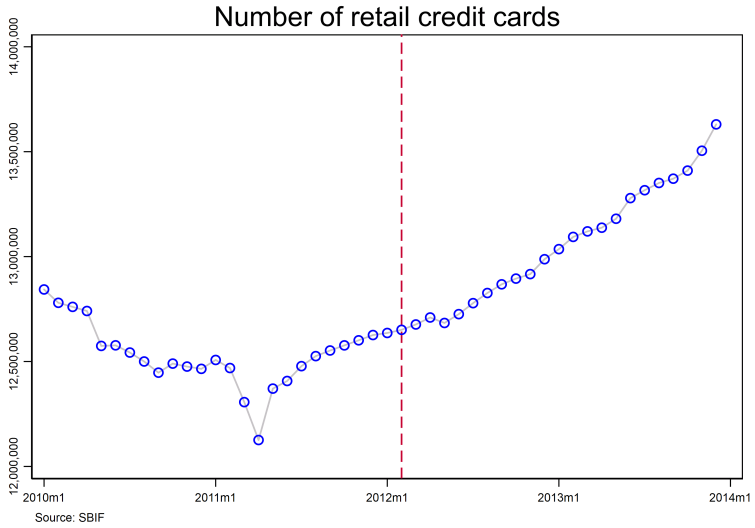
Internet Appendix

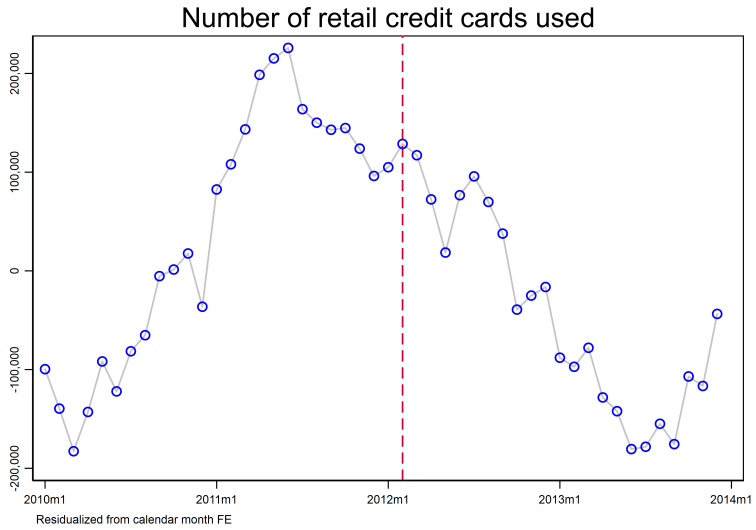
A Additional Results

Figure A1: Stock of retail credit cards over time



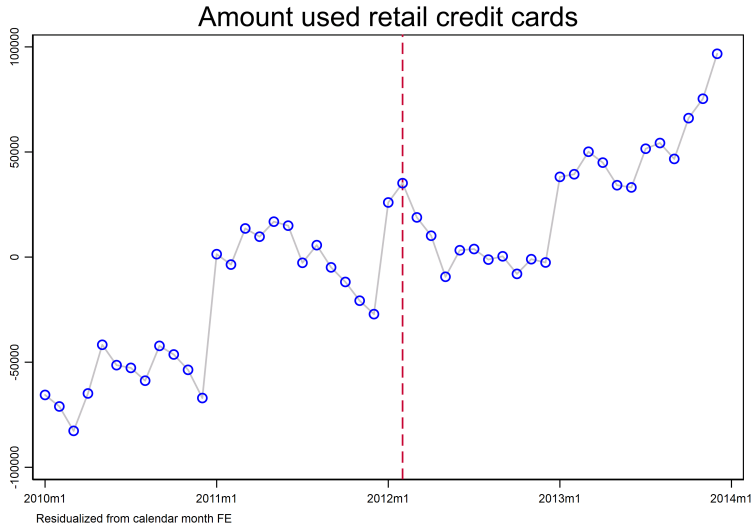
Stock of retail credit cards by month. Time of deletion policy noted with vertical line.

Figure A2: Retail credit cards in use over time



Number of retail credit cards used by month. Time of deletion policy noted with vertical line. Source: SBIF.

Figure A3: Number of retail credit card uses over time



Amount of retail credit purchases by month. Time deletion policy noted with vertical line. Source: SBIF.

