

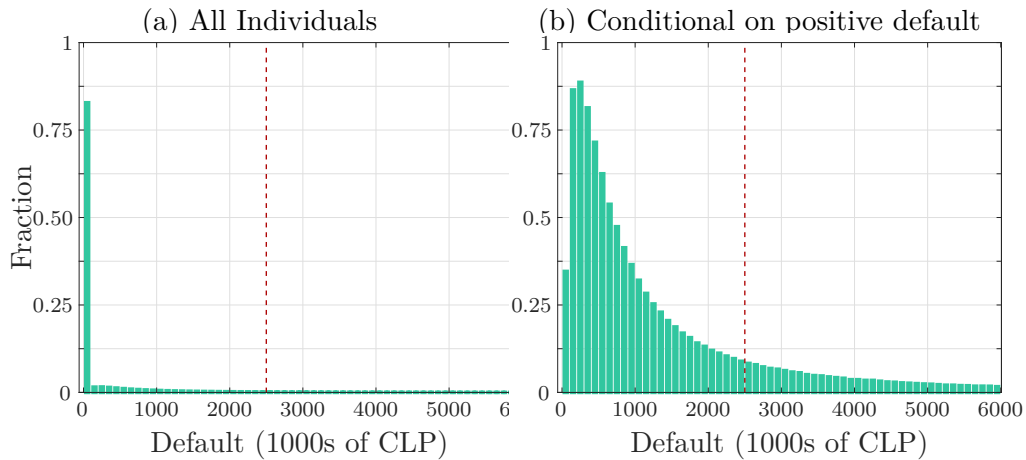
## 1. ONLINE APPENDIX - ADDITIONAL TABLES AND RESULTS

TABLE 1  
DEMOGRAPHICS BY EXPOSURE CATEGORY

	Positive exposure	Zero group	Negative exposure	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

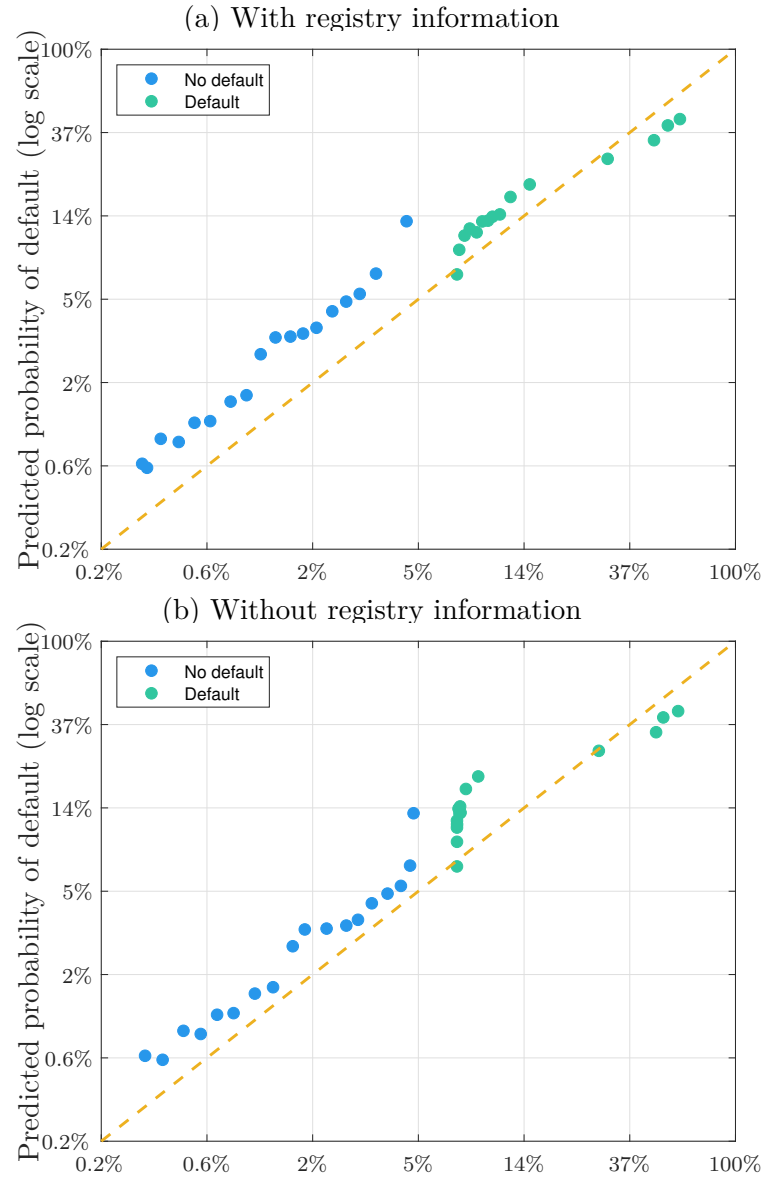
Baseline borrowing and demographic characteristics by exposure-generated market type in July 2011. Rows correspond to features of the sample and columns define market type. ‘Positive default’ is an indicator for whether individuals have positive default balances within the snapshot while ‘Amt. Default’ 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).

