The Equilibrium Effects of Information Deletion: Evidence from Consumer Credit Markets

Andres Liberman Christopher Neilson Luis Opazo Seth Zimmerman

November 2018

Abstract

This paper exploits a large-scale natural experiment to study the equilibrium effects of information restrictions in credit markets. In 2012, Chilean credit bureaus were forced to stop reporting defaults for 2.8 million individuals (21% of the adult population). We show that the effects of information deletion on aggregate borrowing and total surplus are theoretically ambiguous and depend on the pre-deletion demand and cost curves for defaulters and non-defaulters. Using panel data on the universe of bank borrowers in Chile combined with the deleted registry information, we implement machine learning techniques to measure changes in lenders' cost predictions following deletion. Deletion reduces (raises) predicted costs the most for poorer defaulters (non-defaulters) with limited borrowing histories. Using a difference-in-differences design, we find that individuals exposed to increases in predicted costs reduce borrowing by 6.4%, while those exposed to decreases raise borrowing by 11.8% following the deletion, for a 3.5% aggregate drop in borrowing. Using the difference-in-difference estimates as inputs into the theoretical framework, we find evidence that deletion reduced aggregate welfare under a variety of assumptions about lenders' pricing strategies.

Keywords: Information asymmetry, consumer credit
JEL codes: G20, D14, D82

*Andres Liberman is at New York University, email: aliberma@stern.nyu.edu. Christopher Neilson is at Princeton University, email: cneilson@princeton.edu. Luis Opazo is at ABIF, email: lopazo@abif.cl. Seth Zimmerman is at University of Chicago Booth School of Business, email: seth.zimmerman@chicagobooth.edu. Previous drafts of this paper were circulated under the title "The Equilibrium Effects of Asymmetric Information: Evidence from Consumer Credit Markets." We thank Andrew Hertzberg, Amir Kermani, Neale Mahoney, Holger Mueller, Christopher Palmer, Philipp Schnabl, Johannes Stroebel, and numerous seminar participants for comments and suggestions. Sean Hyland and Jordan Rosenthal-Kay provided excellent research assistance. This research was funded in part by the Fama-Miller Center for Research in Finance and the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business. We thank Sinacofi for providing the data. All errors and omissions are ours only. First version: October 2017. Online Appendix available at http://faculty.chicagobooth.edu/seth.zimmerman/research/papers/LNOZ_Online_Appendix.pdf

1 Introduction

Many countries have institutions that limit the information available to consumer lenders. For example, in 2007, over 90% of countries with credit bureaus also had provisions that erased defaults after set periods of time (Elul and Gottardi 2015). Other forms of information limits include restrictions on the types of past borrowing outcomes and demographic variables that can be used to inform future lending decisions, and one-time purges of default records. The stated motivation for these policies is often that allowing lenders access to certain kinds of information unfairly reduces borrowing opportunities for individuals with past defaults (Miller 2003, Steinberg 2014), who may be disproportionately drawn from disadvantaged groups or have suffered from a negative past shock such as a natural disaster, an economic downturn, or a health event.

Several recent empirical studies confirm that deleting default records increases borrowing for beneficiaries (Bos and Nakamura 2014, González-Uribe and Osorio 2014, Herkenhoff, Phillips and Cohen-Cole 2016, Liberman 2016, Dobbie, Goldsmith-Pinkham, Mahoney and Song 2016).[1] However, theory emphasizes that the implications of these institutions for aggregate lending and the distribution of access to credit depend not just on how they affect the beneficiaries of deletion, but on the information asymmetries they induce in consumer credit markets and the equilibrium responses by lenders (Akerlof 1970, Jaffee and Russell 1976, Stiglitz and Weiss 1981). Individuals whose credit information is deleted benefit if lenders perceive them as more willing or able to repay their loans. But this gain may come at a cost to the non-defaulters with whom defaulters are pooled. The effects of information-limiting institutions depend on the tradeoff between these two groups.

This paper develops a simple framework for analyzing the equilibrium effects of information deletion in markets with asymmetric information and uses it to study a policy change that forced Chilean credit bureaus to cease reporting information on defaults for most consumer defaulters. To begin, we show that the effects of information deletion on aggregate borrowing and welfare outcomes are theoretically ambiguous and depend on a) the demand and average cost curves in the pre-deletion equilibrium and b) the changes in lenders' cost beliefs that result from deletion. We then take the model to the data. We use machine learning techniques to show that deletion reduces predicted costs for defaulters and raises predicted costs for non-defaulters, with the largest increases for low-income non-defaulters who resemble defaulters but have good credit records. Our cost predictions form the basis for a difference-in-differences design com-

[1] See also Musto (2004) and Brown and Zehnder (2007).

paring outcomes for winners and losers from the policy to a "control group" for whom cost predictions are unchanged. We find that increases in borrowing for defaulters are more than offset by losses for non-defaulters, so that on net deletion reduced borrowing by 3.5% in both markets combined. Taking the difference-in-difference estimates as inputs to our theoretical framework, we conclude that deletion reduced welfare under a variety of assumptions about lenders' pricing strategies.

Underlying our analysis is a large-scale change in the information available to lenders in Chilean consumer credit markets. In February 2012, the Chilean Congress passed Law 20,575 (henceforth, the "policy change"), which contained a provision that forced all credit bureaus operating in the country to stop reporting individual-level information on certain defaults.[2] The policy change affected information for all individuals whose defaults as of December 2011 added up to less than 2.5 million Chilean pesos (CLP; roughly USD $5,000). This group made up 21% of all Chilean adults and 84% of all bank borrowers in default at the time of implementation. After the deletion, credit bureau information no longer distinguished individuals with deleted records from those with no defaults. The policy change was a one-time deletion and did not affect how subsequent defaults were recorded. Three years after the one-time deletion, the count of individuals reported as in default in the credit bureau had nearly returned to its pre-deletion level and was still rising.

We study this change using an empirical framework that takes an unraveling model in the style of Akerlof (1970) and Einav, Finkelstein and Cullen (2010) as a baseline. We extend this framework by considering the effects of pooling multiple submarkets following the deletion of differentiating information. Focusing on a simple case with average cost pricing and linear demand and cost curves in markets for defaulters and non-defaulters (henceforth, high- and low-cost markets), we show that the welfare and borrowing effects of pooling the two markets are ambiguous. The key tradeoff is that, when the deleted information predicts costs, deletion reduces welfare losses from adverse selection for the high-cost group but increases these losses for the low-cost group. The size of these effects depends only on borrowers' demand and cost curves. The benefit of the framework is that it yields a simple mapping between these relatively easy to estimate objects and welfare.

We construct empirical estimates of cost predictions, demand curves, and cost curves using panel data on the universe of bank borrowers in Chile. These data include the information deleted following the policy change and made unavailable to lenders. Our empirical analysis proceeds in two main steps. First, we use machine learning tech-

[2]See http://www.leychile.cl/Navegar?idNorma=1037366 for the complete text.

niques to construct cost predictions with and without the deleted credit information. This helps us understand how lenders' cost expectations shift following deletion. Second, we use variation in cost expectations generated by information deletion to estimate the slope of borrowers' demand and cost curves. The second step employs a difference-in-differences identification strategy that looks at at how borrowing changes following deletion for individuals for whom cost expectations rise and fall.

Our administrative data track defaults and borrowing outcomes between 2009 and 2015, the period surrounding the policy change. Before the policy change, all banks had access to two types of information on lending outcomes for individuals. The first was a database of borrowing and default outcomes for bank borrowing only.[3] This dataset was shared by all banks through the Chilean banking regulator and allowed banks to observe bank borrowing and default even for loans originated at other banks. The second was the credit bureau, which contained default amounts for both bank and non-bank loans. In addition to bank defaults, records in the credit bureau included defaults on retail (department store) credit cards, as well as other missed payments such as bounced checks. After the policy change, banks lost access to credit bureau data but kept access to bank data. The effect of the policy was thus to cut off banks' information on non-bank defaults for borrowers dropped from the credit bureau. We have access to both bank borrowing and credit bureau data, including credit bureau records from before the deletion to which banks were subsequently denied access.

We combine these data with a machine learning approach to show how deletion of credit bureau data affects the predictions banks can make about borrowers' costs (Mullainathan and Spiess 2017). We use a random forest algorithm to construct two sets of predictions about borrower costs. The first uses both bank borrowing data and credit bureau records, while the second uses only the bank borrowing data and not deleted credit bureau records. Eliminating credit bureau data reduces both in- and out-of-sample log likelihoods of observed values given predictions and leads to systematic overestimates of default probabilities for borrowers without defaults and underestimates for borrowers with defaults.

We define exposure to the policy as percent increase in predicted costs following deletion. Because non-defaulters outnumber defaulters, exposure is positive (i.e., predicted costs rise) for 61% of the population. The individuals with the largest exposure borrow small amounts and do not have bank or non-bank defaults. They are on average poorer and less likely to own homes. These individuals resemble the borrowers for whom costs fall most dramatically, except that they do not show up on the credit

[3] See Liberman (2016) and Cowan and De Gregorio (2003) for details on Chilean credit bureaus.

bureau as in default. In contrast, predicted costs for individuals who borrow large amounts with higher rates of bank default do not change after deletion.

We use a difference-in-differences analysis to unpack the causal effect of the deletion of information across the exposure distribution. This analysis compares changes in borrowing (and costs) for individuals exposed to increases and decreases in lenders' cost predictions to changes in borrowing for individuals for whom the deleted default records are uniformative about costs. To do this, we use snapshots of borrower and credit bureau data at six month intervals leading up to and including the December 2011 snapshot to identify groups of borrowers who would have been exposed to positive, negative, and zero changes in cost predictions had deletion taken place at that time. We use interactions between the predicted exposure variables and a dummy equal to one for cohorts exposed to the actual deletion policy–the December 2011 snapshot–to estimate the effects of deletion in the positive- and negative-exposure group.

There are two assumptions underlying this analysis. The first is that trends in the positive- and negative-exposure groups would have evolved in parallel to the zero-exposure group in the absence of the deletion policy. We evaluate this assumption using a standard analysis of pretrends. The second is that deletion does not affect outcomes for individuals whose cost predictions do not change, i.e., the zero-exposure group. We evaluate this assumption using a supplemental difference-in-differences analysis that compares borrowing for defaulters in the zero-exposure group above the deletion cutoff whose information was not deleted to borrowing for below-threshold borrowers in the zero-exposure group whose information was deleted. Within this supplemental analysis, we find that deletion did not affect borrowing for the no-change group.

We find that quantities borrowed by the negative- and positive-exposure groups move in parallel to the zero exposure group during the pre-deletion period. Following deletion, borrowing jumps up by 11.7% for the group exposed to cost decreases (on a baseline mean of $141,000 CLP) and falls by 6.4% for the group exposed to cost increases (on a baseline mean of $215,000 CLP). Lenders' cost predictions fall by 29% in the former group and rise by 22% in the latter, corresponding to elasticities of lending to predicted costs of -0.40 and -0.29 in the positive and negative exposure groups, respectively.

Because more borrowers are exposed to increases in predicted costs than to decreases, these estimates mean that the aggregate effect of deletion across the two groups was to reduce borrowing by 3.5%. The total value of the reduction in borrowing is about $20 billion CLP over a six-month period, or $40 million USD. Aggregate declines are largest as a share of borrowing for lower-income borrowers: borrowing drops by 4.2% for lower-income individuals and by 3.7% for individuals without mortgages.

Though deletion reduces borrowing, it could still raise total surplus if the individuals for whom borrowing rises value that borrowing more relative to costs than those for whom it falls. This could occur if borrowers exposed to decreases in predicted costs suffer more from adverse selection than borrowers exposed to increases in predicted costs at baseline. Repeating our difference-in-difference analysis with actual costs as the dependent variable suggests that realized costs change only slightly as quantity varies in both markets. The signs of our estimated cost effects are consistent with adverse selection at the margin in both markets, but we cannot reject effects of zero.

We formalize the welfare analysis by mapping the market for borrowers who faced decreases and increases in predicted costs following the deletion to the high and low-cost markets described in our framework, respectively. We use the estimated quantity, predicted cost, and realized cost effects from our difference-in-difference analysis as inputs into the model. Together with baseline mean values, these quasi-experimental effect estimates identify welfare effects under the assumption of average cost pricing. We find that pooling increases welfare losses from adverse selection by 66% relative to the no-pooling equilibrium. The high-cost market switches from underprovision in the unpooled equilibrium to overprovision in the pooled market, with smaller welfare losses in the latter case. Underprovision increases in the low-cost market.

One limitation of the benchmark welfare analysis is that it imposes average cost pricing. In practice lenders likely mark up prices over costs.[4] We do not observe interest rates for individual loans. However, our qualitative findings of a welfare loss hold over a wide range of possible markup values in the low- and high-cost markets. Another limitation is that our measure of default is only a proxy for lenders' total cost of making a particular loan. However, our findings do not change when we use alternate plausible cost measures. Finally, deletion may have dynamic welfare effects or welfare effects outside of the credit market. For example, periodic information deletion may help insure against the ex ante 'reclassification' risk of defaulting and losing access to credit markets (Handel, Hendel and Whinston 2015), or may induce externalities in labor markets (Bos, Breza and Liberman 2018, Herkenhoff et al. 2016, Dobbie et al. 2016).[5] One may view our findings on the borrowing and welfare effects of deletion as measures of the costs of providing insurance and benefits outside the credit market.

In the final section of the paper, we use our procedure to study the effects of two counterfactual policies that limit information available to lenders: deleting bank default records in addition to credit bureau default records, and deleting information

[4]E.g., Ausubel (1991) shows evidence of lack of competition in the US credit card market.

[5]See also Clifford and Shoag (2016), Bartik and Nelson (2016) and Cortes, Glover and Tasci (2016).

on gender (Munnell, Tootell, Browne and McEneaney 1996, Blanchflower, Levine and Zimmerman 2003, Pope and Sydnor 2011). Deleting additional default information increases the spread of changes in predicted costs, with bigger gains for winners and losses for losers than in the policy as implemented. Deleting information on gender increases costs disproportionately for women. The theme here is that the costs of deletion fall mostly on individuals observably similar to the intended beneficiaries.

This paper contributes to a broader literature on the empirics of asymmetric information. Our finding that deleting information reduces overall borrowing and that costs fall most heavily on non-defaulters who resemble defaulters is similar to Agan and Starr (2017), which shows that restricting information on criminal records in job applications reduces callback rates for black applicants. We show how a machine learning approach can identify individuals affected by deletion policies, and develop a framework that can be used to evaluate welfare effects. From a methods perspective, we innovate by bringing "sufficient statistics" models common in public finance (Chetty 2009, Hackmann, Kolstad and Kowalski 2015) to a consumer finance setting. Several papers estimate structural models of consumer credit markets and use them to simulate the effects of policy changes (Adams, Einav and Levin 2009, Einav, Jenkins and Levin 2013, Einav, Jenkins and Levin 2012). In contrast, we study a simple model of credit demand and identify elasticities governing welfare outcomes using a natural experiment.

Finally, we contribute to a literature that uses machine learning to explore treatment effect heterogeneity given access to many possible mediating variables (Athey and Imbens 2016, Athey and Wagner 2017), and to generate counterfactuals that allow for causal inference where no credible experiment exists (e.g., Burlig, Knittel, Rapson, Reguant and Wolfram (2017)).[6] In contrast to this work, we focus on theoretically-motivated measures of predicted average costs as the key determinant of heterogeneous treatment effects. This reduces the set of causal parameters required to apply our approach in other settings from a potentially large number of heterogeneous effects defined across interactions of mediator variables to a single set of elasticities. Our approach complements the 'big data' that is increasingly prevalent in credit markets and other settings (Petersen and Rajan 2002, Einav and Levin 2014).

[6] See Varian (2016) or Mullainathan and Spiess (2017) for a review. Several other papers employ machine learning techniques to study credit markets. These include Huang, Chen and Wang (2007), Khandani, Kim and Lo (2010) and Fuster, Goldsmith-Pinkham, Ramadorai and Walther (2017). These papers focus on using machine learning techniques to improve cost prediction. In contrast, we use ML techniques to study the effects of actual and counterfactual policy changes on borrowing.

2 Economic Framework

2.1 Model

This section presents a framework to help interpret policies that limit credit information and to motivate our empirical approach. Our focus is on understanding how deletion affects welfare and borrowing outcomes through adverse selection, not moral hazard. This is consistent with the empirical application we study here, a one-time deletion based on characteristics that were predetermined at the time of policy announcement. Other applications for which it might be relevant are restrictions on the use of personal characteristics such as race or age in lending and insurance markets. With the goal of transparent empirical implementation, we focus on a simple model of market unraveling in the style of Akerlof (1970).

Consider a consumer credit market where lenders set prices on the basis of observable borrower characteristics but borrowers may have private information on their own costs. Assume for simplicity that the lending market is competitive, so that in equilibrium prices are equal to average costs. This market structure parallels and builds on the Einav et al. (2010) model of insurance under adverse selection in that lenders set prices and quantities are endogenously determined.