Table A1: Surplus 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 0.10 51.94%	0.19 −0.19 0.10 51.40%	0.19 −0.19 0.10 50.87%	0.19 −0.20 0.10 49.27%	0.19 −0.22 0.10 47.95%	0.20 −0.25 0.09 44.09%
10	0.21 −0.26 0.10 42.35%	0.21 −0.27 0.10 41.47%	0.21 −0.27 0.10 41.77%	0.21 −0.29 0.10 39.16%	0.22 −0.33 0.09 37.19%	0.23 −0.40 0.08 31.24%
25	0.27 −0.48 0.10 26.58%	0.27 −0.49 0.10 26.05%	0.27 −0.51 0.09 24.66%	0.28 −0.57 0.09 22.28%	0.30 −0.66 0.08 18.25%	0.32 −0.86 0.05 11.01%
50	0.37 −0.84 0.09 15.01%	0.38 −0.88 0.09 14.52%	0.38 −0.92 0.08 13.43%	0.40 −1.04 0.07 10.28%	0.43 −1.25 0.04 6.02%	0.49 −1.68 −0.01 −0.72%
100	0.57 −1.56 0.08 7.23%	0.59 −1.65 0.07 6.46%	0.60 −1.74 0.06 5.33%	0.64 −2.00 0.03 2.49%	0.71 −2.46 −0.02 −1.59%	0.85 −3.38 −0.12 −7.79%
200	0.98 −3.00 0.06 2.81%	1.01 −3.20 0.04 1.85%	1.03 −3.39 0.02 0.71%	1.13 −3.97 −0.04 −1.81%	1.28 −4.94 −0.15 −5.57%	1.61 −6.88 −0.35 −11.18%

This table describes changes in changes in surplus 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 surplus loss in the low cost, high cost, mean change in surplus loss across both markets, and percent change in surplus loss relative to baseline loss the pooled market following deletion.

Table A2: Difference-in-difference predictions using long run default measures

	Positive exposure			*Negative exposure*		
	Predicted Default	Average Cost	New Borrowing	Predicted Default	Average Cost	New Borrowing
Jun. 2010	0.01	0.00	−7.09*	0.03	0.03	−5.68^{+}
	(0.02)	(0.02)	(3.05)	(0.05)	(0.05)	(3.23)
Dec. 2010	0.01	0.01	−2.11	0.02	0.01	0.30
	(0.02)	(0.02)	(3.52)	(0.05)	(0.05)	(3.25)
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	0.25***	0.12***	−13.28**	−0.30***	0.04	17.98***
	(0.02)	(0.02)	(4.21)	(0.04)	(0.04)	(3.47)
Elasticity		0.48	−0.24		−0.12	−0.36
Dep. Var. Base Period Mean	0.08	0.08	214.70	0.14	0.14	165.09
N Clusters	307	307	307	299	299	300
N Obs.	2,929,133	4,961,674	13,163,613	1,486,567	2,519,339	8,117,207
N Individuals	1,844,615	2,394,399	4,373,700	1,104,246	1,571,258	3,422,263
N Exposed Individuals	452,132	765,941	1,967,865	79,572	134,306	589,628

Significance: $^{+}$ 0.10 * 0.05 ** 0.01 *** 0.001. Difference and difference estimates from equation 1. Table is identical to Table 4 but uses a one-year ahead measure of default to compute predicted default rates. See section 3 for details. 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. We take the log of 'Predicted default' for estimation but report the base period mean in levels. 'Elasticity' is borrowing effect scaled by base period outcome mean and predicted default effect. '*N* exposed individuals' reports the number of individuals not in the zero 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.

