Gathering Information: Forecasting Recoveries in Debt Collection[*]

Johannes Kriebel[¶]
University of Muenster

Kevin Yam[§]
Seghorn AG
University of Marburg

January 15, 2018

Abstract

Forecasting third-party debt collection recoveries is characterized by low levels of debtor information and an absence of collateral. We identify drivers behind these collection recoveries on a unique proprietary data set of more than 400,000 collection claims. (1) Our analysis of information, initially disclosed to the collection agency, shows that large exposures and old accounts recover less. Available contact information positively relates to recoveries. However, these characteristics have a limited information content. (2) Pieces of acquired information gathered by the collection agency (e.g. credit bureau scores, regional scores and collections on other accounts) substantially increase the prediction accuracy. (3) Considering information from the pre-collection process, we find the handover policy to affect the collection success and the quality of contact information to be a signal of undisclosed debtor quality besides its practical value in use. – While initial debtor information is limited, the quality of predictions largely increases by information gathered in the third-party collection process.

Keywords: LGD, Trade Credit, Debt Collection, Collection Rate.
JEL Classification: G20, G21, G23, G30

[*] We are deeply indebted to Seghorn AG for providing the data set and a wide range of advice on the debt collection industry that enabled this study. We are further grateful to the participants of the HypoVereinsbank-Seminar 2017 in Bochum, the participants of the GOR Financial Management and Financial Institutions Workshop 2017 in Magdeburg, the participants of the European Conference on Data Analysis 2017 in Wroclaw, Christopher Hudyma, Henning Cordes and the members of the Finance Center Muenster for valuable comments that helped improving this study.

[¶] Johannes Kriebel, Finance Center Muenster, University of Muenster, Universitaetsstr. 14-16, 48143 Muenster, Germany, phone +49-251-83-22692, johannes.kriebel@wiwi.uni-muenster.de.

[§] Kevin Yam, Seghorn AG, Legienstrae 1, 28188 Bremen, Germany, yam@seghorn.de.

1 Introduction

Trade credit and the management of accounts receivable have a vital role for the balance sheets of many non-financial companies (Walter et al., 2017). It is common practice in many industries (e.g. insurance, telecommunication and mail-order services) to commission specialized collection agencies to collect distressed receivables. Likewise, banks tend to resort to collection agencies in difficult cases (Thomas et al., 2012). According to industry studies, collection firms managed a total of €60bn of receivables in Germany at the end of 2015 (Buelow, 2016) and over $750bn in the United States at the end of 2013 (Ernst & Young, 2014).

Surprisingly, only a few studies (e.g. Walter et al., 2017 and Thomas et al., 2012) have investigated how collection agencies manage accounts successfully and predict collection rates. This represents a decisive research gap as the business model of third-party debt collection has several particularities. Three key features of the industry should be highlighted to the reader: (1) The debtor information is generally sparse (especially compared to credit application data due to leaner acceptance processes in the case of trade credit and the data transmission from the creditor to the collection agency).[1] In many cases, information such as the employment status, the residential status and the regular income among others are not available. There is further usually no collateral in third-party debt collection. (2) As a collection agency is commissioned at a later stage[2], the repayment behavior depends on the previous in-house collection process and can further be argued to also depend on characteristics of the customer relationship before becoming distressed. (3) In contrast to the original creditor that will usually only attempt to recover distressed receivables as a sideline, establishing processes to work out these claims efficiently is at the center of a collection agency's business model.

We aim at analyzing what characteristics a debt collection agency can make use of to make successful predictions on collection recoveries. We particularly address the question what additional information besides the mentioned sparse transferred data can be sourced to improve predictions. These results are of practical use for collection agencies and original creditors for assessing the risk of trade credit, bank loans and consumer finance. Our study makes use of a unique proprietary data set of more than 400,000 distressed insurance receivables of individual and corporate debtors from a German collection agency. Our research approach is two-fold: In a first step, we present an assessment of the account characteristics transferred by the original creditor and characteristics gathered by the collection agency. We contrast these results to findings from the existing loss given

[1] We use the term customer or debtor when referring to the individual or company that is overdue on one or more claims. The initial holder of the claim is referred to as original creditor. One customer or debtor could be linked to one or more (ongoing or past) accounts.

[2] The first stage is usually the in-house collection.

default/recovery rate (LGD/RR) and debt collection literature and evaluate their variable importance based on their contribution to the models' adjusted R^2. In a second step, we look deeper into some of the predictive characteristics that allow a more fine-grained understanding of the in-house collection process and a more detailed understanding of additional undisclosed information available to the original creditor but not externally.

In the first step of the analysis, we find the collection rate to be negatively related to the exposure size and the age of the account. Accounts with telephone contact details display higher collection rates. Considering additional information gathered by the collection agency, good credit bureau scores (SCHUFA[3]), recoveries on other accounts and good areas of residence have higher collection rates. We assess the influence of the predictive characteristics on the prediction accuracy by calculating the change in the adjusted R^2 for including the gathered characteristics. The adjusted R^2 is considerably increased from between 10.7 and 14.3% to between 15.7 and 43.1%. Accordingly, the credit bureau score and the collection success on other accounts constitute major drivers of the predictions. The quality of the residence area can be shown to relate mainly to regional economic conditions.

In the second part of the analysis, we assess the influence of the exposure size and the telephone contact details more closely, as these characteristics allow to draw conclusions of how the transfer policy and undisclosed information influence the collection rate. Original creditors tend to keep larger exposures longer in in-house collection compared to smaller exposures. This aligns with the notion that original creditors tend to work out the more profitable collection cases in in-house collection and tend to hand over the less profitable ones. In addition, we even find evidence of a preselection in a way that large exposures with a particularly low credit assessment are handed over. The telephone details are hypothesized to convey information on the quality of the debtors in the in-house collection. Debtors with missing telephone details are generally customers of lower quality given the credit bureau score levels. We test the alternative hypothesis that telephone details help to accelerate the collection process by enabling an easier contact to the debtor and therefore receive payments earlier. Surprisingly, we do not find evidence for this hypothesis.