Individual borrowers are denoted by i. Lenders partition markets using two types of borrower observable characteristics. The first type, X_i, are always observable to lenders. The second, $Z_i \in \{0, 1\}$, are variables that will be deleted from the lender's information set, e.g., by the policy change. In what follows we supress X_i notation. One can think of this analysis as taking place within subgroups of borrowers defined by $X_i = x$. We model $Z_i = 1$ as being a default flag that predicts higher costs. There are a unit measure of borrowers in the market, of whom a fraction α have $Z_i = 0$ and a fraction $1 - \alpha$ have $Z_i = 1$. Demand and cost functions may vary across values of Z_i. Let $q_z(R)$, $MC_z(R)$, and $AC_z(R)$ denote the demand for credit, marginal cost, and average cost functions for type $Z_i = z$ as a function of the lender's (gross) offer rate R. $q_z(R)$ denotes the *average* quantity of credit purchased for individuals in the market, so that total market quantity is given by $\alpha q_0(R)$ for $Z_i = 0$ and $(1 - \alpha)q_1(R)$ for $Z_i = 1$. To guarantee unique equilibria, we assume that the (inverse) demand curve crosses the marginal cost curve from above exactly once in each market. For analytic tractability, we further assume that the demand and cost curves are linear.

2.1.1 Pre-deletion equilibria

When lenders observe Z_i, equilbria are defined by the intersection of inverse demand and average cost curves in each market. Letting $R_z(q)$ represent the inverse demand curve in each market, equilibrium quantities q_z^e are determined by $R_z(q_z^e) = AC_z(R_z(q_z^e))$. Let $AC_z^e = AC_z(R_z(q_z^e))$ denote the equilibrium average cost in each market.

Following Einav et al. (2010), the slopes of the cost and demand curves determine the losses in total surplus from asymmetric information. We focus on the empirically relevant case where there is adverse selection in both markets; i.e., where marginal cost curves are downward sloping. We illustrate this in Figure 1. The surplus-maximizing quantity q_z^* is determined by $R_z(q_z^*) = MC_z(q_z^*)$. We denote the surplus-maximizing rate as $R_z^* = R_z(q_z^*)$. Deadweight loss due to asymmetric information in market z is the area of the shaded triangle (denoted by "A" in the high cost market and "B" in the low-cost market in Figure 1, respectively), with total welfare loss in each market given by the formula:

$$DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC_z^e)\right) \times \left(AC_z^e - MC_z(AC_z^e)\right). \tag{1}$$

2.1.2 Deletion policy

In the pooling equilibrium lenders no longer observe Z_i. Demand in the pooled market at price R is given by $q(R) = q_0(R) + q_1(R)$, and the pooled market average cost is $AC(R) = s(R)AC_0(R) + (1 - s(R))AC_1(R)$, where the low-cost share $s(R)$ is defined as $s(R) = \frac{\alpha q_0(R)}{\alpha q_0(R)+(1-\alpha)q_1(R)},$. The equilibrium price/average cost AC^e and quantity q^e are determined by $AC^e = AC(R(q^e))$. The changes in average borrowing from pooling in each market are then given by:

$$\Delta q_z = q_z(AC^e) - q_z(AC_z^e),$$

and the average welfare loss by:

$$DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC^e)\right) \times \left(AC^e - MC_z(AC^e)\right).$$

Changes in total surplus from pooling are determined by the relationship between the group-specific demand and cost curves and the pooled average costs. For individuals with $Z_i = 0$ at baseline, rising rates due to pooling increase surplus losses due to underprovision of credit. These additional losses are denoted by "D" in the left panel of Figure 1, the low-cost market. For individuals with $Z_i = 1$, the effects of pooling on total surplus are ambiguous. If $AC^e > R_1^*$, then the effects of the policy for this group are unambiguously positive, as pooling reduces the underprovision of credit due to adverse selection. If $AC^e < R_1^*$, then the effects are unclear. Losses from overprovision in the pooled market may outweigh losses from underprovision in the segregated market. Figure 1 illustrates the latter case, with surplus losses from overprovision equal to the area of triangle C in the left panel.

2.1.3 Measuring the effects of pooling

The equilibrium borrowing and welfare effects of pooling are determined by the slopes of the demand and cost curves in the high- and low-cost markets. Given observations of unpooled quantities q_z^e, costs AC_z^e, and slopes $\frac{dq_z}{dR}$ and $\frac{dAC_z}{dR}$, pooled equilibrium average costs and quantities are given by the solution to the system of equations

$$
\begin{aligned}
AC^p &= \frac{\alpha q_0^p}{\alpha q_0^p + (1-\alpha) q_1^p} AC_0^p + \frac{(1-\alpha) q_1^p}{\alpha q_0^p + (1-\alpha) q_1^p} AC_1^p \\
q_z^p &= q_z^e + \frac{dq_z}{dR}(AC^p - AC_z^e) \text{ for } z \in \{0,1\} \\
AC_z^p &= AC_z^e + \frac{dAC_z}{dR}(AC^p - AC_z^e) \text{ for } z \in \{0,1\}
\end{aligned}
$$

There are five equations and five unknowns, yielding an analytic solution for each value. Multiple equilibria are possible but, as we discuss below, not empirically relevant in the setting we consider here.

Computing welfare effects requires knowledge of the levels and slopes of marginal cost curves in addition to the demand and average cost curves. Here we exploit the observation that the equilibrium value of marginal cost $MC_z^e = \frac{dAC_z}{dq} q_z^e + AC_z^e$, and that with linear average cost curves $\frac{dMC_z}{dq} = 2\frac{dAC_z}{dq}$.

In Appendix Figure A1, we simulate equilibrium outcomes from pooling a low-cost and high-cost market under different assumptions of the slopes of the demand and cost curves in each market. The figure illustrates how the effects of pooling on aggregate borrowing and total welfare are ambiguous, even in this simple model, and how they relate to the slopes of demand and costs.

2.2 Mapping the model to data

The key empirical challenge when taking this model of pooling to the data is estimating the slopes of demand and cost curves in the high- and low-cost markets. In principle, one could estimate the slopes of these curves using any exogenous shock to price in each market. We use shocks to lenders' beliefs about borrowers' average costs due to information deletion. Under an average cost pricing policy, these shocks translate directly into prices. Let $AC^e_z(x,z)$ denote the value of AC^e_z for individuals with characteristics $X_i = x$ and $Z_i = z$, and $AC^e(x)$ be the value of AC^e_z for individuals with $X_i = x$. We define exposure to the deletion policy E_i for individuals i in markets defined by $X_i = x$ and $Z_i = z$ as

$$E(x,z) = \log AC^e(x) - \log AC^e_z(x,z).$$

We then map the high-cost and low-cost markets in the framework to the markets that face a reduction and an increase in predicted costs, i.e., the markets with negative and positive exposure, respectively. Although our model's predictions depend on differences in the level of costs, this stems from the assumption of linearity. In our empirical implementation we focus instead on the difference in log costs because it better reflects lenders' beliefs about changes in risk, and thus, exposure to the policy. For example, under our proposed measure of exposure, a 4 percentage point increase in the predicted default rate for borrowers whose baseline predicted default is 4% is comparable to a 10 percentage point increase in default for borrowers whose baseline is 10%.

The distribution of pooling effects across borrowers depend on the distribution of $E(x,z)$. Groups for whom deleting information on Z_i has little effect on cost predictions will be less affected by pooling, either positively or negatively. Our empirical strategy is to look across markets with different values of $E(x,z)$ and observe how borrowing and cost outcomes change following deletion. We use a machine learning approach to choose the conditioning sets of covariates X_i and Z_i. We describe the empirical setting and our approach to analyzing it in the next two sections.

2.3 Alternative modeling approaches

The framework described here is one of several plausible approaches to modeling the Chilean consumer credit market and the policy change. Most notably, we assume that lenders set prices rather than offering contracts consisting of price-quantity pairs (Rothschild and Stiglitz 1976), and that the form of contract does not change following policy implementation. This rules out separating equilibria where lenders screen borrowers based on their contract choice (e.g., see Bester (1985)). In a simple screening model equilibrium, good types–non-defaulters– would have less credit than in the full information setting, while bad types–defaulters– would not have more credit. Because there is no counteracting positive effect for bad types, deletion reduces welfare.

3 Empirical setting

3.1 Formal consumer credit and credit information in Chile

In Chile, formal consumer credit is supplied by banks and by other non-bank financial intermediaries, most notably department stores. As of December 2011 there were 23 banks operating in Chile, including one state owned and 11 foreign owned institutions, which had issued approximately $23 billion in non-housing consumer credit (i.e., credit cards, overdraft credit lines, and unsecured term loans).[7] As of the same month, the 9 largest non-banking lenders (all department stores) had a total consumer credit portfolio of approximately $5 billion. Although banks issue more credit, the number of department store borrowers is larger (14.7 million active non-bank credit cards, of which 5.4 million recorded a transaction during that month, versus 3.8 million consumer credit bank borrowers).[8]

Banks (and non-bank lenders) rely on defaults reported in the credit bureau to run credit checks of potential borrowers (Cowan and De Gregorio 2003, Liberman 2016). Defaults reported to the credit bureau include bank and non-bank debt, as well as other obligations such as bounced checks and utility bills. Importantly, banks are required by law to disclose their borrowers' outstanding balance and defaults to the banking regulator (SBIF), who then makes this information available only to banks. As a result, banks may learn a borrower's total bank debt and bank defaults, but may only observe reported defaults from non-banks (i.e., cannot access non-bank debt balances). In turn, non-banks can only learn an individuals' bank and non-bank defaults from the credit bureau, but not the level of bank or non-bank consumer credit.

3.2 The policy change

In early 2012, the Chilean Congress passed Law 20,575 to regulate credit information.[9] The bill included a one-time "clean slate" provision by which credit bureaus would stop

[7] All information in this paragraph is publicly available through the local banking regulator's website, www.sbif.cl.

[8] Chile's population is approximately 17 million.

[9] See http://www.leychile.cl/Navegar?idNorma=1037366.

sharing information on individuals' delinquencies that were reported as of December 2011. This provision affected only borrowers whose total defaults, including bank and non-bank debts, added up to at most 2.5 million pesos. According to press reports, the provision was a way to alleviate alleged negative consequences of the February 2010 earthquake, which had caused large damage to property and had ostensibly forced a number of individuals into financial distress. The Chilean Congress had already enacted a similar law that forced credit bureaus to stop reporting information on past defaults in 2002. Nevertheless, this new "clean-slate" was marketed as a one-time change, and indeed, all new defaults incurred after December 2011 were subsequently subject to the regular treatment and reported by credit bureaus.

Following the passage and implementation on February 2012 of Law 20,575, credit bureaus stopped sharing information on defaults for roughly 2.8 million individuals, approximately 21% of the 13 million Chileans older than 15 years old.[10] In effect, this means that individuals who were in default on any bank or non-bank credit as of December 2011 for a consolidated amount below 2.5 million pesos appeared as having no defaults after the passage of the law. This is shown in Figure 2, where we plot the time series of the number of individuals in our data with any positive default reported through credit bureaus as of the last day of each semester (ending in June or December).[11] The figure shows a large reduction in the number of individuals with any defaults as of June 2012, after the policy change, relative to December 2011.[12] Interestingly, the figure shows a sharp increase in the number of affected individuals in the following semesters until December 2015, the last semester in our data. This is consistent with the fact that the policy was a one-time change, as future defaults were recorded and reported by credit bureaus, as well as with the fact that many individuals whose defaults were no longer reported did default on new obligations.

The policy change modified the information that lenders, bank and non-bank, could obtain on defaults at other lenders. After the policy change, non-bank lenders could no longer verify any type of defaults, while banks could not observe whether individuals had defaulted on non-bank debt. However, banks could still verify whether an individual had bank defaults because the banking regulator's data was not subject to the policy change. Thus, the policy change induced a sharp information asymmetry between the banking industry as a whole and its borrowers, rather than creating asymmetries in the

[10] Figure taken from press reports of the "Primer Informe Trimestral de Deuda Personal", U. San Sebastian.

[11] Due to data constraints, our data is limited to individuals who were present in the regulatory banking dataset prior to the passage of Law 20,575.

[12] There is no evidence of an aggregate increase in defaults following the February 2010 earthquake.

information available to each bank with respect to its borrowers.

The median interest rate charged to small borrowers rose following deletion. Figure 3 plots median interest rates for small and large consumer loans before and after the deletion. We observe a 5.3 percentage point increase in rates in the small loan market, a 20% rise from a base of 26%. Rates continue to rise following the policy change, reaching almost 35% (30% above the base pre-policy rate) by the fourth quarter following implementation. We do not observe changes in rates for larger borrowing amounts, which suggests that the effects we see are not driven by coincident changes in other determinants of borrowing rates. We show below that on average most new borrowing is done by borrowers with no defaults. This means that the median new loan can be thought of as belonging to this market.

3.3 Data and summary statistics

We obtain from Sinacofi, a privately owned Chilean credit bureau, individual-level panel data at the monthly level on the debt holdings and repayment status for the universe of bank borrowers in Chile from April 2009 until 2014. Sinacofi has access to the banking data that is not available to other credit bureaus because Sinacofi's only clients are banks. Sinacofi merged the data to measures of consolidated defaults from the credit registry. We observe registry data at six month intervals, in June and December of each year. As is typical in empirical research on consumer credit, microdata do not include interest rates or other contract terms.

We use these data to build a panel dataset that links snapshots of defaults as reported to the credit bureau to borrowing outcomes. We use the six credit bureau snapshots from December 2009 through December 2011. We link each snapshot to bank borrowing and default outcomes over the six month period beginning two months after the snapshot (i.e., the six month interval beginning in February for the December snapshots, and the six-month interval beginning in August for the June snapshot). This alignment corresponds to the timing of the deletion policy, which took place in February 2012 based on the December 2011 credit bureau default records.

Table 1 reports summary statistics for this data. The first column is the full sample, which includes all individuals who show up in the borrowing data. There are 23 million person-time period observations from 5.6 million individuals in the dataset. 37% of borrowers in our dataset have a positive value of credit bureau defaults, with an average value in default of $554,500 CLP. 31% of the population, or 84% of all defaulters, have a default amount strictly between 0 and $2.5 million CLP, and are eligible for deletion. Figure 4 presents a histogram of the default amount as of December 2011 for all individuals and for individuals with positive defaults. We observe deletion for 29% of all individuals in the December 2011 cohort. The two percent gap between our calculated deletion eligibility rate and observed deletion rate is due to rare default types that are not included in the consolidated measure we observe. Conditional on eligiblity for deletion, the average consolidated amount in default is $172,250 CLP.

The average bank debt balance for consumers is $7.8 million CLP. Unsecured consumer lending accounts for 28% of all debt, for an average of $2.2 million CLP. Mortgage debt accounts for the majority of the remainder. The average bank default balance (defined as debt on which payments are at least 90 days overdue) across all borrowers is $338,090 CLP, or 12% of the overall debt balance. For borrowers eligible for deletion of defaults, this average is $147,460. Comparing bank default balances to credit bureau default balances shows that deletion eliminates banks' access to 15% $(= 100 \times (1 - 147/172))$ of the default amount among individuals whose balances in default falls below the deletion threshold.

We do not directly observe new borrowing or repayment. Thus, we define new consumer borrowing as any increase in an individual's consumer debt balance of at least 10% month over month, and the amount of new consumer borrowing as an indicator for new borrowing times the amount of the increase. In the full sample, 30% of consumers take out at least one new consumer loan in the six month period following each credit snapshot. The average amount of new borrowing is $184,000 CLP. We define new bank defaults analogously using borrowers' bank default balances. 17% of customers have a new bank default, with an average default amount of $37,000 CLP. In our analysis of the effects of information deletion, we focus on new consumer borrowing as the outcome of interest as defaults are most costly to lenders for uncollateralized borrowing.

The average age in our sample is 44, and 44% of borrowers are female. Our data identify borrowers' socioeconomic status for 9% of individuals overall. These data, which were collected by banks, divide individuals into five groups by socioeconomic background. We use these data to generate predictions of socioeconomic status for all individuals in the sample using a machine learning approach. We describe this process in Appendix B. In our empirical analysis we split our sample by this predicted SES categorization. One strong predictor of SES classification is whether or not an individual has a home mortgage. We split by this categorization as well.

The second column of Table 1 describes our main analysis sample. We focus on borrowers who have a positive debt balance six months prior to the credit snapshot and consolidated default of $2.5 million CLP or less, including zero values. This group accounts for 97% of individuals and 95% of observations. The restriction on debt balances allows us to define a consistent sample across time. Without it, the structure of our data generates spurious increases in mean borrowing over time. This occurs because individuals are included in our sample only if they borrow at some point between 2009 and 2014. An individual with a zero debt balance in 2009 must borrow in the future; otherwise, she would not be included in the data. Subsetting on individuals with positive debt balances at baseline addresses this issue.[13] The restriction to consolidated defaults of $2.5 million CLP or less lets us focus on the part of the credit market where available information changed. Lenders were able to observe consolidated defaults above $2.5 million CLP both before and after the cutoff. Demographics and borrowing in the panel sample are similar to the full dataset.