Figure 1: Histogram of ammount in default as of December 2011



Note: The left panel shows a histogram of consolidated defaults as of December 2011, for amounts below \$6 million CLP (approximately \$3,000). The right panel shows a histogram of consolidated defaults for individuals with positive defaults only.

Figure 2: Predictions with and without registry data



Binned means of random forest default predictions (horizontal axis; log scale) vs. out-of-sample observed default outcome (vertical axis, log scale). Top panel uses predictions that include registry information. The bottom panel uses predictions that exclude registry information. Our default outcome measure is 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.

TABLE 2  
DiD BY EXPOSURE, MORTGAGE, AND SOCIOECONOMIC STATUS

Positive Exposure				
	Predicted Defaults		New Borrowing	
By Mortgage Status				
	No Mortgage	Mortgage	No Mortgage	Mortgage
Jun. 2010	0.03 (0.04)	−0.05* (0.02)	−5.21 (3.31)	−3.84 (4.22)
Dec. 2010	0.02 (0.04)	−0.05+ (0.03)	1.04 (3.29)	4.48 (5.66)
Dec. 2011	0.20*** (0.04)	0.22*** (0.04)	−13.22*** (3.72)	−8.85 (6.91)
Elasticity			−0.35	−0.13
Dep. Var. Base Period Mean	0.05	0.03	185.39	318.06
N Clusters	303	292	303	293
N Obs.	2,204,290	706,443	10,148,532	2,945,193
N Individuals	1,432,239	437,433	3,566,538	923,617
N Exposed Individuals	375,676	129,619	1,609,450	522,605
By Socioeconomic Status				
	Low SES	High SES	Low SES	High SES
Jun. 2010	0.04 (0.05)	−0.00 (0.02)	−0.40 (3.59)	−2.78 (3.59)
Dec. 2010	0.02 (0.04)	−0.02 (0.02)	1.61 (3.09)	−1.59 (4.22)
Dec. 2011	0.22*** (0.05)	0.21*** (0.03)	−8.78*** (2.58)	−21.31*** (4.82)
Elasticity			−0.41	−0.30
Dep. Var. Base Period Mean	0.07	0.02	95.12	347.84
N Clusters	303	302	303	302
N Obs.	1,147,411	1,763,322	6,999,869	6,093,856
N Individuals	849,835	1,064,389	2,768,287	2,021,242
N Exposed Individuals	216,450	288,845	1,109,738	1,022,317

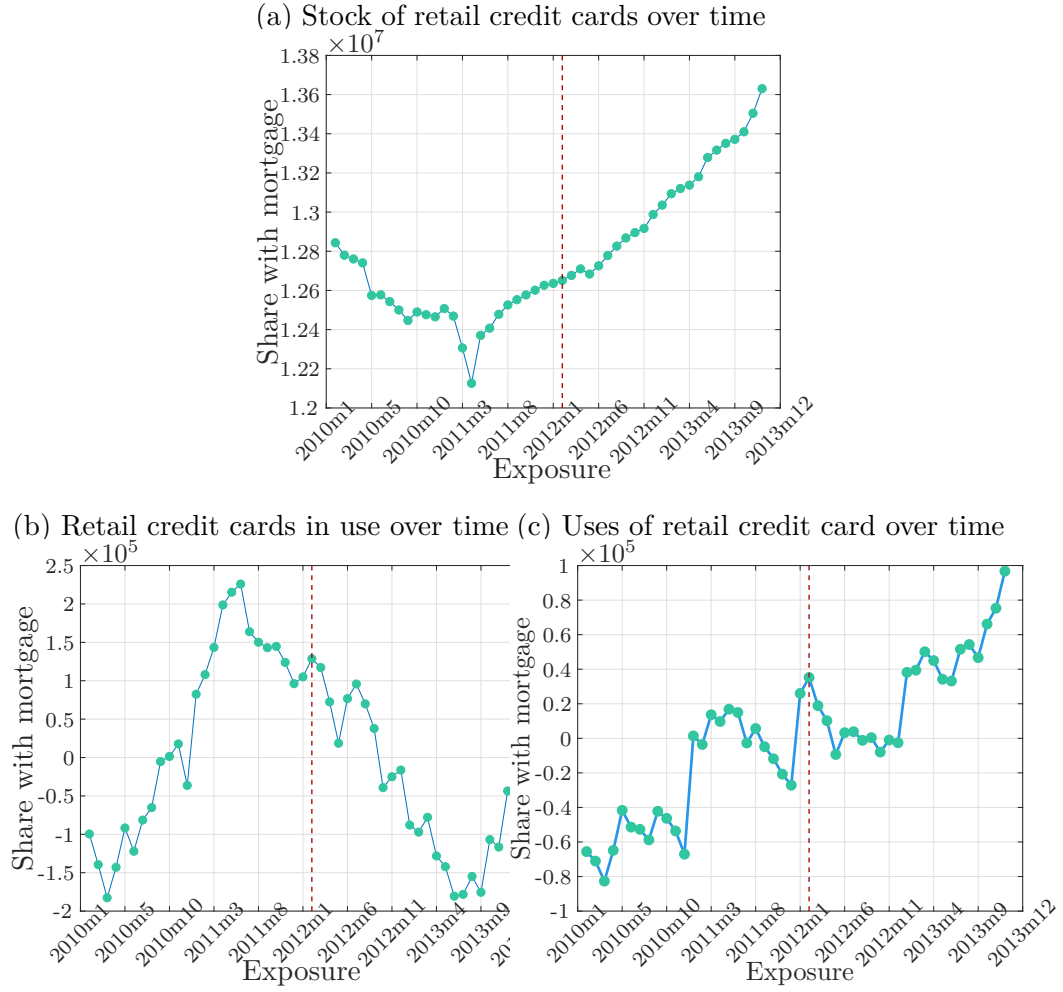
Significance: + 0.10 \* 0.05 \*\* 0.01 \*\*\* 0.001. Difference in difference estimates from equation 2 over defined subsamples. Columns 1 through 4 are predicted default and borrowing diff-in-diff effect estimates in the high exposure market while columns 5 through 8 report estimates in the low exposure market. Column headers report dependent variable at the top and subsample below. Sample in specifications where default is an outcome conditions on positive borrowing (see text for details). ‘Elasticity’ is borrowing effect scaled by base period outcome mean and predicted default effect within each market-subsample. We take the log of ‘Predicted Default’ for estimation but report the base period mean in levels. ‘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 3  
DiD BY EXPOSURE, MORTGAGE, AND SOCIOECONOMIC STATUS