Our findings contribute to at least three important strands of literature. First, while there are discussions on dealing with the recovery risk of bonds and bank loans for quite a while[4], this discussion rarely covers recovery risk in trade credit (refer to Walter et al., 2017). Thus, there is little knowledge of the drivers of trade credit recoveries. In addressing this research gap, we provide fruitful evidence on what debtor information is useful in assessing the trade credit recovery risk. In a more general reference to the trade credit literature, we provide evidence that creditors tend to transfer smaller or more difficult

[3] The Schufa Holding AG is a large supplier of individual and corporate credit assessments in Germany.

[4] Refer to Section 2 for a brief overview of relevant results from the LGD/RR literature.

accounts to a third party debt collection agency. This is in line with the notion that collection agencies are commissioned due to economics of scale and specialization gains (Mian and Smith, 1992 and Mian and Smith, 1994).

Second and to the best of our knowledge, we are one of the first papers to investigate the management and risk assessment of accounts in third-party debt collection. Our work identifies drivers of collection recoveries, which is relevant to collection agencies and beyond that to original creditors for assessing and communicating the inherent risk of their portfolios. Moreover, the perspectives on gathered information and the decisions and processes in the earlier in-house collection add a crucial new perspective that is missing in the collection rate literature and is not covered in the bank loan and consumer finance LGD/RR literature due to differences in business models.

Third, we identify the credit bureau score calculated by SCHUFA to be an extremely strong predictor of recoveries on the German market. This is likely to extend to recoveries in bank loans and consumer finance. We further find regional economic differences to influence collection rates. This provides evidence to the discussion, whether the LGD/RR is influenced by changes in the economic conditions (e.g. Bellotti and Crook, 2012, Caselli et al., 2008, Calabrese, 2014 and Leow et al., 2014).

The remaining part of the paper proceeds as follows: In the second section, we present a short introduction to the available literature on trade credit, collection agencies and some relevant findings from the LGD/RR literature. The third section introduces our data set and descriptive statistics. The fourth section presents the research design, and the regression results from the first step of our analysis. The fifth section assesses some more detailed aspects of the transfer policy and undisclosed debtor information. The sixth section comprises various checks of robustness. The seventh section concludes.

2 Literature Review: Trade Credit, Debt Collection Agencies and the LGD/RR

Trade credit

Trade credit is known to be an important source of funding for many non-financial companies (e.g. Mian and Smith, 1992, Ng et al., 1999 and Petersen and Rajan, 1997). There is an extensive literature on the existence and the management of trade credit (refer to Schwartz, 1974, Ferris, 1981, Biais and Gollier, 1997, Aktas et al., 2011, Burkart and Ellingsen, 2004, Cuñat, 2007, Boissay and Gropp, 2013, Mian and Smith, 1994, Long et al., 1993, Deloof and Jegers, 1996, Ng et al., 1999 or Brennan et al., 1988). One central question is why there is trade credit overall. This relates to questions on what incentivizes suppliers to provide trade credit, why customers choose trade credit as a source of funding and why trade credit can be competitive against professional lenders. The role of debt collection as a part of the management of trade credit is discussed in some places in the literature. Mian and Smith (1994) list it as one of five management functions in trade credit.[5] These functions could be performed either in-house or by a service contractor. Reasons to make use of third-party contractors, among others, include economics of scale for standardized goods and specialization gains at the professional debt collector (e.g. local knowledge of the legal system, Mian and Smith, 1992, Mian and Smith, 1994). In contrast, payments for more specialized goods are in the same work hypothesized to be rather collected in-house as the supplier is more skillful to repossess and resell the good.

Debt collection agencies

The literature on the functioning of third-party debt collection and the management of accounts in collection agencies is much more sparse compared to the trade credit literature. From a macroeconomic perspective, Fedaseyeu (2015) and Fonseca et al. (2017) study the relationship between the availability of collection services and the supply of credit. Their findings suggest that stricter state rules for debt collectors in the United States negatively affect the credit supply. Fonseca et al. (2017) further find that this effect is particularly strong for debtors with weak credit ratings.

Fedaseyeu and Hunt (2015) draw from the intuition that collection agencies can sometimes use harsher collection actions than banks can, as they are more dependent on their reputation in order to attract customers. In their model, the introduction of a collection agency can improve customer welfare in an economy with a high level of debtors that are not willing to pay without harsh collection methods. Their implications are more impaired in an economy with a low level of opportunism.

The number of publications that deal with the management of debt collection accounts is equally small. Walter et al. (2017) use data from a German collection agency including 78 different suppliers from three different industries. They find a strong bi-modal distribution of collection recoveries with a mean at about 65%, the exposure at default and prior collection rates to be positively related to the collection success and the age of the account as well as prior experience with the debtor to be negatively related to the collection success.

Thomas et al. (2012) study differences in the characteristics of in-house and third-party collection using a data set of 11,000 consumer loans from a UK financial institution and 70,000 loans in third-party collection. The extent of information is lower in the second case (e.g. debtor income, credit scores, payment history). There were only small recoveries in third-party debt collection as opposed to the financial institution. Third-party collection accounts were older and in some cases, the debtor moved or hid intentionally. In in-house

[5] The others being credit risk assessment, credit-granting decision, trade credit financing, and bearing the default risk.

collection, their empirical model predicts higher recovery rates for lower loan amounts, higher loan lifetimes, higher application scores, more time in arrears over the previous 12 months and loans longer in arrears. In third-party collection, their model predicts higher payment probabilities for available telephone contact details and particularly low exposure amounts (less than £100).

Hoechstoetter et al. (2012) study a very large data set of almost 10 million receivables from a German collection agency. This data set is particularly interesting, as it comprises collection claims from nine different industries (e.g. bank loans, telecommunication and mail order services). The mean collection rate differed considerably over the nine industries. In differentiating full-payers from non-payers, not having an address and higher exposure sizes decreased the payment probability consistently over all nine samples. A credit bureau score was likewise consistently informative. The results for the out-of-sample accuracy of prediction models differed over the nine different samples.

Hoyer (2011) uses a data set of about 30,000 claims from a German collection agency to build a rating system via survival analysis for collection agency claims. The author finds several variables including the exposure size, the gender, a dummy for corporate customers, the industry of the mandate firm, several spatial variables and several spatial macroeconomic variables to be significantly related to recoveries.