Table A3: Distribution of deletion effects using long run default measures

	Separate	Pooled	Difference
Positive exposure			
Predicted cost	0.065	0.081	0.016
Average cost	0.065	0.073	0.008
New borrowing (1000s CLP)	234.779	222.246	−12.533
Surplus loss (1000s CLP)	1.711	2.138	0.427
Aggregate new borrowing (Bns CLP)	447	424	−24
Aggregate surplus loss (1000s CLP)	3,261,672	4,075,579	813,908
			24.95%
N individuals	1,905,946	1,905,946	1,905,946
Negative exposure			
Predicted cost	0.120	0.081	−0.039
Average cost	0.120	0.125	0.005
New borrowing (1000s CLP)	112.490	132.079	19.589
Surplus loss (1000s CLP)	0.140	1.128	0.988
Aggregate new borrowing (Bns CLP)	67	78	12
Aggregate surplus loss (1000s CLP)	83,086	668,656	585,570
			704.77%
N individuals	592,732	592,732	592,732
Combined			
Average cost	0.072	0.081	0.008
New borrowing (1000s CLP)	205.770	200.857	−4.913
Surplus loss (1000s CLP)	1.339	1.899	0.560
			41.84%
Aggregate new borrowing (Bns CLP)	514	502	−12
Aggregate surplus loss (1000s CLP)	3,344,758	4,744,236	1,399,478
			41.84%
N individuals	2,498,678	2,498,678	2,498,678

This table describes changes in key metrics before and following deletion, with inputs to the theoretical framework using the long-run cost measure, assuming a 0% markup.

B Detail on the machine learning procedure

We generate cost predictions by regressing an indicator for new default against a large selection of features using a random forest algorithm. We create four sets of predictions trained on 10% of the data with new borrowing within each snapshot – approximately 8% of the overall data. Predictions are trained and predicted either contemporaneously, within each 6-month post-December snapshot (PD^{post}), or only in the December 2009 snapshot (PD^{pre}). The random forests for each type are constructed with or without registry information. We use `python`'s `sklearn` package to perform our machine learning tasks (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot and Duchesnay 2011).

Our random forest regression design constructs regression trees using a feature vector of the following observable characteristics of each observation: a gender indicator, and one and two period lags of innovations in borrowing, innovations in total debt, total borrowing, total debt, average costs, and credit line information. We additionally include the default history deleted from the credit registry in some of the trees. In total, these trees have either thirteen or fourteen predictor variables.

We scale our features by binning their nonzero values into quartiles. This reduces noise in the feature vector and creates parsimonious regression trees. In our dataset, we find that this additionally decreases the time necessary to construct a random forest. Finally, we subset over only new borrowers in each period so that our cost estimates reflect costs conditional on borrowing.

To generate our PD^{pre} predictions, we train a model only using observations in the December 2009 snapshot. PD^{post} predictions are generated using a training sample from each snapshot; these predictions are actually generated using a suite of models each tied to a particular snapshot.

We use three-fold cross validation combined with a grid search to pick parameters for each model. The parameters over which we search are the minimum number of observations in a terminal node (*minleaf*) and the number of features over which each tree can sample. We set the number of trees in a forest to 150. Predictive power is not sensitive to choices in this range. See figures B1 and B2 to see outcomes from this procedure.

Constructing random forests is (generally) a supervised learning task. Breiman (2001) defines a random forest as a set of regression trees, $h_k = h(x, \Theta_k)$ where h is a tree and Θ_k is a random selection of observations and features from the training data,

where each tree "votes" on the output given an observation. We pick splits in the data to reduce mean-squared error, as is common with regression tasks. We use this loss function and a regression task, despite our target variable existing only in $\{0, 1\}$, to ensure that our outputs are continuous on $[0, 1]$ and reflect probabilities. Our predictions are best thought of as a weighted average of default rate in pools of observations clustered together by similarity along a set of their covariates.

We additionally estimate a regression tree[21] to bin borrowers into smaller markets. We define a market as a set of observations M such that $h(x_i, \Theta)$ returns a prediction stemming from the same terminal node for all $i \in M$. We use this method to cluster borrowers into borrowers with similar features and default rates. These clusters therefore represent infered groups in the data at the level which we believe the treatment is applied and are analagous to the clusters defined in each tree in the forest.

Finally, we recreate the analysis above, exchanging the random forest algorithm for two other machine learning procedures that return classification probabilities. These are a naive Bayes classifier and a logistic LASSO. Our naive Bayes classifier first bins nonzero values along the feature vector into quartiles. Under the naive assumption of independence of features in the feature vector, the classifier constructs $P(\text{default}|X)$ using Bayes' formula under the assumption that $P(X|\text{default})$ is Gaussian, though this is functionally irrelevant due to binning.

For the logistic LASSO, we take the log of nonzero values of continuous features, dummying out zero values using indicator variables. We perform a logistic regression with a λ penalty term of the sum absolute value of the coefficients and use three-fold cross validation to pick λ for each model.

