Sc. old system) indexes coefficients for a given group k, and the y-axis (labeled C.Sc. new system) indexes coefficients for a given group j. Positive estimates of β^{kj} are shown in blue, whereas negative estimates are shown in red. $\beta^{14}, \beta 15$ are not defined because there is no individual in the sample in those groups of the joint distribution of credit scores.

Our estimates show that credit changes over the joint distribution are consistent with our conceptual framework. First, we can compare estimates *vertically*. This implies comparing individuals with similar levels of the old system credit scores and different values of new system credit scores. We observe that changes in credit are increasing as the groups of new system credit scores increase. We can also compare our estimates *horizontally*. These comparisons are across individuals with similar levels of new system credit scores but different levels in the old system. Our estimates suggest that changes in credit decrease horizontally. This suggests a *catch-up* of groups towards their new creditworthiness assessment.

Lastly, we can compare estimates *diagonally*. By doing so, we are comparing changes in credit across groups that had similar changes in the value of the credit score. Our estimates increase as individuals go from lower to higher credit scores, consistent with our proposition outlined in the conceptual framework. To further visualize these comparisons, Figure A9 plots linear fits between coefficients over *vertical*, *horizontal*, and *diagonal* comparisons, and Table A2 shows the estimates of linear fits between coefficients.

To pin down values for the *vertical*, *horizontal*, and *diagonal* comparisons, we make parametric assumptions. In particular, we assume write changes in credit as a linear function of these signals: $h_l(s_i, s_i') = a \cdot s_i + b \cdot s_i'$. We estimate $h_l(s_i, s_i')$ with OLS through the following equation:

$$Y_{it} = \alpha_i + \delta_t + \beta_0 \cdot \text{C. Sc. Old sys.}_i \cdot \text{Post}_t + \beta_1 \cdot \text{C. Sc. New sys.}_i \cdot \text{Post}_t + \varepsilon_{it} \quad (4)$$

where C. Sc. Old sys.$_i$, C. Sc. New sys.$_i$ are the Z-score of credit scores under the old and new systems calculated in the last period before the implementation of the policy.

Our findings are presented in Figure 6. On the left-hand side, the table shows the estimated coefficient and standard deviation of β_0, β_1. On the right-hand-side, the figure shows the predicted changes in credit which consists of $\hat{y}_i = \hat{\beta}_0 \cdot \text{C. Sc. Old sys.}_i \cdot \text{Post}_t + \hat{\beta}_1 \cdot \text{C. Sc. New sys.}_i$.

The linear model reveals that the effects of signal value include a 10.95 thousand Reais increase in credit for each standard deviation increase in positive system credit scores, which represents a 20% increase relative to the average individual in the sample. In terms of precision effects, there is a 5.78 thousand Reais increase in credit with a one standard deviation increase in both positive and negative system credit scores, equating to a 10% increase relative to the average individual in the sample.

In Appendix D, we relax the linearity assumption and estimate the function $h(s_i, s_i')$ non-parametrically with sieve-estimators. This allows us to further test the implications of our conceptual framework and evaluate the quality of the linear fit. We show that the non-parametric estimates propose a qualitatively similar function of changes in credit over the joint distribution of credit scores. Furthermore, average partial derivatives of the non-parametric estimates also suggest the *vertical*, *horizontal*, and *diagonal* patterns described above when comparing changes in credit across groups. Lastly, when calculating partial derivatives at each point of the joint distribution of credit scores, we show that in the majority of the space spanned by $s_i \times s_i'$, the direction of partial derivatives correspond to the patterns predicted by our conceptual framework.

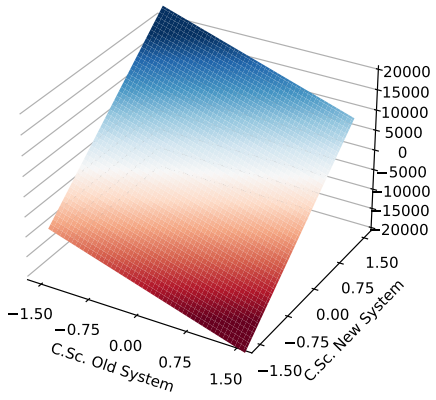
5.3 Who gets more affected and distributional consequences

We next investigate the impacts of revealing information about borrowers on credit inequality and calculate differences in the effects of the policy across demographic groups.

Let C_{i0}, C_{i1} define credit access without and with the policy. Our goal is to compare

Figure 6: Estimates of Changes in Credit over the Joint Distribution with Linear $h(s_i, s_i')$

	(1) Credit
C. Sc. Old Sys.	- 5,172.3
	(273.66)
C. Sc. New Sys.	10,951.58
	(370.70)
Observations	2875942



This Figure shows the estimates of changes in credit with linear restriction. In the Table on the left-hand side, we show our estimates for the coefficient and the standard deviation of β_0, β_1 from equation 4. On the right-hand-side, the figure shows the predicted changes in credit which consists of $\hat{y}_i = \hat{\beta_0} \cdot \text{C. Sc. Old sys.}_i \cdot \text{Post}_t + \hat{\beta_1} \cdot \text{C. Sc. New sys.}_i$.

$\mathcal{G}_0(C_0)$ and $\mathcal{G}_1(C_1)$, which represent the counterfactual distribution of credit access in the absence of the policy and the distribution of credit under the actual policy. Using the estimates of the effects of the policy on credit access described above, we first compute for each individual their credit access with and without the policy at any given period as:

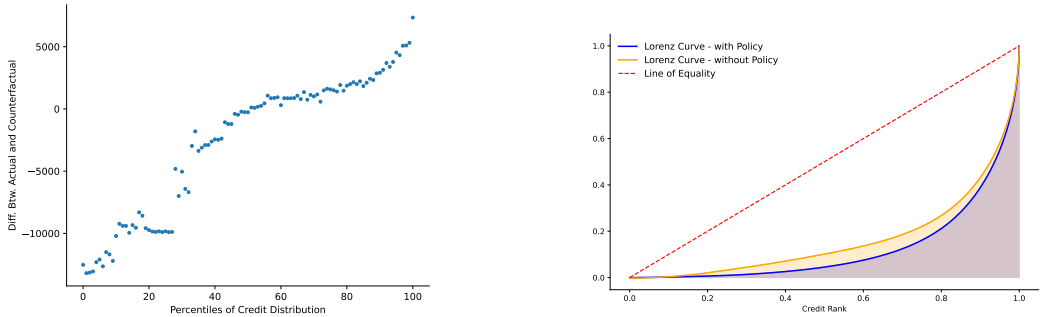
$$C_{it0} = \hat{\alpha_i} + \hat{\delta}_t$$
$$C_{it1} = \hat{\alpha_i} + \hat{\delta}_t + \hat{\beta}^{kj} \cdot D_i^k \cdot D_i^j \cdot Post_t$$

For periods before the policy implementation, we see that $C_{it0} = C_{it1}$. Thus, we focus only on periods after the policy implementation and use the average value of C_{it0}, C_{it1} across periods for each individual as their credit access with and without the policy (C_{i0}, C_{i1}).

Different statistics allow us to characterize the differences between both distributions. First, we calculate the variance of both distributions. We find that under the policy, variance of the distribution of credit is 2.5% higher than without the policy[22]. We then conduct two different exercises with results shown in Figure 7. In Panel (a), we plot the average difference $C_{i1} - C_{i0}$ across percentiles of credit distribution. In Panel (b), we show Lorenz curves of the actual and counterfactual distributions of credit.

[22] $$\frac{\text{Var}(G_1(C_1)) - \text{Var}(G_0(C_0))}{\text{Var}(G_0(C_0))} = 0.025$$

Figure 7: Comparisons between Distributions of Credit



(a) Difference in Credit across Distribution

(b) Lorenz Curve of Credit Access

This Figure shows comparisons between distributions of credit with and without the policy $\mathcal{G}_0(C_0)$ and $\mathcal{G}_1(C_1)$. In Panel (a), we plot the average difference $C_{i1} - C_{i0}$ across percentiles of credit distribution. In Panel (b), we show Lorenz curves of the actual and counterfactual credit distributions.

We observe that changes in credit are almost monotonically increasing on the percentiles of credit distribution. The Lorenz curve shows that the distribution of credit without the policy is closer to the line of equality, indicating a more unequal credit distribution with the policy.

Comparisons by Demographic Characteristics: We can also characterize how the policy affected individuals differently according to their demographic characteristics. Our analysis is based on two dimensions in an Oaxaca-Blinder spirit. First, estimated treatment effects can be different according to each group. Second, the distribution of groups can vary across the joint distribution of credit scores. Thus, even if effects are homogeneous across different groups, the policy can have different aggregate effects for different groups due to the composition component.

Empirical Strategy: Let $G_i \in \{0, 1\}$ represent a binary variable that defines a specific demographic group, such as male vs. female or white vs. nonwhite. Additionally, C_{i0} and C_{i1} still represent individual credit access before and after the implementation of the policy, respectively.

We are interested in comparing the effect of the policy on credit access across different demographic groups. Specifically, we analyze the difference in the expected change in credit access between two groups. Formally, we make comparisons of the following type:

$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 1] - \mathbb{E}[C_{i1} - C_{i0}|G_i = 0].$$

To estimate the differences in credit access before and after the policy for each group,

we rely on our semi-parametric approach, exploring the 5 groups of the old credit score $\mathcal{K} \in \{1,2,3,4,5\}$ and the 5 groups of the new system credit scores $\mathcal{J} \in \{1,2,3,4,5\}$. This generates the following estimating equation:

$$Y_{it} = \alpha_i + \delta_t + \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} \beta_{g0}^{kj} \cdot D_i^k \cdot D_i^j \cdot \text{Post}_t + \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} \beta_{g1}^{kj} \cdot D_i^k \cdot D_i^j \cdot \text{Post}_t \cdot G_i + \varepsilon_{it}. \quad (5)$$

where α_i represents individual-specific effects, δ_t represents time-specific effects, D_i^k and D_i^j represent credit score categories, and Post_t is a dummy variable indicating the post-policy period. The coefficients β_{g0}^{kj} capture the impact of the policy for the baseline group, while the interaction terms $\beta_{g1}^{kj} \cdot G_i$ capture the differential effect for group $G_i = 1$.

Given these estimates, we define the expected change in credit access for each group as:

$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 0, D_i^k = 1, D_i^j = 1] = \hat{\beta}_{g0}^{kj}$$
$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 1, D_i^k = 1, D_i^j = 1] = \hat{\beta}_{g0}^{kj} + \hat{\beta}_{g1}^{kj}$$

This allows us to express the expected change in credit access for each group as follows:

$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 0] = \mathbb{E}[\hat{\beta}_{g0}^{kj}]$$
$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 1] = \mathbb{E}[\hat{\beta}_{g0}^{kj} + \hat{\beta}_{g1}^{kj}]$$

To calculate the value of expected changes by group we weight the coefficient change by the amount of people of the given group in each of the points of our grid that divides the joint distribution of credit scores. This is equivalent to:

$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 0] = \mathbb{E}[\hat{\beta}_{g0}^{kj}] = \frac{\sum_i (1 - G_i) \cdot D_i^k \cdot D_i^j \hat{\beta}_{g0}^{kj}}{\sum_i (1 - G_i)}$$
$$\mathbb{E}[C_{i1} - C_{i0}|G_i = 1] = \mathbb{E}[\hat{\beta}_{g0}^{kj} + \hat{\beta}_{g1}^{kj}] = \frac{\sum_i G_i \cdot D_i^k \cdot D_i^j (\hat{\beta}_{g0}^{kj} + \hat{\beta}_{g1}^{kj})}{\sum_i G_i}$$

Thus, we can write the difference in policy effects between the two groups as:

$$\begin{aligned}\mathbb{E}[C_{i1} - C_{i0}|G_i = 1] - \mathbb{E}[C_{i1} - C_{i0}|G_i = 0] = &\frac{\sum_i G_i \cdot D_i^k \cdot D_i^j (\hat{\beta}_{g0}^{kj} + \hat{\beta}_{g1}^{kj})}{\sum_i G_i} \\ &- \frac{\sum_i (1 - G_i) \cdot D_i^k \cdot D_i^j \hat{\beta}_{g0}^{kj}}{\sum_i (1 - G_i)}\end{aligned} \quad (6)$$

The expression suggests that to understand which group is more affected by the policy,

Table 2: Effects on Credit Differences Across Groups

	(1) Composition	(2) Total
Gender		
Women - Men	317.88	-745.18
Race		
Nonwhite - White	-2034.85	-4629.11
Education		
High School - Less than H.S.	1167.26	2659.22
Some College - Less than H.S.	6089.81	17462.14
Age		
$<$ 40 y.o. - $\geq$ 40 y.o.	-3002.64	-1888.03

This Table shows our estimates of how difference in credit access between groups change due to the policy. Values correspond to our estimates of equation 6 with different comparisons in gender, race, education, and age. When looking at education, we add a second set of coefficients β_{g2}^{kj} and estimate equation 5 using the full sample. Column (1) shows the compositional role on the total effect by fixing treatment effects for the average population estimate (β^{kj} from equation 3.

we must consider two components. First, the groups are potentially differently distributed across the joint distribution of credit scores. Second, the heterogeneous treatment effects, captured by the values of $\hat{\beta}_{g1}^{kj}, \hat{\beta}_{g0}^{kj}$.

To disentangle between both effects, we can also make an Oaxaca-Blinder type decomposition, replacing the coefficients above by $\hat{\beta}^{kj}$ from equation 3. This would give us just the effects of the differential composition of individuals by group. [23]

Results: We focus our analysis on comparisons across four key demographic dimensions: gender, race, education, and age. Table 2 summarizes our findings. In Column (1), we show the difference with homogeneous treatment effects, while in column (2), we show the differences allowing for heterogeneous treatment effects by group (i.e., equation 6)

We observe no substantial difference in credit access by gender. When looking only at the composition side, women have a slightly higher credit change than men. This implies that women are slightly overrepresented among groups with positive changes in credit access.

[23] This essentially means calculating the following formula

$$\frac{\sum_i G_i \cdot D_i^k \cdot D_i^j(\hat{\beta}^{kj})}{\sum_i G_i \cdot D_i^k \cdot D_i^j} - \frac{\sum_i (1-G_i) \cdot D_i^k \cdot D_i^j \hat{\beta}^{kj}}{\sum_i (1-G_i) \cdot D_i^k \cdot D_i^j}$$

where $\hat{\beta}^{kj}$ are the coefficients estimated from equation 3 and presented in Figure 5

However, this flips when observing the Total Effects. This is because the estimates of credit increases are slightly lower for positively exposed women than men. This can be seen in Panel (b) of Figure A10, where we show our estimates of β_{g1}^{kj}.

When comparing effects by race, we observe that the policy enhanced race inequality in credit. This is due both to compositional effects, i.e. nonwhite individuals are over-represented in negatively exposed parts of the joint distribution of credit scores, and to heterogeneous treatment effects. In panel (c) of Figure A11, we illustrate the compositional aspect by showing the share of nonwhite individuals in each of the 23 groups of the joint distribution. We observe that they are overrepresented at groups with lower values of credit scores. At the same time, by looking at panel (b), we see that increases in credit are substantially lower for nonwhite individuals than for white ones. Thus, through both compositional and treatment effects, our estimates suggest that the policy increased the racial inequality in credit access.[24]

These findings are a counterpoint to Blattner and Nelson (2021), who argue that reducing the gap in credit score noise between white and black individuals in the U.S. could substantially reduce the racial inequality in mortgage markets. Our *policy* differs from theirs as it is theoretically race-neutral (borrowers' information increases for both white and nonwhite individuals). In our case, the revelation of information amplifies overall credit inequality. This disproportionately affects nonwhite and less educated individuals as they are disproportionately represented in the less favored parts of the joint distribution of credit scores. Thus, race-neutral increases in information can increase racial inequality in credit access.

6 Effects on Default Rates

In this section, we investigate the implications of the revelation of information about borrowers on the quality of loans. So far, we have demonstrated that credit is reallocated along the joint distribution of credit scores. However, we have not yet addressed the outcomes resulting from these changes. Next, we will explore whether the policy affected the quality of credit supplied by examining the key question: did revealing information reallocate credit

[24]The same pattern occurs when looking at education levels. Individuals with less than high school education have a decrease in credit relative to more educated groups. Both results in the table and in Figure A12 show that composition and treatment effects play a role in this. When looking at age, we observe that older individuals benefit more from the policy, despite treatment effects being more favorable to younger individuals, as we can see in Figure A14.For visualization purposes in case the 3D graphs are not straightforward, Figure A15 shows heatmaps of the distribution of demographic characteristics across the joint distributions of credit scores.

to more or less risky loans? To investigate this, we will delve into defaults, commonly used in the literature as an approximation for the costs of providing loans to lenders (Liberman et al., 2018; DeFusco et al., 2022).

The ideal empirical setting to analyze default rates would involve observing loan-specific defaults to estimate the effects on the defaults of new loans. In our approach, we approximate default rates by calculating the ratio between an individual's total *financial delinquency* and their total amount of credit. We start by estimating the effects of the policy on overall financial delinquency. Next, we assess how default rates vary across the joint distribution of credit. We then calculate the default rates for credit reallocated by the policy. Finally, we compare the default rates of credit increases to those of credit decreases.

6.1 Effects on Financial Delinquency

We begin our analysis by estimating the effects on total financial delinquency, which refers to the amount reported by financial institutions to credit bureaus when individuals fail to meet their financial obligations on time. Typically, the lender issues a notice regarding the debt and sets a deadline for regularization. If the deadline is not met, the overdue debt may be marked as delinquent and reported to the credit bureau.

We proceed with the same research design as used in the previous section, estimating the effects of the signal's value (equation 1), the effects of the signal's precision (equation 2), and the effects over the joint distribution of credit scores (equation 3). Now, instead of total credit at a given period as the outcome, we use the total amount of financial delinquency at a given period. This includes 0s whenever individual i at a given period t does not have any financial delinquency reported. Figure 8 shows our findings for the three different empirical strategies[25].