Loss Given Default and Recovery Rates

While there have been studies on the LGD/RR of defaulted bonds for a longer time, it was only after the advent of Basel II and Basel III that a more extensive interest in predicting the LGD/RR of bank loans and consumer finance developed.

Calabrese and Zenga (2010), Loterman et al. (2012), Zhang and Thomas (2012), Tong et al. (2013), Gürtler and Hibbeln (2013) and Bijak and Thomas (2015) discuss the choice of prediction models. Bellotti and Crook (2012), Caselli et al. (2008), Calabrese (2014) and Leow et al. (2014) discuss the relation of the LGD to the economic cycle. Qi and Yang (2009), Leow and Mues (2012) and Ingermann et al. (2016) discuss the influence of collateral on the LGD. Matuszyk et al. (2010), De Almeida Filho et al. (2010) and Makuch et al. (1992) examine how workout processes can be designed and operated to improve the recovery success. Han and Jang (2013) examine how workout processes information can improve recovery rate predictions. Davydenko and Franks (2008) find that differences in bankruptcy laws might cause differences in the recovery rates in a cross-country study.

As the class of distressed receivables in third-party collection might incorporate bank loans but is not limited to this specific group, results on recoveries of defaulted bank loans are not necessarily fully transferable. However, results from the LGD/RR literature are likely to be relevant for collection claims in terms of predictive characteristics and prediction models.

3 Data Description

Our data set is provided by Seghorn AG, a major German collection agency. Seghorn AG mainly offers collection services in the insurance, banking and mail order industry. The data set contains claims that resulted from insurance contracts. The claims were transferred on a commission basis. The collection agency receives a compensation for the collection service, while the debtor payments and the claim itself remain in the possession of the initial holder. The collection agency guarantees a minimum average collection rate to the holder of the claim.

The data set contains three subsamples, A, B and C. Sample A and sample B resulted from the same insurance product. Sample C resulted from a different insurance product. Sample A, B and C were each transferred by different insurance companies and were independent before the transfer to the collection agency in a way that claims from the same insurance company cannot appear in more than exactly one sample. Sample A contains more than 250,000, sample B more than 30,000 and sample C more than 200,000 claims. The claims were transferred to the collection agency over the years 2012 to 2016 on a regular basis. All claims are towards debtors located in Germany. The data set contains both individual and corporate debtors. We analyze these groups collectively, as robustness checks find the same driving characteristics for both groups. (Refer to Section 6.)

Our data set contains a comprehensive body of contract, debtor and customer relationship characteristics that is provided by the collection agency. We further collect spatial and macroeconomic information from the following sources: (1) Unemployment rates are taken from the Federal Employment Agency (Bundesagentur fuer Arbeit). (2) Postal codes (Postleitzahl) are matched to counties (Landkreis) and independent cities (kreisfreie Staedte) via a list available at OpenGeoDB. (3) Geographic coordinates of postal code areas are taken from OpenGeoDB as well. (4) The area and population of counties and cities is taken from the federal bureau of statistics (Statistisches Bundesamt).

3.1 Construction of the Dependent Variable

The data set contains information on the monthly collection payments for each individual account starting from the date of transfer to the collection agency. The monthly payments are recorded until 2016 when we received the data.

Following Dermine and de Carvalho (2006), we analyze payments over a standardized repayment horizon. Only accounts with the minimum number of months with available payment information are considered. This is done in order to be consistent over all claims and to be cautious towards time-varying effects.[6] As a standardized payment horizon, we use a period of less than two years. The exact number of months of the repayment horizon and the level of collections are not explicitly stated here on request of the collection agency. The choice of the repayment horizon results from a trade-off between not losing too many recent claims and having a long repayment period. (Our results in Section 4.2 are robust to using varied other time periods spanning up to more than four years. Refer to Section 6). As sample B contains more recent accounts, we use a repayment horizon that is half as long as in sample A and C.[7] The final data set contains 182,880 claims in sample A, 16,623 claims in sample B and 126,615 claims in sample C.[8]

The collection rate is then calculated by dividing the sum of monthly payments over t time periods for an account i by its exposure at the time of transfer $t = 0$ (see Equation 3.1). Workout expenses and collection fees paid by the debtor to the collection agency do not enter the calculation. The initial claim is serviced first. Fees and possible late payment interest imposed by the collection agency are serviced last and are accounted separately.

$$CR_{i,t} = \frac{\sum_{j=1}^{t} Payment_{i,j}}{Initial\ exposure_{i,t=0}} \tag{3.1}$$

The mean collection rate over varying levels of the repayment horizon t is presented in Figure 1.[9] The exact collection rate numbers and repayment horizons are intentionally left blank. It is apparent from the graph that the collection rate is higher for sample A and B that originate from the same product compared to sample C. In all three samples, the major part of payments is received over the first year of the collection process. The repayment horizons used in the calculation of the dependent variable are plotted as dashed vertical lines (right: sample A and C, left: sample B).

The distribution of cumulative collections for each account over the repayment horizons is presented in Table 1 for all three samples. The distribution has strong concentrations at the boundaries. About 95% of the accounts either pay the full exposure or do not pay anything at all. About 5% (A: 5.88%, B: 4.21%, C: 4.37%) of all accounts have a partial payment and lie strictly between zero and one.

[6] More recent accounts will tend to have smaller collection rates on average as there are less payment months. This could bias the coefficients for time-varying explanatory variables such as macroeconomic characteristics. Moreover, using only accounts that are "closed" would introduce a bias as successful accounts are "closed" by definition whereas unsuccessful accounts could remain "open" over several years.

[7] Checks of robustness are as well available in Section 6 and in the Appendix.

[8] Claims amounting to less than €5 and accounts with missing exposure sizes are further excluded. This applies to 594 accounts.

[9] The collection rate is calculated over all accounts that have at least the respective number of payment period information.