Finally, we classify observations' socioeconomic status by training a random forest classifier on observations for whom the bank defined socioeconomic status group. Our three-fold cross validation procedure indicates that we are able to do this with approximately 35% accuracy using a random forest composed of 100 trees and built on a feature vector consisting of continuous measures of consumer debt, mortgage amount, debt balance, credit line, bank default, average cose, age, total default amount, and indicators for gender, new borrowing, and having positive borrowing cap. See figure B3 for cross-validation output.

[21] We estimate CART-style regression trees that split using variance reduction (Breiman, Friedman, Stone and Olshen 1984).

Figure B1: Cross-validation output forPD^{pre} random forest predictions

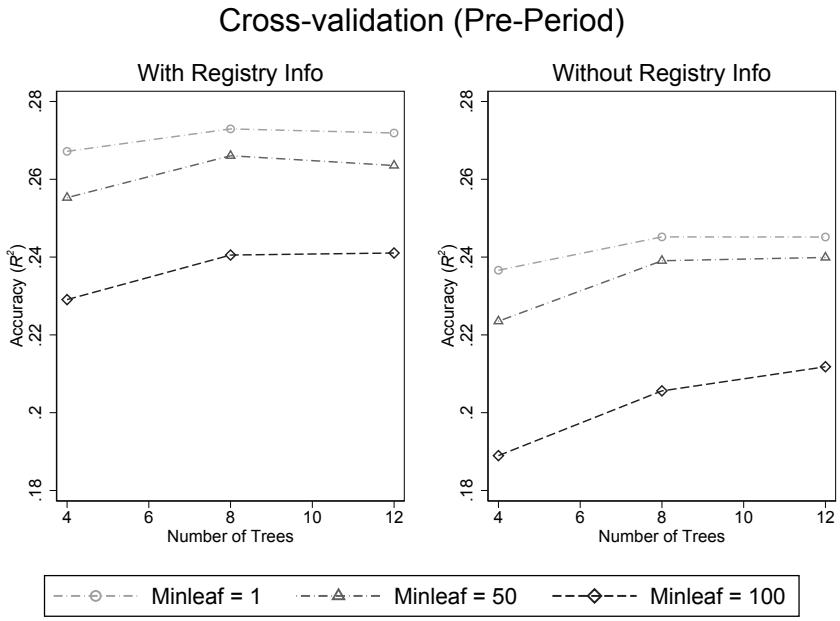


Figure B2: Cross-validation output for PD^{post} random forest predictions

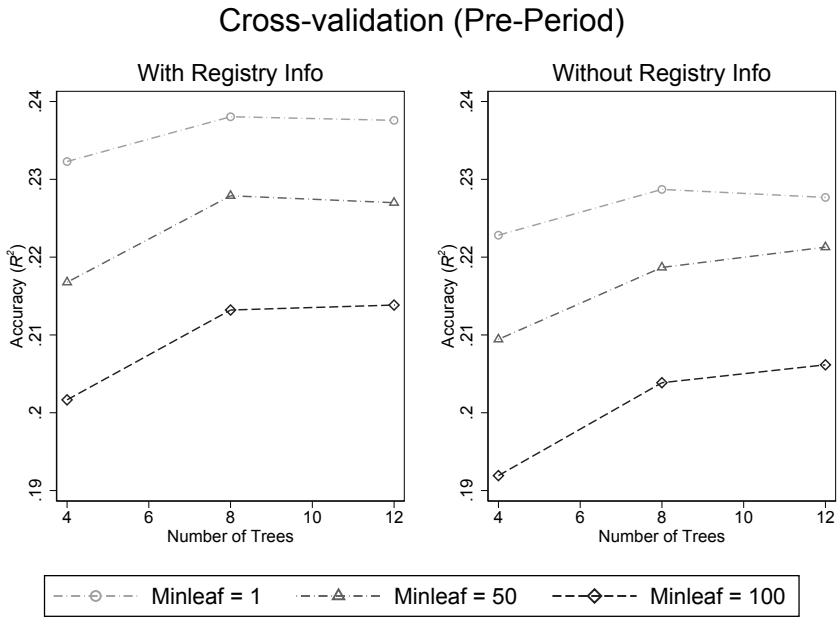
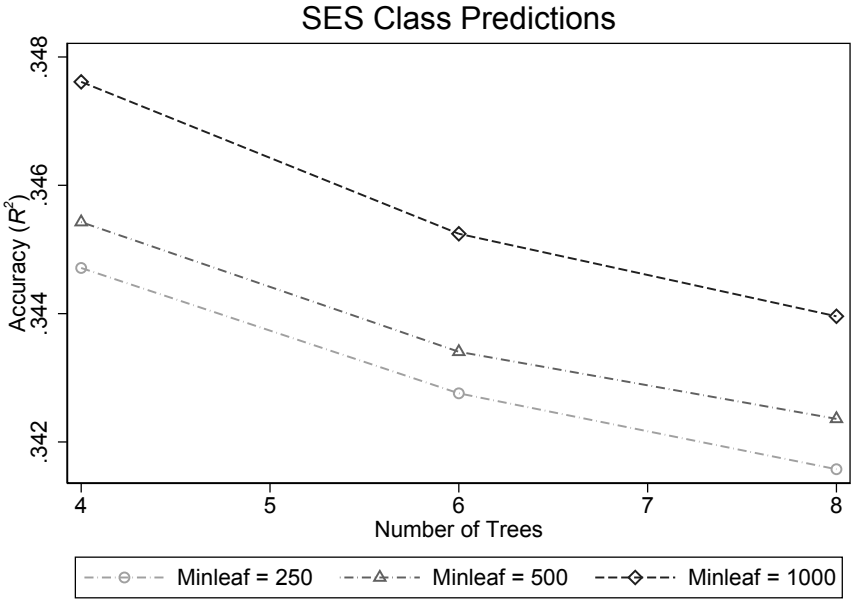