Panel (a) shows difference in differences estimates comparing over time, financial delinquency of individuals with credit score increases and decreases with those that did not have substantial changes in their scores with the policy. We observe that prior to the policy, there are no differences in the trend of financial delinquency across groups, which we take as evidence that our identifying assumption holds. After the policy, we observe that those who were positively affected, increase their total financial delinquency by almost 3 thousand BRL 2 years after the policy. On the other hand, negatively exposed individuals decrease financial delinquency by around 1 thousand BRL.

We find similar results when looking at individuals who did not change their credit

[25]Figure A16 shows a 2-dimensional visualization in a heatmap of the estimates from Panel (c) of Figure 8.

scores but were distant from the population average. Those estimates are plotted in Panel (b). We see that positively exposed individuals had on average 1800 BRL more in financial delinquency two years after the policy. We observe no significant change for those negatively affected.

When looking at the effects of the policy on financial delinquency over the joint distribution of credit scores, we observe that most of the changes in financial delinquency are concentrated at the lower part of the distribution of the old system credit scores. Although the patterns are less clear than in the credit results, our estimates also suggest the *vertical*, *horizontal* and *diagonal* patterns. Financial delinquency increases when comparing changes across groups *vertically*, except for those in the first group of old system credit scores. Estimates on average also increase *diagonally*, which implies that for groups with similar differences between credit scores in both systems, delinquency changes increase along the distribution of credit scores. When comparing estimates *horizontally*, we observe slightly different results. For levels below the average of new system credit scores, estimates are increasing horizontally.[26]

6.2 Default Rates

Financial delinquency tends to rise among groups that experience an increase in credit access. However, as the availability of credit rises, the amount of money over which an individual can be delinquent increases as well. A more policy-relevant question in this context is to determine the proportion of the credit extended that ultimately results in financial delinquency. To understand this, we first calculate average default rates over the joint distribution of credit scores. We define default rates as the ratio between the financial delinquency of an individual at a given period and their total credit at the same period:

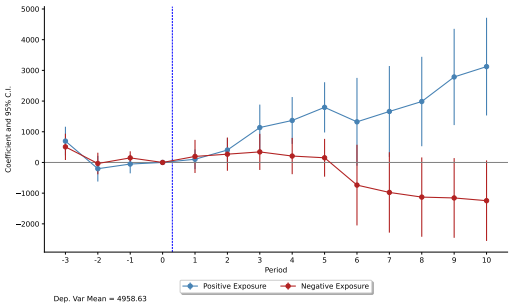
$$\text{Default Rate} = \frac{\text{Financial Delinquency}}{\text{Credit}}$$

Since both financial delinquency and credit endogenously change with the information revealed by the policy, we need to be careful in estimating the effects on changes in Default Rates. In particular, since the denominator can be zero for a given individual, that generates cases where default rates are not defined.

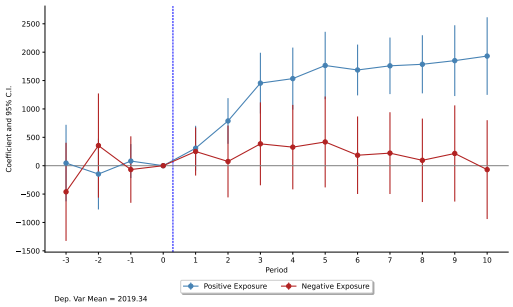
We overcome this using our estimates of financial delinquency and credit access separately to construct Default Rates in the presence and in the absence of the policy. It is

[26] Similarly to our analysis on credit, Figure A17 plots linear fits between coefficients over *vertical*, *horizontal*, and *diagonal* comparisons, and Table A3 shows the estimates of linear fits between coefficients.

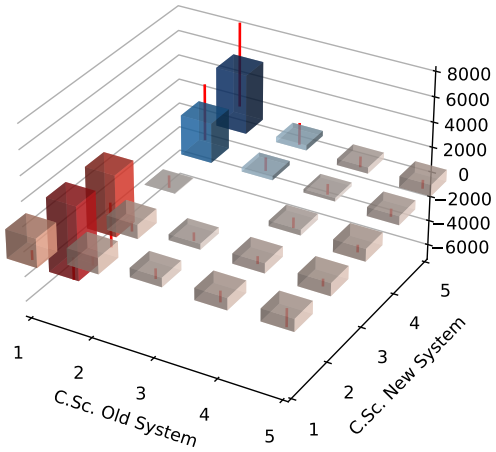
Figure 8: Effects on Financial Delinquency



(a) Signal Value Effects



(b) Precision Effects



(c) Effects over the Joint Distribution

This Figure shows estimates of the effect of the policy on Total Financial Delinquency. Individuals with no financial delinquency at a given period are included and have $Y_{it} = 0$. Panel (a) estimates the effects of the signal value, comparing individuals with $\Delta_i \in (-0.25, 0.25)$ with those positively affected $\Delta_i \in (1, 1.5)$ and negatively affected $\Delta_i \in (-1.5, -1)$. The sample is restricted to those three groups. Panel (b) estimates the effects of the signals' precision, comparing individuals with $s_i \in (-0.25, 0.25)$ with those positively affected $s_i > 1$ and negatively affected $s_i < -1$. The sample is restricted to individuals in those three groups and $\Delta_i \in (-0.25, 0.25)$. Panel (c) shows estimates of equation 3, using the full sample of individuals.

useful to write down both objects under a Potential Outcomes notation as follows:

$$DR^0_{it} = \text{Default Rate}_{it}|\text{no policy} \approx \frac{E[\text{Financial Delinquency}_{it}|\text{no policy}]}{E[\text{Credit}_{it}|\text{no policy}]}$$

$$DR^1_{it} = \text{Default Rates}_{it}|\text{with policy} \approx \frac{E[\text{Financial Delinquency}_{it}|\text{with policy}]}{E[\text{Credit}_{it}|\text{with policy}]}$$

We use our estimates of the policy effects on financial delinquency and credit from our semi-parametric approach over the full joint distribution of credit scores to construct the expected values of both measures with and without the policy.[27]

For each outcome, the corresponding expectations with and without the policy are calculated as follows:

$$\begin{aligned} E[Y_{it}|\text{no policy}] &= \hat{\alpha}_i + \hat{\delta}_t \\ E[Y_{it}|\text{with policy}] &= \hat{\alpha}_i + \hat{\delta}_t + \textstyle\sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} \hat{\beta^{kj}} \cdot D^k_i \cdot D^j_i \cdot Post_t \end{aligned}$$

To visualize how default rates change with the policy, we aggregate them to the groups defined over the joint distribution of credit $D^k(s_i)D^j(s'_i)$. To do so, we take the average of $E[Y_{it}|\text{no policy}]$ and $E[Y_{it}|\text{with policy}]$ in each group across periods after the implementation of the policy, and calculate the ratio between expected values for financial delinquency and credit.[28]

We show our findings for default rates in Figure 9. In panel (a), we plot the default rates of each group in the presence of the policy, whereas panel (b) shows counterfactual default rates in its absence.[29] We can look at the patterns similarly to our analysis on credit. In both Panels (a) and (b), default rates are increasing *vertically* and decreasing *horizontally*. *Vertically*, we see that they are substantially higher for groups with lower credit scores in the new system (group 1), and decrease as the new system credit scores increase for all initial levels of old system credit scores. At the same time, *horizontally* we observe that for a fixed group in the new system credit score, default rates are decreasing as we move from worse to better groups of the old system credit scores. In turn, Default Rates decrease as we move across groups *diagonally*, suggesting that, for the same difference in credit scores across the new and old systems, individuals with higher credit scores individuals have lower default rates. When looking at the differences in default rates between actual and counterfactual

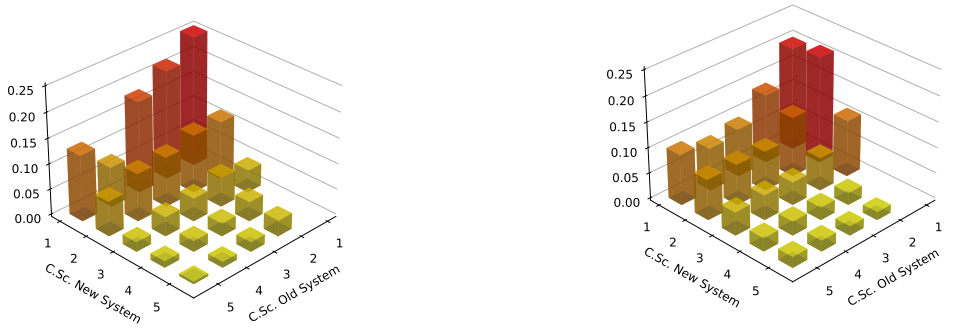
[27] That is, we estimate for both outcomes equation 3, which corresponds to $Y_{it} = \alpha_i + \delta_t + \sum_{k\in\mathcal{K}}\sum_{j\in\mathcal{J}} \beta_{kj} \cdot D^k_i \cdot D^j_i \cdot \text{Post}_t + \varepsilon_{it}$

[28] In periods before the policy $E[Y_{it}|\text{with policy}] = E[Y_{it}|\text{no policy}]$ since $Post_t = 0$

[29] We change the angle from the previous Figures for better visualization, but Figure A20 shows the same estimates in the previous angle.

estimates, we do not observe a clear pattern across the distribution. These results are shown in Figure A19.

Figure 9: Default Rates Across the Joint Distribution of Credit Scores



(a) Default Rates Post Policy

(b) Counterfactual Default Rates in the Absence of the Policy

This Figure shows average Default Rates in the periods after the policy. Panel (a) plots our estimates of default rates in the presence of the policy, whereas Panel (b) plots our estimates of Default rates in the absence of the Policy.

6.3 Default Rates of Reallocated Credit

Having established our definition of default rates, and shown how they vary over the joint distribution of credit scores, we can estimate default rates of the reallocated credit. Under the assumption that the Cadastro Positivo did not affect default rates of credit that would have been observed in the absence of the policy, we can write the default rate of the marginal credit as the ratio between the change in financial delinquency and changes in credit(Angrist and Imbens, 1995)[30]. To show this, it is useful to introduce more potential outcomes notation. C_i^0, C_i^1 are still representing the credit access for individual i in the absence of the policy in its presence. DR_i^1, DR_i^0 are the default rates in the presence and absence of the policy, respectively.

We can write $C_i^1 \cdot DR_i^1 = (C_i^1 + C_i^0 - C_i^0) \cdot DR_i^1$, by adding and subtracting C_i^0 in the right hand side. We define the object $\Delta C_i = C_i^1 - C_i^0$, the change in credit caused by the policy. Thus $C_i^1 \cdot DR_i^1 = (C_i^0 + \Delta C_i) \cdot DR_i^1$. Our assumption implies that among the credit given with the policy, the C_i^0 part would have default rate DR_i^0, whereas the

[30]This is the same assumption as in Card and Hyslop (2005) and subsequent papers that evaluate outcomes for marginally treated observations (compliers) in the absence of an empirical setting that allows more precise identification of outcome tests.

marginal credit (ΔC_i) would have their own default rate DR^{Re}. Thus, we can rewrite $C_i^1 \cdot DR_i^1 = C_i^0 \cdot DR_i^0 + \Delta C_i DR^{Re}$. From this, it is straightforward to see that

$$DR^{Re} = \frac{\mathbb{E}[C_i^1 \cdot DR_i^1 - C_i^0 \cdot DR_i^0]}{\mathbb{E}[\Delta C_i]} = \frac{\mathbb{E}[\Delta Delinquency_i]}{\mathbb{E}[\Delta C_i]}$$

Thus, having consistent estimates of the effects of the policy on financial delinquency and credit allows us to calculate the default rate of the marginal credit. We use our estimated coefficients from our semi-parametric analysis on delinquency and credit as approximations of the effects for each individual. We can write this as:

$$\Delta \text{Delinquency}_i = \mathbb{E}[\Delta \text{Delinquency}_i | D_i^k = 1, D_i^j = 1] = \hat{\beta_{Del}^{kj}}$$

$$\Delta \text{C}_i = \mathbb{E}[\Delta \text{Credit}_i | D_i^k = 1, D_i^j = 1] = \hat{\beta_{C}^{kj}}$$

where $\hat{\beta}_{Del}^{kj}, \hat{\beta}_{C}^{kj}$ are the estimated coefficients of equation 3 using financial delinquency and credit as outcomes respectively, and D_i^k, D_i^j are dummies indicating that individual i belongs to groups k, j of the new and old system credit scores.

To compare default rates over credit that was given because of the policy, counterfactual credit that would have been given in its absence, but was not given because of the policy, we divide our sample into those with positive change in credit and those with negative change in credit. We then compute the average changes in financial delinquency and credit weighted by the number of observations in each bin of joint distribution of credit scores. This can be written as:

$$\begin{aligned} Positive &= \frac{\sum_i D_i^k \cdot D_i^j \cdot \hat{\beta_{Del}^{kj}}}{\sum_i D_i^k \cdot D_i^j \cdot \hat{\beta_{C}^{kj}}} \quad \text{for k, j with} \quad \mathbb{E}[\Delta \text{Credit} | D_i^k = D_i^j = 1] > 0 \\ Negative &= \frac{\sum_i D_i^k \cdot D_i^j \cdot \hat{\beta_{Del}^{kj}}}{\sum_i D_i^k \cdot D_i^j \cdot \hat{\beta_{C}^{kj}}} \quad \text{for k, j with} \quad \mathbb{E}[\Delta \text{Credit} | D_i^k = D_i^j = 1] < 0 \end{aligned} \tag{7}$$

where *Positive* shows our estimates of the default rates of credit that was observed because of the policy, and *Negative* shows our estimates of the default rate of credit that would have been observed in the absence of the policy, but was not observed because of the policy.

We summarize our findings about default rates in Table 3. First, we see in column (1) that the default rate of credit that was positively reallocated was around 12p.p. lower compared to the credit that would have been given in the absence of the policy.

Table 3: Default Rates of Credit Reallocated with the Policy

	(1) Default Rate Reallocated Credit
Default Rate of Credit that was allocated because of the Policy	0.03
Default Rate of Credit that was not allocated because of the policy	0.15

This Table shows our estimates of the default rate of credit that was reallocated positively and negatively because of the policy. These estimates were calculated using our findings for changes in credit and financial delinquency over the joint distribution of credit scores using our expression in Equation 7.

Discussion on the Type of Selection in this Market

Classical discussions about the consequences of imperfect information in credit markets focus on the type of selection present in the market. Stiglitz and Weiss (1981) demonstrate that asymmetric information leads to adverse selection, where the marginal borrower is less risky than supra-marginal ones, resulting in credit rationing. In contrast, De Meza and Webb (1987) argue the opposite, showing conditions under which asymmetric information leads to advantageous selection, where the marginal borrower is less profitable than supra-marginal ones, resulting in overinvestment.

Recent papers have utilized default rates to interpret the type of selection in credit markets. Liberman et al. (2018); DeFusco et al. (2022); Jansen et al. (2022), borrowing insights from the insurance markets literature of Einav et al. (2010), examine how changes in default rates following credit expansion can serve as a measure of the quality of marginal borrowers, often finding evidence of adverse selection. If we interpret our results through their frameworks, our findings would imply advantageous selection in credit markets. This occurs because, on average, marginal credit is riskier than the credit that would have been extended without the policy. Liberman et al. (2018) specifically interpret absolute delinquency as their measure of default, and if we adopt their interpretation, our findings would suggest adverse selection in the credit market.

7 Effects on Entrepreneurial Activity

In this section, we examine if changes in credit allocation had consequences beyond credit market outcomes, focusing on Entrepreneurial activity. We begin by describing entrepreneurship in our context, with statistics on firms and entrepreneurs' characteristics for the uni-

verse of new formal firms in São Paulo. We then use our sample from the credit bureau matched with firm records to understand if the reallocation of credit across the population had consequences in firm creation (extensive margin of entrepreneurship) and firm quality (intensive margin). Lastly, we assess whether the reallocation of personal credit led to an improved allocation of resources to entrepreneurial activity.

Additional Sample: To gain precision in the analysis, we increase the sample of individuals by around 560 thousand randomly selected individuals from the same pool of adults in the state of Sao Paulo used in our credit market analysis. For these additional individuals, we only observe credit scores in the new and old systems for this additional sample. Thus, we could not obtain individual-level credit and default information for them (this is why this sample is not included in previous exercises). As expected, given that the sample is randomly selected, characteristics of both samples are extremely similar. In Table A4 we show summary statistics of both samples. We then match credit scores data with the universe of formal firm ownership in the State of São Paulo. Therefore, we observe if and when each of the individuals created their first firm. Using the firm identification number, we then match them to firm-level data provided by our partner institution. Appendix C fully details the sample construction procedures.

7.1 What are these firms?

Before we start our empirical analysis, describing the characteristics of firms and entrepreneurs we observe in our data is useful.

Our firm records are restricted to the formal sector, implying that informal firms are excluded from our analysis. In Appendix F, we detail the differences between informal and formal entrepreneurs using household survey data that encompasses both formal and informal sectors. We show that more than 90% of individuals that self-declare as *employers* have a formal business registration[31]. This share reduces to around 40% when we look at *self-employed* workers[32]. But high informality levels in the latter group are particularly

[31]These Figures are strikingly different than those in the well-known Ulyssea (2018) paper, whose analysis point to 70% of employing firms being informal. This is due to three reasons. First, we are looking specifically at the state of São Paulo, the country's richest state, which has substantially lower informality rates than the rest of the country. Second, we look at data 20 years after his analysis. Brazil had a large informality reduction during this period (Haanwinckel and Soares, 2021). Lastly, we use PNAD instead of ECINF. PNAD is supposed to be representative of the full workforce, whereas ECINF is a survey restricted to firms with up to five employees. As shown by Ulyssea (2018), there is a positive gradient between formalization and firm size.

[32]The informality rate of self-employed workers reduced substantially since 2010 because of a new tax system referred to as MEI, which was designed specifically for the formalization of these workers. See Hsu Rocha and de Farias (2021) for a complete description of the effects of these new tax systems on firm

driven by the construction and transportation sectors, with almost 80% of informality. In contrast, retail has, for example, less than 50% of informal self-employment.