Negative Exposure				
	Predicted Defaults		New Borrowing	
By Mortgage Status				
	No Mortgage	Mortgage	No Mortgage	Mortgage
Jun. 2010	0.11 (0.08)	−0.10 (0.06)	−6.08* (2.56)	−0.78 (4.33)
Dec. 2010	0.07 (0.08)	−0.09+ (0.05)	0.46 (2.81)	5.59 (4.16)
Dec. 2011	−0.27*** (0.07)	−0.42*** (0.05)	15.73*** (4.06)	19.78*** (5.11)
Elasticity			−0.46	−0.23
Dep. Var. Base Period Mean	0.1	0.09	127.19	204.06
N Clusters	278	266	281	272
N Obs.	1,028,499	244,872	6,135,611	1,358,357
N Individuals	800,061	193,751	2,649,628	606,131
N Exposed Individuals	70,162	14,584	497,783	110,446
By Socioeconomic Status				
	Low SES	High SES	Low SES	High SES
Jun. 2010	0.12 (0.09)	−0.04 (0.05)	−1.32 (2.89)	−6.32 (4.15)
Dec. 2010	0.08 (0.08)	−0.05 (0.04)	−1.03 (2.55)	6.47 (4.53)
Dec. 2011	−0.30*** (0.07)	−0.32*** (0.05)	9.27** (3.05)	18.78*** (5.47)
Elasticity			−0.41	−0.24
Dep. Var. Base Period Mean	0.16	0.05	75.44	243.48
N Clusters	274	282	279	285
N Obs.	555,634	717,737	4,617,114	2,876,854
N Individuals	471,664	532,229	2,021,269	1,378,643
N Exposed Individuals	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 default and borrowing diff-in-diff effect estimates in the high exposure market while columns 5 through 8 report estimates in the low exposure market. Column headers report dependent variable at the top and subsample below. Sample in specifications where default is an outcome conditions on positive borrowing (see text for details). ‘Elasticity’ is borrowing effect scaled by base period outcome mean and predicted default effect within each market-subsample. We take the log of ‘Predicted Default’ for estimation but report the base period mean in levels. ‘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.

Figure 3: Retail credit cards over time



Panel (a) Stock of retail credit cards by month. Time of deletion policy noted with vertical line; Panel (b) Number of retail credit cards used by month. Time of deletion policy noted with vertical line. Source: SBIF; Panel (c) Amount of retail credit purchases by month. Time deletion policy noted with vertical line. Source: SBIF.

TABLE 4  
LOG LIKELIHOODS OF VARIOUS ALGORITHMS

	Pre-period		Contemporaneous	
	Training	Testing	Training	Testing
Naive Bayes				
<i>With registry info</i>	−0.412	−0.682	−0.398	−0.633
<i>Without registry info</i>	−0.324	−0.516	−0.300	−0.458
Logistic LASSO				
<i>With registry info</i>	−0.176	−0.324	−0.176	−0.335
<i>Without registry info</i>	−0.180	−0.337	−0.182	−0.348
Random Forest				
<i>With registry info</i>	−0.176	−0.278	−0.173	−0.295
<i>Without registry info</i>	−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 (pre-period) and within each snapshot (contemporaneous). ‘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 registry 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 A for details on the transformation of these predictors and the structure of each algorithm.