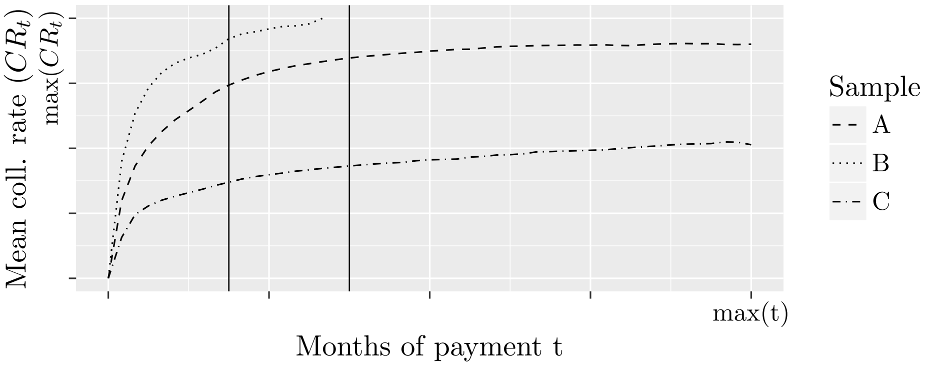

Figure 1: Mean collection rate for varying numbers of payment months

Table 1: Distribution of the collection rate to boundary and non-boundary cases

		Sample		
Interval		**A**	**B**	**C**
Full or no payment	$(CR = 0 \ \& \ CR = 1)$	0.9412	0.9579	0.9563
Partial payment	$(0 < CR < 1)$	0.0588	0.0421	0.0437

3.2 Independent Variables

The sample contains a comprehensive set of independent variables. We categorize the variables in four groups of characteristics: Contract, debtor, relationship and spatial characteristics. Descriptive statistics for the continuous characteristics are presented in Table 2. For the dichotomous variables, relative frequencies of the variable values are presented in Table 3. The relative frequencies of credit bureau score values are presented in Table 4.

Contract Characteristics

The exposure (EXP) is given by the amount that the insurance company initially reports as overdue. For sample A and B (Table 2) the median exposure is at around €100 and more than three quarters of the accounts have an exposure size of less than €200. However, some of the exposures reach up to several ten thousand euros.[10] The exposures in sample C (Table 2) are generally slightly higher with a median exposure of €114.1 and more than three quarters of the accounts lying below €350.

[10] Our results in Section 4.2 are robust toward taking the log of the exposure (Refer to Section 6.).

Table 2: Summary statistics - continuous independent variables

				Sample A				
	N	**Mean**	**St. Dev.**	**Min**	**Pctl(25)**	**Median**	**Pctl(75)**	**Max**
EXP	182,880	200.996	762.768	6.410	61.250	99.310	185.460	>100,000
AGE	172,118	44.927	12.798	15	35	45	53	110
L_ORIG_ACQU	182,698	145.112	93.378	−21	87	120	170	4,456
UNEMPL	182,880	7.226	2.970	1.100	5.000	6.600	9.200	18.500
INHAB_DENS	182,880	607.845	835.483	36.273	128.231	246.900	729.064	3,947.639
CR_C_COUNT_ind	182,880	1,211.821	1,094.363	7	427	838	1,681	4,443
CR_P_COUNT_ind	182,880	81.648	76.030	0	25	57	117	402
C_CASES_PER_CAP	182,880	0.004	0.002	0.0002	0.002	0.003	0.006	0.010
DISTANCE	182,880	242.356	129.270	0.000	143.868	221.938	312.163	674.780
				Sample B				
	N	**Mean**	**St. Dev.**	**Min**	**Pctl(25)**	**Median**	**Pctl(75)**	**Max**
EXP	16,623	173.578	298.542	5.000	57.810	98.580	197.550	>15,000
AGE	6,979	45.309	13.279	18	35	45	54	96
L_ORIG_ACQU	10,501	131.809	62.911	35	105	108	144	1,904
UNEMPL	16,623	6.872	2.912	1.100	4.500	6.500	8.800	16.500
INHAB_DENS	16,623	934.453	1,081.417	36.273	166.722	402.119	1,355.618	3,947.639
CR_C_COUNT_ind	16,623	119.695	186.301	1	31	59	131	859
CR_P_COUNT_ind	16,623	4.622	4.151	0	2	4	7	27
C_CASES_PER_CAP	16,623	0.0002	0.0001	0.00002	0.0002	0.0002	0.0003	0.001
DISTANCE	16,623	306.414	145.171	0.000	206.862	295.732	408.768	674.780
				Sample C				
	N	**Mean**	**St. Dev.**	**Min**	**Pctl(25)**	**Median**	**Pctl(75)**	**Max**
EXP	126,015	352.285	828.646	5.000	39.765	114.100	330.170	>50,000
AGE	116,174	42.376	12.441	15	32	41	51	101
L_ORIG_ACQU	102,449	170.367	212.956	−146	66	106	234	7,879
UNEMPL	126,015	7.333	3.177	1.100	4.700	7.000	9.800	18.500
INHAB_DENS	126,015	966.922	1,132.510	36.273	154.323	402.119	1,525.172	3,947.639
CR_C_COUNT_ind	126,015	1,068.138	2,016.586	45	237	389	820	8,408
CR_P_COUNT_ind	126,015	36.553	31.047	0	15	29	49	248
C_CASES_PER_CAP	126,015	0.002	0.001	0.001	0.001	0.002	0.002	0.007
DISTANCE	126,015	307.505	146.713	0.000	205.750	303.264	408.810	674.780

Debtor Characteristics

The debtor information contains the age (*AGE*, Table 2) of the customer at the time of transfer. The *AGE* ranges from a low of 15 years up to rare cases above 100 years. A small number of customers with an age smaller than 15 were set to missing but kept in the samples (A: 7, B: 0, C: 15). About half of the debtors have an age between 35 and 55 years. The age is generally missing in all cases with corporate debtors but was transferred missing in some other cases as well.