The formal firms we observe in our sample encompass a diverse set of businesses. Most are small and medium enterprises concentrated in the retail and service sectors. Add Descriptive statistics of our sample and discuss them here

Despite not being likely to be *disruptive* entrepreneurs, these firms are still responsible for the majority of job creation. In Appendix E, we combine our entrepreneurship data with matched employer-employee data that covers the entirety of the formal labor market[33] and replicate exercises from Haltiwanger et al. (2013) and Decker et al. (2014) that show the importance of entrepreneurs in job creation in the U.S. We find that firms in their three first years are responsible for around 20% of new jobs and over 80% of *net job creation* highlighting the economic relevance of these new firms.

7.2 Effects on Firm Creation

Next, we investigate if the policy changed the probability that an individual creates a new business.

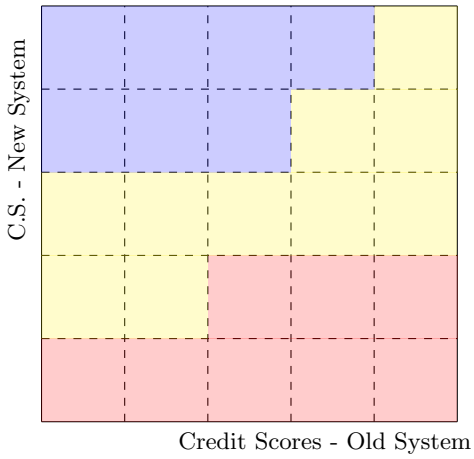
To empirically assess the effects of the policy on business creation, we slightly change our empirical strategy from the previous sections in two ways. First, we group individuals into three groups instead of using the 25 groups defined by the joint distribution of credit scores. We define the three groups from our estimates of the effects in credit access given the policy. The diagram of the joint distribution of credit scores illustrates our comparison groups. Those who lie in the blue area had an increase in their credit access, while those in the red area had a decrease in their credit access. Meanwhile, individuals in the yellow area are considered our control group, as our results show that they did not have substantial changes in credit[34].

creation and informality.

[33] We use the well-known RAIS data combined with our entrepreneurship records to do this exercise. Important to highlight that this was done outside of our partner institution environment, as we did not share RAIS information with them.

[34] Using our definitions of the 25 groups and the same notation as before, we can define the three groups as $D^{+} = \{D^{14}, D^{15}, D^{24}, D^{25}, D^{34}, D^{35}, D^{45}\}$ and $D_i^{-} = \{D^{11}, D^{21}, D^{22}, D^{31}, D^{32}, D^{41}, D^{42}, D^{51}, D^{52}\}$, where the first superscript refers to the group in the old system and the second the group in the new system of credit scores.

Diagram of Groups to be Compared in our Empirical Analysis



The second difference from the previous sections is that we move away from the OLS estimation of the difference in differences to a survival model, using hazard models with time-dependent covariates to estimate the probability of opening a business. We do so because individuals who establish a firm at time t are unlikely to do so again at $t+1$. Our previous approach, using an indicator variable if the individual i opened a firm at period t as Y_{it} would incorrectly measure those at risk of creating a firm.

We focus on the probability of individual i creating their first firm at period t. We define as our *risk set*, those individuals who had never owned a firm up to 3 years before the policy. Excluding individuals who had already created firms before the 3 year window of our analysis implies our final analysis sample comprises around 650 thousand individuals. In Table A5, we summarize the demographic characteristics of our analysis sample in comparison with those who were excluded. The remaining sample is similar in age and gender composition but slightly less educated and more likely to be nonwhite.

Using our analysis sample, we construct an unbalanced panel to estimate the hazard model with time-dependent covariates. In the data, each observation corresponds to an individual x quarter. It covers six years, between the first quarter of 2018 to the last quarter of 2023. The panel is unbalanced because if an individual creates a firm, they are excluded from subsequent periods.

Our estimating equation can be written as follows:

$$\lambda(t|X(t)) = \lambda_0(t)\exp\left(\beta_0^- \cdot D_i^+ + \beta_1^+ \cdot D_i^+ \cdot Post_t + \beta_0^- \cdot D_i^- + \beta_1^- \cdot D_i^- \cdot Post_t + \Gamma X_i\right) \quad (8)$$

where D_i^+ is an indicator that individual i is in the blue area in the diagram above, D_i^-

indicates that they were in the red area, $Post_t$ is an indicator that the respective period is after the implementation of the policy. We include a vector of controls X_i that consists of fixed effects by "cell" created interacting education, gender, race and age.

The coefficients β_1^+, β_1^- capture how the probability of creating a firm changes after the policy for individuals in the positively and negatively exposed groups, relative to those in the control group who on average did not experience substantial changes in their credit access.

We show our estimates in Table 4. Both coefficients β_1^+, β_1^- are extremely close to zero and not statistically significant. Looking at hazard ratios in column (2), we show that our estimates can reject changes in the probability of opening a business as big as 5%.

Table 4: Effects on the Probability of Creating a Firm

	(1)	(2)
	Created a Firm	
	Coefficient	Hazard Ratio
Positive Exposure x Post	0.004104	1.004
	(0.0169)	[0.971,1.038]
Negative Exposure x Post	-0.0179	0.982
	(0.0119)	[0.959,1.006]
Number of Individuals	645040	
Observations	13170363	

This table shows coefficients and hazard ratio estimates from equation 8. In column (1), we show coefficient estimates and standard errors in parenthesis; in column (2), we show the hazard ratio and 95% confidence interval in brackets. The sample comprises all individuals we observe who were never a firm owner until three years before the policy implementation. A row in the data corresponds to sub-spells of a quarter for each individual. If the individual creates a firm, they are not observed in subsequent periods.

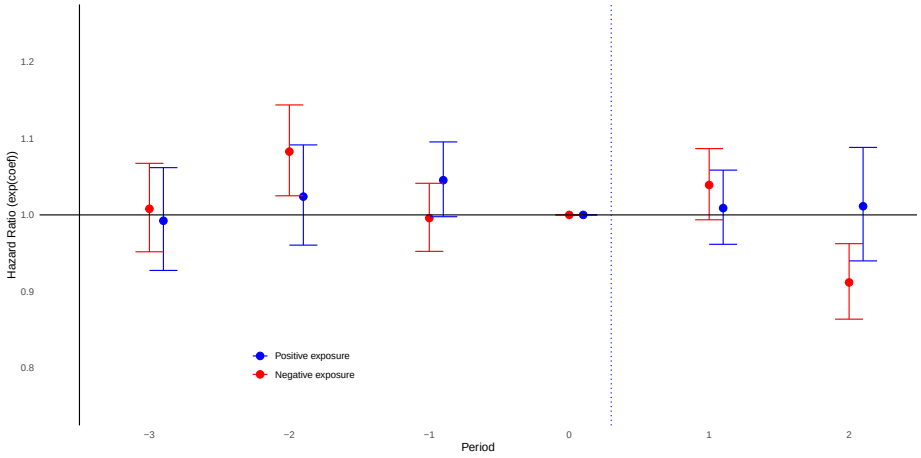
In addition to the pre-post analysis, we can also estimate coefficients for each year relative to the policy. This can be summarized in equation 9 below:

$$\lambda(t|X(t)) = \lambda_0(t) \exp\Big(\beta_0^+ \cdot D_i^+ + \beta_2 \cdot D_i^- + \sum_{year \in \{-3,2\}} \beta_{year}^+ \cdot D_i^+ \cdot \delta_{year(t)} + \sum_{year \in \{-3,2\}} \beta_{year(t)}^- \cdot D_i^- \cdot \delta_{year(t)} + \Gamma X_i\Big) \tag{9}$$

where $\delta_{year(t)}$ is an indicator function that the corresponding sub-spell indexed by t is in the respective year relative to the policy implementation. We then recover coefficients $\beta^{+}_{year}, \beta^{-}_{year}$ for each year before and after the policy in our analysis.

We show our estimates of $\beta^{+}_{t}, \beta^{-}_{t}$ in Figure 10. For the positively exposed group, our estimates indicate a precise zero in the change in the probability of opening a business. At the same time, for those negatively exposed, we find that estimates vary more across years. Despite having a negative effect 2 years after the policy (in line with what a credit constraints story would suggest), we show that one year after the policy, estimates actually suggest a positive effect of the policy for negatively exposed individuals. Pooling those estimates gets us the zero effects described above.

Figure 10: Effects on the Probability of Creating a Firm by Year



This Figure shows how the policy affected firm survival. In Panel (a), we show the estimates of coefficients $\{\beta^{+}_{t}, \beta^{-}_{t}\}$ from equation 9. In panel (b), we show the survival rates of these new firms with and without the policy. The sample comprises all individuals we observe who were never a firm owner until three years before the policy implementation. A row in the data corresponds to sub-spells of a quarter for each individual. If the individual creates a firm, they are not observed in subsequent periods.

7.3 Effects on Firm's Outcomes

We next explore if the policy affected the outcomes of firms created. Since we find no effects on firm creation, we return to the difference in differences analysis using linear models. We restrict our sample only to eventual entrepreneurs. We then compare the outcomes of firms created by positively and negatively exposed individuals, before and after the policy, with those created by individuals in our control who had no changes in credit access.

Using our sample of entrepreneurs, we estimate the following equation:

$$Y_i = \delta_{t(i)} + \beta^+ \cdot D_i^+ \cdot Post_{t(i)} + \beta^- \cdot D_i^- \cdot Post_{t(i)} + \Gamma X_i + \varepsilon_i \tag{10}$$

where $\delta_{t(i)}$ are fixed effects of the quarter in which individual i created their firm. D_i^+, D_i^- are indicator variables that take value one if individual i belongs to groups positively or negatively affected. $Post_{t(i)}$ is an indicator variable that takes value one if the firm was created after the policy. We include a set of control variables X_i, which includes gender, race, age, education, and fixed effects of the group of the joint distribution of credit individual i belong to.

Therefore, our coefficients of interest $\{\beta^+, \beta^-\}$ capture the difference in outcomes of firms created by positively (negatively) exposed individuals before and after the policy, under the assumption that in the absence of the policy, their difference would behave similarly to the difference for control individuals.

Firm Survival: We look at firm survival one and two years after firm creation. We define a firm surviving as having an active registry in *Receita Federal*, the Brazilian equivalent of the IRS. Using public data, we can observe firm survival until August 2024 for all formal firms.[35] To make it a consistent outcome, we restrict our sample to firms created until August 2022, around six quarters after the policy implementation.

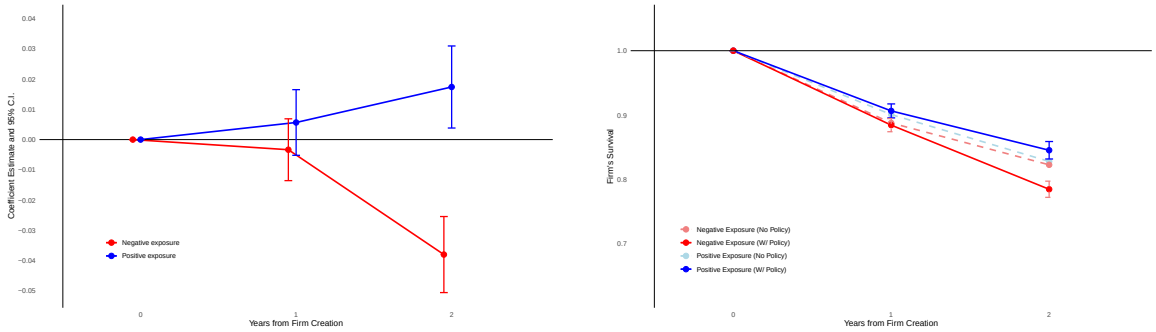
We find that more (less) credit access increases (decreases) the likelihood of firm survival two years after their creation. We show these results in Figure 11. In Panel (a) we plot our estimates and 95% confidence intervals of the coefficients β^+, β^- from equation 10. We estimate two different regressions using as outcome firm survival one and two years after firm creation. Our estimates indicate that firms created by positively exposed individuals increase by 1.8 p.p. their likelihood of survival two years after their creation because of the policy. In contrast, those created by negatively exposed individuals have a 3.8p.p. smaller likelihood of survival because of the policy.

This represents a 2.1% increase in the likelihood of firm survival and a 4.6% decrease for firms created by positively and negatively exposed individuals, respectively. We can see this in Panel (b) of Figure 11. We plot the average survival of firms created by positively and negatively exposed individuals with and without the policy. Survival without the policy is calculated as the average likelihood of firms created before the policy by the respective group. Survival with the policy is defined as that value plus the coefficient estimate from

[35] We use CNPJ public data to construct firm survival information. This data is a snapshot of all active and inactive registered firms in the country, identified by their registry number (often referred to as CNPJ number). A firm is defined as inactive if it has *situação cadastral* different from *Ativa* (active). This includes firms defined as *nula, suspensa, inapta or baixada.* In the data, we also see the date the firm became inactive, which allows us to construct our outcome variables.

Panel (a). The percent increase is calculated as the ratio between the firm survival with the policy by the firm survival without the policy.

Figure 11: Effects on Firm Survival



(a) Effects on Firm Survival **(b)** Survival Rates

This Figure presents our estimates of the effects of the policy on the survival of firms created by individuals exposed to changes in credit access. The sample comprises all individuals who created their first firm three years before and two years after the implementation of the policy. In Panel (a), we show the estimates and 95% confidence intervals of coefficients $\{\beta^+, \beta^-\}$ from equation 10. In Panel (b), we show firm survival rates with and without the policy for each group. Firm survival rates without the policy are calculated as the average survival of firms created by individuals in each group before the policy. Coefficients for one-year survival and two years survival are estimated in separate regressions. Controls included are cells of education, gender, race, and year of birth, and dummies for the group of the joint distribution of credit scores the individual belonged to.

7.4 Consequences on Average Firm Quality

This is the section I have to finish with the stuff that was in the cluster and I couldn't access because it went down these last 10 days.

What I already have. Differences in the average "quality" of the firms by each group. The positively exposed, are on average on more productive industries, pay better wages, and employ slightly more.

What I was also doing. Estimating the diff in diff with these quality measures. (Credit, number of emps, industry)

Cluster is coming back.

8 Conclusion

This paper studies the effects of revealing borrowers' information to lenders through credit scores. We explore a unique policy that took place in Brazil, which changed the information that financial institutions must share with credit bureaus from a financial delinquency registry to a complete registry of borrowers' credit history.

Our empirical analysis shows that the effects of revealing information about borrowers on credit can be rationalized by a simple conceptual framework that takes into account changes in both the value of the signal and its precision.

We show that revealing information generates an equity-efficiency trade-off. The policy increases overall inequality in credit access and enhances the racial gap in credit between white and nonwhite individuals. At the same time, we show that the policy increases efficiency in the market as credit gets reallocated from more to less risky loans.

We then show that changes in credit had effects beyond credit markets by investigating the effects of the policy on entrepreneurial activity. We identify that despite not increasing the likelihood of an individual creating a firm, an increase in credit access increases the revenue and investments of firms.

Bibliography

Aigner, Dennis J and Glen G Cain, "Statistical theories of discrimination in labor markets," *Ilr Review*, 1977, *30* (2), 175–187.

Angrist, Joshua D and Guido W Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 1995, *90* (430), 431–442.

BACEN, Brazilian Central Bank, "Análise dos Efeitos do Cadastro Positivo," *Policy Note*, 2021.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman, "Six randomized evaluations of microcredit: Introduction and further steps," *American Economic Journal: Applied Economics*, 2015, *7* (1), 1–21.

Banerjee, Abhijit V and Andrew F Newman, "Occupational choice and the process of development," *Journal of political economy*, 1993, *101* (2), 274–298.

Bank, The World, *Global financial development report 2013: Rethinking the role of the state in finance*, The World Bank, 2012.

Behr, Patrick and Simon Sonnekalb, "The effect of information sharing between lenders on access to credit, cost of credit, and loan performance–Evidence from a credit registry introduction," *Journal of Banking & Finance*, 2012, *36* (11), 3017–3032.

Black, Sandra E and Philip E Strahan, "Entrepreneurship and bank credit availability," *The Journal of Finance*, 2002, *57* (6), 2807–2833.

Blattner, Laura and Scott Nelson, "How costly is noise? Data and disparities in consumer credit," *arXiv preprint arXiv:2105.07554*, 2021.

Bos, Marieke and Leonard I Nakamura, "Should defaults be forgotten? Evidence from variation in removal of negative consumer credit information," 2014.

_ , Emily Breza, and Andres Liberman, "The labor market effects of credit market information," *The Review of Financial Studies*, 2018, *31* (6), 2005–2037.

Brooks, Benjamin, Alexander Frankel, and Emir Kamenica, "Comparisons of signals," *American Economic Review*, 2022.

_ , _ , and _ , "Information hierarchies," *Econometrica*, 2022, *90* (5), 2187–2214.

Cahn, Christophe, Mattia Girotti, and Augustin Landier, "Entrepreneurship and information on past failures: A natural experiment," *Journal of financial economics*, 2021, *141* (1), 102–121.

Card, David and Dean R Hyslop, "Estimating the effects of a time-limited earnings subsidy for welfare-leavers," *Econometrica*, 2005, *73* (6), 1723–1770.

Chambers, Christopher P and Paul J Healy, "Updating toward the signal," *Economic Theory*, 2012, *50*, 765–786.

Decker, Ryan, John Haltiwanger, Ron Jarmin, and Javier Miranda, "The role of entrepreneurship in US job creation and economic dynamism," *Journal of Economic Perspectives*, 2014, *28* (3), 3–24.

DeFusco, Anthony A, Huan Tang, and Constantine Yannelis, "Measuring the welfare cost of asymmetric information in consumer credit markets," *Journal of Financial Economics*, 2022, *146* (3), 821–840.

Diaconis, Persi and Donald Ylvisaker, "Conjugate priors for exponential families," *The Annals of statistics*, 1979, pp. 269–281.