TABLE 5  
DIFFERENCE IN DIFFERENCE ESTIMATES ON REALIZED DEFAULT

Positive Exposure			
	Pooled	No Mortgage	Have Mortgage
Jun. 2010	0.02 (0.03)	0.03 (0.04)	-0.05* (0.02)
Dec. 2010	0.00 (0.03)	0.01 (0.04)	-0.05+ (0.03)
Dec. 2011	0.02 (0.03)	0.01 (0.04)	0.03 (0.03)
Elasticity	0.10	0.05	0.12
Dep. Var. Base Period			
Mean	0.04	0.05	0.03
N Clusters	303	303	292
N Obs	4,930,411	3,734,294	1,196,117
N Individuals	2,385,366	1,894,374	558,811
N Exposed Individuals	855,928	636,066	219,862
Negative Exposure			
	Pooled	No Mortgage	Have Mortgage
Jun. 2010	0.07 (0.08)	0.11 (0.08)	-0.10 (0.06)
Dec. 2010	0.04 (0.07)	0.06 (0.08)	-0.09+ (0.05)
Dec. 2011	-0.04 (0.06)	-0.03 (0.07)	-0.08 (0.05)
Elasticity	0.12	0.11	0.18
Dep. Var. Mean	0.1	0.1	0.09
N Clusters	284	281	268
N Obs	2,156,891	1,742,719	414,172
N Individuals	1,433,629	1,167,470	284,204
N Exposed Individuals	143,165	118,441	24,724

Significance: + 0.10 \* 0.05 \*\* 0.01 \*\*\* 0.001. Difference and difference estimates from equation 2 where the dependent variable is realized default '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 'Realized default' when estimating the regressions but report the base period mean in levels. 'Elasticity' is the realized default effect scaled by the predicted default 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 REALIZED DEFAULT

Positive Exposure		
	Low SES	High SES
Jun. 2010	0.04 (0.05)	-0.00 (0.02)
Dec. 2010	0.01 (0.04)	-0.02 (0.02)
Dec. 2011	0.01 (0.04)	0.03 (0.03)
Elasticity	0.06	0.16
Dep. Var. Base Period		
Mean	0.07	0.02
N Clusters	303	302
N Obs	1,943,879	2,986,532
N Individuals	1,201,766	1,347,265
N Exposed Individuals	366,368	489,560
Negative Exposure		
	Low SES	High SES
Jun. 2010	0.12 (0.09)	-0.04 (0.05)
Dec. 2010	0.07 (0.08)	-0.05 (0.04)
Dec. 2011	-0.06 (0.07)	-0.02 (0.05)
Elasticity	0.20	0.05
Dep. Var. Mean	0.16	0.05
N Clusters	278	283
N Obs	941,603	1,215,288
N Individuals	721,292	755,356
N Exposed Individuals	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 realized default '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 'Realized default' when estimating the regressions but report the base period mean in levels. 'Elasticity' is the realized default effect scaled by the predicted default 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  
DIFFERENCE-IN-DIFFERENCE PREDICTIONS USING LONG RUN DEFAULT MEASURES

	<i>Positive exposure</i>			<i>Negative exposure</i>		
	Predicted Default	Average Cost	New Borrowing	Predicted Default	Average Cost	New Borrowing
Jun. 2010	0.01 (0.02)	0.00 (0.02)	−7.09* (3.05)	0.03 (0.05)	0.03 (0.05)	−5.68 <sup>+</sup> (3.23)
Dec. 2010	0.01 (0.02)	0.01 (0.02)	−2.11 (3.52)	0.02 (0.05)	0.01 (0.05)	0.30 (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*** (0.02)	0.12*** (0.02)	−13.28** (4.21)	−0.30*** (0.04)	0.04 (0.04)	17.98*** (3.47)
Elasticity		0.48	−0.24		−0.12	−0.36
Dep.Var. Mean	0.08	0.08	214.70	0.14	0.14	165.09
<i>N</i> Clusters	307	307	307	299	299	300
<i>N</i> Obs.	2,929,133	4,961,674	13,163,613	1,486,567	2,519,339	8,117,207
<i>N</i> Individuals	1,844,615	2,394,399	4,373,700	1,104,246	1,571,258	3,422,263
<i>N</i> Exposed Ind.	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 2. Table is identical to Table II but uses a one-year ahead measure of default to compute predicted default rates. See section 4 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 8  
DISTRIBUTION OF DELETION EFFECTS USING LONG RUN DEFAULT MEASURES