The third column of Table 1 describes the sample of individuals with positive borrowing. As we discuss in the next section, this is the sample we use for constructing cost predictions. They tend to be richer, and have much lower current default balances relative to overall borrowing (0.01 vs 0.09 in the full panel). Their rates of future default are also somewhat lower (0.05 vs. 0.08 in the full panel).

4 Machine learning cost predictions and registry deletion

4.1 Constructing cost predictions

The effect of deletion policies is to change the predictions lenders can make about costs for different kinds of borrowers. We take a machine learning approach that describes changes in cost predictions using a random forest (Mullainathan and Spiess 2017). The intuition underlying this approach is that banks make lending decisions by dividing potential borrowers into groups based on observable characteristics (Agarwal, Chomsisengphet, Mahoney and Stroebel 2018). We have access to borrowers' observable characteristics but do not observe banks' grouping choices. The random forest repeatedly chooses sets of possible predictor variables at random and constructs a regression tree using those predictors. Each tree iteratively splits by the explanatory variables, choosing splits to maximize in-sample predictive power. The random forest obtains predictions by averaging over predictions from each tree. One way to think about this process in our context is as averaging over different guesses about which variables banks might use to classify borrowers.

[13] An alternate approach would be to take the population of all Chileans, irrespective of borrowing, as the sample. We do not have access to data on non-borrowers.

We do not observe banks' true costs. Instead, we focus our discussion on a simple measure of costs: an indicator variable equal to one if a borrower adds to his default balance in the six month period following each registry snapshot. Although this measure is not comprehensive, it is likely to be correlated with banks' supply decisions and ex ante profits. For example, Dobbie, Liberman, Paravisini and Pathania (2018) show that banks focus more on default than other measures of costs due to agency concerns with loan officers. When predicting default outcomes we focus on the sample of individuals who have new borrowing over that same period. We make this restriction because the goal of the exercise is to recover cost predictions for market participants.

We build each tree in our random forest by choosing variables at random from a set of 15 possible predictors. These consist of two lags (relative to the time of policy implementation) of new quarterly consumer borrowing, new quarterly total borrowing, consumer borrowing balance, secured debt balance, average cost, and available credit line, as well as a gender indicator. For pre-policy predictions, the set of variables also includes the credit bureau default data. We set the number of trees in a forest to 150. Predictive power is not sensitive to other choices in this range. We choose other model parameters (how many variables to select for inclusion in each tree and the minimum number of observations in a terminal node in the tree) using a cross-validation procedure. For comparison, we also construct predictions using two alternate methods: a logistic LASSO and a naive Bayes classifier. See Appendix B for details on these approaches.

For each method, we construct two sets of predictions. The first set uses training data from the same registry cross-section as the outcome data. These predictions correspond to the best guess a lender can make about default outcomes using data available to them at the time of the loan. We use the contemporaneous predictions as our measure of lenders' best guess about average costs. For this set of predictions, differences between predicted values with and without the default information depend on differences in the average costs in each submarket in the separate market equilibrium, potentially time-varying shocks to credit demand that move individuals with different covariate values along their cost curves, and endogenous responses to pooling (in the post-pooling time period).

To estimate the slopes of the demand and cost curves, we need to isolate variation in costs due to supply-side price shocks. Our second set of predictions helps us do this. This set of predictions uses training data from the December 2009 credit bureau default cross section to generate predictions for all other cross sections. Conditional on covariates, these predictions do not vary across cohorts in the remaining data, and therefore do not reflect the effects of time-varying demand shocks. They use only data from before pooling took place, so they do not reflect endogenous reponses to information deletion. Our empirical analysis splits borrowers into positive-, negative-, and zero-exposure groups based on the second set of predictions, and tracks how contemporaneous cost predictions (and quantities borrowed) change in these groups following deletion. We construct both types of predictions using a training sample consisting of 10% of the observations in the relevant snapshot. We exclude the December 2009 data from our difference-in-differences analysis in all specifications, and exclude training data from our cost outcome analysis.

Table 2 compares in- and out-of-sample log likelihood measures for the random forest to those from other prediction methods. We present separate estimates for predictors trained in the pre-period (labeled AC^{pre}) and those trained contemporaneously (labeled AC^{post}). The contemporaneous random forest predictions have in-sample (out-of-sample) log likelihood values of -0.173 (-0.295) when including registry information. Without registry information, these values fall to -0.177 (-0.305). The pre-period random forest predictions have slightly higher log likelihoods in both the training and testing sample, with a similar percentage decline from dropping registry information. Random forest predictions outperform the naive Bayes and logistic LASSO predictions.

4.2 The distribution of exposure to changes in predicted costs

In addition to reducing explanatory power, deletion affects the distribution of cost predictions across defaulters and non-defaulters. We describe these changes in Figures 5 and 6. We focus on predictions trained in pre-period data, but results are very similar using the predictions based on contemporaneous data.

The upper panel of Figure 5 shows the means of predictions made without default information within bins defined by values of the predictions that include default information. We split the sample by prior default status. For individuals without prior defaults, deletion increases predicted costs on average (points are above the 45-degree line). For individuals with prior defaults, deletion reduces cost predictions (points are below the 45-degree line).

The lower panel of Figure 5 shows that predictions with and without deleted default information both track observed costs across the distribution of realized cost outcomes, on average. The cost predictions slightly underpredict costs at the bottom and middle of the cost distribution, and overpredict at the top. As shown in the lower-left panel of the graph, differences in observed outcomes between borrowers with and without defaults tend to be small conditional on the full-information prediction. There are almost no borrowers with defaults at the bottom of the full-information predicted cost distribution, and few borrowers without defaults at the very top. In the deleted information predictions (right panel), defaulters shift towards the bottom of the distribution and non-defaulters towards to the top. Conditional on the predicted costs, defaulters have higher costs going forward.

Figure 6 explores the distribution of changes in predicted values from deletion in more detail. For each individual, we define exposure E_i as the percentage change in cost prediction caused by deletion. The upper panel of Figure 6 plots the density of E_i by default status using predictions from the pre-period training set. For non-defaulters, predicted costs rise for 89% of borrowers, with an average increase of 29%. For defaulters, predicted costs fall for 95% of borrowers, with an average drop of 32%. The exposure distribution for defaulters is bimodal, with one mode at zero and the other centered near a decline of 75%. More borrowers are non-defaulters than defaulters, so predicted costs increase for a majority (63%) of borrowers in the market. The lower panel shows a similar distribution of exposure using the contemporaneous training set.

Table 3 describes how observable attributes of borrowers vary by exposure. We split borrowers into three groups: the 'low-cost market', defined as individuals for whom cost predictions rise by at least 15% following deletion, the 'high-cost market,' defined as individuals for whom cost predictions fall by at least 15%, and the 'zero group,' defined as individuals for whom cost predictions change by less than 15% in either direction. An important point from this table is that most borrowers are exposed to cost increases from deletion: 52% of observations fall into the low-cost category, compared to 32% in the zero-change group and 16% in the high-cost group. Almost all borrowers in the high-cost group have bank defaults, while almost no borrowers in the low-cost group do.

Though the individuals in the low-cost market are more likely to come from high-SES backgrounds and have mortgages, the borrowers whose cost predictions rise most following deletion are those who resemble high-cost borrowers along these dimensions. Figure 7 plots binned means of indicators for holding some mortgage debt at baseline (left panel) and coming from a high-SES background (right panel). Both graphs have upside-down V shapes. About 20% of borrowers in both the top and bottom deciles of the exposure distribution hold mortgage debt, compared to a maximum of about 30% for borrowers with modest positive exposure. Similarly, about 25% of borrowers in the top and bottom deciles of the exposure distribution come from high-SES backgrounds, compared to a maximum of over 60% for individuals with exposed to slight increases in cost predictions. Intuitively, the borrowers who benefit most from the policy are those who are difficult to distinguish from non-defaulters without access to the deleted information. In contrast, borrowers who are relatively unaffected by the policy are those for whom more accurate information about costs is available outside of the deleted registry.

5 Equilibrium effects

5.1 Empirical strategy

We isolate the effects of changes in lenders' cost beliefs on borrowing outcomes using a difference-in-differences approach. Intuitively, we compare changes in borrowing outcomes before and after deletion for individuals exposed to increases (and decreases) in cost beliefs to those for individuals with near-zero exposure. We construct cohorts of borrowers at six month intervals leading up to the policy change, including the month of the policy change itself. We then compare the effects of exposure to changes in cost expectations in the treated cohort to the effects of exposure in pre-treatment placebo cohorts.

Consider a sample of individuals who are either not exposed to changes in lender beliefs to deletion, or who are exposed to increases (decreases) in predicted costs. Within this sample, we estimate specifications of the form:

$$Y_{ic} = \gamma_c + \tau_c D_{ic} + X_{ic}\Psi_c + e_{ic}. \tag{2}$$

Y_{ic} is borrowing for individual i in cohort c, γ_c are cohort fixed effects, and X_{ic} are a set of individual covariates that include age, gender, and lagged borrowing and default outcomes. D_{ic} is an indicator equal to one if an individual is in the group exposed to increased (decreased) costs.

The coefficients of interest are the τ_c, which capture cohort-specific estimates of the effects of exposure to increases in lender cost predictions on borrowing. We normalize τ_c to be zero in the cohort immediateley prior to deletion. If deletion reduces borrowing for exposed individuals, we expect τ_c to be flat in the cohorts leading up to treatment, and then to become negative in the deletion cohort. We measure exposure using random forest predictions trained in the December 2009 pre-period. We define the zero-exposure group to be the set of individuals for whom $|E_{ic}| < 0.15$. Our findings are not affected by raising or lowering the cutoff for inclusion in the zero group. When computing exposure we winsorize values in the bottom 5% of the predicted cost distributions with and without registry data to avoid classifying very small differences in predicted cost levels as very large log differences. Our findings are not affected by raising or lowering the winsorization threshold.

This type of specification can recover the total effect of deletion on borrowing under two assumptions. The first is the standard difference-in-differences assumption that borrowing in the non-zero exposure groups follows parallel trends to the zero exposure group. We can evaluate this assumption by looking at pre-trends in the τ_c. The second assumption is that deletion of credit bureau defaults does not affect borrowing outcomes for individuals in the zero-exposure group. If the deletion raised (lowered) borrowing in the zero-exposure group, our estimates will understate (overstate) the gains in borrowing attributable to deletion. We revisit this assumption below using a supplementary difference-in-differences approach.

We also use the difference-in-differences specifications to estimate the effects of deletion on prices and realized costs. Under the assumption that lenders price based on their best predictions of average costs, we define $P_{ic} = AC(X_{ic}, Z_{ic})$ for pre-deletion values of c, and $P_{ic} = AC(X_{ic})$ for post-deletion values. We define average costs as $AC_{ic} = AC(X_{ic}, Z_{ic})$ across all values of c. Consistent with the idea that costs and price estimates should capture the best information available to lenders at the time of purchase, we use contemporaneously-trained random forest predictions to generate these outcome variables. Because the goal of these estimates is to capture how prices and costs change for loans following deletion, we estimate these specifications in the sample of individuals with positive borrowing.

Statistical inference is not straightforward in this setting. We would like to allow for correlation in error terms within the categories that lenders use to set prices, but we do not observe what these categories are. We use an auxiliary machine learning step to identify interactions of covariates within which individuals have similar cost values (i.e., each of these interactions identifies smaller "markets" where borrowers look similar to lenders). We then cluster standard errors in our regressions within groups defined by these interactions. There are 330 such groups in the full sample. Inference is robust to changes in the coarseness of these groupings.

5.2 The effects of deletion for defaulters relative to non-defaulters

We first report how borrowing, predicted cost, and observed cost outcomes change for individuals with deleted credit bureau default records relative to individuals without deleted records. Using the full sample of borrower data in each credit bureau snapshot, we estimate difference-in-differences specifications that interact cohort relative to default with an indicator variable for a positive default on the credit bureau snapshot. Panel A of Figure 8 and column 1 of Table 4 report estimates of this specification when the dependent variable is the log of predicted costs. This variable is equal to the (log) prediction using credit bureau defaults records in the pre-deletion period and the prediction that excludes these records in the post-deletion period. The log difference in cost predictions for defaulters relative to non-defaulters is steady in the year leading up to deletion, then falls by 0.66 after deletion, corresponding to a 52% decline in in lenders' cost expectations for defaulters relative to non-defaulters. This is consistent with evidence from Figures 5 and 6 that deletion reduces predicted costs for defaulters and raises them for non-defaulters.

The right panel of Figure 8 and column 2 of Table 4 report estimates when the dependent variable is new consumer borrowing. Borrowing is steady in the year leading up to deletion. In the six months following deletion borrowing for defaulters rises by just over \$41,000 CLP relative to borrowing for non-defaulters. This is 46% of the base-period borrowing of \$88,000 CLP for defaulters. Credit bureau deletion of defaults raised borrowing for the beneficiaries of deletion relative to non-beneficiaries. However, relative gains reflect a combination of gains for defaulters and losses for non-defaulters, so they do not imply that the policy raised borrowing overall

5.3 Effects of deletion by exposure to changes in predicted costs

Figure 9 and the right two panels of Table 4 report estimates of equation 2. These estimates recover effects for borrowers exposed to positive and negative shocks to cost predictions relative to the group where cost predictions do not change following deletion. Lenders' cost expectations for both groups are flat in the year leading up to deletion. In what follows, we refer to the group exposed to increases in predicted costs (positive exposure) as the low-cost market, and the group exposed to decreases in predicted costs (negative exposure) as the high-cost market. These labels allow us to map the framework developed above with the welfare calculations below. At the time of deletion, log cost predictions rise by 0.22 in the positive exposure group and fall by 0.29 in the negative exposure group. Pre-trends in borrowing are also flat for both groups in the year leading up to deletion. Following deletion, borrowing falls by \$14,000 CLP in the positive exposure group, equal to 6.4% of pre-period mean for that group. Borrowing rises by \$17,000 CLP for the negative exposure group, equal to 11.8% of the pre-deletion mean. The implied elasticity of borrowing with respect to changes in cost predictions is

-0.29 (-0.40) in the positive (negative) exposure group.

These estimates indicate that the net effect of deletion was to reduce borrowing. The group exposed to increases in predicted costs consists of 2.1 million individuals. At an average loss of $14,000 CLP per person, the total loss is just under $30 billion CLP, or $60 million USD at an exchange rate of 500 CLP per dollar. The group exposed to decreases in predicted costs consists of 608,000 individuals, with an average gain of $17,000 CLP per person and a total gain of $10 billion CLP or $20 million USD. The net effect of deletion across the two markets was thus to reduce borrowing by $20 billion CLP, or 3.5% of the total borrowing across the two groups. To the extent the goal of deletion policy was to increase access to credit, it appears to have been counterproductive.

The effects of deletion are largest for the low-SES borrowers who are most exposed to changes in predicted costs. Table 5 repeats the analysis from Table 4, subsetting by whether borrowers have a mortgage at baseline, and by our predicted measure of socioeconomic status. Individuals without mortgages and lower-SES individuals are more responsive to changes in lenders' cost expectations, and experience larger percentage changes in borrowing. For individuals without mortages, exposure to increased (decreased) costs lowers (raises) borrowing by 7.1% (12.3%) of baseline values. For individuals with mortgages, the percent decline (rise) in quantity borrowed is 2.8% (9.7%). For low-SES individuals, the percent decrease (increase) in quantity borrowed is 9.2% (12.4%) compared to 6.1% (7.7%) for high-SES individuals.

5.4 Net effects of deletion on welfare

5.4.1 Benchmark estimates

The deletion policy reduced overall borrowing, with declines for borrowers exposed to increases in predicted costs more than offseting gains for borrowers exposed to decreases in predicted costs. However, the policy may still have raised welfare if it transferred borrowing from individuals who value credit less relative to costs to individuals who value it more. To explore the welfare effects of pooling, we first assess the potential for market failure in each market using the slopes of the average cost curves. We then tie this analysis explicitly to welfare using the framework from section 2.

Table 6 reports the effects of deletion on realized average costs in the low-cost (columns 2-6) and high-cost markets (columns 7-11), in the full sample and split by mortgage and SES categories. At baseline, the average cost for borrowers in the low-cost market is 0.04, and the average cost in the high-cost market is 0.10. Deletion has little effect on average costs in either market. It slightly raises average costs for borrowers in the low-cost group and lowers average costs in the high-cost group. Because quantities fall in the low-cost group and rise in the high-cost group, the signs of these point estimates are consistent with downward-sloping average cost curves in both markets. However, in neither case can we reject an effect of zero at conventional levels of significance. These findings suggest that adverse selection is not large conditional on the information available to borrowers before deletion takes effect, and that welfare losses due to asymmetric information may be limited prior to deletion. Findings across subgroups are similar.