Djankov, Simeon, Caralee McLiesh, and Andrei Shleifer, "Private credit in 129 countries," *Journal of financial Economics*, 2007, *84* (2), 299–329.

Dobbie, Will, Paul Goldsmith-Pinkham, Neale Mahoney, and Jae Song, "Bad credit, no problem? Credit and labor market consequences of bad credit reports," *The Journal of Finance*, 2020, *75* (5), 2377–2419.

Einav, Liran, Amy Finkelstein, and Mark R Cullen, "Estimating welfare in insurance markets using variation in prices," *The quarterly journal of economics*, 2010, *125* (3), 877–921.

Evans, David S and Boyan Jovanovic, "An estimated model of entrepreneurial choice under liquidity constraints," *Journal of political economy*, 1989, *97* (4), 808–827.

Feinmann, Javier, M Lauletta, and R Rocha, "Employer-employee collusion and payments under the table: Evidence from Brazil," *University of California, Berkeley*, 2022.

Gonzalez, Lauro, João Pedro Haddad, and Julio Leandro, "Evolução do crédito para pessoas físicas no Brasil e suas distorções," Technical Report, Centro de Estudos em Microfinanças e Inclusão Financeira FGV 2023.

Green, Jerry R and Nancy L Stokey, "Two representations of information structures and their comparisons," *Technical Report*, 1978, *271.*

Gross, Tal, Matthew J Notowidigdo, and Jialan Wang, "The marginal propensity to consume over the business cycle," *American Economic Journal: Macroeconomics*, 2020, *12* (2), 351–384.

Haanwinckel, Daniel and Rodrigo R Soares, "Workforce composition, productivity, and labour regulations in a compensating differentials theory of informality," *The Review of Economic Studies*, 2021, *88* (6), 2970–3010.

Haltiwanger, John, Ron S Jarmin, and Javier Miranda, "Who creates jobs? Small versus large versus young," *Review of Economics and Statistics*, 2013, *95* (2), 347–361.

Herkenhoff, Kyle, Gordon M Phillips, and Ethan Cohen-Cole, "The impact of consumer credit access on self-employment and entrepreneurship," *Journal of financial economics*, 2021, *141* (1), 345–371.

Hertzberg, Andrew, Jose Maria Liberti, and Daniel Paravisini, "Public information and coordination: evidence from a credit registry expansion," *The Journal of Finance*, 2011, *66* (2), 379–412.

Hurst, Erik and Annamaria Lusardi, "Liquidity constraints, household wealth, and entrepreneurship," *Journal of political Economy*, 2004, *112* (2), 319–347.

Jansen, Mark, Fabian Nagel, Anthony Lee Zhang, and Constantine Yannelis, "Data and welfare in credit markets," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2022, (2022-88).

Jappelli, Tullio and Marco Pagano, "Information sharing, lending and defaults: Cross-country evidence," *Journal of Banking & Finance*, 2002, *26* (10), 2017–2045.

Karlan, Dean and Jonathan Zinman, "Expanding credit access: Using randomized supply decisions to estimate the impacts," *The Review of Financial Studies*, 2010, *23* (1), 433–464.

_ **and** _ , "Microcredit in theory and practice: Using randomized credit scoring for impact evaluation," *Science*, 2011, *332* (6035), 1278–1284.

Liberman, Andres, Christopher Neilson, Luis Opazo, and Seth Zimmerman, "The equilibrium effects of information deletion: Evidence from consumer credit markets," Technical Report, National Bureau of Economic Research 2018.

McKenzie, David, "Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition," *American Economic Review*, 2017, *107* (8), 2278–2307.

Mel, Suresh De, David McKenzie, and Christopher Woodruff, "Returns to capital in microenterprises: evidence from a field experiment," *The quarterly journal of Economics*, 2008, *123* (4), 1329–1372.

_ **,** _ **, and** _ , "Are women more credit constrained? Experimental evidence on gender and microenterprise returns," *American Economic Journal: Applied Economics*, 2009, *1* (3), 1–32.

Meza, David De and David C Webb, "Too much investment: A problem of asymmetric information," *The quarterly journal of economics*, 1987, *102* (2), 281–292.

Musto, David K, "What happens when information leaves a market? Evidence from postbankruptcy consumers," *The Journal of Business*, 2004, *77* (4), 725–748.

Pagano, Marco and Tullio Jappelli, "Information sharing in credit markets," *The journal of finance*, 1993, *48* (5), 1693–1718.

Phelps, Edmund S, "The statistical theory of racism and sexism," *The american economic review*, 1972, *62* (4), 659–661.

Robb, Alicia M and David T Robinson, "The capital structure decisions of new firms," *The Review of Financial Studies*, 2014, *27* (1), 153–179.

Rocha, Roberto Hsu and Alison de Farias, "Formality Cost, Registration and Development of Microentreprenuers: Evidence From Brazil," *Available at SSRN 3969404*, 2021.

Schmalz, Martin C, David A Sraer, and David Thesmar, "Housing collateral and entrepreneurship," *The Journal of Finance*, 2017, *72* (1), 99–132.

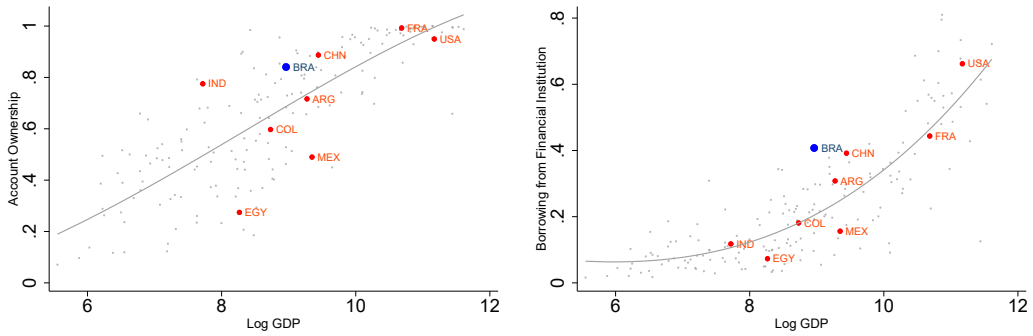
Spence, Michael, "Job Market Signaling," *The Quarterly Journal of Economics*, 1973, *87* (3), 355–374.

Stiglitz, Joseph E and Andrew Weiss, "Credit rationing in markets with imperfect information," *The American economic review*, 1981, *71* (3), 393–410.

Ulyssea, Gabriel, "Firms, informality, and development: Theory and evidence from Brazil," *American Economic Review*, 2018, *108* (8), 2015–2047.

A Additional Figures and Tables

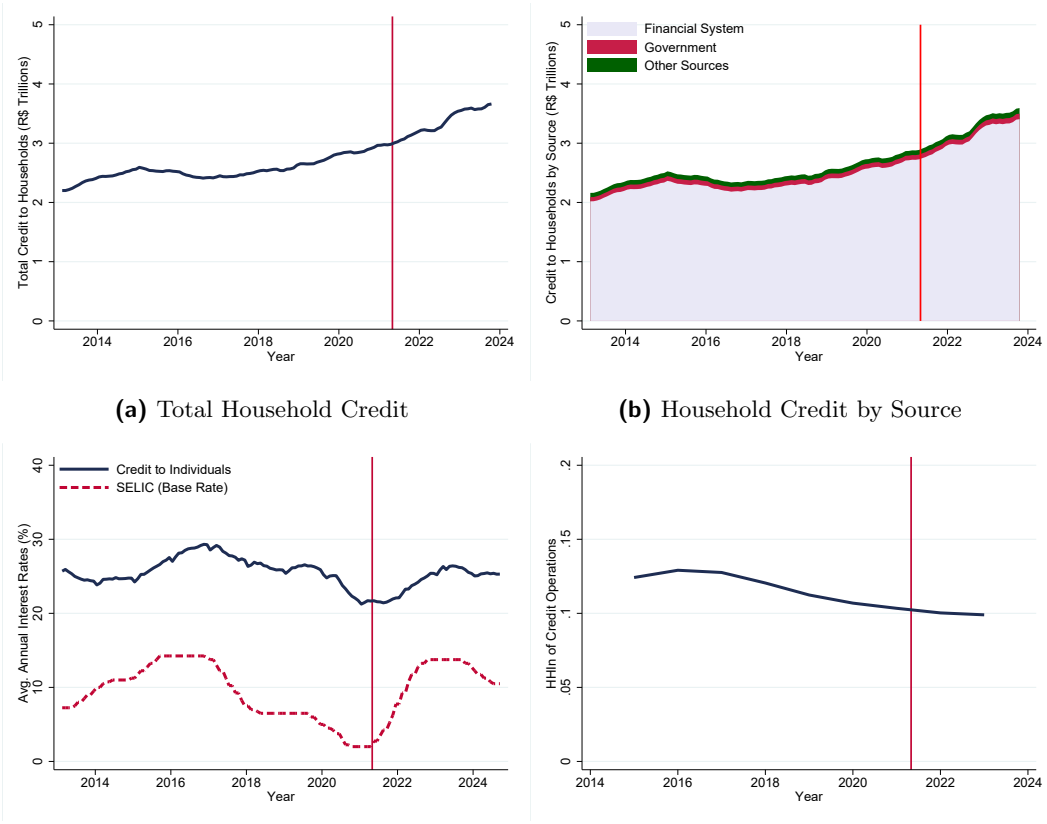
Figure A1: Credit Characteristics by Country and Income



(a) Has Bank Account **(b)** Loan From Financial Institutions

This Figure shows the credit characteristics of countries relative to the Log of their Per capita GDP. Figures were constructed using 2021 data from the World Bank Global Financial Development Database.

Figure A2: Aggregate Credit Statistics Before and After Cadastro Positivo



(a) Total Household Credit

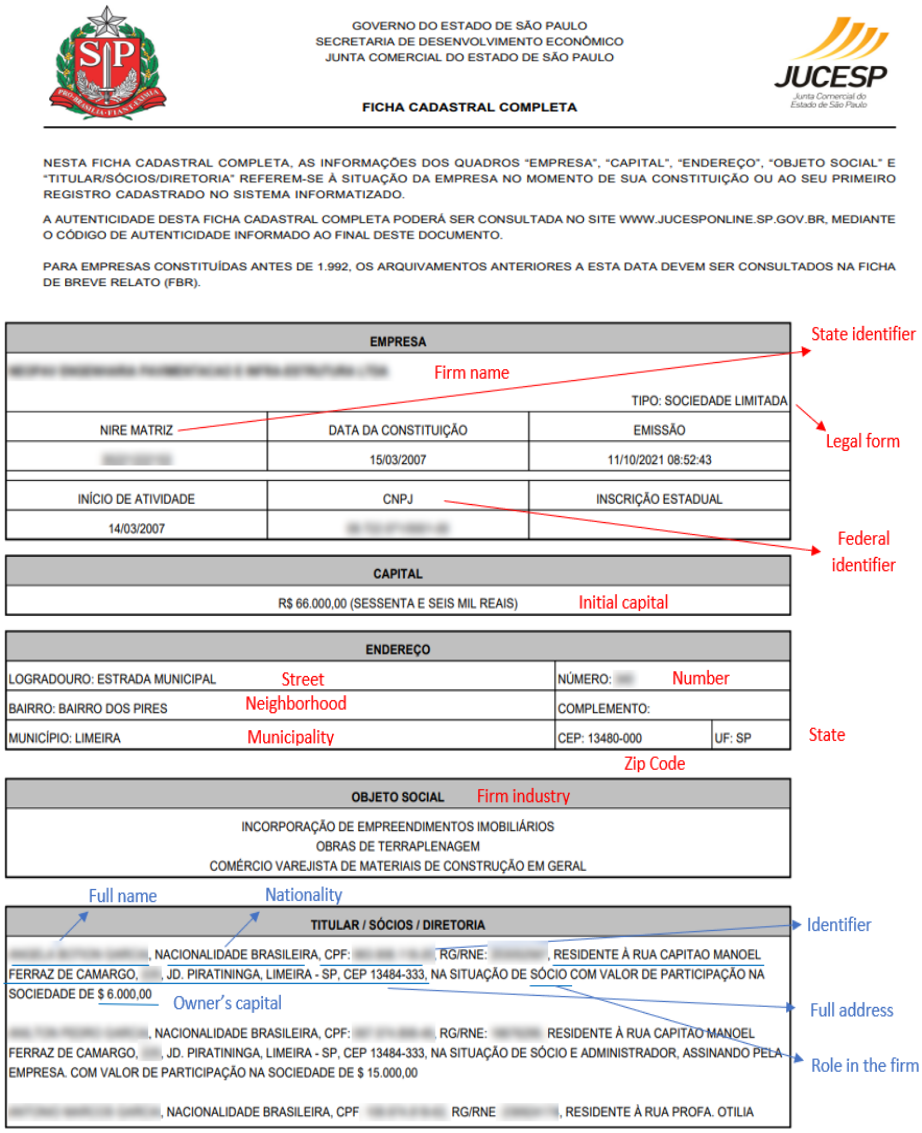
(b) Household Credit by Source

(c) Average Annual Interest Rates

(d) Concentration (HHIn)

This Figure shows aggregate credit statistics in the periods before and after *Cadastro Positivo* implementation, indicated by the red vertical lines. In Panel (a), we show Total Household Credit in the country. Panel (b) shows the same measure decomposed by the source of that given credit. Other Sources included credit cooperatives and institutions non directly tied to the financial system. Panel (c) plots the average interest rate of credit given to households. Panel (d) plots normalized Herfindahl-Hirschman Index. The index is obtained by summing the square of the market share (in decimal form) of each financial institution in the considered market: HHIn = $(IF1)^2 + (IF2)^2 + ... + (IFj)^2$, resulting in a number between 0 and 1. Normalization choice was made by the Brazilian Central Bank. Panels (a), (b), and (c) use data available in the Brazilian Central Bank's Time Series Management System. Panel (d) uses data collected from the Brazilian Central Bank's report of banking activity.

Figure A3: Example of Firm Record from Sao Paulo's Trade Board



GOVERNO DO ESTADO DE SÃO PAULO
SECRETARIA DE DESENVOLVIMENTO ECONÔMICO
JUNTA COMERCIAL DO ESTADO DE SÃO PAULO

JUCESP
Junta Comercial do Estado de São Paulo

FICHA CADASTRAL COMPLETA

NESTA FICHA CADASTRAL COMPLETA, AS INFORMAÇÕES DOS QUADROS "EMPRESA", "CAPITAL", "ENDEREÇO", "OBJETO SOCIAL" E "TITULAR/SÓCIOS/DIRETORIA" REFEREM-SE À SITUAÇÃO DA EMPRESA NO MOMENTO DE SUA CONSTITUIÇÃO OU AO SEU PRIMEIRO REGISTRO CADASTRADO NO SISTEMA INFORMATIZADO.

A AUTENTICIDADE DESTA FICHA CADASTRAL COMPLETA PODERÁ SER CONSULTADA NO SITE WWW.JUCESPONLINE.SP.GOV.BR, MEDIANTE O CÓDIGO DE AUTENTICIDADE INFORMADO AO FINAL DESTE DOCUMENTO.

PARA EMPRESAS CONSTITUÍDAS ANTES DE 1.992, OS ARQUIVAMENTOS ANTERIORES A ESTA DATA DEVEM SER CONSULTADOS NA FICHA DE BREVE RELATO (FBR).

EMPRESA

[redacted] Firm name

TIPO: SOCIEDADE LIMITADA

NIRE MATRIZ	DATA DA CONSTITUIÇÃO	EMISSÃO
[redacted]	15/03/2007	11/10/2021 08:52:43

INÍCIO DE ATIVIDADE	CNPJ	INSCRIÇÃO ESTADUAL
14/03/2007	[redacted]	

State identifier
Legal form
Federal identifier

CAPITAL

R$ 66.000,00 (SESSENTA E SEIS MIL REAIS) Initial capital

ENDEREÇO

LOGRADOURO: ESTRADA MUNICIPAL Street		NÚMERO: [redacted] Number	
BAIRRO: BAIRRO DOS PIRES Neighborhood		COMPLEMENTO:	
MUNICÍPIO: LIMEIRA Municipality		CEP: 13480-000	UF: SP

State
Zip Code

OBJETO SOCIAL Firm industry

INCORPORAÇÃO DE EMPREENDIMENTOS IMOBILIÁRIOS
OBRAS DE TERRAPLENAGEM
COMÉRCIO VAREJISTA DE MATERIAIS DE CONSTRUÇÃO EM GERAL

Full name Nationality

TITULAR / SÓCIOS / DIRETORIA

Identifier

[redacted], NACIONALIDADE BRASILEIRA, CPF: [redacted] RG/RNE: [redacted], RESIDENTE À RUA CAPITAO MANOEL FERRAZ DE CAMARGO, [redacted], JD. PIRATININGA, LIMEIRA - SP, CEP 13484-333, NA SITUAÇÃO DE SÓCIO COM VALOR DE PARTICIPAÇÃO NA SOCIEDADE DE $ 6.000,00

Owner's capital

Full address

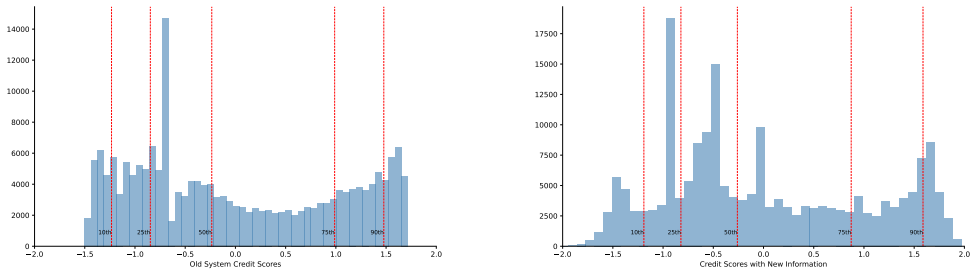
[redacted], NACIONALIDADE BRASILEIRA, CPF: [redacted] RG/RNE: [redacted] RESIDENTE À RUA CAPITAO MANOEL FERRAZ DE CAMARGO, [redacted], JD. PIRATININGA, LIMEIRA - SP, CEP 13484-333, NA SITUAÇÃO DE SÓCIO E ADMINISTRADOR, ASSINANDO PELA EMPRESA. COM VALOR DE PARTICIPAÇÃO NA SOCIEDADE DE $ 15.000,00

Role in the firm

[redacted], NACIONALIDADE BRASILEIRA, CPF [redacted] RG/RNE [redacted], RESIDENTE À RUA PROFA. OTILIA

This Figure shows an example of a firm record from São Paulo's trade board. Despite being publicly available records, we blurred identifiable information of individuals and firms to protect their confidentiality. All records are available in *Serviços Online* under the option *Pesquisar Empresas*. It is necessary to have a Brazilian social security number (CPF) to log in the website. Additional information such as ownership and capital change are also available in these records despite not being apparent in this example.