	Separate	Pooled	Difference
<i>Positive exposure</i>			
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%
<i>N</i> individuals	1,905,946	1,905,946	1,905,946
<i>Negative exposure</i>			
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%
<i>N</i> individuals	592,732	592,732	592,732
<i>Combined</i>			
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%
<i>N</i> individuals	2,498,678	2,498,678	2,498,678

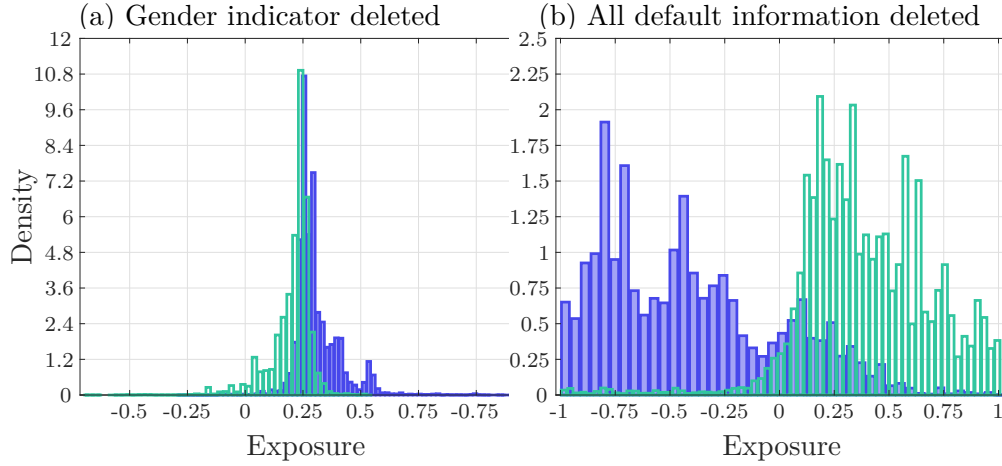
This table describes changes in key metrics before and following deletion, with inputs to the theoretical framework using the long-run cost measure, assuming a 0% markup.

TABLE 9  
SURPLUS CHANGES BY MARKUP

		<i>Additional high cost market markup (%)</i>					
		0	5	10	25	50	100
<i>Low cost market markup (%)</i>	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%

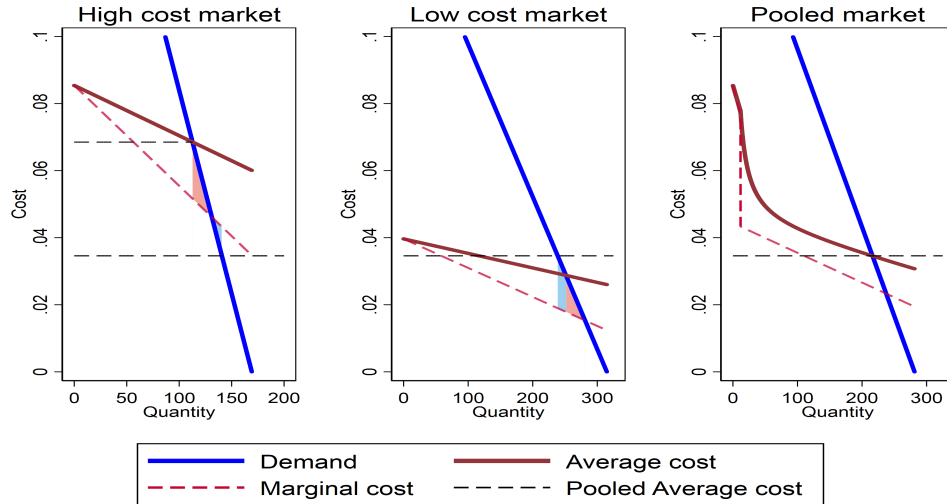
This table describes changes in changes in surplus loss before and following deletion. Cells are additional markups (columns, in percent terms) relative to a given markup rate in the low cost market (rows). Within each cell, rows are level changes in surplus loss in the low cost, high cost, mean change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

Figure 4: Distribution of exposure under counterfactual deletion policies



Histograms of exposure under counterfactual deletion policies. On top: log difference in predicted defaults ('exposure') 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 amount. See text for details. Predictions trained within each month.

Figure 5: Empirical estimates of different markets



Empirical estimate of figure 7 using difference-in-difference estimates of slopes, assuming average cost pricing in both markets. See Section 4 for details.