To assess the welfare effects of information deletion, we consider the following thought experiment: for a market at the average value of pooled average costs $AC(x)$, what is the welfare effect of moving from an equilibrium where lenders can condition prices on the credit bureau default flag z to one where they cannot? The mean value of $AC(x)$ is 0.050. Conditional on log $AC(x)$, costs are 43% lower for the low-cost group, exposed to increase in predicted costs, and 36% higher for the high cost group, exposed to decreases in predicted costs, for level values of $AC(x, z)$ of 0.029 and 0.069 in the low- and high-cost markets respectively. We estimate the slope of the demand curve in each market using results from Table 4 and the slope of the average cost curve in each market using results from Table 6.

Panels A and B of Figure 10 show the demand, average cost, and marginal cost curves in the low-cost and high-cost markets, respectively. Demand curves reflect *average* quantity borrowed by an individual in each market. The pre-deletion equilibrium in each market is determined by the intersection of the demand and average cost curves. Equilibrium (q, p) pairs are $(113, 0.069)$ and $(252, 0.029)$ in the high- and low-cost markets, respectively. The average quantity borrowed across both markets is 220 and the average price is 0.033. Average cost curves slope down in both markets, leading to underprovision relative to the efficient quantity.

Demand is less elastic in the high-cost market than the low-cost market. This means that for some common offer price p in both markets, the share of high cost types in market rises with p. In our linear parameterization, the share of high-cost types in the market is equal to one for $p > 0.14$.

Panel C of Figure 10 shows the pooled demand, average cost, and marginal cost curves. The demand curve is piecewise linear, with the slope becoming flatter when the low-cost types enter the market at lower prices. The pooled average cost curve is the quantity-weighted average of the average cost curves in the low- and high-cost markets. The marginal cost curve follows the high-cost curve at very high prices, then shifts rapidly downward as the low-cost types enter the market. Equilibrium price and average quantity in the pooled market are given by the intersection of the pooled demand curve and the pooled average cost curve, with $(q, p) = (215, 0.035)$. Quantity borrowed declines on average, and prices rise.

The welfare effect of pooling differs in the high- and low-cost markets. In the low-cost market, pooling exacerbates welfare losses from underprovision. In the high-cost market, the pooled price is below the intersection of the demand and marginal cost curves, so welfare losses in the pooling equilibrium come from overprovision.

Table 7 summarizes the quantitative implications of this analysis. In the low-cost market, the equilibrium price rises from 0.029 before deletion to 0.035 afterward, while average costs do not meaningfully change. Quantity borrowed declines by an average of $13,000 CLP per person, or a total of $26.4 billion CLP. The welfare loss relative to the efficient quantity rises by 106% of the baseline value. In contrast, prices in the high-cost market drop from 0.069 to 0.035, and borrowing rises by $28,000 CLP per person, or $17 billion CLP in aggregate. Welfare losses in this market decline by 73%. Aggregating across markets, borrowing falls by $9 billion CLP, and welfare losses rise by an amount equal to 66% of the welfare loss relative to the optimum at baseline.

5.4.2 Longer run default outcomes

Changes in lender beliefs about six-month-ahead defaults have substantial predictive power for the effects of deletion on borrowing. But measuring cost curves using six-month-ahead default rates will tend to understate lender average costs if many defaults take place over a longer timeline. In the context of our benchmark model for welfare analysis that equates average costs and prices in equilibrium, underestimates of average costs imply underestimates of prices and borrower valuations.

Appendix Table A1 repeats the analysis from Table 4 but uses one-year-ahead bank default measures rather than six-month-ahead bank default measures to proxy for costs. As part of this analysis, we rerun our cost prediction procedure to obtain one-year-ahead predictors. As expected, one-year-out cost measures are higher than six-month-out cost measures in both markets. Proportionally, the increase is bigger in the low cost market (0.08 vs. 0.04) than the high-cost market (0.14 vs. 0.10), suggesting that the short-term focus of our benchmark analysis understates valuations of low-cost borrowers (whose borrowing is reduced by deletion) compared to high-cost borrowers. Estimated effects of deletion on borrowing levels are close to unchanged relative to the benchmark analysis. As reported in Appendix Table A2, the higher costs lead to larger estimates of welfare losses in the low-cost market. We also observe welfare losses from pooling in the high-cost market in this setting due to overborrowing. Percentage increases in welfare loss in aggregate are larger in levels but smaller in percentage terms (42%) due to larger estimates of welfare losses at baseline.

We prefer our benchmark estimates because using one-year-ahead default measures means that some defaults attributed to loans originated in the pre-deletion period occur following deletion, which does not occur with the six-month-ahead measure. What our findings here illustrate is that underestimates of cost levels due to the short measurement timeline do not affect findings about borrowing quantities and suggest larger welfare losses in aggregate.

5.4.3 Markups over average cost

Our analysis of welfare effects thus far assumes that lenders do not mark up rates over costs. If borrowers face imperfect competition and are able to mark up prices relative to our cost measures, our analysis of welfare effects will systematically underestimate how much consumers value borrowing. Further, if borrowers in the high- and low-cost markets face *different* markups at baseline, we will mismeasure their relative valuations.

We explore how different assumptions about markups in the high- and low-cost markets affect our welfare analysis by augmenting the model from Section 2 with markups relative to average cost. We consider the effects of raising markups overall, and of raising markups in the pre-deletion high-cost market relative to the low-cost market.

Recall that in benchmark case, the pre-deletion equilibrium quantity q_z^e and price $R_z^e = R_z(q_z^e)$ in market z were determined by $R_z(q_z^e) = AC_z(R_z(q_z^e))$. We now add a market-specific markup term m_z for prices relative to average costs, so equilibrium is determined by $R_z(q_z^e) = (1 + m_z) \times AC_z(R_z(q_z^e))$ for $z \in \{0, 1\}$. In the pooled market we allow a markup of value m_p, so that equilibrium price $R^e = R(q_e)$ and quantity q^e are determined by $R(q^e) = (1 + m_p) \times AC(R(q^e))$.

Within this framework we conduct the following exercise. We fix the low-cost market markup m_0 at a value μ_0, and set the high-cost market markup m_1 to $m_1 = \mu_0 \times (1 + \mu_1)$. We cycle through combinations of μ_0 and μ_1, in each case setting m_p to the quantity-weighted average markup in the pre-deletion period so that deletion does not affect the average markup in the market.

Table 8 shows percentages changes in welfare loss relative to baseline value in both markets combined for different combinations of μ_0 and μ_1. Each cell of the table has four numbers: the welfare loss from pooling for the average individual in the low cost market, the welfare loss for the average individual in the high-cost market, the average welfare loss across both markets, and the percentage change in welfare loss.

Welfare losses persist as we raise markups in both markets equally (reading down the leftmost column). As markups rise, both losses in the low-cost market and gains in the high-cost market rise in absolute value. This makes sense: higher markups mean that the consumers in both markets place a higher value on borrowing, leading to higher welfare stakes. Net losses rise in levels but fall in percentage terms due to a larger denominator.

Augmenting the markup in the high-cost market relative to the low-cost baseline tends to reduce the welfare losses from pooling (moving left to right within a row). Again, this makes sense. Higher markups for high-cost borrowers mean that those individuals value borrowing more. At baseline markup levels up to 25%, welfare losses persist for additional high-cost markups of up to 100%. The welfare effects of pooling become zero or modestly positive in percentage term when markups are very high overall, *and* there are large additional markups in the high-cost market. For example, we find that pooling breaks even in welfare terms when the low-cost markup is 50% and the additional high-cost markup is 100%, and reduces welfare losses relative to the efficient outcome by 11% when the low-cost market markup is 200% and the high-cost market markup is an additional 100%. For the pooling policy to break even in welfare terms requires both a) large markups overall, and b) large additional markups in the high cost market relative to the low cost market.

The assumption underlying this analysis– that pooling does not affect the average markup– seems reasonable given price and cost data we observe, in which prices and costs increase proportionally following deletion. Figure 3 shows that after the deletion the median consumer credit rate increases by 5.3 percentage points from a base of 26%, a 20% increase. The increase in the median rate is similar to the estimated 22% increase in predicted costs for the low-cost market (the median borrower is not in default, i.e. low cost), shown in Table 4, column 3.

Additional simulations (available upon request) suggest that if lenders are able to increase markups on average following the deletion, then the policy would would induce a larger reduction in welfare than those reported here. Incumbent lenders may be able to increase markups following the deletion if, for example, borrowers reveal part of their unobserved propensity to default through actions that are not shared via credit bureaus but are observed by the incumbent lender. Then, following deletion, the act of switching banks may send a negative signal about the likelihood of repayment, allowing incumbent lenders to increase markups. We note that we cannot test this hypothesis directly as we do not observe micro-level data for non-bank lenders.

5.5 Effects for the zero-exposure group

5.5.1 Comparison to no-deletion group

We test the assumption of no effect on the zero-exposure group using two strategies. First, we exploit the 2.5 million pesos policy cutoff in a difference-in-differences test. We test for differential changes in new consumer borrowing for individuals whose defaults added up to less than 2.5 million pesos, who were exposed to the policy change, relative to individuals whose defaults added up to more (or equal) than 2.5 million pesos, who were not exposed to the policy change. To control non-parametrically for differences in new borrowing along the distribution of amount in default, we restrict our analysis to a bandwidth of 250 thousand pesos around the policy cutoff.[14] We compute this change in new borrowing for the three cohorts prior to the policy change (June 2010, December 2010, and June 2011) and the cohort exposed to the policy change (December 2011).

For each cohort we divide the sample in two groups defined by our machine learning cost predictions: high-cost (negative-exposure) individuals, for whom predicted costs drop by more than 15%, and the zero-exposure group. There are no individuals exposed to an increase in predicted costs in this sample of individuals with high amounts in default.[15] We run the following specification differentially for the two groups:

$$Y_{ic} = \gamma_c + \tau_c \times 1[Default_{ic} < 2,500,000] + e_{ic}, \tag{3}$$

where, again, Y_{ic} is borrowing for individual i in cohort c. The γ_c are cohort fixed effects. $1[Default_{ic} < 2,500,000]$ is an indicator equal to one if total defaults for individual i in cohort c add up to less than 2.5 million pesos. The τ_c are the effects of interest, capturing how borrowing changes after registry deletion in 2011 for individuals whose amount in default is less than the policy cutoff of 2.5 million pesos.

This test recovers the causal effect of the policy change for the zero-exposure and negative-exposure groups under the assumption of no differential trends for individuals above and below the cutoff, which we examine visually with pre-trends. If our assumption that deletion does not affect borrowing for the zero-exposure group is correct, we should see no change in outcomes for this group following deletion. An in-

[14] Our findings are robust to widening or narrowing this bandwidth, although standard errors grow due to small sample sizes at very narrow bandwidths. We obtain near-identical findings in RD-DD specifications that allow for separate linear trends in default amount above and below the cutoff value in each cohort relative to policy change. These results are available upon request.

[15] To compute predicted costs for the above-threshold group under the information deletion policy we apply the predicted values from the machine learning exercise described in section 4 based on observable covariates X_{ic}.

crease in borrowing for the negative-exposure group would help make the zero-group test more compelling by showing that the deletion policy and our measures of exposure to that policy are good predictors of outcomes not just overall but within the subgroup of relatively large defaulters.

We present the findings in Figure 11. The coefficients of interest of equation (3) for the zero-group are indistinguishable from zero before the policy change, indicating no pre-trends, and indistinguishable from zero after the policy change, which is consistent with the identification assumption for our main analysis. The graph also shows a large increase in borrowing for high-cost individuals, exposed to decreases in predicted costs, whose defaults are less than the cutoff after the policy change. This rules out that the absence of an effect for the zero-group after the policy change is driven by a lack of power to identify any effects of the policy change among high-default individuals and is consistent with the main findings in this paper.

5.5.2 Cross-time comparison

Second, we implement difference-in-differences specification that exploits variation within borrower cohort by time relative to deletion. Let t index six-month periods relative to the period beginning in February of calendar year c. Within the zero exposure groups, we estimate equations of the form:

$$Y_{ict} = \gamma_c + \theta_t + \tau_t \times 1[c = c_T] + e_{ict}, \tag{4}$$

where Y_{ict} is borrowing for individual i in cohort c at time relative to deletion t. The γ_c and θ_t are cohort and event-time fixed effects, respectively. $1[c = C_T]$ is an indicator equal to one if c is the treated cohort c_0. Here, the τ_t are the effects of interest, capturing how borrowing changes after registry deletion in 2011 relative to changes at the same time of year in previous years.

This specification will capture unbiased estimates of the effect of deletion of credit bureau defaults on borrowing for the zero-exposure group if time-of-year effects are the same in the 2011 and earlier borrowing cohorts. It differs from the main approach in section 5.1 in the requirements for unbiased estimation. In particular, our main approach differences out time-varying shocks that affect all borrowers by measuring outcomes relative to the zero-exposure group. This supplementary specification requires the strong asssumption that seasonal effects be constant across years.

We present our findings in Figure 12. We follow borrowing outcomes for a year before and after deletion, divided into six month windows. Borrowing grows more rapidly in the pre-deletion period for the 2011 cohort than it did in earlier cohorts, suggesting that seasonal effects may differ from year-to-year. Following deletion, the trend reverses, and borrowing falls for the cohort treated with information deletion relative to the control cohort. That is, following deletion borrowing falls relative to the pre-deletion baseline and even more relative to the pre-deletion trend for the zero-exposure group. Though the presence of pre-deletion trends argues for caution in interpretation, these findings are hard to reconcile with a claim that information deletion *raised* borrowing in the zero exposure group. It follows that our main estimates of the effects of deletion *underestimate* the decline in borrowing from the deletion policy, if anything.

5.6 Additional evidence: borrowing from non-banks

The effects of deletion on aggregate borrowing could be reduced if individuals subject to higher prices for bank credit shift towards non-bank borrowing. The largest non-bank lenders in Chile are department stores that issue credit cards. We explore how borrowing changed at these institutions using aggregate data on retail credit card lending. Appendix Figures A2, A3, and A4 show no distinct breaks in the total stock of retail credit cards, the number of retail credit cards used, or the amount transacted at the time of deletion.

These findings are consistent with the hypothesis that deletion reduced aggregate borrowing. Deletion effects in the retailer-issued credit card market may be smaller than in the consumer bank lending market because low-risk individuals are very unlikely to borrow in that market both before and after deletion. Median interest rates for retailer credit card lending are 75% higher than for consumer bank lending just before deletion (45% vs. 26% in November 2011) and remain higher following deletion (e.g. 45% vs. 31% in February 2012). That few individuals subsitute from consumer credit to credit card borrowing is consistent with the observation that prices remained lower in the consumer credit market following the deletion.

In fact, the deletion may have induced larger information asymmetries among non-banks than banks. While banks continued to observe bank defaults (at all other banks) following deletion, the deleted credit bureau information was the only default information available to non-bank lenders. Because there is no micro-level data for non-bank lenders, we cannot directly calculate how exposure to the policy affects non-bank lending, but our results for bank lending suggest there may be welfare losses here as well. In the next section we use our empirical strategy to evaluate the effects on bank lending of a counterfactual policy change that would delete all bank defaults, which is similar to the informational change for non-banks after the policy change.

6 Evaluation of counterfactual deletion policies

The methodology used above to study the effects of the large-scale deletion of credit bureau defaults provides a framework through which policymakers can predict the distributional and aggregate effects of changes in any type of credit information. In this section we apply this methodology to two hypothetical changes in the credit information available to lenders. The first is a deletion of information about gender. The idea of eliminating the use of demographic information has parallels in US anti-discrimation laws as applied to credit markets (Munnell et al. 1996, Blanchflower et al. 2003, Pope and Sydnor 2011). The second is deletion of banks' internal and external default records across all banks in addition to the credit bureau defaults. This is a more radical version of the original policy.

In each case, we can simulate the effects of counterfactual policies using the following procedure. First, we compute each individual's (log) exposure to the policy by estimating predicted costs with and without the deleted information. We then take our estimates of exposure to cost changes, and scale them by an estimated elasticity of borrowing with respect to costs. For example, we can use the elasticity estimates from Table 4.

We present the analysis in Table 9, which mimics Table 3 for our baseline analysis. For each of the two counterfactual policies, we split the sample into individuals whose costs increase by 15% or more, individuals whose costs decrease by 15% or more, and the zero change group, which groups everyone else. This follows the procedure from our analysis of the observed deletion policy.

The top panel presents the first counterfactual policy, deletion of the gender indicator. Three things emerge from the analysis. First, most individuals (87% of the sample) belong to the zero change group. This is because the distribution of changes in costs is much tighter than in our baseline analysis, as is evident in the histogram of exposures shown in Figure 13. Second, as expected, gender is a strong predictor of cost changes: 98% of individuals exposed to cost increases are female, while females only represent 16% of those exposed to cost decreases. Thus, women would experience average increases in predicted costs following a deletion of the gender flag. Third, individuals whose costs increase or decrease have no registry defaults, and little variation in socio-economic status. These variables have little explanatory power for changes in banks' expected costs following deletion of the gender flag, which is consistent with the fact that costs do not change much when gender is deleted.