A small number of accounts are linked with debtors that were insolvent before the workout process (*INSOLV_ACQU*) or that became insolvent during the workout process (*INSOLV_PROC*; Table 3). The proportion of insolvent accounts is about 4% in sample

Table 3: **Frequency table - dichotomous variables**

	Sample A	**Sample B**	**Sample C**
	True	**True**	**True**
INSOLV_ACQU	0.019	0.020	0.049
INSOLV_PROC	0.023	0.017	0.035
FIRM	0.046	0.069	0.060
MALE	0.647	0.658	0.642
TEL	0.624	0.419	0.396
END_missing	0.033	1.000	0.849

Table 4: **Frequency table - credit bureau score**

	SCORE (#obs[%])												
	NA	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**K**	**L**	**M**
A	0.729	0.010	0.011	0.012	0.010	0.021	0.019	0.018	0.020	0.029	0.038	0.048	0.035
B	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
C	0.393	0.039	0.032	0.030	0.026	0.044	0.032	0.032	0.034	0.049	0.071	0.102	0.116

A and B and about 8% in sample C. About half of the insolvent accounts became insolvent before the workout process and the remaining accounts became insolvent during the workout process. (Both individual and corporate debtors can be insolvent.)

A small proportion of the debtors are corporate debtors ($FIRM$, Table 3). The proportion of corporate accounts is 4.6% in sample A, 6.9% in sample B and 6% in sample C. Close to 65% of all three samples are accounts of male debtors ($MALE$, Table 3). Debtors are recorded as male, female or corporation. A small number of accounts, where this information is missing, is removed from the data set (A: 13, B: 6, C: 7).

27.1% and 60.7% of debtors in sample A and C is provided with credit bureau scores ($SCORE$, Table 4). Obtaining the credit bureau score is costly and is conducted conditional on internal decision rules which could not be outlined here. In sample B, no credit bureau score was obtained due to internal processes that could not be outlined here as well. The levels of the score range from 'A' which is the best score value to 'M' which is the worst score value. Bad levels of the score are comparably more frequent than good levels.

For more than 60% of the accounts in sample A, there are telephone details (TEL, Table 3) known to the collection agency at the time of data query. In sample B and C, about 40% of the accounts have available telephone contact information. The contact information was conveyed by the insurance companies at the time of transfer in most cases. It is, however, possible that the contact information was obtained later in some cases. For debtors with more than one account, the information could stem from a previous workout process.

Relationship Characteristics

The age of the account (*L_ORIG_ACQU*, Table 2) is calculated as the difference between the beginning of the contract period and the time of transfer to the collection agency. In the case of sample A and B, the insurance product generally requires an insurance fee payment before the lifetime of the insurance contract. Payments are therefore at least in arrears when the payment was not received until the beginning of the insurance period. In sample C, fee payments could generally be acceptable shortly after the beginning of the insurance period. The payments are, however, expected to arrive at least closely after the beginning of the insurance period. A claim from sample C is therefore considered to be in arrears when the payment is not received shortly after the beginning of the insurance term. In sample A and B, more than three quarters of the accounts are transferred to the collection agency within less than half a year after the beginning of the contract. In sample C, more than three quarters of accounts are transferred within eight months. However, some accounts have been transferred after several years. The values of the age of the account can be negative in some rare cases, when the claim was transferred to the collection agency before the beginning of the insured period.

For some of the accounts an end date of the insurance contract is specified by the insurance company. This could result either from being in a particular customer segment or from a termination of the contract by the insurance company or the customer. The variable is only available in sample A and C (*END_missing*, Table 3). In sample A, the vast majority of accounts have a specified end of the contract (3.3% missing). In sample C, an end of the contract is mostly missing (84.9%).

In cases, in which there is more than one account linked to the same debtor, it is possible to calculate the mean collection rate from other accounts (*CR_other*). This variable is calculated on all other accounts from the data set that are linked with the same debtor. The variable is intentionally not included in Table 2 to conceal the specific level of collection.

Spatial Characteristics

The summary statistics for the spatial characteristics are reported in Table 2. The unemployment rate (*UNEMPL*) for counties and cities is added for the time of transfer. It ranges from values as low as 1.1% up to values of more than 18%. The population density (*INHAB_DENS*) is calculated by dividing the population of counties and cities by the respective area. It spans from close to 36 inhabitants per square kilometer up to close to 4,000 inhabitants per square kilometer.

The number of accounts from the same county or city (*CR_C_COUNT_ind*) for each claim is calculated over all accounts not linked with the same debtor. It ranges from a few cases (A: 7, B: 1, C: 45) to several thousand cases (A: 4,456, B: 859, C: 8,408). The numbers are considerably lower for the level of postal code areas ($CR\_C\_COUNT\_ind$). We also calculated the mean collection rate over the accounts of other debtors in a county/city ($CR\_C\_CODE\_ind$) and in a postal code area ($CR\_C\_CODE\_ind$). These are intentionally not included in Table 2.

The accounts per capita in a county/city are calculated by dividing $CR\_C\_COUNT\_ind$ by the respective population. For sample A and C, the values range from two to ten cases per 1,000 inhabitants and one to seven cases per 1,000 inhabitants. For sample B, the values range from 2 to 10 cases per 10,000 inhabitants. The air distance ($DISTANCE$) of the collection agency to a debtor is calculated from the geographic coordinates of the postal code areas extracted from OpenGeoDB. The distance ranges from zero to almost 700 kilometers.

4 Analysis

4.1 Predictive Characteristics

As outlined in Section 1, we aim at identifying the drivers of collection recoveries in third-party debt collection. In Section 3, we distinguish four types of predictive characteristics: Contract, debtor, relationship and spatial characteristics. We make a second distinction here. The characteristics that are available in our data differ in the way they are acquired by the collection agency. We first proceed by assessing the quality and direction of predictive characteristics that are initially disclosed by the original creditor at the time of transfer of the claim. We subsequently add variables that could be acquired by the collection agency externally or over the course of the customer relationship. This is done, primarily, to assess how important these characteristics are for improving the quality of predictions. We contrast our results with findings from the LGD/RR and collection rate literature.