TABLE 10  
DISTRIBUTION OF DELETION EFFECTS

	Separate	Pooled	Difference
<i>Positive exposure</i>			
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
Surplus loss (1000s CLP)	0.161	0.331	0.170
Aggregate new borrowing (Bns CLP)	516	490	−26
Aggregate surplus loss (1000s CLP)	330,480	679,717	349,238 105.68%
<i>N</i> individuals	2,051,138	2,051,138	2,051,138
<i>Negative exposure</i>			
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
Surplus loss (1000s CLP)	0.156	0.041	−0.114
Aggregate new borrowing (Bns CLP)	69	86	17
Aggregate surplus loss (1000s CLP)	95,456	25,307	−70,149 −73.49%
<i>N</i> individuals	612,737	612,737	612,737
<i>Combined</i>			
Average cost	0.033	0.035	0.001
New borrowing (1000s CLP)	219.624	216.168	−3.455
Surplus loss (1000s CLP)	0.160	0.265	0.105 65.52%
Aggregate new borrowing (Bns CLP)	585	576	−9
Aggregate surplus loss (1000s CLP)	425,936	705,025	279,089 65.52%
<i>N</i> individuals	2,663,875	2,663,875	2,663,875

This table describes changes in key metrics before and following deletion. Prices and surplus calculations assume average cost pricing. See text for details. ‘Positive exposure’ panel is individuals whose predicted defaults rise following deletion; ‘Negative exposure’ is individuals whose predicted defaults fall. ‘Combined’ panel averages over both markets for prices, average cost, new borrowing, and surplus measures, while summing for aggregate borrowing/surplus measures. ‘New borrowing’ in 1000s of CLP. Aggregate new borrowing is in billions of CLP.

TABLE 11  
EFFECTS OF COUNTERFACTUAL EXPOSURE POLICIES

	Exposed to pred. default increases	Zero group	Exposed to pred. default decreases	Pooled
<i>Gender deleted</i>				
Exposure 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
<i>N</i>	171,878	4,111,244	166,565	4,721,885
<i>All info. deleted</i>				
Exposure 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
<i>N</i>	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).

## A. ONLINE SUPPLEMENT A - ADDITIONAL PREDICTION MODEL DETAILS

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 A1 and A2 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  $\Theta_k$  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<sup>27</sup> 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 inferred 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

---

<sup>27</sup>We estimate CART-style regression trees that split using variance reduction (Breiman, Friedman, Stone, and Olshen, 1984).



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, dummifying out zero values using indicator variables. We perform a logistic regression with a  $\lambda$  penalty term of the sum absolute value of the coefficients and use three-fold cross validation to pick  $\lambda$  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 A3 for cross-validation output.

Figure A1: Cross-validation output for  $PD^{pre}$  random forest predictions

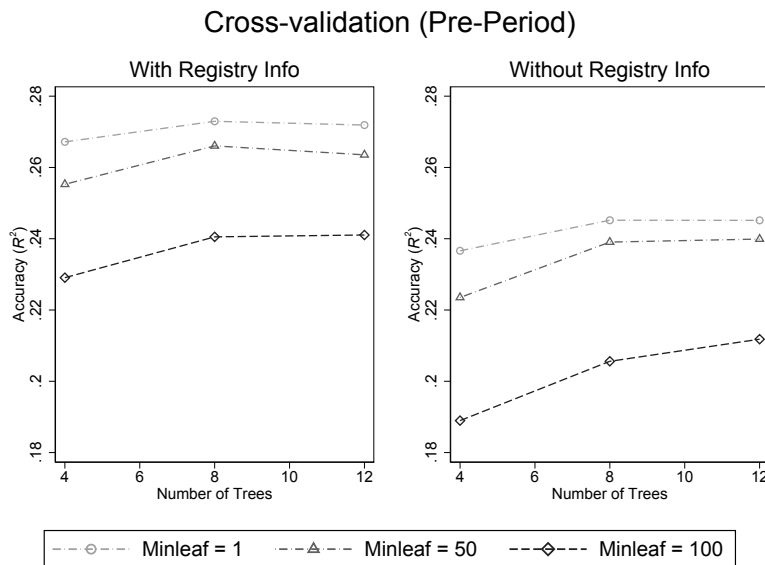
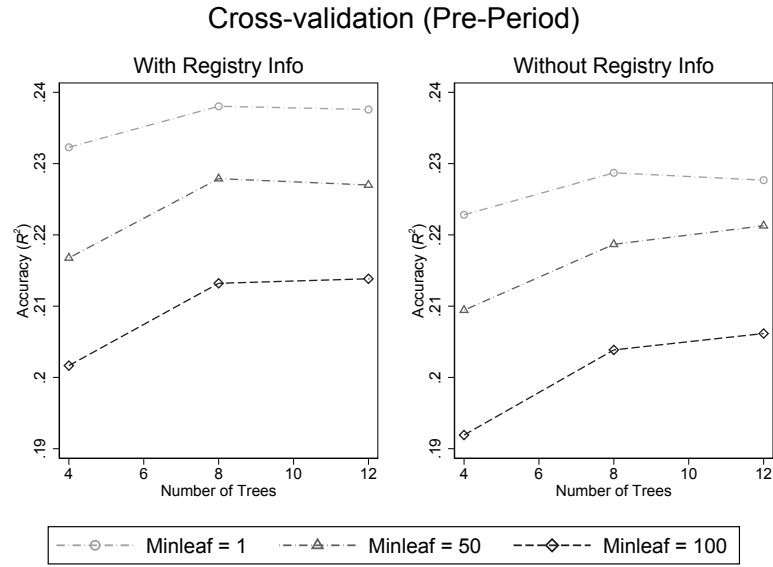
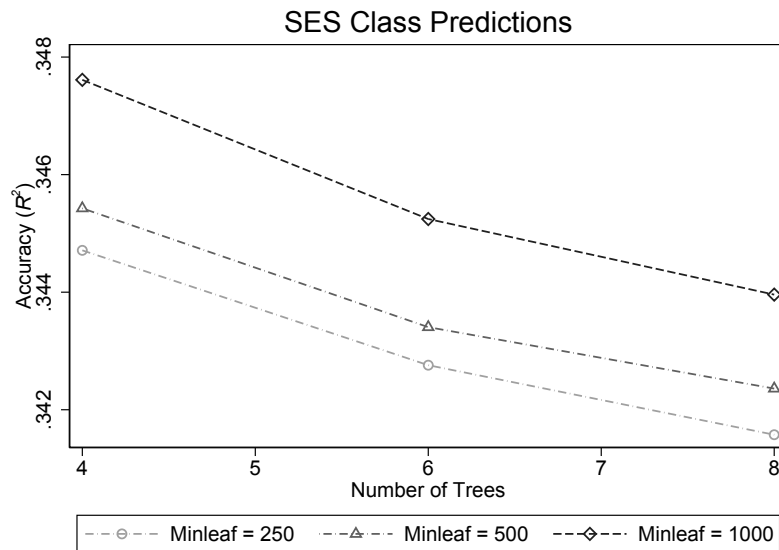