The bottom panel shows the second counterfactual policy, deletion of banks' internal default records in addition to consolidate default. Unsurprisingly, the more radical deletion option leads to larger changes in predicted costs than the actual deletion policy, as only 13% of the distribution is concentrated in the zero change group. This point is also shown in Figure 13. This suggests that the measure of defaults is highly predictive of future bank costs. Second, gender is uncorrelated with changes in costs following deletion of bank defaults, while bank defaults are, unsurprisingly, highly correlated with changes in predicted costs. Finally, socio-economic status is also correlated with changes in predicted costs: individuals exposed to reductions in costs are about 20 percent more likely to belong to a low socio-economic status group than those exposed to increases.

If one is willing to assume that elasticities of borrowing with respect to changes in average costs are the same as what we observe in the analysis of the observed deletion policy, we can go beyond the analysis of changes in the predicted cost distribution and predict the effects of these counterfactual deletion policies on borrowing. For example, if we take an estimated elasticity of -0.29 from Table 3 and multiply by the mean measures of exposure to the gender deletion in each group, we get that groups exposed to increases in costs see a 7 percent decline in new borrowing, a decline of $4,400 CLP per borrower, while groups exposed to decreases in costs see a 7.3 percent increase in new borrowing, an increase of $5,600 CLP per borrower. Multiplying each effect by the number of individuals in each group implies a near-zero change in aggregate new borrowing. The counterfactual deletion of banks' default records leads to a 18% drop in lending for individuals exposed to increases in costs and a 25% increase in lending for individuals exposed to decreases in costs. These effects aggregate to a drop in lending of $42 billion CLP over a six month period, roughly twice the size of the $20 billion CLP net effect of the observed deletion policy.

7 Conclusion

This paper explores the equilibrium effects of information asymmetries on credit markets in the context of a large-scale policy change that forced credit bureaus to stop reporting past defaults for the majority of defaulters in the Chilean consumer credit market. We develop a simple economic framework that illustrates how the demand and cost curves in separate high- and low-cost markets determine the effects of pooling those markets on borrowing and welfare. The effects of deletion on both of these outcomes are ambiguous in sign.

After using a machine learning approach to construct cost predictions and document the distribution of exposure to increases and decreases in predicted costs from deletion, we estimate the demand and cost curves using a difference-in-differences design. Our core empirical finding is that losses from information deletion are regressive and outweigh gains in this setting: consumer borrowing falls by 3.5% after the policy change, with the largest losses for lower-income individuals with smaller borrowing balances. There is no evidence that the winners from the policy value borrowing sufficiently more than the losers to offset these losses.

Our findings suggest that although policies that limit information availability in credit markets can raise welfare, they should be deployed cautiously. Even if deletion lead to increased borrowing for defaulters, it may reduce lending over all. A feature of deletion policies is that the biggest losers tend to resemble the biggest winners on all characteristics observable to the lender other than the deleted information, so policies implemented with the goal of helping disadvantaged populations also have greatest risk of negative effects for these populations.

Our findings motivate a simple procedure by which policymakers can predict the distributional consequences of a proposed change in credit information. The procedure is to construct cost predictions before and after the change, and identify the individuals with the biggest gains and losses in predicted costs. These estimates can be used alone to classify likely winners and losers, can be paired with estimates of demand elasticities to predict changes in quantity borrowed, or can be combined with estimates of demand and cost elasticities to predict changes in welfare. This approach can also be applied to understanding how existing information-restricting institutions such as sunset provisions affect lending. We leave this exercise for future research.

References

Adams, William, Liran Einav, and Jonathan Levin, "Liquidity Constraints and Imperfect Information in Subprime Lending," *American Economic Review*, 2009, *99* (1), 49–84.

Agan, Amanda and Sonja Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics*, 2017, *133* (1), 191–235.

Agarwal, Sumit, Souphala Chomsisengphet, Neale Mahoney, and Johannes Stroebel, "Do Banks Pass Through Credit Expansions to Consumers who Want to Borrow?," *Quarterly Journal of Economics*, 2018, *133* (1).

Akerlof, George A, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 1970, *84* (3), 488–500.

Athey, Susan and Guido Imbens, "Recursive Partitioning for Heterogeneous Causal Effects," *Proceedings of the National Academy of Sciences*, 2016, *113* (27), 7353–7360.

____ **and Stefan Wagner**, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Working Paper*, 2017.

Ausubel, Lawrence M., "The Failure of Competition in the Credit Card Market," *The American Economic Review*, 1991, *81* (1), 50–81.

Bartik, Alexander W. and Scott Nelson, "Credit Reports as Resumes: The Incidence of Pre-Employment Credit Screening," *Working Paper*, 2016.

Bester, Helmut, "Screening vs. Rationing in Credit Markets with Imperfect Information," *American Economic Review*, 1985, *75* (4), 850–55.

Blanchflower, David G, Phillip B Levine, and David J Zimmerman, "Discrimination in the small-business credit market," *The Review of Economics and Statistics*, 2003, *85* (4), 930–943.

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," *Review of Financial Studies*, 2018, *31* (6), 2005–2037.

Breiman, Leo, "Random Forests," *Machine Learning*, 2001, *45*, 5–32.

——, **Jerom Friedman, Charles J. Stone, and R.A. Olshen**, *Classification and Regression Trees*, Chapman and Hall/CRC, 1984.

Brown, M. and C. Zehnder, "Credit Reporting, Relationship Banking, and Loan Repayment," *Journal of Money, Credit and Banking*, 2007, *39* (8), 1883–1918.

Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram, "Machine Learning From Schools About Energy Efficiency," *NBER Working Paper*, 2017, (w23908).

Chetty, Raj, "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods," *Annual Review of Economics*, 2009, *1*, 451–488.

Clifford, Robert and Daniel Shoag, ""No More Credit Score" Employer Credit Check Banks and Signal Substitution," *Working Paper*, 2016.

Cortes, Kristle, Andrew Glover, and Murat Tasci, "The Unintended Consequences of Employer Credit Check Bans on Labor and Credit Markets," *Working Paper*, 2016.

Cowan, Kevin and Jose De Gregorio, "Credit Information and Market Performance: The Case of Chile," in Margaret J. Miller, ed., *Credit Reporting Systems and the International Economy*, Vol. 4, Cambridge, MA: MIT Press, 2003, pp. 163–201.

Dobbie, Will, Andres Liberman, Daniel Paravisini, and Vikram Pathania, "Measuring Bias in Consumer Lending," Working Paper 24953, National Bureau of Economic Research August 2018.

——, **Paul Goldsmith-Pinkham, Neale Mahoney, and Jae Song**, "Bad Credit, No Problem? Credit and Labor Market Consequences of Bad Credit Reports," Technical Report, National Bureau of Economic Research 2016.

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.

—— **and Jonathan Levin**, "Economics in the age of big data," *Science*, 2014, *346* (6210), 1243089.

——, **Mark Jenkins, and Jonathan Levin**, "Contract Pricing in Consumer Credit Markets," *Econometrica*, 2012, *80* (4), 1387–1432.

——, ——, and ——, "The impact of credit scoring on consumer lending," *The RAND Journal of Economics*, 2013, *44* (2), 249–274.

Elul, Ronel and Piero Gottardi, "Bankruptcy: Is It Enough to Forgive or Must We Also Forget?," *American Economic Journal: Microeconomics*, November 2015, *7* (4), 294–338.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," 2017.

González-Uribe, Juanita and Daniel Osorio, "Information Sharing and Credit Outcomes: Evidence from a Natural Experiment," 2014.

Hackmann, Martin B, Jonathan T Kolstad, and Amanda E Kowalski, "Adverse Selection and an Individual Mandate: When Theory Meets Practice," *American Economic Review*, 2015, *105* (3), 1030–66.

Handel, Ben, Igal Hendel, and Michael D Whinston, "Equilibria in health exchanges: Adverse selection versus reclassification risk," *Econometrica*, 2015, *83* (4), 1261–1313.

Herkenhoff, Kyle, Gordon Phillips, and Ethan Cohen-Cole, "The impact of consumer credit access on employment, earnings and entrepreneurship," Technical Report, National Bureau of Economic Research 2016.

Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jan Wang, "Credit Scoring with a Data Mining Approach Based on Support Vector Machines," *Expert Systems with Applications*, 2007, *33* (4), 847–856.

Jaffee, Dwight M and Thomas Russell, "Imperfect Information, Uncertainty, and Credit Rationing," *The Quarterly Journal of Economics*, 1976, pp. 651–666.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo, "Consumer Credit-Risk Models via Machine-Learning Algorithms," *Journal of Banking & Finance*, 2010, *34* (4), 2767–2787.

Liberman, Andres, "The Value of a Good Credit Reputation: Evidence from Credit Card Renegotiations," *Journal of Financial Economics*, 2016, *120* (3), 644–660.

Miller, Margaret J, *Credit reporting systems and the international economy*, Mit Press, 2003.

Mullainathan, Sendhil and Jann Spiess, "Machine Learning: An Applied EconometricAapproach," *Journal of Economic Perspectives*, 2017, *31* (2), 87–106.

Munnell, Alicia H, Geoffrey MB Tootell, Lynn E Browne, and James McEneaney, "Mortgage lending in Boston: Interpreting HMDA data," *The American Economic Review*, 1996, pp. 25–53.

Musto, David K, "What Happens when Information Leaves a Market? Evidence from Postbankruptcy Consumers," *The Journal of Business*, 2004, *77* (4), 725–748.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 2011, *12*, 2825–2830.

Petersen, Mitchell A and Raghuram G Rajan, "Does Distance Still Matter? The Information Revolution in Small Business Lending," *The Journal of Finance*, 2002, *57* (6), 2533–2570.

Pope, Devin G and Justin R Sydnor, "What's in a Picture? Evidence of Discrimination from Prosper. com," *Journal of Human Resources*, 2011, *46* (1), 53–92.

Rothschild, Michael and Joseph E Stiglitz, "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *The Quarterly Journal of Economics*, 1976, *90* (4), 630–49.

Steinberg, Joseph, "Your privacy is now at risk from search engines– even if the law says otherwise," *Forbes*, June 2014.

Stiglitz, J.E. and A. Weiss, "Credit Rationing in Markets with Imperfect Information," *The American Economic Review*, 1981, *71* (3), 393–410.

Varian, Hal, "Causal Inference in Economics and Marketing," *Proceedings of the Natural Academy of Sciences*, 2016, *113* (27), 7310–7315.

Figures and Tables

Figure 1: Equilibria for high- and low-cost markets and under pooling

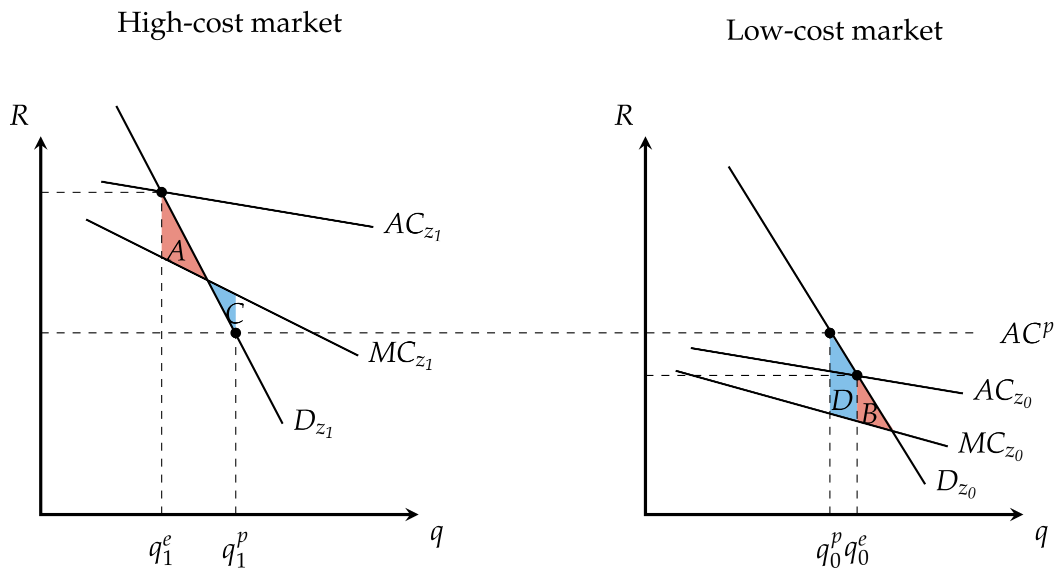
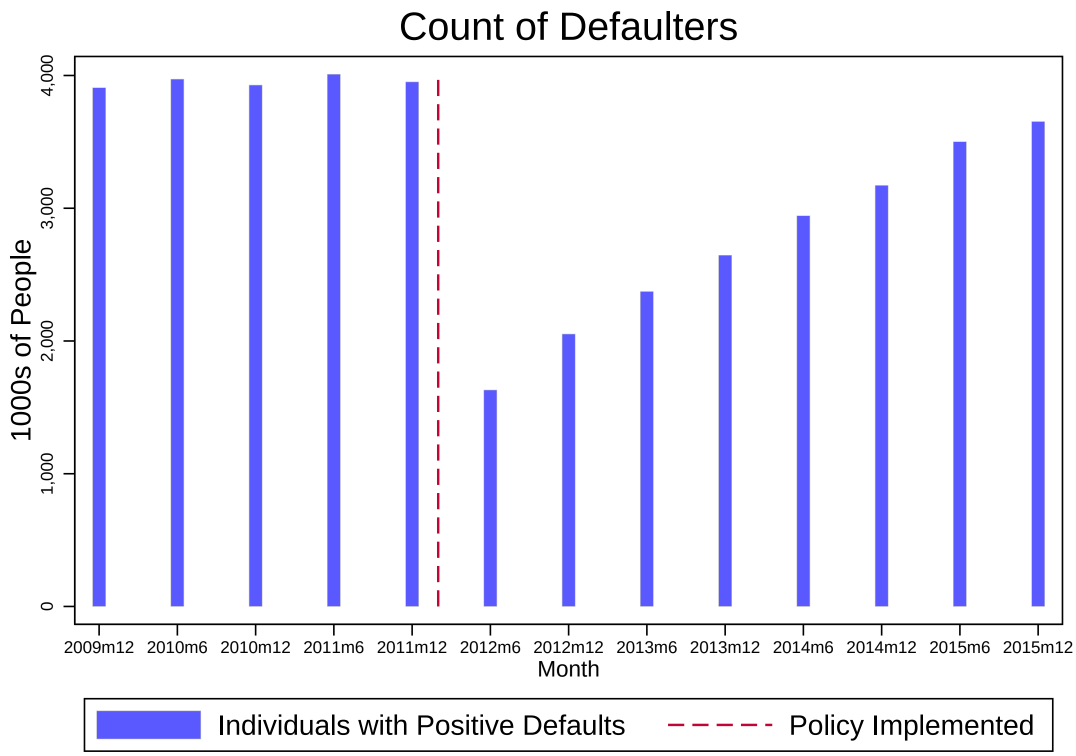


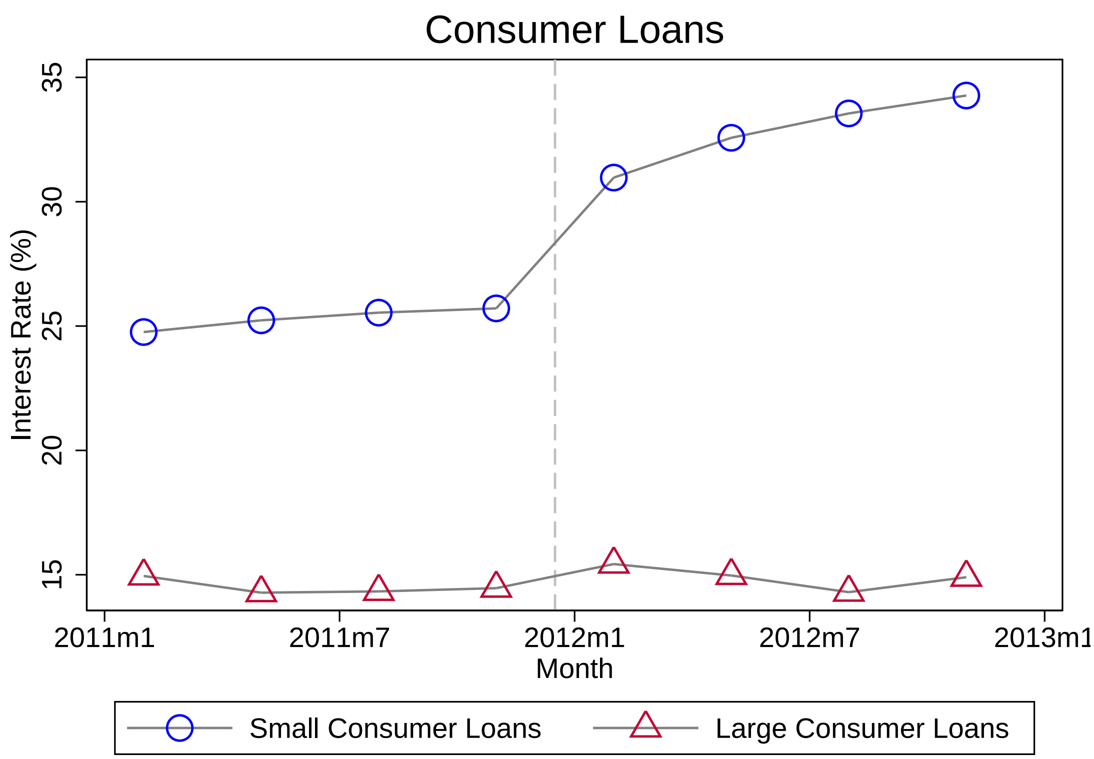
Diagram illustrating the economic framework. Left panel describes the high-cost market; right panel describes the low-cost market.