Contract Characteristics

The exposure size (EXP) is a common variable both in the LGD/RR (Bellotti and Crook, 2012; Ingermann et al., 2016; Tong et al., 2013) and in the collection rate literature (Walter et al., 2017; Thomas et al., 2012; Hoechstoetter et al., 2012). The relation to recoveries is, however, ambiguous. Bellotti and Crook (2012), Tong et al. (2013), Thomas et al. (2012) and Hoechstoetter et al. (2012) find a negative relation. Ingermann et al. (2016) find a negative relation for retail customers and a positive relation for corporate customers. Further, Walter et al. (2017) find a positive relation to the collection rate. There are several lines of argument to explain these results. According to Bellotti and Crook (2012) and

Ingermann et al. (2016), higher exposures are generally more difficult to repay and thereby negatively influence recoveries. Walter et al. (2017) support this argument insofar as they assume small exposures to be generally paid fast in order to circumvent complications with the collection agency. They state, however, that this could also be in line with a positive relation, when this effect applies mainly to exposures in in-house collection, leaving the more difficult small exposures for the third-party collector. Walter et al. (2017) further argue that a collector has a higher incentive to invest time and effort in recovering large exposures, resulting in a positive relation between the exposure and the collection rate.

Debtor Characteristics

For the AGE of the debtor there is no formal ex-ante expectation of how the age relates to the collection rate. Bellotti and Crook (2012) expect a higher age of the debtor to relate to a lower credit risk and therefore higher recoveries. Their findings are, however, contrary to this assumption. This is in line with the findings of Thomas et al. (2012) and Zhang and Thomas (2012). Hoechstoetter et al. (2012) find an inconsistent relation over nine different collection samples.

We do not expect a specific relation to the gender dummy ($MALE$). Hoechstoetter et al. (2012) include this variable and find inconsistent results. We still include the variable to control for potential gender specific differences. Ingermann et al. (2016) argue that for commercial customers ($FIRM$) there are broader options for restructuring and amicable agreements while the business remains in operation. They, therefore, assume a positive relation to the recovery rate which is in line with their findings. Hoechstoetter et al. (2012), again, find inconsistent relations over different samples. Walter et al. (2017) find a positive relation as well. They note that this is not in line with their intuition that corporations have lower collection rates due to the absence of personal liabilities. For the debtors that became insolvent during ($INSOLV\_ACQU$) or after the workout process ($INSOLV\_PROC$) we expect considerably smaller or no recoveries.

Including the availability of contact information (TEL) could be argued to be of particular interest in third-party collection where the lack of debtor information is a typical issue. It is usually not included in the LGD/RR literature. Thomas et al. (2012) find that better contact information has a negative relation to losses in third-party debt collection. Their intuition is that better contact information makes the collection process more efficient. In line with this argument, Hoechstoetter et al. (2012) find a lower probability of a full payment for missing addresses in debt collection. From the perspective of a collection agency, one could further argue that more thoroughly maintained customer information might indicate a better and less problematic customer relation with the originator of the claim and therefore signals higher collection recoveries.

We generally expect higher collection rates for better levels of the credit bureau score (*SCORE*). This is in line with Bellotti and Crook (2012) and Thomas et al. (2012). Hoechstoetter et al. (2012) find that while intermediate credit bureau score values do not always have a monotonous influence in regression, high and low levels are consistently informative. Besides the straightforward interpretation that better credit bureau scores indicate a higher payment capability, higher recoveries could result from the collection agency investing more effort in debtors with a high score. We are going to shortly discuss this issue in Section 4.4.

Relationship characteristics

The age of the account is found to be negatively related to the collection rate in Walter et al. (2017). Walter et al. (2017) argue that a long time period in in-house collection indicates that a debtor has proven to be unwilling or unable to repay in the in-house collection process, as potentially more effort has been invested and did not render a sufficient payment. In line with this argument, we expect the age of the account (*L_ORIG_ACQU*) to be negatively related to the collection success.

In our data, we have information on whether the underlying insurance contract had a specified end of the contract indicated (*END_missing*). This could be informative in at least two important ways. First, the variable could capture differences between customer segments. This could therefore result in a positive, negative or no relation to the collection rate. Second, this could also indicate that the insurance company or the customer have ended the contract. It might result in lower collection rates for accounts with a specified contract end. We include the variable in order to account for these effects.

Contacts at other instances between the debtor and the collection agency could generally convey different types of information, as Walter et al. (2017) point out. On the one hand, previous contacts indicate that a debtor was distressed at an earlier time already, which could be argued to be a negative sign. On the other hand, Walter et al. (2017) argue that a collector could gain contact and personal information about the debtor from earlier collection processes. That way, prior contact might lead to a more efficient collection process. Thomas et al. (2012) further argue that a debtor that overcame difficulties in the past may be more likely to overcome future difficulties. Given that the exposures in our data are relatively small and in line with this interpretation, multiple exposures might indicate that a debtor only failed to pay the exposure due to obliviousness rather than serious difficulties. This would result in higher collection rates. We include a dummy for single contacts with the collection agency (*D_NO_MULTIPLE*) and the mean collection rate from other accounts of the same debtor (*CR_other*) in line with Walter et al. (2017).

Spatial characteristics

In order to capture spatial differences in the ability to pay, we include the mean collection rate in a county or city $CR\_C\_CODE\_ind$ as a proxy for regional differences in the capability to pay. Our rationale is that this proxy should capture several possible regional differences such as economic strength as well as differences in the efficiency of the collection process. We assess the driving factors behind possible differences more closely in Section 4.3. Bellotti and Crook (2012) categorize housing areas into council/poor housing, suburban/wealthy, rural and other areas. They find a positive effect for the first category and a negative effect for the latter two categories. Walter et al. (2017) include the unemployment rate and the GDP growth rate on the level of the federal state and find a negative relation to the collection rate for both. This is intuitive for the unemployment rate but less intuitive for the GDP growth. Apart from the regional differences in macroeconomic conditions, there is some evidence that nationwide macroeconomic conditions are predictive for the LGD/RR (Bellotti and Crook, 2012, Caselli et al., 2008).

4.2 Regression Results

We assess the relation of the predictive variables outlined in Section 4.1 using the fractional logit model of Papke and Wooldridge (1996). The fractional logit model is a generalized linear model with a dependent variable within the interval $[0; 1]$. This is a common approach in modeling the LGD/RR (Ingermann et al., 2016, Dermine and de Carvalho, 2006). Walter et al. (2017) further use this approach to model collection rates.

All tables state heteroskedasticity robust standard errors. The coefficients are calculated as the conditional partial effects, given by the change in the outcome of the collection rate for a change of one standard deviation for continuous and one unit for dichotomous variables. The partial effects are calculated for the continuous variables being at their median and the dichotomous variables being zero.