Figure A4: Histogram of Credit Scores under both Systems

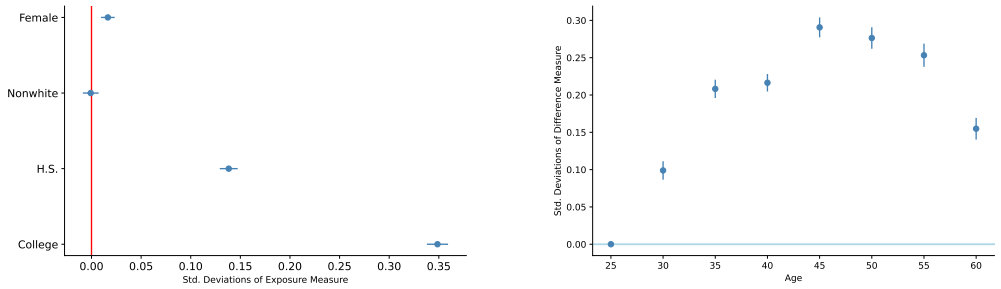


(a) Negative Registry Information Only

(b) Positive and Negative Registry Information

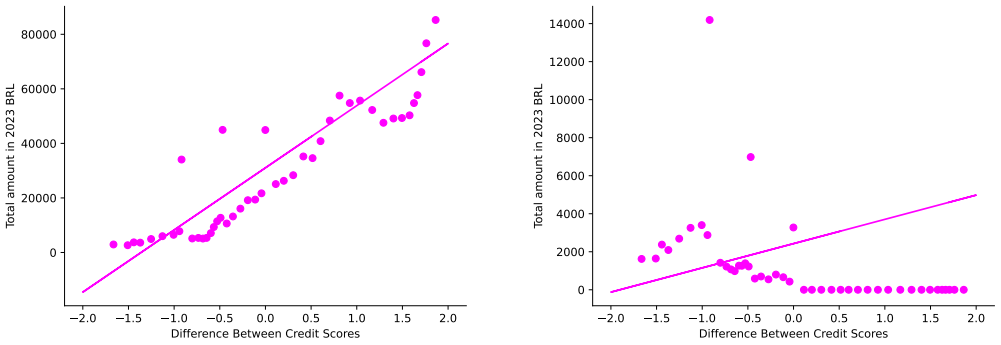
This Figure shows the histogram of credit scores calculated with information from both positive and negative registries, and credit scores calculated with information from the negative registry. The sample is restricted to the last period before the implementation of the policy.

Figure A5: Correlation of the Difference between Credit Scores with Observable Characteristics



(a) Gender Race and Educ.

(b) Age Groups

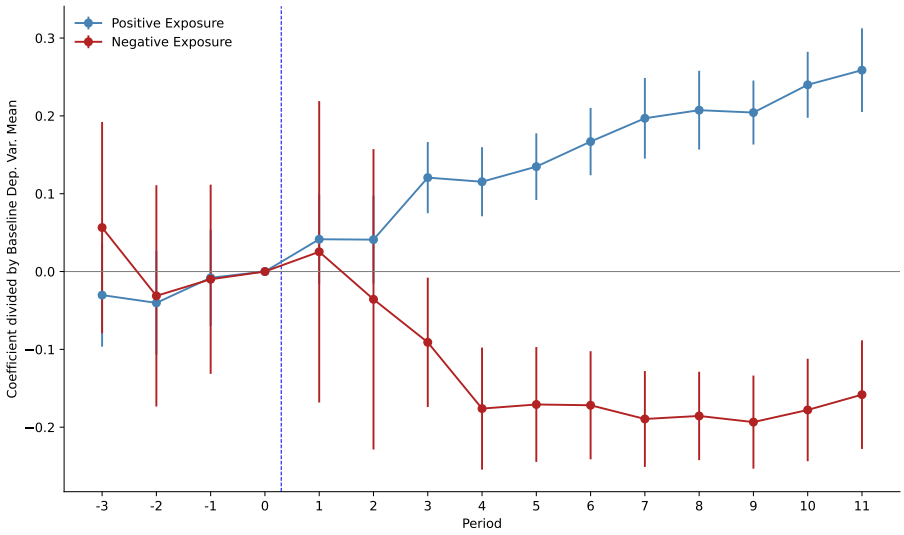


(c) Total Credit

(d) Payments not made

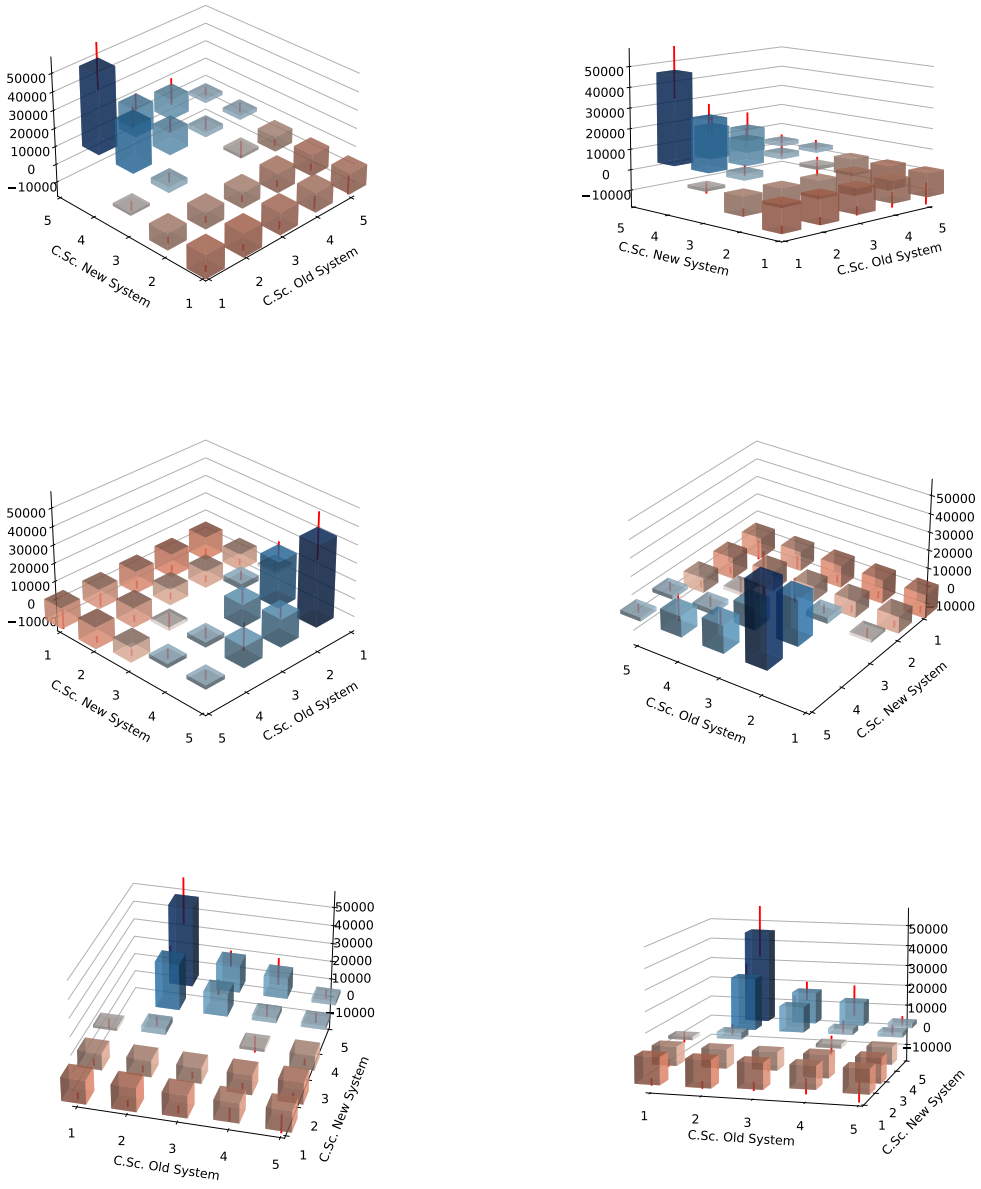
Panels (a) and (b) plot the coefficients of a regression of Δ_i on observable characteristics. The sample is restricted to the last period before the implementation of the policy. Coefficients in panels (a) and (b) are estimated in the same regression that includes dummies for gender, race, education groups, and age groups. We omit from the regression white men with less than high school education in the youngest age group. In panels (c) and (d), we show binscatters of Credit and Default with respect to Δ_i. The sample is restricted to the last period before the implementation of the policy.

Figure A6: Changes in Credit Divided by Predicted Credit w/o Policy



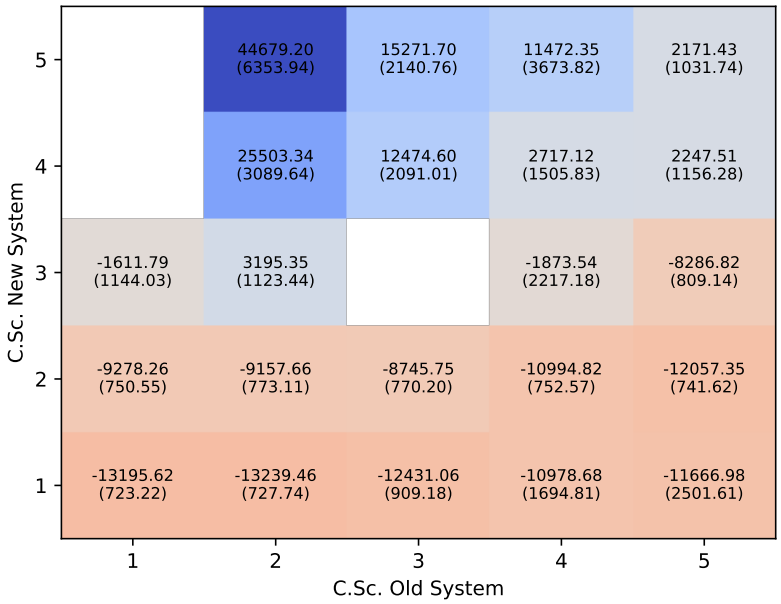
This figure plots the $\{\beta_t^+, \beta_t^-\}$ estimates from equation 1 normalized. We divide the coefficient estimates by the baseline level of the respective group in period=0 summed with the time fixed effects (for positive exposure $\frac{\hat{\beta_t^+}}{E[Y_{it}|t=0,D_i^+]+\hat{\delta_t}}$, and for negative exposure $\frac{\hat{\beta_t^-}}{E[Y_{it}|t=0,D_i^-]+\hat{\delta_t}}$). Positive exposure is defined as individuals with $\Delta_i \in [0.75, 1.25]$ and negative exposure as $\Delta_i \in [-1.25, -0.75]$.

Figure A7: Effects on Credit over the Joint Distribution of Credit Scores - Rotated Bars



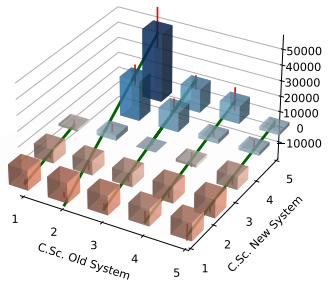
This Figure shows the estimates of coefficients $\{\beta^{kj}\}$ from equation 3. It shows the same results as in Figure 5 but over different angles. Each bar corresponds to a given coefficient, with 95% confidence intervals plotted in the red lines. Standard Errors are clustered at the individual level. Bars are organized such that the x-axis (labeled C. Sc. old system) indexes coefficients for a given group k, and the y-axis (labeled C.Sc. new system) indexes coefficients for a given group j. Positive estimates of β^{kj} are shown in blue, whereas negative estimates are shown in red. $\beta^{14}, \beta 15$ are not defined because there is no individual in the sample in those groups of the joint distribution of credit scores.

Figure A8: Effects of the Policy on Credit over the Joint Distribution of Credit Scores

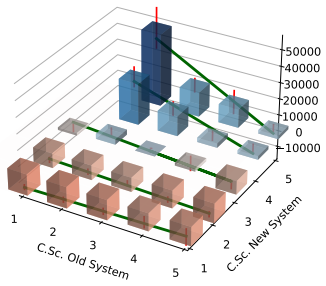


Notes: This Figure shows estimates of coefficients $\{\beta^{kj}\}$ from equation 3. Standard errors in parenthesis are clustered at the individual level. Coefficients are organized such that the x-axis (labeled C. Sc. old system) indexes coefficients for a given group k, and the y-axis (labeled C.Sc. new system) indexes coefficients for a given group j.

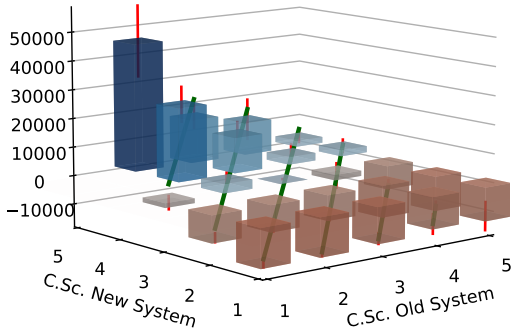
Figure A9: Comparisons Between Changes in Credit over the Joint Distribution of Credit Scores



(a) *Vertical Comparisons*



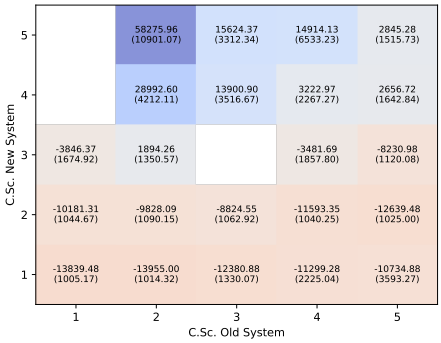
(b) *Horizontal Comparisons*



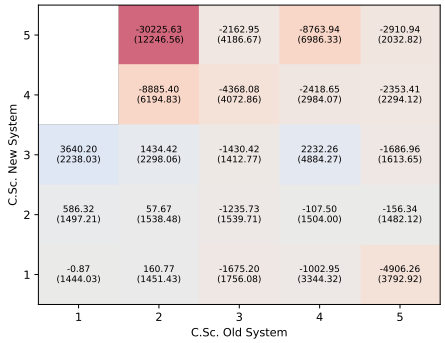
(c) *Diagonal Comparisons*

This Figure shows the estimates of coefficients $\{\beta^{kj}\}$ from equation 3 with the same specifications as Figure 5. The difference is the green lines that correspond to the linear fit between coefficients. A linear fit is computed by regressing values of estimates along each coordinate of the corresponding comparison without weighting. Panel (c) shows the same estimates but rotated to a different angle to better visualize the diagonal comparisons. Table A2 shows the estimates of the linear fits across coordinates.

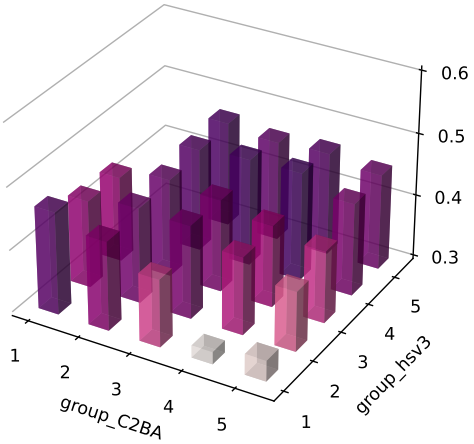
Figure A10: Gender by Group of the Joint Distribution of Credit Scores



(a) Male β_{g0}^{kj}



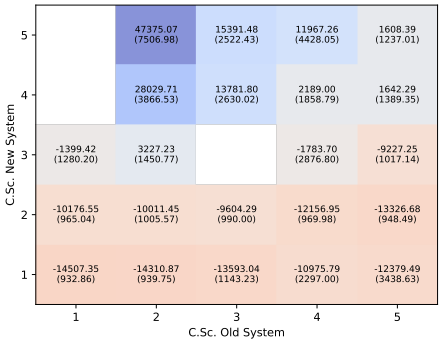
(b) β_{g1}^{kj} (Female - Male)



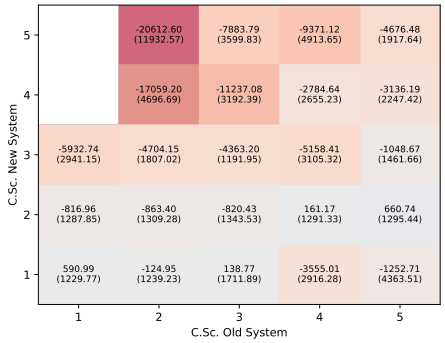
(c) Share of Women across the Joint Distribution of Credit Scores

In Panels (a) and (b), we plot our coefficient estimates from Equation 5. $G_i = 1$ indicates that indiviual i is a woman. These are the values used to calculate our expression in equation 6. Standard errors are shown in parenthesis. In Panel (c) we show the share of women across the joint distribution of credit scores.