Figure 2: Individuals with positive past defaults over time



Each bar represents the count of individuals in the credit registry with positive default values at six month intervals. The vertical line represents the implementation of the registry deletion policy.

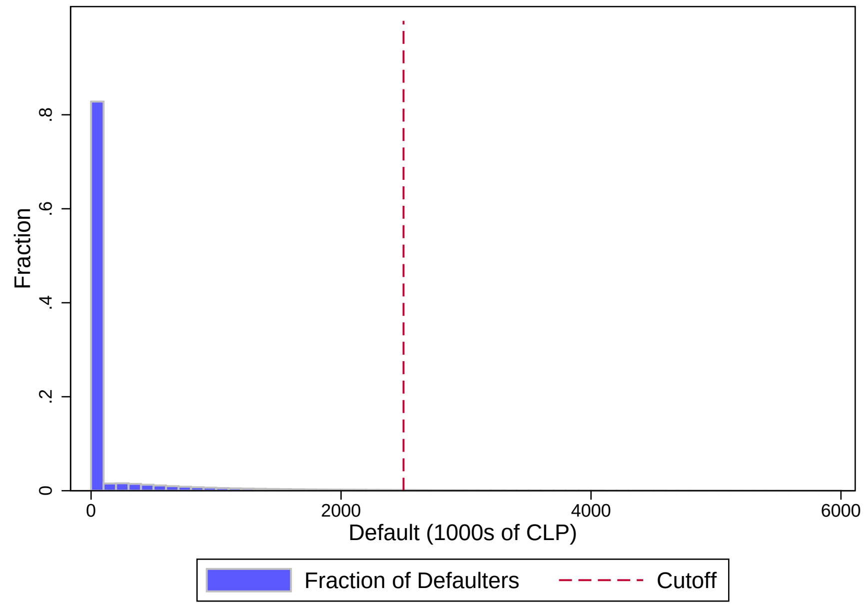
Figure 3: Interest rates



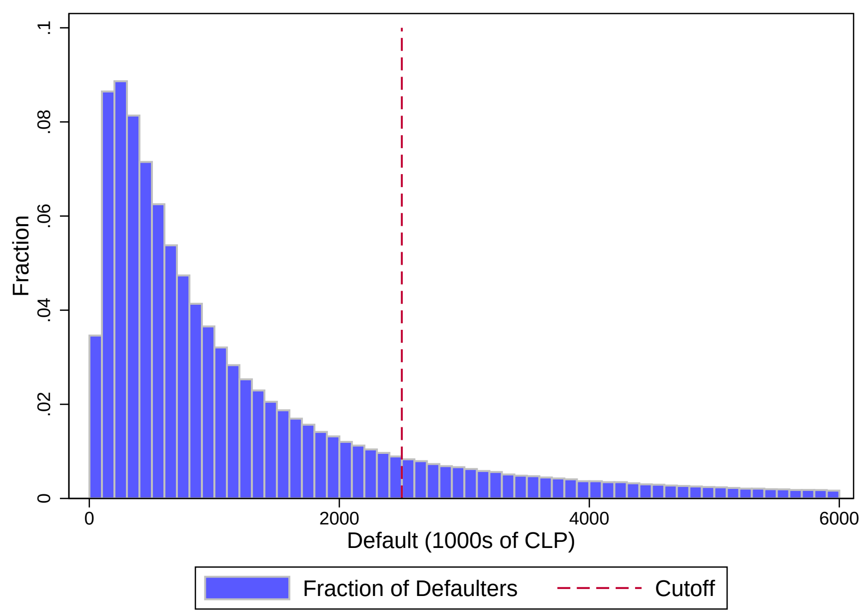
End of period median interest rates for small (top) and large (bottom) consumer loans issued by banks, by quarter relative to December 2011-February 2012. Information on rates obtained from website of Superintendencia de Bancos e Instituciones Financieras, www.sbif.cl.

Figure 4: Histogram of amount in default as of December 2011

Panel A: all individuals

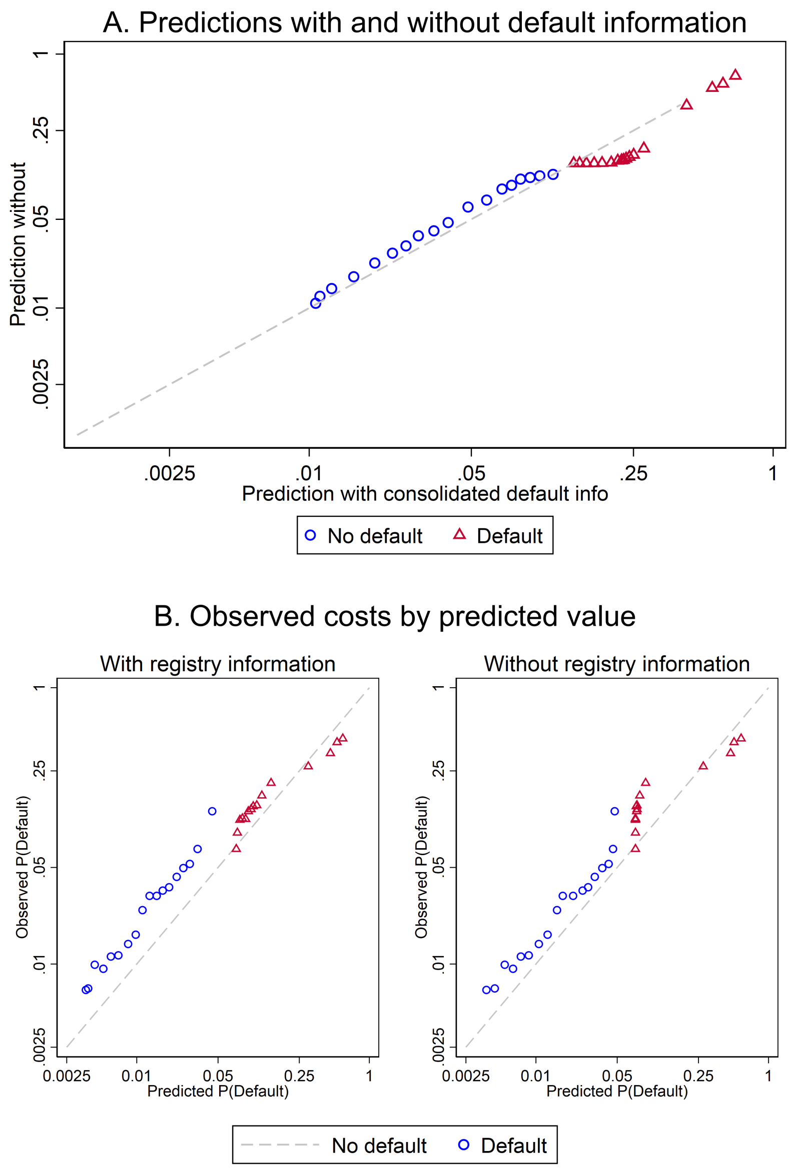


Panel B: conditional on positive default



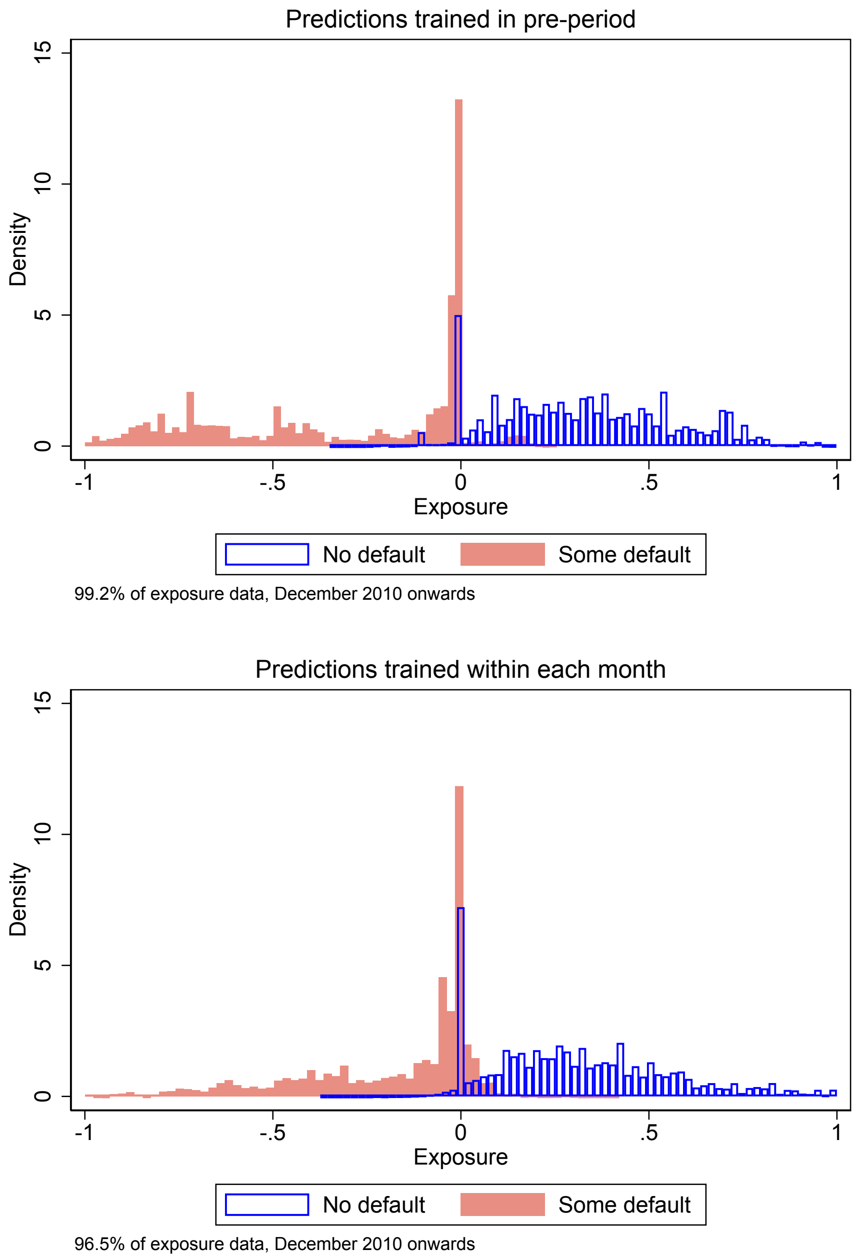
Panel A: Histogram of consolidated defaults as of December 2011, for amounts below \$6 million CLP (approximately \$3,000). Panel B: Histogram of consolidated defaults for individuals with positive defaults only.

Figure 5: Predictions with and without registry data



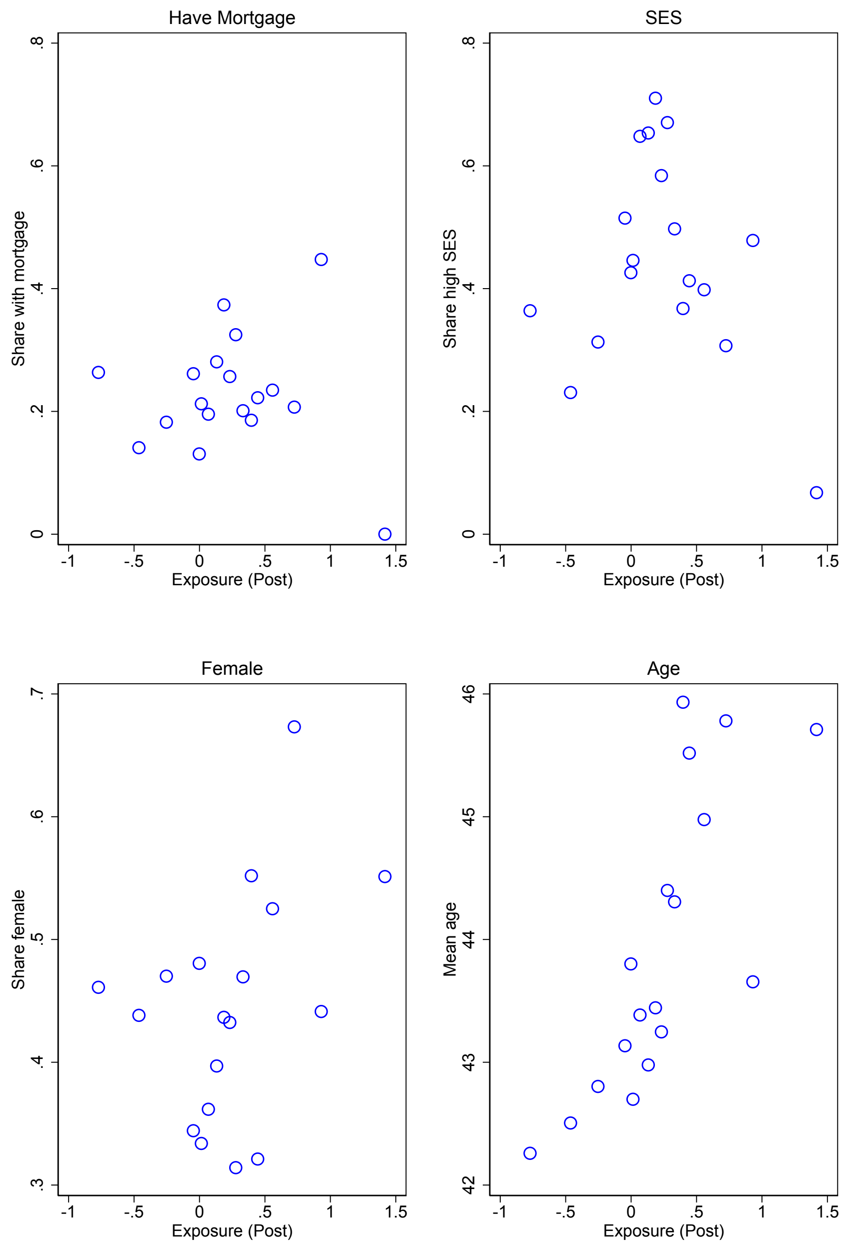
Upper panel: binned means of random forest cost predictions made without using registry data (vertical axis, log scale) by predicted value including registry data (horizontal axis, log scale). Bins are 20 quantiles of the distribution of full-information predictions for the no prior default and some prior default groups. 45-degree line plotted for convenience. Note that binned means are above teh 45-degree line for no default group and below the line for default group. Lower panel: Binned means of random forest cost predictions (horizontal axis; log scale) vs. out-of-sample observed cost outcomes (vertical axis, log scale). Left panel uses predictions that include registry information. Right panel uses predictions that exclude registry information. Our cost outcome measure is simply an indicator variable for at least one new default in the six month period beginning in February 2012, the date of registry deletion. Predictions are constructed using registry and borrowing data from December 2009 and June 2010. See text for details.

Figure 6: Density of log exposure to information deletion



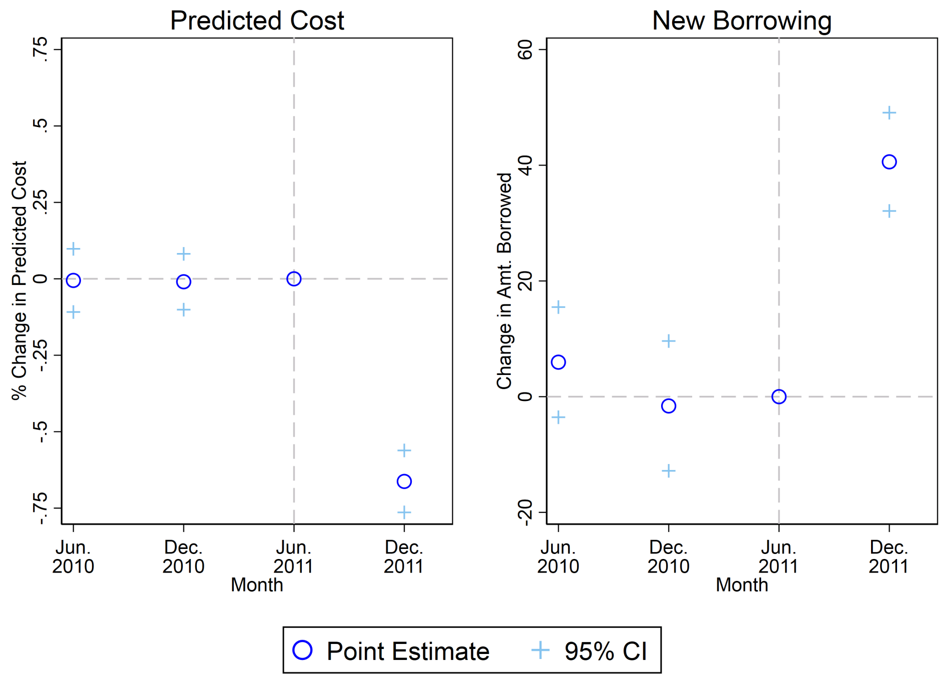
Histogram of exposure to predicted log cost changes by default status. Top: exposure generated from AC^{pre} predictions. Bottom: exposure generated from AC^{post} predictions. Red bars is exposure for defaulters, blue for non-defaulters. Defaulter mean pre-period (post) exposure is -0.32 (-0.17) and non-defaulter mean exposure is 0.33 (0.34). Graphs show exposure distribution between -1 and 1 for each group. Sample: borrower panel from December 2010 through December 2011.