Figure B3: Cross-validation output for PD^{post} logistic LASSO predictions



C Model details

C.1 Model setup

This Appendix presents the details of main text Section 4. Model setup is as in the text. Let there be 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.

C.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. We focus on the empirically relevant case where there is adverse selection in both markets; i.e., where marginal cost curves are downward sloping. 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 11, respectively), with total surplus loss in each market given by the formula:

$$DWL_z = \frac{1}{2}\left(q_z^* - q_z(AC_z^e)\right) \times \left(AC_z^e - MC_z(AC_z^e)\right). \tag{4}$$

C.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 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 right panel of Figure 11, 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 11 in the main text illustrates the latter case, with surplus losses from overprovision equal to the area of triangle C in the left panel.

C.1.3 Measuring the effects of pooling

The effects of pooling on equilibrium borrowing and surplus 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 effects on surplus requires knowledge of the levels and slopes of marginal

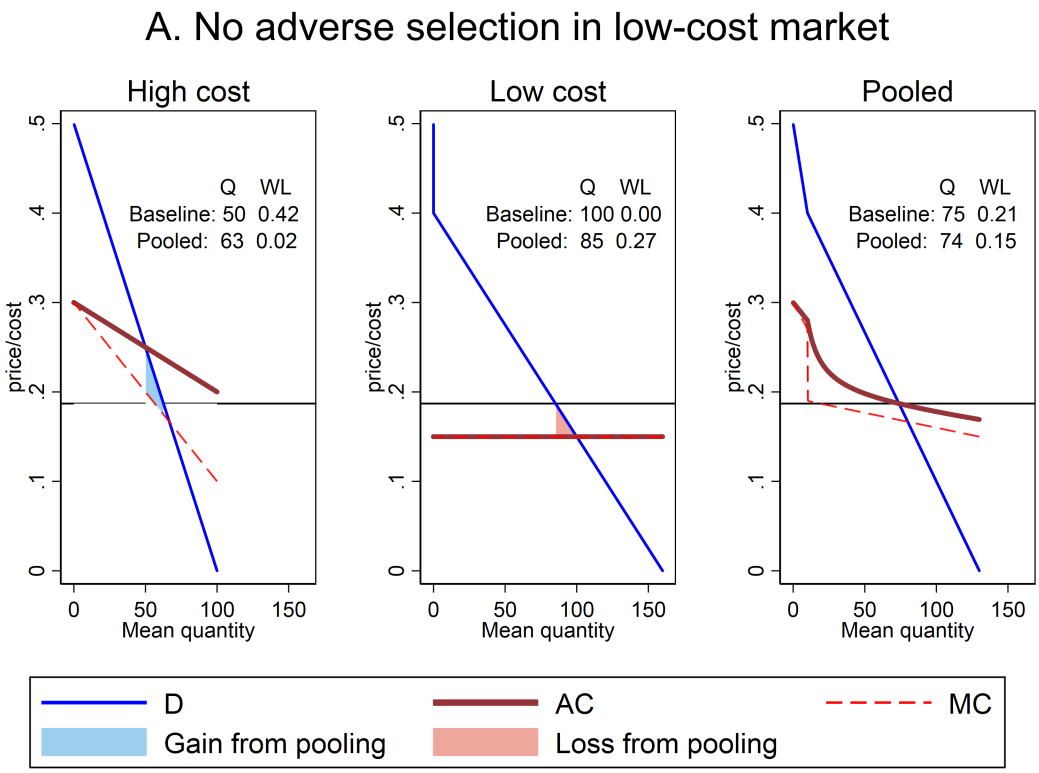
cost curves in addition to the demand and average cost curves. Here we exploit the observation that the equilibrium value of marginal cost $MC_z^e = \frac{dAC_z}{dq} q_z^e + AC_z^e$, and that with linear average cost curves $\frac{dMC_z}{dq} = 2\frac{dAC_z}{dq}$.