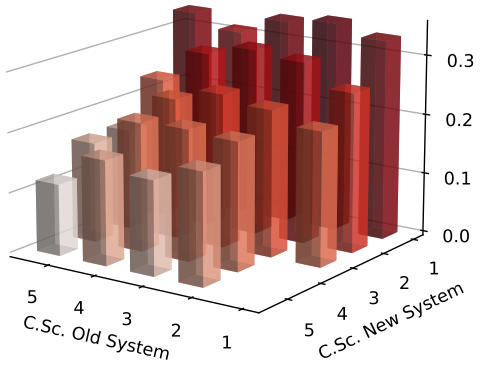
Figure A11: Changes in Credit by Race by Group of the Joint Distribution of Credit Scores



(a) White β_{g0}^{kj}



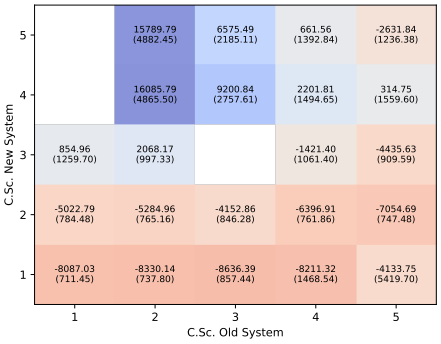
(b) β_{g1}^{kj} (Nonwhite - White)



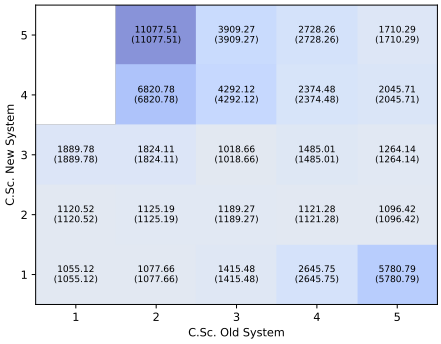
(c) Share of Nonwhite by Group

In Panels (a) and (b), we plot our coefficient estimates from Equation 5. $G_i = 1$ indicates that individual i is a nonwhite individual. These are the values used to calculate our expression in equation 6. Standard errors are shown in parentheses. In Panel (c), we show the share of nonwhite individuals across the joint distribution of credit scores.

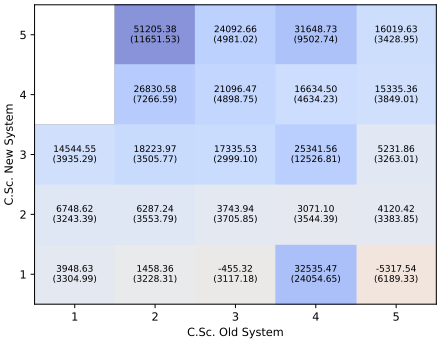
Figure A12: Education by Group of the Joint Distribution of Credit Scores



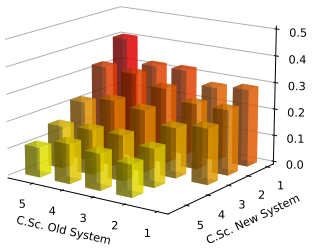
(a) Less than H.S. β_{g0}^{kj}



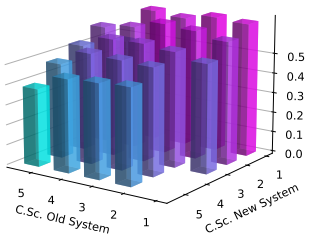
(b) β_{g1}^{kj} (H.S. - < H.S.)



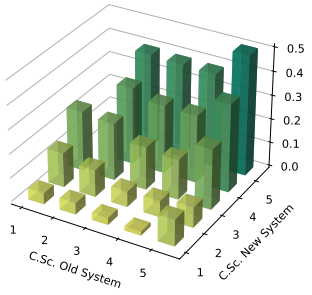
(c) β_{g2}^{kj} (Coll. - < H.S.)



(d) Share with Less than H.S.



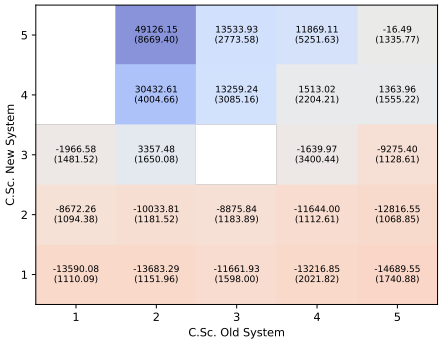
(e) Share with High School



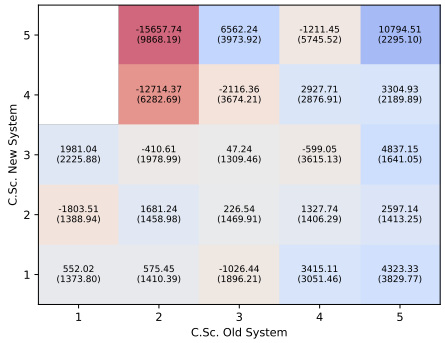
(f) Share with Some College

Figure A13: In Panels (a) and (b), we plot our coefficient estimates from Equation 5. We add a second heterogeneity group, leaving at base individuals with less than high school. These are the values used to calculate our expression in equation 6. Standard errors are shown in parentheses. In Panel (c), we show the share of individuals with each educational attainment across the joint distribution of credit scores.

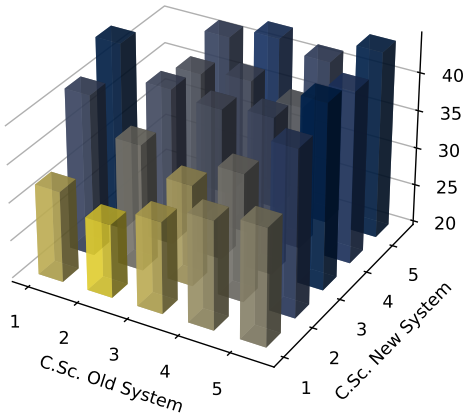
Figure A14: Changes in Credit by Age Group by Group of the Joint Distribution of Credit Scores



(a) 40 y.o. or more β_{g0}^{kj}



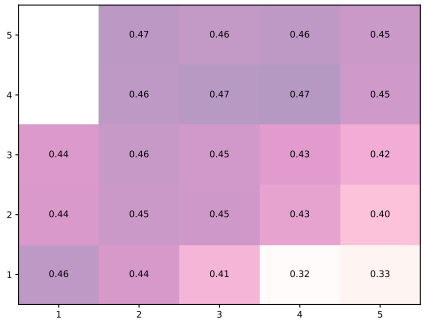
(b) β_{g1}^{kj} (Below 40 - 40 y.o. or more)



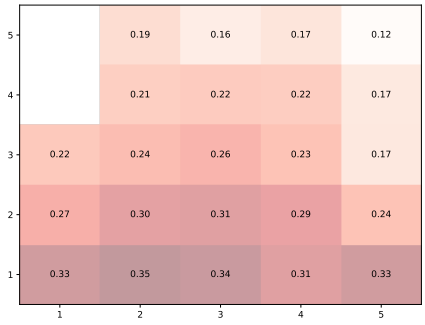
(c) Avg. Age by Group

In Panels (a) and (b), we plot our coefficient estimates from Equation 5. $G_i = 1$ indicates that individual i is less than 40 years old. These are the values used to calculate our expression in equation 6. Standard errors are shown in parenthesis. In Panel (c) we show the average age of individuals across the joint distribution of credit scores.

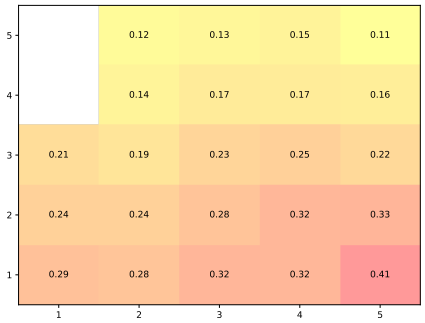
Figure A15: Education by Group of the Joint Distribution of Credit Scores



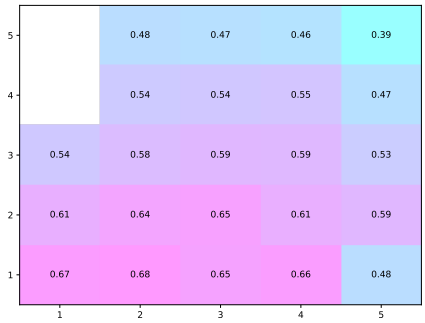
(a) Share Women



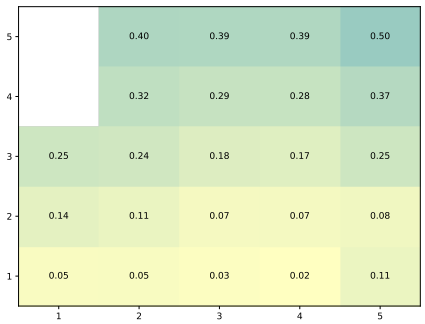
(b) Share Nonwhite



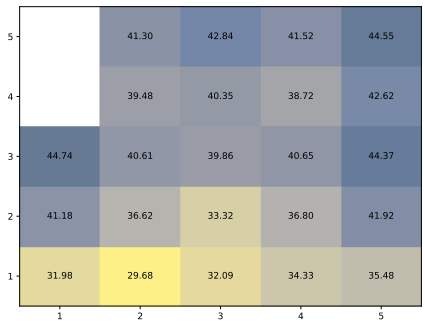
(c) Share with Less than H.S.



(d) Share with High School



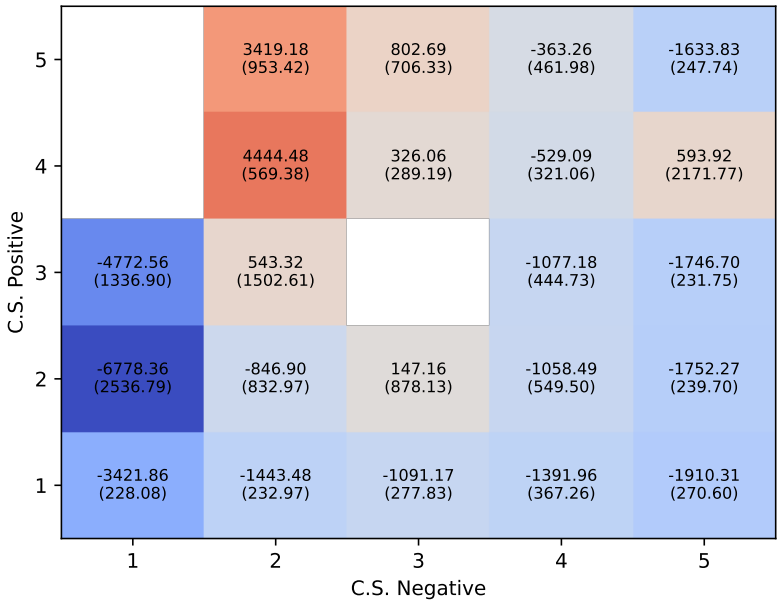
(e) Share with Some College



(f) Avg. Age

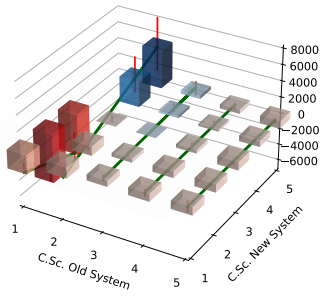
This Figure shows heatmap plots of how demographic characteristics vary across the joint distribution of credit scores. These are the same results from Panel (c) in Figures A10,A11,A14 and Panels (d), (e), (f) of Figure A12.

Figure A16: Effects of the Policy on Financial Delinquency over the Joint Distribution of Credit Scores

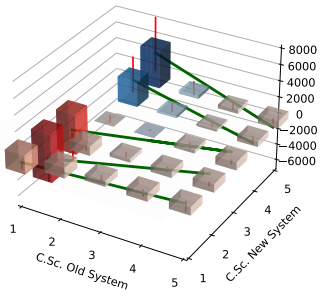


Notes: This Figure shows estimates of coefficients $\{\beta^{kj}\}$ from equation 3 with total financial delinquency from individual i at period t as the outcome. Standard errors in parenthesis are clustered at the individual level. Coefficients are organized such that the x-axis (labeled C. Sc. old system) indexes coefficients for a given group k, and the y-axis (labeled C.Sc. new system) indexes coefficients for a given group j.

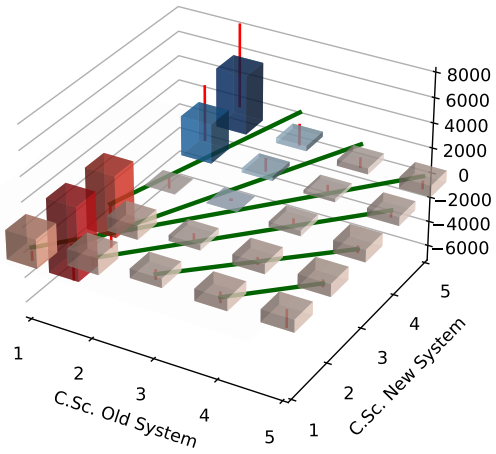
Figure A17: Comparisons Between Changes in Financial Delinquency over the Joint Distribution of Credit Scores



(a) *Vertical Comparisons*



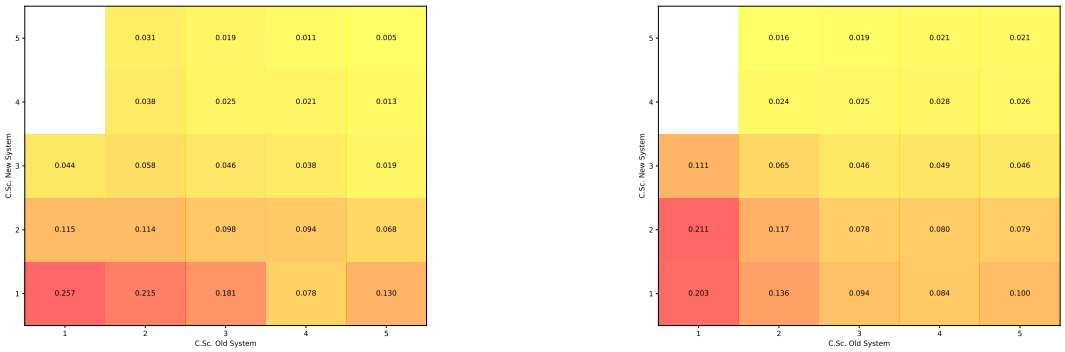
(b) *Horizontal Comparisons*



(c) *Diagonal Comparisons*

This Figure shows the estimates of coefficients $\{\beta^{kj}\}$ from equation 3 with the same specifications as Panel (c) in Figure 8. The difference is the green lines that correspond to the linear fit between coefficients. A linear fit is computed by regressing values of estimates along each coordinate of the corresponding comparison without weighting. Panel (c) shows the same estimates but rotated to a different angle to better visualize the diagonal comparisons. Table A2 shows the estimates of the linear fits across coordinates.

Figure A18: Estimates of Default Rate

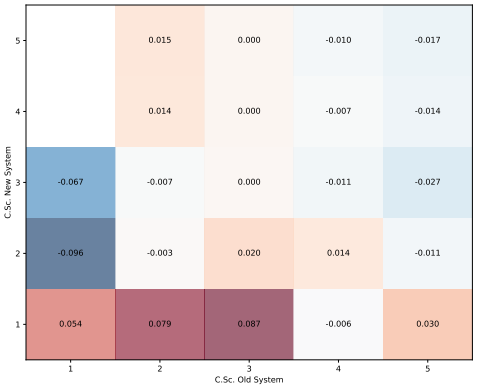


(a) Default Rates - with Policy

(b) Default Rates - no Policy

This Figure shows average Default Rates in the periods after the policy. These are the same results from Panels (a) and (b) of Figure 9 in a 2Dimensional perspective. Panel (a) plots our estimates of default rates in the presence of the policy, whereas Panel (b) plots our estimates of Default rates in the absence of the Policy.

Figure A19: Changes in Overall Default Rate



This Figure shows changes in overall default rate for each group. This consists of the difference between average default rate for each group with policy DR^1_{it} in the periods after the policy and the counterfactual average default rate without the policy in the periods after the policy implementation.

Figure A20: Estimates of Default Rate

(a) Default Rates - with Policy

(b) Default Rates - no Policy

This Figure shows average Default Rates in the periods after the policy. Panel (a) and (b) are the same results from Figure 9 from a different angle. Panel (a) plots our estimates of default rates in the presence of the policy, whereas Panel (b) plots our estimates of Default rates in the absence of the Policy.

Table A1: Correlation between Credit Scores and Demographic Characteristics

	Negative System	Positive System	Difference in C.Scs.
	(1)	(2)	(3)
Female	-0.059	-0.043	0.016
	(0.004)	(0.004)	(0.004)
Nonwhite	-0.170	-0.171	-0.001
	(0.005)	(0.005)	(0.004)
High School	0.087	0.226	0.138
	(0.005)	(0.005)	(0.005)
College	0.561	0.909	0.349
	(0.007)	(0.006)	(0.005)
30	0.186	0.285	0.099
	(0.007)	(0.008)	(0.006)
35	0.269	0.477	0.208
	(0.007)	(0.008)	(0.006)
40	0.418	0.634	0.216
	(0.007)	(0.007)	(0.006)
45	0.356	0.646	0.291
	(0.008)	(0.008)	(0.007)
50	0.453	0.729	0.276
	(0.009)	(0.009)	(0.007)
55	0.574	0.827	0.253
	(0.010)	(0.009)	(0.008)
60	0.780	0.935	0.155
	(0.009)	(0.008)	(0.007)
Intercept	-0.475	-0.824	-0.349
	(0.007)	(0.007)	(0.006)
Observations	195179	195179	195179
R^2	0.103	0.196	0.036

This Table presents the correlation between Credit Scores and demographic characteristics of individuals. Coefficients are estimates of a linear regression of $\frac{cs_i - \bar{cs}}{sd(cs_i)}$ on observable characteristics. The sample is restricted to the last period before the implementation of the policy. Column (1) shows results using credit scores from the negative system, and Column (2) from the positive system. The outcome in Column (3) is Positive System credit scores minus Negative System credit scores. Robust Standard errors in parenthesis.