Figure 7: Borrower SES and share of mortgage holders by exposure to information deletion



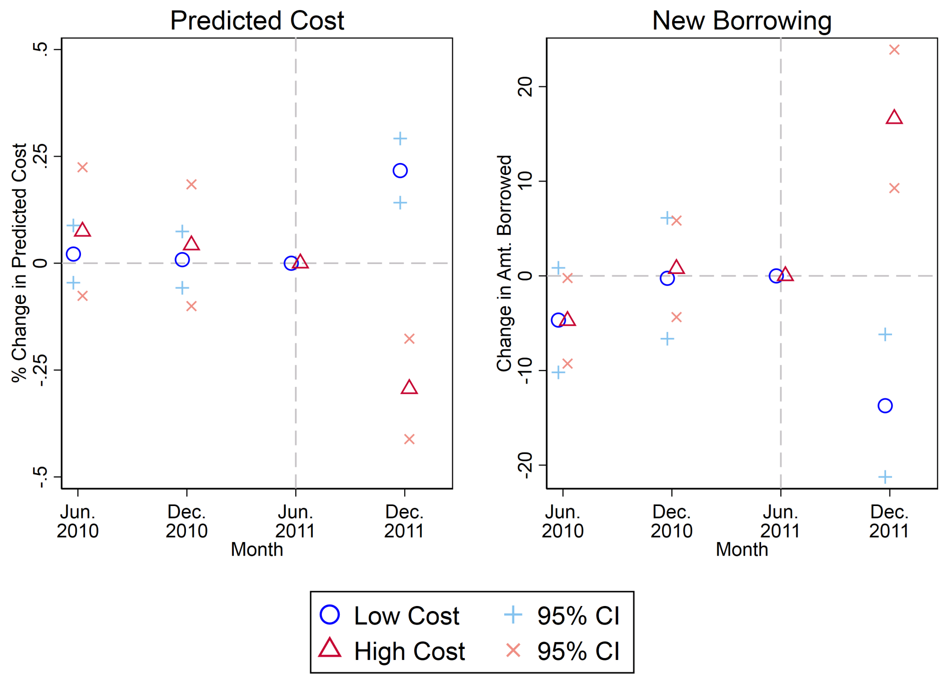
Binned means of indicators for having outstanding mortgage debt (left panel) and coming from a low-SES background (right panel) by decile of exposure distribution. Horizontal axis is log change in cost prediction from deletion. ML predictions come from pre-period training dataset.

Figure 8: Effects of registry deletion on defaulters relative to non-defaulters



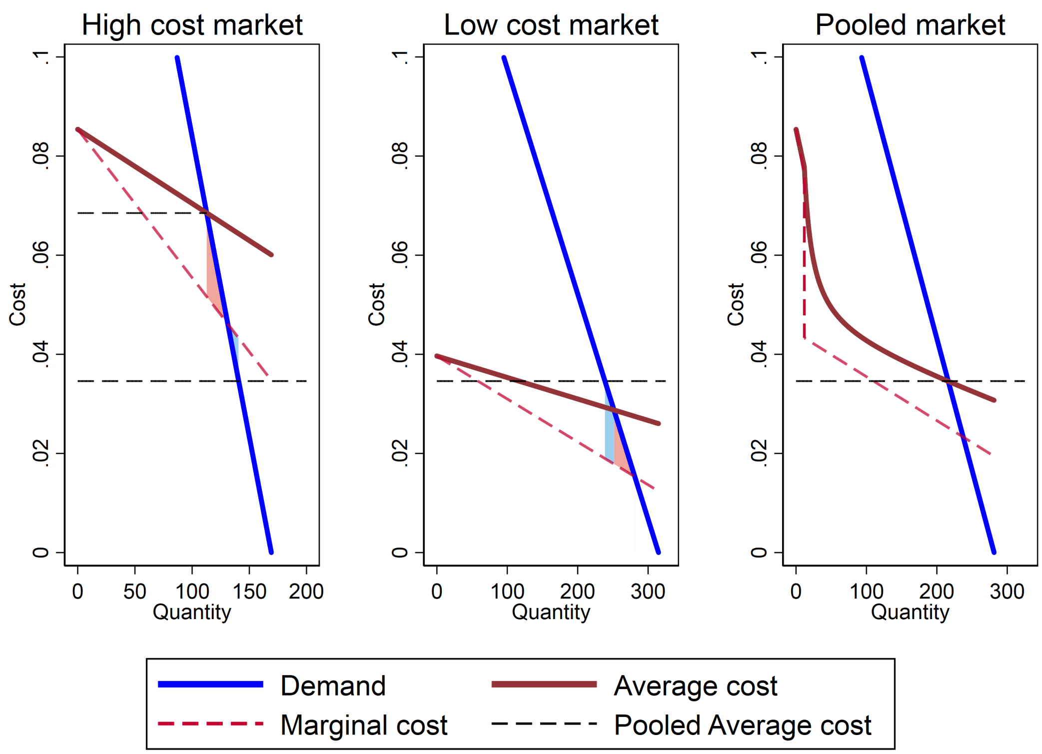
Difference-in-difference estimates and 95% CIs of the effects of prior default on predicted costs (left panel) and observed borrowing (right panel) using equation 2. Borrowing is measured over six month intervals with $t = 0$ in the six month period following deletion in February 2012. Consistent with the implementation of the deletion policy, default status is determined using registry snapshot three months prior to the start of each interval. Standard errors clustered at market level. See text for details.

Figure 9: Effects of registry deletion by exposure to changes in predicted costs



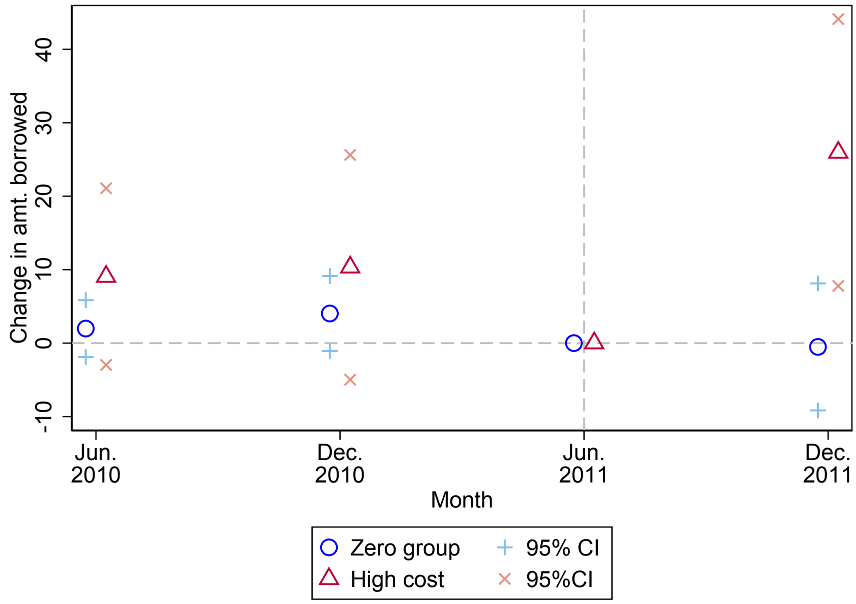
Difference-in-difference estimates and 95% CIs of the effects of exposure to changes in predicted costs on cost predictions (left panel) and new borrowing (right panel) using equation 2. Each panel splits the sample into individual with positive (high exposure) and negative (low exposure) changes in predicted costs. Effects for each group are measured relative to the omitted category of no exposure to changes in predicted costs, defined as the bottom ten percent of the distribution of the absolute value of cost changes. Standard errors clustered at market level. See text for details.

Figure 10: Empirical estimates of different markets



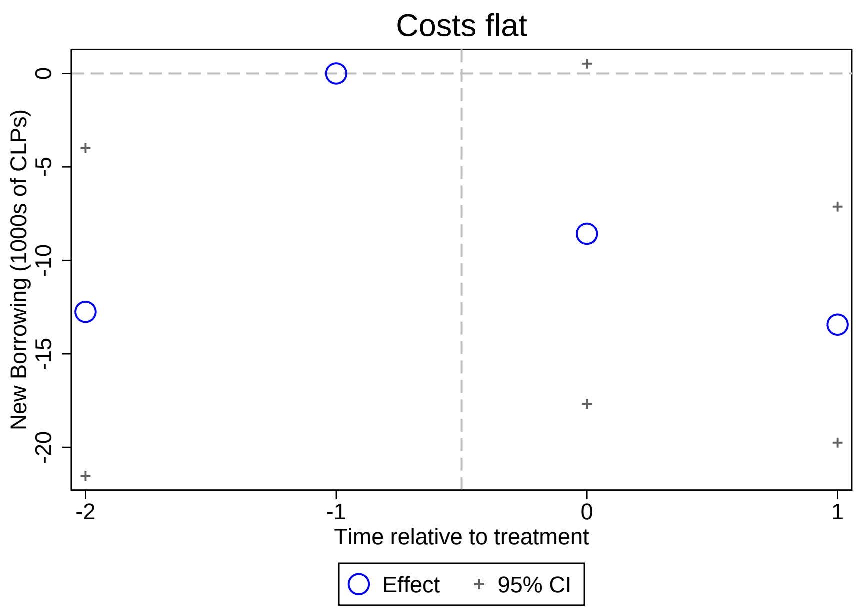
Empirical estimate of figure 1 using difference-in-difference estimates of slopes, assuming average cost pricing in both markets. See Section 5 for details.

Figure 11: Effects of registry deletion at the policy cutoff



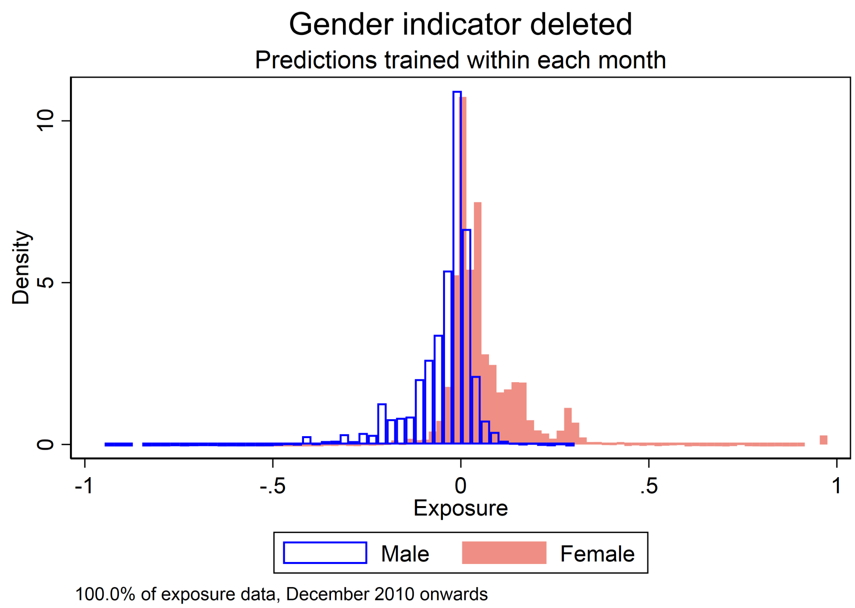
Difference-in-difference estimates and 95% confidence intervals for effects of the policy change at the policy cutoff of 2.5 million pesos using equation 3 for the exposure-defined 'zero group' and 'high-cost'. Horizontal axis in each graph is time in six month intervals relative to the February 2012 deletion policy. These estimates compare new borrowing for individuals whose defaults are less than the cutoff relative to those whose defaults are higher than the cutoff, before and after the policy change, for the high-cost and zero groups. Standard errors clustered at market level. See Section 5 for details.

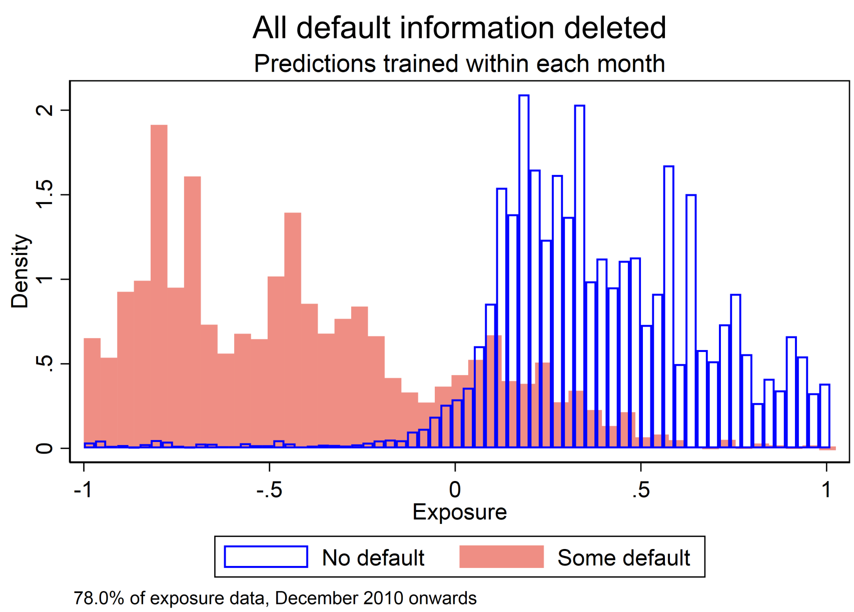
Figure 12: Effects of registry deletion by exposure and time relative to deletion



Difference-in-difference estimates and 95% confidence intervals for effects of exposure to changes in borrowing using equation 4 for the exposure-defined 'zero group' only. Horizontal axis in each graph is time in six month intervals relative to the February 2012 deletion policy. These estimates work by comparing changes in borrowing pre- and post-February 2012 to changes pre- and post-February 2011. The 'Costs flat' or zero group is the bottom 15% of the distribution of absolute values in cost changes. Exposure is measured using December 2011 registry data in the 'treatment' sample and in December 2010 in the 'control' sample. Standard errors clustered at market level. See text for details.

Figure 13: Distribution of exposure under counterfactual deletion policies





Histograms of exposure under counterfactual deletion policies. On top: log difference in prediced costs excluding and including a gender indicator variable, split by gender. Below: exposure defined when all default information is deleted from the credit registry, split by default amout. See text for details.

Table 1: Sample description

	All	In Panel	In Panel, Positive Borrowing
Any registry default	0.37	0.33	0.14
Deletion eligible	0.31	0.33	0.14
Observed deletion	0.29	0.30	0.17
Registry default amt.	554.50	182.00	54.45
Reg. default amt | reg. $<$2.5m	172.25	182.00	54.45
Debt balance	7,768	7,675	13,075
Consumer borrowing balance	2,172	2,097	2,634
Have mortgage	0.19	0.19	0.24
Mortgage balance	4,343	4,387	8,192
Any bank default	0.17	0.14	0.03
Bank default amt.	338.09	155.81	31.06
Bank default amt | reg. $<$2.5m	147.46	155.81	31.06
Average costs	0.12	0.09	0.01
New consumer borrowing	0.31	0.32	1.00
New consumer borrowing amt.	184	190	650
New bank default	0.08	0.08	0.05
New bank default amt.	36.57	27.28	14.55
Age	44.12	44.08	43.40
Female	0.44	0.45	0.45
Have SES	0.10	0.10	0.13
SES A	0.25	0.25	0.36
SES B	0.29	0.29	0.27
SES C	0.25	0.25	0.20
SES D & E	0.22	0.22	0.17
N of observations	23,001,337	21,769,213	4,593,511
N of clusters	330	330	330
N of individuals	5,577,605	5,433,403	2,314,786

Descriptive statistics on borrowing sample. Observations are at the person by half-year level. Data run from August 2009 through July 2012. Six-month snapshots run from February-July and August-January. Borrowing outcomes from each six month interval are linked to credit registry data from two months prior to the start of the interval (December and June, respectively). We refer to time periods by the registry month. Columns define samples. 'All' column is all Chilean consumer bank borrowers. 'In panel' is the set of borrowers with a positive balance six months prior to a given month. 'In panel, positive borrowing' is the subset of borrowers who additionally have new borrowing in the snapshot – a 10% random sample of this subset defines our machine learning training set, which we exclude from the main panel. See text for details. 'Positive default' and 'Default (amt)' are dummies for positive registry defaults and mean default amount conditional on some positive value, respectively. 'Borrowing' is mean consumer borrowing balance. 'New borrowing' is an indicator variable equal to one if quarterly consumer balance expands by 10%, and 'New borrowing, amt' is that indicator multiplied by the observed balance change. 'Debt,' 'New debt,' and 'New debt (amt)' are defined analogously but for all debt, including secured debt. SES categories are internal categorizations used by banks. 'Average costs' are the share of debt at least 90 days overdue divided by the total debt balance.