AGE_NA is a dummy for missing age values. All missing values for the AGE are set to the mean. The same is done for the missing values of L_ORIG_ACQU and CR_other (dummy for missing CR_other: $D\_NO\_MULTIPLE$).[11]

The regression results are presented in Table 5. The table has three separate sections containing the results for the three samples A, B and C. There are five columns relating to different sets of information. The first column contains information that is initially available. The second and third column includes information that could be acquired by the collection agency by external sources or by aggregating spatial information. The

[11] As a check of robustness, we estimate models in this section when excluding accounts with missing AGE, L_ORIG_ACQU and CR_other. The results remain qualitatively unchanged. These results are available on request.

fourth and fifth column include information that is gathered over the relationship with the individual debtor.

Initially disclosed information

The exposure size (EXP) has a negative significant coefficient for all three samples. This is in line with the assumption that large exposures are more difficult to repay. For sample A and sample B, an increase of the exposure of one standard deviation results in a two and 4.2 percentage points lower expected collection rate. In sample C, the collection rate decreases by 25.8 percentage points for an increase of the exposure by one standard deviation. The effect of the exposure size is thus stronger in sample C. The effect in sample A and B are rather small given that the majority of exposure sizes is small compared to the magnitude of the standard deviation (refer to Table 2). The relation between the exposure and the collection rate might therefore mainly result from an influence of the larger exposures in these two samples.

There is no specific assumption about the relation of the age of the customer (AGE) to the collection rate. The coefficient is insignificant in sample B and C and only significant in sample A. A debtor that is older by one standard deviation (refer to Table 2) has a collection rate that is lower by about two percentage points.

The dummy for male debtors ($MALE$) is significant in sample B and sample C. The signs are, however, inconsistent. In sample C, there is a positive sign. In sample B, there is a negative sign. In sample C, male debtors have an about 1.5 percentage points higher collection rate. In sample B, the collection rate is lower by 1.7 percentage points. Corporate debtors ($FIRM$) have a significant negative coefficient in sample A and C. Debtors insolvent before the transfer ($INSOLV\_ACQU$) consistently have a far lower collection rate.

The availability of telephone contact information (TEL) has a considerably large positive relation to the collection rate. The effect is at around 25 percentage points in all three samples.

The age of the account (L_ORIG_ACQU) has a significant negative coefficient in all three samples. The coefficient is considerably higher in the first two samples but rather low in sample C. Overall, it seems, that a high age of the account is informative of a worse debtor quality. The coefficient for an unspecified end of the contract ($END\_missing$) is significantly negative for sample A and C, where this information was provided to the collection agency. The coefficient is particularly high in sample A.

The adjusted R^2 of the models built on the initially disclosed set of information is between 10.7 and 14.3%. Low R^2 values are a typical feature of LGD/RR forecasts.[12] Bellotti

[12] The adjusted R^2 is considerably higher in studies including information on collateral (refer to Ingermann et al., 2016 or Qi and Yang, 2009).

Table 5: Regression results - initially disclosed information with a stepwise addition of gathered characteristics