Table A2: Linear Fit between Estimates of β^{kj} from Equation 3 with Credit as Outcome

Vertical		**Horizontal**		**Diagonal**	
(1)		(2)		(3)	
Fit Between Coordinates		Coordinates		Coordinates	
(1,1) - (1,3)	5791.9119	(1,1) - (5,1)	531.8042	(1,1) - (5,5)	4260.8871
	(1082.2779)		(184.4172)		(957.9481)
(2,1) - (2,5)	15049.8305	(1,2) - (5,2)	-739.5318	(4,1) - (5,2)	-1078.6651
	(2091.4539)		(291.3136)		(.)
(3,1) - (3,5)	7662.5857	(1,3) - (5,3)	-1841.8956	(3,1) - (5,3)	2072.1172
	(895.8384)		(1100.7821)		(367.1220)
(4,1) - (4,5)	5861.3999	(2,4) - (5,4)	-7952.4951	(2,1) - (5,4)	5333.3116
	(850.7122)		(2030.8364)		(367.4760)
(5,1) - (5,5)	4198.1683	(5,4) - (5,5)	-13132.2642	(1,3) - (3,5)	8441.7446
	(1023.4545)		(3865.9973)		(10781.0828)
				(1,2) - (2,5)	7153.1092
					(2188.8593)

This Table shows coefficients and standard errors in parenthesis of linear fit between coefficients β^{kj} from equation 3. Superscript k corresponds to groups of old system of credit scores, and superscript j corresponds to groups of the new system of credit scores. The coordinates $(x_1, y_1) - (x_2, y_2)$ in parenthesis represent the corresponding start and end points of each linear fit. For example, (1,1) - (1,3) corresponds to the linear fit between our estimates of coefficients $\{\beta^{11}, \beta^{12}, \beta^{13}\}$. In turn, (1,1) - (5,1) corresponds to the line between $\{\beta^{11}, \beta^{21}, \beta^{31}, \beta^{41}, \beta^{51}\}$

Table A3: Linear Fit between Estimates of β^{kj} from Equation 3 with Financial Delinquency as outcome

Vertical		**Horizontal**		**Diagonal**	
	(1)		(2)		(3)
Fit Between Coordinates		Coordinates		Coordinates	
(1,1) - (1,3)	-1373,5292	(1,1) - (5,1)	299.6956	(1,1) - (5,5)	367.4869
	(1372.4108)		(150.9276)		(254.7191)
(2,1) - (2,5)	1633.5986	(1,2) - (5,2)	1045.7595	(4,1) - (5,2)	-61.7526
	(302.1776)		(611.7093)		(.)
(3,1) - (3,5)	311.7982	(1,3) - (5,3)	776..0316	(3,1) - (5,3)	-129.2677
	(36.3450)		(642.3734)		(40.3109)
(4,1) - (4,5)	142.0436	(2,4) - (5,4)	-1224.3008	(2,1) - (5,4)	174.0391
	(57.9121)		(403.3562)		(177.7955)
(5,1) - (5,5)	46.5992	(5,4) - (5,5)	-1788.8731	(1,3) - (3,5)	2748.9782
	(37.5222)		(650.1486)		(3170.6027)
				(1,2) - (2,5)	1712.5215
					(1133.2098)

This Table shows coefficients and standard errors in parenthesis of linear fit between coefficients β^{kj} from equation 3 with financial delinquency as the outcome. Superscript k corresponds to groups of old system of credit scores, and superscript j corresponds to groups of the new system of credit scores. The coordinates $(x_1, y_1) - (x_2, y_2)$ in parenthesis represent the corresponding start and end points of each linear fit. For example, (1,1) - (1,3) corresponds to the linear fit between our estimates of coefficients $\{\beta^{11}, \beta^{12}, \beta^{13}\}$. In turn, (1,1) - (5,1) corresponds to the line between $\{\beta^{11}, \beta^{21}, \beta^{31}, \beta^{41}, \beta^{51}\}$

Table A4: Comparisons with Additional Sample

	(1) Credit Sample	(2) Additional Sample	(3) Difference
Female	0.446701	0.458609	0.011908
Nonwhite	0.235102	0.236513	0.001411
Less than H.S.	0.214555	0.201845	-0.012709
High School	0.561083	0.566405	0.005321
Some College	0.224362	0.231750	0.007388
Avg. Age	40.193400	39.146562	-1.046838
Old Sys. C. Sc.	467.149813	473.890127	6.740313
New Sys. C. Sc.	552.889273	548.229299	-4.659974
N. Observations	194235	560172	

This Table shows summary statistics of the sample used for credit analysis and the additional sample included for the entrepreneurship analysis. Credit Score values are from the last period before the implementation of the policy.

Table A5: Comparisons Entrepreneurship Analysis Sample (Risk Set)

	(1) Excluded	(2) Analysis Sample	(3) Difference
Female	0.455	0.455	0.0005
Nonwhite	0.165	0.248	0.0831
Less than H.S.	0.142	0.215	0.0731
High School	0.558	0.566	0.0080
Some College	0.299	0.218	-0.0812
Age	40.86	39.17	-1.696
Old Sys. C. Sc.	459.61	474.28	14.6
New Sys. C. Sc.	565.50	546.70	-18.79
N. Observations	109367	645040	

This Table shows summary statistics of the sample used in entrepreneurship analysis and the excluded sample. Excluded individuals are those who had already created a firm at any moment before three years of the implementation of the policy. Analysis sample includes the remaining individuals. Credit Score values are from the last period before the implementation of the policy.

B Proof of Propositions

Proposition: *For any given pair of old and new signals, s_i, s_i', and a given constant c, rationalizable changes in credit should follow*:

i. $h(s_i, s_i' + c) - h(s_i, s_i') > 0$

ii. $h(s_i + c, s_i') - h(s_i, s_i') < 0$

Given $B^*()$ is increasing

(i)

$$\begin{aligned} h(s_i, s_i' + c) - h(s_i, s_i') &= \Big(B^*(E[\theta|s_i' + c]) - B^*(E[\theta|s_i])\Big) - \Big(B^*(E[\theta|s_i']) - B^*(E[\theta|s_i])\Big) \\ &= B^*(E[\theta|s_i' + c]) - B^*(E[\theta|s_i']) \\ &> 0 \end{aligned}$$

(ii)

$$\begin{aligned} h(s_i + c, s_i') - h(s_i, s_i') &= \Big(B^*(E[\theta|s_i']) - B^*(E[\theta|s_i + c])\Big) - \Big(B^*(E[\theta|s_i']) - B^*(E[\theta|s_i])\Big) \\ &= B^*(E[\theta|s_i]) - B^*(E[\theta|s_i + c]) \\ &< 0 \end{aligned}$$

Proposition 2: *If the solution of the lenders problem $B^*(E[\theta|s_i'])$ is non-concave, for any given pair of old and new signals, s_i, s_i', and a given positive constant c($c \in \mathbb{R}, c > 0$), rationalizable changes in credit should follow:*

iii. $h(s_i + c, s_i' + c) - h(s_i, s_i') > 0$

we can rewrite

$$\begin{aligned} h(s_i + c, s_i' + c) - h(s_i, s_i') &= \Big(B^*(E[\theta|s_i' + c]) - B^*(E[\theta|s_i + c])\Big) - \Big(B^*(E[\theta|s_i']) - B^*(E[\theta|s_i])\Big) \\ &= \Big(B^*(E[\theta|s_i' + c]) - B^*(E[\theta|s_i'])\Big) - \Big(B^*(E[\theta|s_i + c]) - B^*(E[\theta|s_i])\Big) \end{aligned}$$

We know from UTS that:

$$E[\theta|s_i' + c] - E[\theta|s_i'] > E[\theta|s_i + c] - E[\theta|s_i]$$

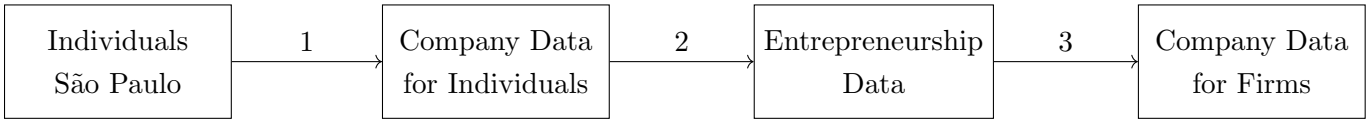
Thus, if $B^*()$ is not concave:

$$h(s_i + c, s_i' + c) - h(s_i, s_i') > 0$$

C Sample Construction Details

In this appendix, we describe our sample construction procedure. We use Figure A21 to guide us in the process.

Figure A21: Data Construction Diagram



We begin with a random sample of 200,000 individuals who live in Sao Paulo. This initial pool of individuals is selected from the universe of formal employees in Sao Paulo between 2015 and 2021, summed with the universe of entrepreneurs (including microentrepreneurs registered as MEI) in the state between 2015 and January 2024. I.e. if an individual had at least one formal job contract in the state of Sao Paulo in the period described above, they are in the pool of potential individuals in the sample. The end product of this step is a list of social security identifiers. In Brazil, the *Cadastro de Pessoa Fisica* (CPF) is the equivalent of Social Security Numbers.

We merge this data with a panel of credit information provided by SERASA. The end product of this step is a panel of individuals with their respective credit profiles.

In the next step, we merge this panel with Entrepreneurship data constructed by scrapping firm records from JUCESP and performing text analysis on them. This data has both the CPF of the entrepreneurs as well as the firm identification number (CNPJ) of the firm they own. We also add gender, age (in bins of five years), education, and race of entrepreneurs, as well as an indicator variable if the firm had multiple owners if the given individual was a founder or joined the company later, and an indicator if the firm is in the MEI tax system.

Due to data construction limitations, we make the entrepreneurship data set uniquely identified at the individual level. This implies that if an individual who owns multiple firms cannot have their CPF-CNPJ pair for all of them. We proceed with the following order of priorities to choose the firm we keep: First, we prioritize non-MEI firms, second we prioritize firms created after 2016, third, we prioritize firms employing someone, then we prioritize if the individual was a founder of that firm, lastly we take the oldest firm among the remaining ones.

In the last step, we merge firm-level information provided by SERASA about the firms created by the entrepreneurs in our sample. As a firm can be owned by multiple individuals, this procedure comprises a "many-1" merge, as potentially a group of individuals in our

sample owns the same firm.

After these steps, the CPF and CNPJ identifiers are masked, implying that all information provided by the company was not observed by non-company-affiliated researchers with their identifiers. It is important to highlight that the data provided by the researchers not affiliated with SERASA is not uniquely identified in any of the columns in our sample. That ensures that the researchers not affiliated with the company had no way to identify any of the individuals in our sample following the procedures of Brazilian law.

In Table A6 we summarize how the different pieces of data were merged.

Table A6: Merge Descriptions

	Master Base ("Left df')	Using Base ("Right df")	Key Variable Merge (key)	Match Type	"How"
→ 1	Sample of Individuals SP	Credit Panel of Individuals	CPF	1-many	Keep only the ones that match with how=='inner' in Python
→ 2	Individuals' Panel with credit information	Data on Entrepreneurship	CPF	many-1	Keep all from the master base , drop the ones from using that did not merge with how=='left' in Python
→ 3	Individuals in SP + Enriched + Entrepreneur.	Credit and Balance Panel Firms	CNPJ	many-1	Keep all from the master base , drop the ones from using that did not merge with how=='left' in Python

Notes: This table provides a summary of how the data was constructed combining information from the company with datasets provided by the non-company affiliated researchers.

D Non Parametric Estimation of $h(s_i, s_i')$

In addition to the linear projection of changes in credit into the s_i, s_i' presented in Figure 6, we can also estimate $h(s_i, s_i')$ non-parametrically using sieve-estimators. This is useful as it gives us an additional test of our framework's empirical propositions and also assesses the quality of the linear fit.

To estimate $h(s_i, s_i')$ non parametrically we write assume

$$h(s_i, s_i') \approx h_{np}(s_i, s_i') = \phi(s_i) + \phi(s_i')$$

where $\phi()$ are flexible functions of their arguments. We then estimate the following equation using our full sample:

$$Y_{it} = \alpha_i + \delta_t + \phi(\text{C. Sc. Old System}_i) \cdot Post_t + \phi(\text{C. Sc. New System}_i) \cdot Post_t + \varepsilon_{it} \quad \text{(D.11)}$$

where α_i, δ_t are individual and time fixed effects, and $Post_t$ is a dummy that takes value one for observations after the implementation of the policy. We approximate $\phi()$ using fifth-order polynomials. As in our previous analysis, credit Scores in both old and new systems are normalized measures with mean 0 and standard deviation 1 (Z-scores).

Our coefficient estimates for both linear and non-parametric models are presented in Table A8. But the visualization in Figure A22 is more intuitive. In Panels (a) and (b), we observe the fit of both the nonparametric and linear estimates over the space spanned by s_i, s_i'. In Panel (c), we show how the fit differs between both estimates over the same space.

Visually, the restricted model under the linear assumption is similar to the non-parametric estimation. The latter also predicts that changes in credit are increasing *vertically* and *diagonally*, and decreasing *horizontally* (we further show this below). By looking at Panel (c), we observe that the linear model does a particularly good job in predicting changes around the (0,0) point and a poorer job in the extremities. This is somewhat by construction as $h(0,0) = 0$ in both models. As discussed, this assumption comes from our conceptual framework. Despite the visual comparison between both estimates suggesting that they are similar, as somewhat expected given the size of our sample, tests of the linear restriction reject that the linear model is sufficient.

On top of the estimated changes in credit, we can also test the implications of our conceptual framework by looking at the derivatives from the estimated shape of the function. Given our estimated coefficients of function $\phi()$ we can recover numerical values for $\frac{\partial h_{np}(s_i, s_i')}{\partial s_i}$ and $\frac{\partial h_{np}(s_i, s_i')}{\partial s_i'}$.

We show the values of partial derivatives over the space spanned by (s_i, s_i') in Figure A23. In Panel (a), we show $\frac{\partial h_{np}(s_i, s_i')}{\partial s_i'}$, which we label as *vertical* changes consistent with our

terms in the paper. Points in which the surface is colored blue indicate positive estimates for the partial derivative, whereas negative values are colored in red. We can observe that partial derivatives are positive in almost every point of the distribution of credit scores, consistent with the Proposition in our conceptual framework.

Our framework also predicts that changes in credit should decrease *horizontally*. We show our estimates for $\frac{\partial h_{np}(s_i,s_i')}{\partial s_i}$ in Panel (b). We can see that red values dominate in the figure, indicating that in the majority of the distribution, credit changes decrease as we increase the value of the old system credit score.

Figure A23 also allows us to evaluate the last proposition of our framework by observing how changes in credit scores vary *diagonally*. Panel (c) plots values of $\frac{\partial \hat{h_{np}}(s_i,s_i')}{\partial s_i'} + \frac{\partial \hat{h_{np}}(s_i,s_i')}{\partial s_i}$, therefore representing a marginal change in the 45-degree lines summarized by $s_i = s_i' + c$. Our estimates suggest that in the majority of the joint distribution of credit scores, $h(s_i, s_i')$ increases in diagonal comparisons.

Lastly, we can compute the expected value of the derivatives. We do it in two different ways, each with its own interpretation. First, we compute the average of over a grid of equally 1000 points in a space of $s_i \in (-1.5, 1.5) \times s_i' \in (-1.5, 1.5)$. We can define a distribution $\mathcal{F}(s_i, s_i')$ for this grid with density function $f(s_i, s_i') = 1/1000$ for points s_i, s_i' in the grid and 0 otherwise. Our average derivatives will then be:

Vertical:

$$\mathbb{E}\left[\frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i'}\right] = \int \frac{\partial h_{np}(s_i, s_i')}{\partial s_i'} d\mathcal{F}(s_i, s_i')$$

Horizontal:

$$\mathbb{E}\left[\frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i'}\right] = \int \frac{\partial h_{np}(s_i, s_i')}{\partial s_i} d\mathcal{F}(s_i, s_i')$$

Diagonal:

$$\mathbb{E}\left[\frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i'}\right] = \int \frac{\partial h_{np}(s_i, s_i')}{\partial s_i'} d\mathcal{F}(s_i, s_i') + \int \frac{\partial h_{np}(s_i, s_i')}{\partial s_i'} d\mathcal{F}(s_i, s_i)$$

With this method, we input equal weights for all points in the space spanned by s_i, s_i'. Thus, the moment we estimate could be considered an *unweighted* average of partial derivatives.

In another way, we use our sample of individuals with their s_i, s_i' and compute the average derivatives for each individual using the estimated partial derivatives. We could think of this as a *weighted* average over the joint distribution of signals, with the weights being the empirical distribution of signals. We compute the moments as follows:

Vertical:

$$\mathbb{E}\left[\frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i'}\right] = \frac{1}{N}\sum_i \frac{\partial h_{np}(s_i, s_i')}{\partial s_i'}$$

Horizontal:

$$\mathbb{E}\left[\frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i}\right] = \frac{1}{N}\sum_i \frac{\partial h_{np}(s_i, s_i')}{\partial s_i}$$

Diagonal:

$$\mathbb{E}\left[\frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i} + \frac{\partial \hat{h_{np}}(s_i, s_i')}{\partial s_i'}\right] = \frac{1}{N}\sum_i \frac{\partial h_{np}(s_i, s_i')}{\partial s_i} + \frac{1}{N}\sum_i \frac{\partial h_{np}(s_i, s_i')}{\partial s_i'}$$

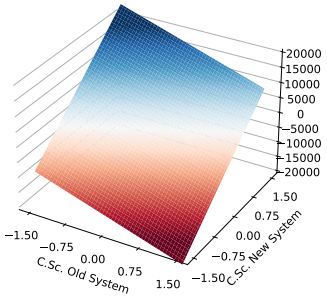
Our estimates for both methods suggest that on the three *directions*, the non-parametric model goes in line with our conceptual framework predictions. Changes in credit increase with respect to the value of the new signal and decrease with respect to the old signal. Furthermore, *vertical* increases are larger than *horizontal* ones, thus indicating that changes in credit also increase *diagonally*.