Table 2: Log Likelihoods of Various Algorithms

	$\mathrm{AC}^{\mathrm{pre}}$		$\mathrm{AC}^{\mathrm{post}}$	
	Training	Testing	Training	Testing
Naive Bayes				
With registry info	−0.412	−0.682	−0.398	−0.633
Without registry info	−0.324	−0.516	−0.319	−0.478
Logistic LASSO				
With registry info	−0.183	−0.327	−0.183	−0.336
Without registry info	−0.187	−0.338	−0.188	−0.348
Random Forest				
With registry info	−0.176	−0.278	−0.173	−0.295
Without registry info	−0.180	−0.284	−0.177	−0.305

Mean binomial log likelihoods for each algorithm. Columns identify the sample in which the log likelihood value is calculated. The 'training' sample is a 10% random sample of borrowers with new borrowing in the July 2009 Snapshot (which trains AC^{pre}) and within each snapshot (AC^{post}). 'Testing' identifies the main sample used in our analysis, from which the training set is dropped. Rows identify prediction methods. Within each prediction method, the 'with registy info' row uses registry information in addition to the other, while the 'without registry info' row does not. See section 4 for the full list of predictors and Appendix B for details on the transformation of these predictors and the structure of each algorithm.

Table 3: Demographics by exposure category

	Low cost market	Zero group	High cost market	Pooled
Positive Default	0.01	0.46	0.99	0.31
Amt. Default	52	696	456	566
New Borrowing	236	175	99	195
New Debt	468	356	156	384
Positive Bank Default	0.04	0.15	0.18	0.10
Low SES	0.50	0.56	0.71	0.55
Have Mortgage	0.25	0.18	0.18	0.22
Age	44.4	43.8	42.5	43.9
Female	0.47	0.41	0.46	0.45
Share of individuals	0.53	0.32	0.16	1
N	2,051,138	1,234,733	612,737	3,898,608

Baseline borrowing and demographic characteristics by exposure-generated market type in July 2011. Rows correspond to features of the sample and columns define market type. 'Positive default' is an indicator for whether individuals have positive default balances within the snapshot while 'Amt. Default' computes the mean default value conditional on having positive default. 'New borrowing' computes mean new borrowing across all individuals, as does new 'New debt.' 'Positive bank default' indicates positives bank default for individuals within the snapshot. 'Low SES' is an indicator flagging bank defined socioeconomic status.' Have mortgage' is an indicator flagging whether individuals have positive mortgage balances in the snapshot. 'Age' reports the mean age of individuals in the snapshot in years. 'Female' is flags gender reported to the bank. Share of individuals computes the share of total individuals in the snapshot contained in each market, while N reports the number of individuals (observations).

Table 4: Difference in differences by default and exposure

	Registry Information		*Low cost market*		*High cost market*	
	Predicted Costs	New Borrowing	Predicted Costs	New Borrowing	Predicted Costs	New Borrowing
Jun. 2010	-0.01	5.97	0.02	-4.67^{+}	0.07	-4.74^{*}
	(0.05)	(4.84)	(0.03)	(2.81)	(0.08)	(2.30)
Dec. 2010	-0.01	-1.60	0.01	-0.25	0.04	0.75
	(0.05)	(5.70)	(0.03)	(3.25)	(0.07)	(2.59)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	-0.66^{***}	40.59^{***}	0.22^{***}	-13.72^{***}	-0.29^{***}	16.60^{***}
	(0.05)	(4.32)	(0.04)	(3.83)	(0.06)	(3.72)
Elasticity				-0.29		-0.40
Dep. Var. Base Period Mean			0.04	215.28	0.10	140.98
N Clusters	329	329	303	303	282	285
N Obs.	3,228,458	15,513,587	2,910,733	13,093,725	1,273,371	7,493,968
N Individuals	2,031,005	4,693,948	1,836,294	4,363,940	986,205	3,212,628
N Exposed Individuals			505,295	2,132,055	84,746	608,229

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2. The first two columns report the difference-in-difference estimated effect of deletion on outcome variables listed in column headers, while the third and fourth estimate the dif-in-dif effect on the different exposure-defined markets. Sample in specifications where cost is an outcome conditions on positive borrowing (see text for details). We take the log of 'Predicted Cost' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted cost effect. 'N exposed individuals' reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Table 5: Difference in differences by exposure, mortgage, and socioeconomic status

	Low cost market				*High cost market*			
	Predicted Cost		New Borrowing		Predicted Cost		New Borrowing	
By Mortgage Status								
	No Mortage	Mortgage	No Mortage	Mortgage	No Mortage	Mortgage	No Mortage	Mortgage
Jun. 2010	0.03	-0.05^{*}	-5.21	-3.84	0.11	-0.10	-6.08^{*}	-0.78
	(0.04)	(0.02)	(3.31)	(4.22)	(0.08)	(0.06)	(2.56)	(4.33)
Dec. 2010	0.02	-0.05^{+}	1.04	4.48	0.07	-0.09^{+}	0.46	5.59
	(0.04)	(0.03)	(3.29)	(5.66)	(0.08)	(0.05)	(2.81)	(4.16)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.20^{***}	0.22^{***}	-13.22^{***}	-8.85	-0.27^{***}	-0.42^{***}	15.73^{***}	19.78^{***}
	(0.04)	(0.04)	(3.72)	(6.91)	(0.07)	(0.05)	(4.06)	(5.11)
Elasticity			-0.35	-0.13			-0.46	-0.23
Dep. Var. Base Period Mean	0.05	0.03	185.39	318.06	0.10	0.09	127.19	204.06
N Clusters	303	292	303	293	278	266	281	272
N Obs.	2,204,290	706,443	10,148,532	2,945,193	1,028,499	244,872	6,135,611	1,358,357
N Individuals	1,432,239	437,433	3,566,538	923,617	800,061	193,751	2,649,628	606,131
N Exposed Individuals	375,676	129,619	1,609,450	522,605	70,162	14,584	497,783	110,446
By Socioeconomic Status								
	Low SES	High SES	Low SES	High SES	Low SES	High SES	Low SES	High SES
Jun. 2010	0.04	-0.00	-0.40	-2.78	0.12	-0.04	-1.32	-6.32
	(0.05)	(0.02)	(3.59)	(3.59)	(0.09)	(0.05)	(2.89)	(4.15)
Dec. 2010	0.02	-0.02	1.61	-1.59	0.08	-0.05	-1.03	6.47
	(0.04)	(0.02)	(3.09)	(4.22)	(0.08)	(0.04)	(2.55)	(4.53)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	0.22^{***}	0.21^{***}	-8.78^{***}	-21.31^{***}	-0.30^{***}	-0.32^{***}	9.27^{**}	18.78^{***}
	(0.05)	(0.03)	(2.58)	(4.82)	(0.07)	(0.05)	(3.05)	(5.47)
Elasticity			-0.41	-0.30			-0.41	-0.24
Dep. Var. Base Period Mean	0.07	0.02	95.12	347.84	0.16	0.05	75.44	243.48
N Clusters	303	302	303	302	274	282	279	285
N Obs.	1,147,411	1,763,322	6,999,869	6,093,856	555,634	717,737	4,617,114	2,876,854
N Individuals	849,835	1,064,389	2,768,287	2,021,242	471,664	532,229	2,021,269	1,378,643
N Exposed Individuals	216,450	288,845	1,109,738	1,022,317	56,279	28,467	421,652	186,577

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference in difference estimates from equation 2 over defined subsamples. Columns 1 through 4 are predicted cost and borrowing diff-in-diff effect estimates in the low cost market while columns 5 through 8 report estimates in the high cost market. Column headers report dependent variable at the top and subsample below. Sample in specifications where cost is an outcome conditions on positive borrowing (see text for details). 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted cost effect within each market-subsample. We take the log of 'Predicted Cost' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted cost effect. 'N exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Table 6: Difference in difference estimates on average costs

	Registry Information	Low cost market Pooled	No Mortgage	Have Mortgage	Low SES	High SES	High cost market Pooled	No Mortgage	Have Mortgage	Low SES	High SES
Jun. 2010	−0.00	0.02	0.03	−0.05*	0.04	−0.00	0.07	0.11	−0.10	0.12	−0.04
	(0.05)	(0.03)	(0.04)	(0.02)	(0.05)	(0.02)	(0.08)	(0.08)	(0.06)	(0.09)	(0.05)
Dec. 2010	−0.01	0.00	0.01	−0.05+	0.01	−0.02	0.04	0.06	−0.09+	0.07	−0.05
	(0.05)	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(0.07)	(0.08)	(0.05)	(0.08)	(0.04)
Jun. 2011	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Dec. 2011	−0.09*	0.02	0.01	0.03	0.01	0.03	−0.04	−0.03	−0.08	−0.06	−0.02
	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.06)	(0.07)	(0.05)	(0.07)	(0.05)
Elasticity		0.10	0.05	0.12	0.06	0.16	0.12	0.11	0.18	0.20	0.05
Dep. Var. Base Period Mean		0.04	0.05	0.03	0.07	0.02	0.10	0.10	0.09	0.16	0.05
N Clusters	329	303	303	292	303	302	284	281	268	278	283
N Obs.	5,468,917	4,930,411	3,734,294	1,196,117	1,943,879	2,986,532	2,156,891	1,742,719	414,172	941,603	1,215,288
N Individuals	2,630,314	2,385,366	1,894,374	558,811	1,201,766	1,347,265	1,433,629	1,167,470	284,204	721,292	755,356
N Exposed Individuals		855,928	636,066	219,862	366,368	489,560	143,165	118,441	24,724	94,855	48,310

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 2 where the dependent variable is average cost. 'Registry information' reports the estimated effect of deletion while other column headers only define subsamples. Columns 2-6 are estimated over the low cost market (as defined by exposure) while columns 7-11 are over the high cost market. We take the log of 'Average cost' when estimating the regressions but report the its base period mean in levels. 'Elasticity' is average cost effect scaled by the predicted cost effect. 'N exposed individuals reports the number of individuals not in the 0 group included in the regression sample in the treatment period. Since some individuals appear in multiple snapshots we report both individuals and observations. Standard errors clustered at market level. See text for details.

Table 7: Distribution of deletion effects

	Separate	Pooled	Difference
Low cost market			
Predicted cost	0.029	0.035	0.006
Average cost	0.029	0.029	0.001
New borrowing (1000s CLP)	251.561	238.714	-12.847
Welfare loss (1000s CLP)	0.161	0.331	0.170
Aggregate new borrowing (Bns CLP)	516	490	-26
Aggregate welfare loss (1000s CLP)	330,480	679,717	349,238
			105.68%
N individuals	2,051,138	2,051,138	2,051,138
High cost market			
Predicted cost	0.069	0.035	-0.034
Average cost	0.069	0.064	-0.004
New borrowing (1000s CLP)	112.713	140.695	27.981
Welfare loss (1000s CLP)	0.156	0.041	-0.114
Aggregate new borrowing (Bns CLP)	69	86	17
Aggregate welfare loss (1000s CLP)	95,456	25,307	$-70,149$
			-73.49%
N individuals	612,737	612,737	612,737
Combined			
Average price	0.033	0.035	0.001
Average cost	0.033	0.035	0.001
New borrowing (1000s CLP)	219.624	216.168	-3.455
Welfare loss (1000s CLP)	0.160	0.265	0.105
			65.52%
Aggregate new borrowing (Bns CLP)	585	576	-9
Aggregate welfare loss (1000s CLP)	425,936	705,025	279,089
			65.52%
N individuals	2,663,875	2,663,875	2,663,875

This table describes changes in key welfare metrics before and following deletion. Prices and welfare calculations assume average cost pricing. See text for details. 'Low cost market' panel is individuals whose predicted costs rise following deletion, 'High cost market' is individuals whose predicted costs fall. 'Aggregate' panel averages over both markets for prices, average cost, new borrowing, and welfare measures, while summing for aggregate borrowing/welfare measures. 'New borrowing' in 1000s of CLP. Aggregate new borrowing is in billions of CLP.

Table 8: Welfare changes by markup

Low cost market markup (%)	*Additional high cost market markup (%)*					
	0	5	10	25	50	100
0	0.17					
	−0.11					
	0.10					
	65.52%					
5	0.19	0.19	0.19	0.19	0.19	0.20
	−0.19	−0.19	−0.19	−0.20	−0.22	−0.25
	0.10	0.10	0.10	0.10	0.10	0.09
	51.94%	51.40%	50.87%	49.27%	47.95%	44.09%
10	0.21	0.21	0.21	0.21	0.22	0.23
	−0.26	−0.27	−0.27	−0.29	−0.33	−0.40
	0.10	0.10	0.10	0.10	0.09	0.08
	42.35%	41.47%	41.77%	39.16%	37.19%	31.24%
25	0.27	0.27	0.27	0.28	0.30	0.32
	−0.48	−0.49	−0.51	−0.57	−0.66	−0.86
	0.10	0.10	0.09	0.09	0.08	0.05
	26.58%	26.05%	24.66%	22.28%	18.25%	11.01%
50	0.37	0.38	0.38	0.40	0.43	0.49
	−0.84	−0.88	−0.92	−1.04	−1.25	−1.68
	0.09	0.09	0.08	0.07	0.04	−0.01
	15.01%	14.52%	13.43%	10.28%	6.02%	−0.72%
100	0.57	0.59	0.60	0.64	0.71	0.85
	−1.56	−1.65	−1.74	−2.00	−2.46	−3.38
	0.08	0.07	0.06	0.03	−0.02	−0.12
	7.23%	6.46%	5.33%	2.49%	−1.59%	−7.79%
200	0.98	1.01	1.03	1.13	1.28	1.61
	−3.00	−3.20	−3.39	−3.97	−4.94	−6.88
	0.06	0.04	0.02	−0.04	−0.15	−0.35
	2.81%	1.85%	0.71%	−1.81%	−5.57%	−11.18%

This table describes changes in changes in welfare loss before and following deletion. Cells are additional markups (columns, in percent terms) relative to a given markup rate in the low cost market (rows). Within each cell, rows are level changes in welfare loss in the low cost, high cost, mean change in welfare loss across both markets, and percent change in welfare loss relative to baseline loss the pooled market following deletion.

Table 9: Effects of counterfactual exposure policies

	Exposed to cost increases	Zero group	Exposed to cost decreases	Pooled
Gender indicator deleted				
Exposure to cost increases	0.24	0.00	-0.25	0.00
Positive Default	0.00	0.34	0.00	0.36
Amt. Default	479	571	71	1,621
New Borrowing	63	184	81	168
New Debt	203	369	106	337
Positive Bank Default	0.02	0.10	0.04	0.10
Low SES	0.18	0.22	0.17	0.22
Have Mortgage	0.08	0.21	0.12	0.20
Age	45.3	43.9	45.5	44.1
Female	0.98	0.44	0.16	0.45
Share of individuals	0.04	0.87	0.04	1
N	171,878	4,111,244	166,565	4,721,885
All default information deleted				
Exposure to cost increases	0.63	0.06	-0.84	0.15
Positive Default	0.07	0.18	0.93	0.36
Amt. Default	460	432	602	1,621
New Borrowing	135	535	77	168
New Debt	307	985	128	337
Positive Bank Default	0.06	0.08	0.20	0.10
Low SES	0.22	0.16	0.26	0.22
Have Mortgage	0.22	0.18	0.18	0.20
Age	44.4	45.2	42.5	44.1
Female	0.46	0.43	0.44	0.45
Share of individuals	0.55	0.13	0.25	1
N	2,615,689	630,130	1,203,868	4,721,885

Baseline borrowing and demographic characteristics by exposure-generated market type in July 2011 under counterfactual policy changes. Panels are separated by counterfactual policy: deleting a gender indicator variable and deleting all default information. Rows correspond to features of the sample and columns define market type. 'Positive default' is an indicator for whether individuals have positive default balances within the snapshot while 'Amt. Default' computes the mean default value conditional on having positive default. 'New borrowing' computes mean new borrowing across all individuals, as does new 'New debt.' 'Positive bank default' indicates positives bank default for individuals within the snapshot. 'Low SES' is an indicator flagging bank defined socioeconomic status.' Have mortgage' is an indicator flagging whether individuals have positive mortgage balances in the snapshot. 'Age' reports the mean age of individuals in the snapshot in years. 'Female' is flags gender reported to the bank. Share of individuals computes the share of total individuals in the snapshot contained in each market, while N reports the number of individuals (observations).