In Appendix Figure C1, 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 surplus are ambiguous, even in this simple model, and how they relate to the slopes of demand and costs.

C.1.4 Alternative modeling approaches

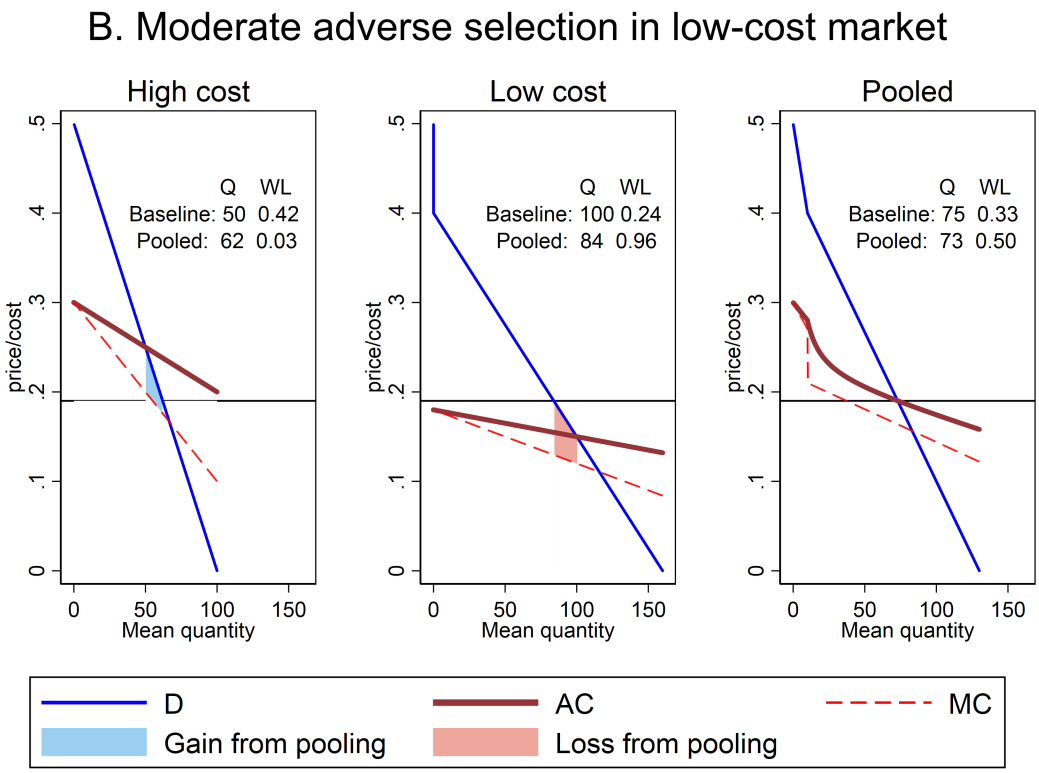
The framework we use to evaluate the consequences of the deletion policy on surplus here is one of several plausible modeling approaches. Most notably, we assume that lenders set prices rather than offering contracts consisting of rate-quantity pairs (Rothschild and Stiglitz 1976), and that the form of contract does not change following policy implementation. This rules out separating equilibria where lenders screen borrowers based on their contract choice (Bester 1985). In a simple screening model equilibrium, however, 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 increases surplus losses.