Table A7: Expected Values of Partial Derivatives in the Non-Parametric Estimation

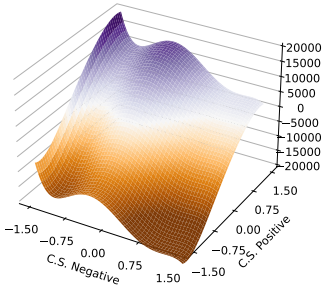
	(1) Over the Grid		(2) Over the Sample	
	Partial Derivative	Value	Partial Derivative	Value
Vertical	$\int \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i'} d\mathcal{F}(s_i, s_i')$	9137.39	$\frac{1}{N}\sum_i \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i'}$	6731.50
Horizontal	$\int \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i} d\mathcal{F}(s_i, s_i')$	-6104.38	$\frac{1}{N}\sum_i \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i}$	-6278.01
Diagonal	$\int \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i'} d\mathcal{F}(s_i, s_i') + \int \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i} d\mathcal{F}(s_i, s_i')$	3033.01	$\frac{1}{N}\sum_i \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i'} + \frac{1}{N}\sum_i \frac{\partial \hat{h}_{np}(s_i,s_i')}{\partial s_i}$	453.49

In this Table we present the estimated values of partial derivatives using non parametric estimation. *Over the grid* estimates assume a 1000 point grid over $s_i \in (-1.5, 1.5) \times s_i' \in (-1.5, 1.5)$ with equal weights to each point. In turn, *Over the Sample* calculates partial derivatives to every individual in our sample and weights each individual equally.

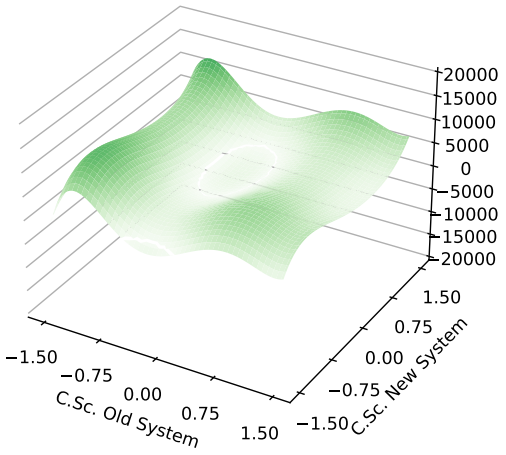
Figure A22: Comparisons Between Linear and Non-Parametric Estimates of Credit Change



(a) Linear Function



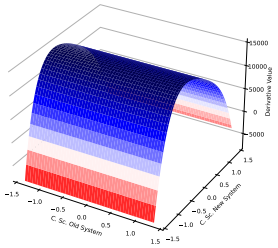
(b) Non-Parametric Function



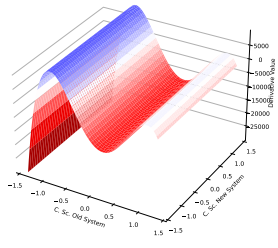
(c) Difference (a) - (b)

Panel (a) and (b) show predicted values of changes in credit over the space spanned by old and new system credit scores using our linear restricted model (equation 4) and the non-parametric model (equation D.11), respectively. Panel (c) shows the differences between estimates from (a) and (b).

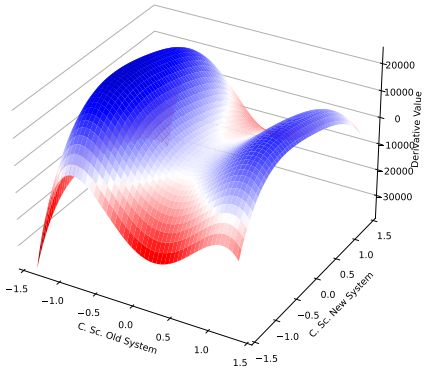
Figure A23: Comparisons Between Linear and Non-Parametric Estimates of Credit Change



(a) $\frac{\partial \hat{h}_{np}(s_i, s_i')}{\partial s_i'}$



(b) $\frac{\partial \hat{h}_{np}(s_i, s_i')}{\partial s_i}$



(c) $\frac{\partial \hat{h}_{np}(s_i, s_i')}{\partial s_i'} + \frac{\partial \hat{h}_{np}(s_i, s_i')}{\partial s_i}$

This Figure plots partial derivatives calculated over the joint distribution of credit scores in the old and new systems. We use our estimates from equation D.11 and calculate partial derivatives over a grid of 100 equally distributed points over the space spanned by s_i, s_i'

Table A8: Coefficients of Linear and Nonlinear Estimates

	(1)	(2)
	Dependent variable: Credit	
	Linear	Polynomials
C. Sc. Old System	-5172.3***	-11026.4486***
	(273.66)	(1471.2169)
C. Sc. Old System^2		-9494.6893***
		(1087.5093)
C. Sc. Old System^3		11295.7751***
		(2009.2653)
C. Sc. Old System^4		3954.4166***
		(507.6831)
C. Sc. Old System^5		-4041.9399***
		(651.5566)
C. Sc. New System	10951.58***	14178.6652***
	(370.70)	(721.0755)
C. Sc. New System^2		-2217.2853***
		(732.6252)
C. Sc. New System^3		-1089.4464**
		(532.9137)
C. Sc. New System^4		636.4236***
		(226.3324)
C. Sc. New System^5		-508.6011***
		(184.7629)
Observations	2875942	2875942

This Table presents coefficient estimates of equations 4 and D.11. Credit Scores in both old and new systems are normalized measures with a mean of 0 and a standard deviation of 1 (Z-scores). Standard errors in parenthesis are clustered at the individual level.

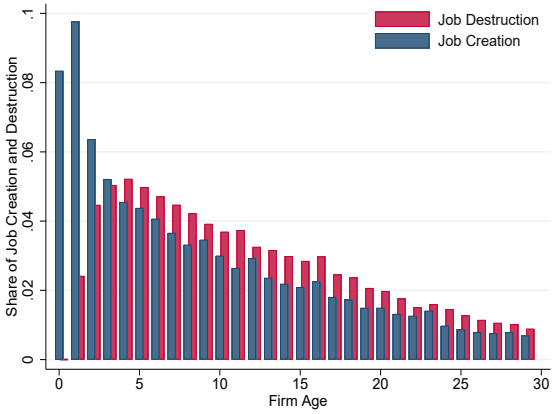
E Appendix - Entrepreneurs and Job Creation in Sao Paulo

Startups in São Paulo mimic behavior that has been documented in the United States. They are responsible for a large share of job creation in the economy. Nevertheless, most startups start small and remain small through time, conditional on survival. In this section, we replicate patterns documented for the U.S. in Decker et al. (2014) combining our entrepreneurship records from JUCESP with matched employer-employee data that covers the universe of formal labor markets (RAIS).

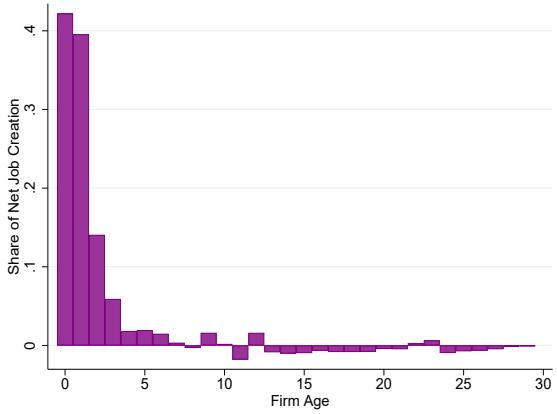
These patterns are illustrated in Figure A24, which are constructed using firm-level data from 2003-2015. We use an older period for this analysis as by the time this paper is being written, more recent RAIS data that allows us to follow firms' for at least 5 years after their creation is not yet available.

Panel A and B reflect job creation and destruction. A job is considered to be created if, for a given firm, one employee was added between year t and year t+1. The opposite holds for job destruction. Panel A shows that firms in their first 5 years of existence are responsible for around 35% of all jobs created in the period. In Panel B, we see that net job creation, calculated as job creation minus job destruction, is especially high in the first two years of a firm's history. This is found mechanically, as firms that have just been opened have no job destruction. In Panel C, we see that the median growth in the number of employees is 0 throughout all the first 15 years of firms' existence. Furthermore, we see that in the firms' first years, there is a higher variance in firm growth, which stabilizes around year 6. Lastly, in Panel D, we observe that firms are usually small in number of employees. The median startup in Sao Paulo starts with 3 employees, and by year 10, among active firms, the median firm still only employs 5 people.

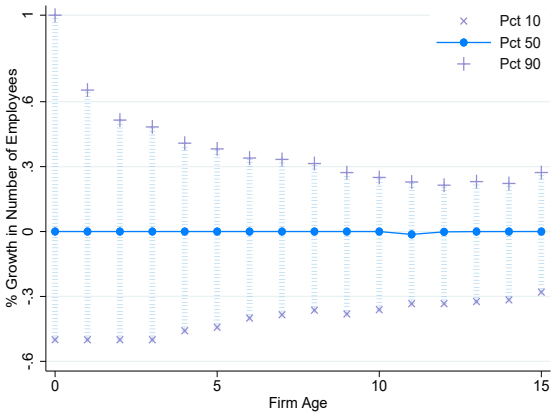
Figure A24: Startups Job Creation and Growth



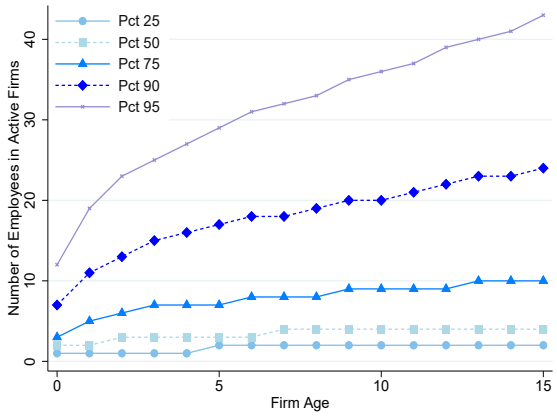
A. Job Creation and Destruction by Firm's Age



B. Net Job Creation by Firm's Age



C. Distribution of Growth by Firms' Age



D. Distribution of Firm Size by Firms' Age

Notes: Write

F Informality

In our main entrepreneurship analysis, our focus lies on formal firms. This is due to the fact that both our firm ownership records and the credit bureau data comprise only information from firms with a registration number in the *Cadastro Nacional de Pessoas Jurídicas* (CNPJ). In this Appendix, we discuss the role of informality in entrepreneurship, providing some descriptive evidence using the National Household Survey (PNAD-C) that encompasses both the formal and informal sectors.

First, we describe our informality definition. Brazil is a setting where this is very clear. For employees, all formal labor relations must have a "signed" *carteira de trabalho*. [36] Although this dichotomous distinction masks informal labor ties between formal workers (see Feinmann et al. (2022) for a full discussion about this issue), it is still useful as it is observable in the household survey. Informal employees represent 15% of all employees in the state of São Paulo and 13% of all non-domestic employees.

In the case of firms, we use as an indicator of formality the registration in the *Cadastro Nacional de Pessoas Jurídicas*. is the Brazilian National Registry of Legal Entities. It is a unique identifier issued by the Brazilian Federal Revenue Service (Receita Federal) to businesses and other legal entities operating in Brazil. The CNPJ number is similar to a tax identification number in other countries and is required for businesses to conduct legal and financial activities within the country. In PNAD-C, self-employed individuals and employers are asked if their business is registered in CNPJ.

Given the definition of formal businesses, we can describe the differences between formal and informal entrepreneurs in this setting. Given the PNAD-C questionnaire, we have to use a broad definition of entrepreneurs, which encompasses employers and self-employed individuals without employees. In our sample, 25% of entrepreneurs are employers.

We summarize the informality levels of entrepreneurs in Table A9. On average, 90.8% of employers' businesses are registered. In turn, 38.5% of self-employed individuals have a registration in CNPJ. When summing employers and self-employed, 51.4% of entrepreneurs have their businesses registered.

We further describe the informality rates of self-employed by economic sectors in Table A10. In Columns (1) and (2) we show the share of self-employed and employers in each sector. In Columns (3) and (4) we show the share of formal businesses in each sector. We observe that construction sector has the lowest formalization rates both among employers and self-employed individuals.

Lastly, we characterize the earnings distributions of formal and informal entrerpeneurs

[36] In the past, this consisted of an actual booklet for each employee, in which employers had to sign and register contract wages and hours. Nowadays, the process can be done fully electronically. Nonetheless, Brazilians still refer to forma job ties as those with "signed" *carteira de trabalho*.

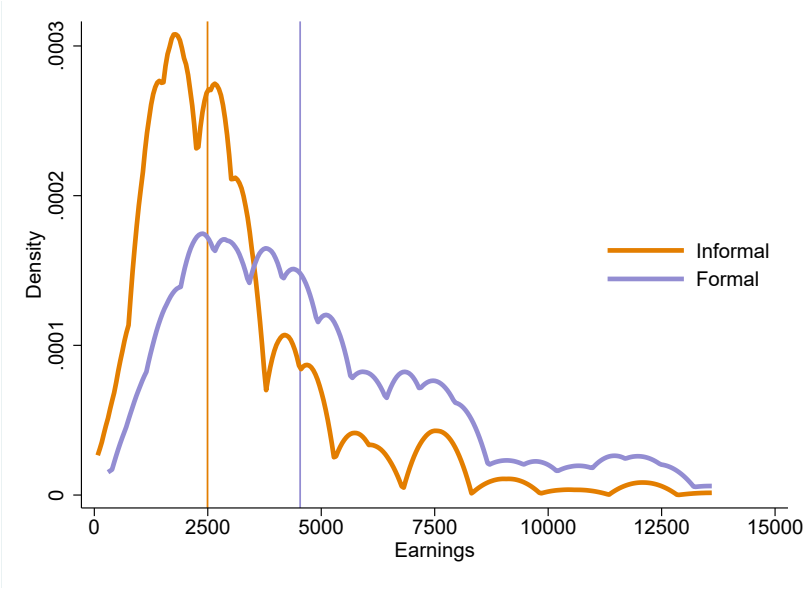
Table A9: Informality among Entrepreneurs

	(1) Informal	(2) Formal
Self-Employed	61.51	38.49
Employers	9.16	90.84
Entrepreneurs	48.63	51.37

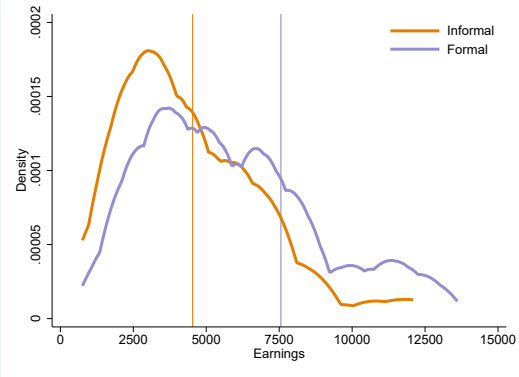
The sample is restricted to those between 18 and 65 years old in the State of São Paulo working at least 30 hours a week as entrepreneurs in their main job. Employers are those who self-report as *empregadores*, and self-employed are those who report as *conta-própria*. Entrepreneurs encompass both groups. Formal is defined as if their business had a registration in the CNPJ.

in Figure A25. We observe that the earnings of informal entrepreneurs are substantially lower than those of formal ones. This is valid for both employers and employees. The median earnings of formal entrepreneurs are almost double that of informal ones.

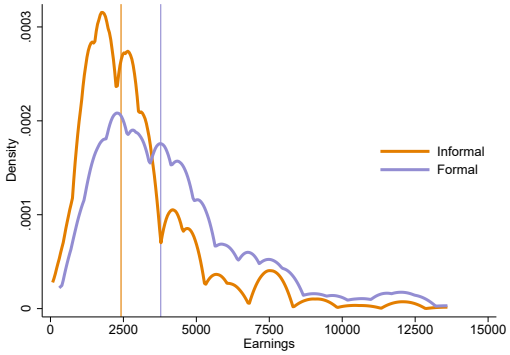
Figure A25: Entrepreneurs Earnings in the Formal and Informal Sector



(a) Entrepreneurs



(b) Employers



(c) Self-Employed

This Figure shows the distribution of labor earnings of entrepreneurs in the State of São Paulo in the last quarter of 2019 built with PNAD-C. The sample is restricted to those between 18 and 65 years old working at least 30 hours a week as entrepreneurs in their main job. Values are adjusted for December 2023 BRL. Employers are those who self-report as *empregadores*, and self-employed are those who report as *conta-própria*. Entrepreneurs encompass both groups. Formal is defined as if their business had a registration in the CNPJ. Vertical lines correspond to median earnings of each group.

Table A10: Share of Business Registration by Economic Sectors

	(1)	(2)	(3)	(4)
	Share in each Sector		Share of Formal Businesses	
	Self-Employed	Employers	Self-Employed	Employers
Agriculture	3.56	4.32	57.32	85.21
Manufacturing	8.61	9.45	32.41	92.39
Construction	17.61	6.59	20.40	71.47
Retail	20.19	31.39	51.62	97.38
Transport & Food	19.88	17.06	28.96	88.10
Professional Services	17.64	25.23	51.84	90.90
Domestic Services	12.51	5.97	38.11	87.05

The sample is restricted to those between 18 and 65 years old in the State of São Paulo working at least 30 hours a week as entrepreneurs in their main job. Employers are those who self-report as *empregadores*, and self-employed are those who report as *conta-própria*. Entrepreneurs encompass both groups. Formal is defined as if their business had a registration in the CNPJ. Sectors are constructed using the first 2-digit of cnae as follows values 1-9 (Agriculture), 2 for 10-39 (Manufacturing), 3 for 41-43 (Construction), 4 for 45-48 (Retail), 5 for 49-56 (Transport & Food), 6 for 58-88 (Professional Services), and 7 for 90-99 (Domestic Services).