Figure A2: Cross-validation output for  $PD^{post}$  random forest predictionsFigure A3: Cross-validation output for  $PD^{post}$  logistic LASSO predictions

## B. ONLINE SUPPLEMENT B - ADDITIONAL MODEL DETAILS

### B.1. *Model setup*

This Appendix presents the details of main text Section 5. Model setup is as in the text. Let there be a unit measure of borrowers in the market, of whom a fraction  $\alpha$  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.

#### B.1.1. *Pre-deletion equilibria*

When lenders observe  $Z_i$ , equilibria 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 7, respectively),

with total surplus loss in each market given by the formula:

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

### B.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:

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

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 7, 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 underprovi-

sion in the segregated market. Figure 7 in the main text illustrates the latter case, with surplus losses from overprovision equal to the area of triangle C in the left panel.

### B.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  $\frac{dMC_z}{dq} = 2 \frac{dAC_z}{dq}$ .

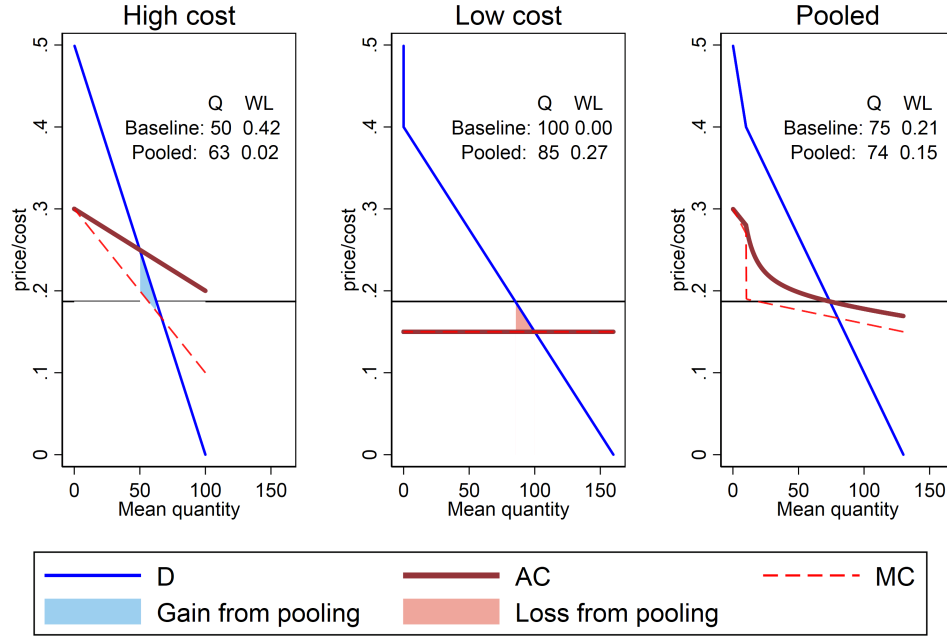
In Appendix Figure B1, we simulate equilibrium outcomes from pooling a low-cost and high-cost market under different assumptions of the slopes of the demand and cost curves in each market. The figure illustrates how the effects of pooling on aggregate borrowing and total surplus are ambiguous, even in this simple model, and how they relate to the slopes of demand and costs.