Figure C1: Simulated separating and pooling equilibria



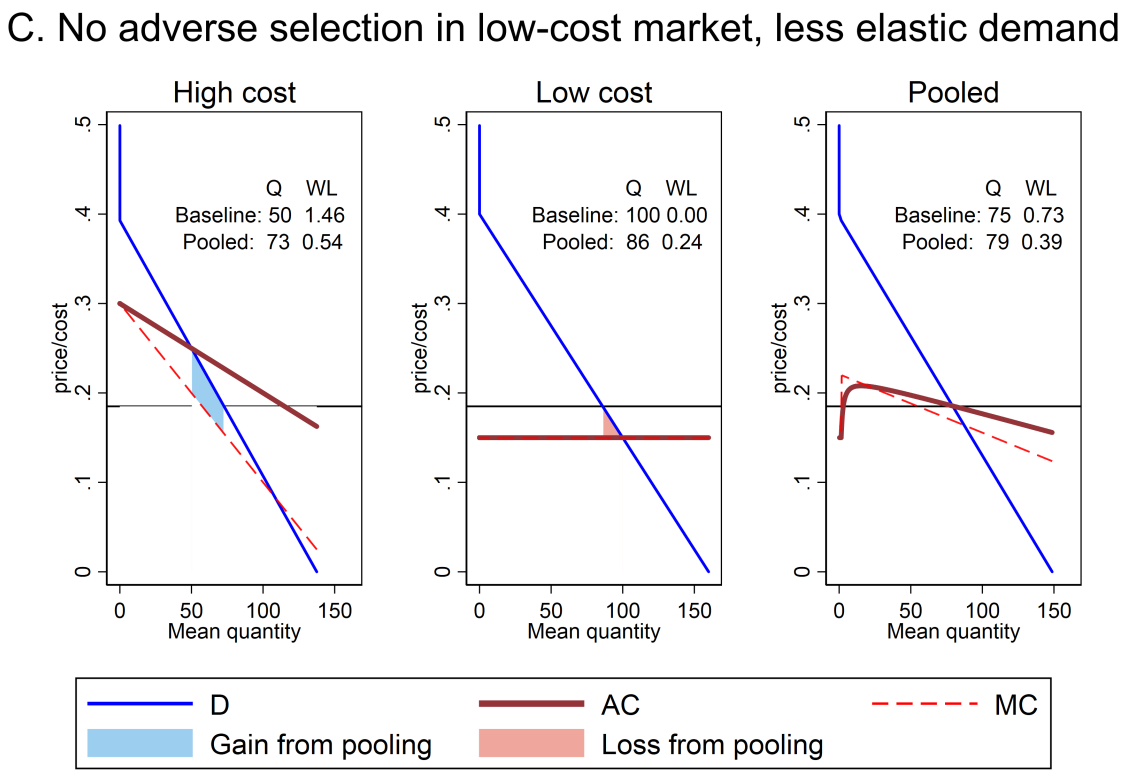
Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

Figure C1: (Cont'd) Simulated separating and pooling equilibria



Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.

Figure C1: (Cont'd) Simulated separating and pooling equilibria



Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted ("Q" column) and welfare loss relative to the efficient quantity ("WL" column) under the separate ("baseline") equilibrium and the "pooled" equlibrium. To see changes in aggregate welfare from pooling compare the "pooled" and "baseline" welfare loss columns in the rightmost panel of each row. Pooling causes welfare to rise in Panel A, fall in Panel B, and rise in Panel C. The number of individuals in high- and low-cost market are normalized to one. Separate equilibrium price and quantity (p, q) are the same in each panel, with $(p, q) = (0.25, 50)$ in the high market and $(p, q) = (0.15, 100)$ in the low-cost market. $\frac{dAC_0}{dq}$ and $\frac{AC_1}{dq}$ are the slopes of average cost curves in the low- and high-cost markets, respectively, with analogous definitions for demand curves. Slopes parameters vary across rows as follows. Panel A: $(0, -0.001, -400, -200)$ Panel B: $(-0.0003, -0.001, -400, -200)$. Panel C: $(0, -0.001, -400, -350)$. See text for model details.