#### B.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 B1: Simulated separating and pooling equilibria

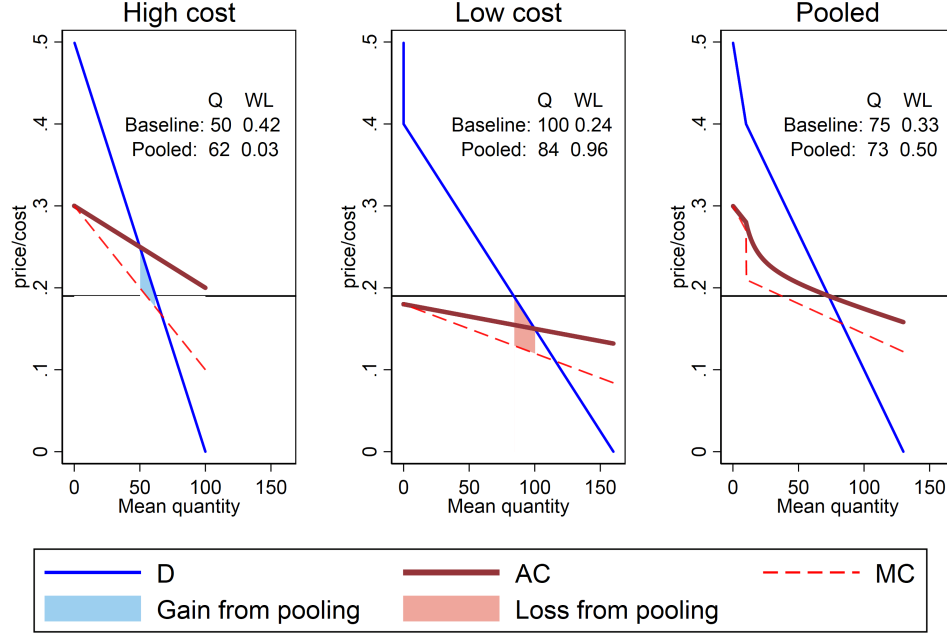
## A. No adverse selection in low-cost market



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” equilibrium. 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  $\frac{dAC_1}{dq}$  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 B2: (Cont'd) Simulated separating and pooling equilibria

## B. Moderate adverse selection in low-cost market

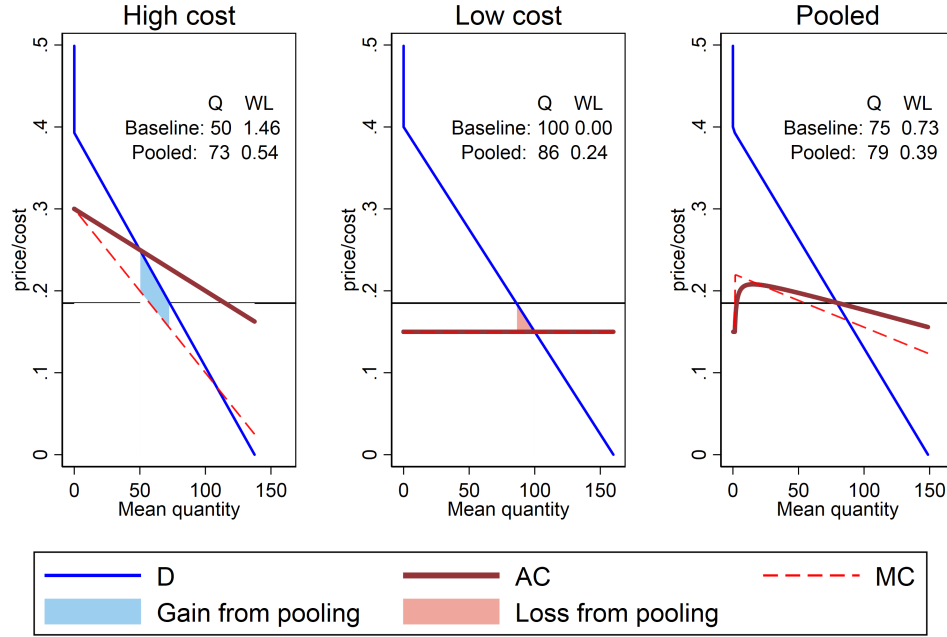


Each panel shows simulated separate-market (left two panels) and pooled market (right panel) equilibria under different assumptions about market sizes and slopes of average cost and demand curves in the high-cost and low-cost market. Text in each panel displays aggregate market quantities transacted (“Q” column) and welfare loss relative to the efficient quantity (“WL” column) under the separate (“baseline”) equilibrium and the “pooled” equilibrium. 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  $\frac{dAC_1}{dq}$  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 B3: (Cont'd) Simulated separating and pooling equilibria

## C. No adverse selection in low-cost market, less elastic demand



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” equilibrium. 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  $\frac{dAC_1}{dq}$  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.