	CR_C_CODE				
	(1)	(2)	(3)	(4)	(5)
			Sample A		
EXP	−0.020*** (0.003)	−0.021*** (0.003)	−0.023*** (0.003)	−0.020*** (0.003)	−0.017*** (0.002)
AGE	0.018*** (0.001)	0.017*** (0.001)	0.007*** (0.001)	0.006*** (0.001)	0.001 (0.001)
AGE_NA	0.020* (0.011)	0.024** (0.011)	−0.090*** (0.009)	−0.090*** (0.009)	−0.035*** (0.007)
MALE	0.003 (0.003)	0.002 (0.003)	0.022*** (0.003)	0.021*** (0.002)	0.010*** (0.002)
FIRM	−0.030** (0.012)	−0.027** (0.012)	−0.011 (0.010)	−0.009 (0.010)	−0.002 (0.008)
INSOLV_ACQU	−0.371*** (0.009)	−0.367*** (0.009)	−0.342*** (0.008)	−0.345*** (0.008)	−0.206*** (0.007)
INSOLV_PROC				−0.489*** (0.008)	−0.325*** (0.007)
TEL	0.252*** (0.003)	0.248*** (0.003)	0.164*** (0.002)	0.159*** (0.002)	0.083*** (0.002)
D_SCORE_A			0.006 (0.013)	0.001 (0.012)	−0.034*** (0.010)
D_SCORE_B			−0.065*** (0.011)	−0.069*** (0.010)	−0.075*** (0.009)
D_SCORE_C			−0.086*** (0.010)	−0.092*** (0.010)	−0.089*** (0.008)
D_SCORE_D			−0.134*** (0.011)	−0.138*** (0.010)	−0.111*** (0.009)
D_SCORE_E			−0.168*** (0.007)	−0.172*** (0.007)	−0.129*** (0.006)
D_SCORE_F			−0.261*** (0.008)	−0.260*** (0.007)	−0.170*** (0.006)
D_SCORE_G			−0.320*** (0.008)	−0.317*** (0.008)	−0.196*** (0.006)
D_SCORE_H			−0.380*** (0.008)	−0.376*** (0.008)	−0.230*** (0.006)
D_SCORE_I			−0.386*** (0.007)	−0.382*** (0.007)	−0.238*** (0.005)
D_SCORE_K			−0.423*** (0.006)	−0.418*** (0.006)	−0.254*** (0.005)
D_SCORE_L			−0.514*** (0.006)	−0.507*** (0.006)	−0.306*** (0.005)
D_SCORE_M			−0.560*** (0.008)	−0.547*** (0.008)	−0.321*** (0.006)
L_ORIG_ACQU	−0.108*** (0.002)	−0.105*** (0.002)	−0.079*** (0.002)	−0.077*** (0.002)	−0.052*** (0.001)
L_ORIG_ACQU_NA	−0.156*** (0.046)	−0.160*** (0.046)	−0.108*** (0.041)	−0.108*** (0.040)	−0.082*** (0.029)
END_missing	−0.332*** (0.008)	−0.323*** (0.008)	−0.251*** (0.008)	−0.246*** (0.007)	−0.130*** (0.006)
D_NO_MULTIPLE					−0.127*** (0.002)
CR_other					0.163*** (0.001)
CR_C_CODE_ind		0.045*** (0.001)	0.035*** (0.001)	0.034*** (0.001)	0.018*** (0.001)
Baseline SCORE	—	NA	NA	NA	NA
R2 adj.	0.128	0.135	0.268	0.29	0.431
Wald Chi2	874.167	932.419	358.762	393.12	1437.846
Wald Chi2 Prob.	0.000	0.000	0.000	0.000	0.000
Observations	182,880	182,880	182,880	182,880	182,880
			Sample B		
EXP	−0.042*** (0.010)	−0.042*** (0.010)	−0.042*** (0.010)	−0.040*** (0.009)	−0.016*** (0.004)
AGE	−0.002 (0.004)	−0.002 (0.004)	−0.002 (0.004)	−0.003 (0.004)	−0.001 (0.002)
AGE_NA	0.067*** (0.011)	0.067*** (0.011)	0.067*** (0.011)	0.047*** (0.011)	0.022*** (0.005)
MALE	−0.017* (0.009)	−0.017* (0.009)	−0.017* (0.009)	−0.016* (0.009)	−0.007* (0.004)
FIRM	0.008 (0.021)	0.009 (0.021)	0.009 (0.021)	0.032 (0.021)	0.015 (0.010)
INSOLV_ACQU	−0.357*** (0.027)	−0.357*** (0.027)	−0.357*** (0.027)	−0.357*** (0.026)	−0.154*** (0.012)
INSOLV_PROC				−0.561*** (0.036)	−0.256*** (0.018)
TEL	0.259*** (0.011)	0.258*** (0.011)	0.258*** (0.011)	0.240*** (0.011)	0.106*** (0.005)
L_ORIG_ACQU	−0.084*** (0.010)	−0.083*** (0.009)	−0.083*** (0.009)	−0.080*** (0.009)	−0.034*** (0.004)
L_ORIG_ACQU_NA	−0.140*** (0.008)	−0.140*** (0.008)	−0.140*** (0.008)	−0.134*** (0.008)	−0.063*** (0.004)
D_NO_MULTIPLE					−0.135*** (0.009)
CR_other					0.039*** (0.002)
CR_C_CODE_ind		0.008** (0.004)	0.008** (0.004)	0.008** (0.004)	0.004** (0.002)
R2 adj.	0.107	0.108	0.108	0.129	0.157
Wald Chi2	39.926	39.611	39.611	62.641	59.62
Wald Chi2 Prob.	0.000	0.000	0.000	0.000	0.000
Observations	16,623	16,623	16,623	16,623	16,623
			Sample C		
EXP	−0.258*** (0.005)	−0.257*** (0.005)	−0.240*** (0.005)	−0.242*** (0.005)	−0.144*** (0.003)
AGE	−0.001 (0.001)	−0.001 (0.001)	−0.012*** (0.002)	−0.012*** (0.002)	−0.008*** (0.001)
AGE_NA	0.095*** (0.010)	0.095*** (0.010)	0.011 (0.011)	0.009 (0.011)	0.001 (0.007)
MALE	0.015*** (0.003)	0.014*** (0.003)	0.055*** (0.003)	0.055*** (0.003)	0.032*** (0.002)
FIRM	−0.121*** (0.012)	−0.120*** (0.012)	−0.109*** (0.013)	−0.108*** (0.013)	−0.014* (0.008)
INSOLV_ACQU	−0.125*** (0.007)	−0.125*** (0.006)	−0.061*** (0.008)	−0.070*** (0.008)	−0.054*** (0.005)
INSOLV_PROC				−0.257*** (0.010)	−0.160*** (0.006)
TEL	0.273*** (0.003)	0.271*** (0.003)	0.257*** (0.003)	0.260*** (0.003)	0.146*** (0.002)
D_SCORE_A			0.223*** (0.008)	0.223*** (0.008)	0.120*** (0.005)
D_SCORE_B			0.195*** (0.008)	0.195*** (0.008)	0.105*** (0.005)
D_SCORE_C			0.149*** (0.008)	0.148*** (0.008)	0.081*** (0.005)
D_SCORE_D			0.092*** (0.009)	0.091*** (0.009)	0.050*** (0.006)
D_SCORE_E			0.003 (0.007)	0.001 (0.007)	−0.001 (0.004)
D_SCORE_F			−0.081*** (0.008)	−0.083*** (0.008)	−0.047*** (0.005)
D_SCORE_G			−0.133*** (0.009)	−0.135*** (0.009)	−0.074*** (0.006)
D_SCORE_H			−0.236*** (0.009)	−0.238*** (0.009)	−0.133*** (0.006)
D_SCORE_I			−0.271*** (0.008)	−0.273*** (0.008)	−0.153*** (0.005)
D_SCORE_K			−0.323*** (0.007)	−0.326*** (0.007)	−0.180*** (0.005)
D_SCORE_L			−0.485*** (0.008)	−0.492*** (0.008)	−0.270*** (0.005)
D_SCORE_M			−0.448*** (0.007)	−0.453*** (0.007)	−0.237*** (0.004)
L_ORIG_ACQU	0.0001 (0.002)	−0.001 (0.002)	−0.008*** (0.002)	−0.008*** (0.002)	−0.003** (0.001)
L_ORIG_ACQU_NA	−0.007* (0.004)	−0.005 (0.004)	−0.006 (0.004)	−0.007 (0.004)	−0.007*** (0.003)
END_missing	−0.063*** (0.004)	−0.063*** (0.004)	−0.041*** (0.005)	−0.039*** (0.005)	−0.014*** (0.003)
D_NO_MULTIPLE					0.152*** (0.003)
CR_other					0.090*** (0.001)
CR_C_CODE_ind		0.027*** (0.001)	0.027*** (0.002)	0.027*** (0.002)	0.014*** (0.001)
Baseline SCORE	—	NA	NA	NA	NA
R2 adj.	0.143	0.146	0.258	0.263	0.327
Wald Chi2	1556.862	1961.971	102.519	105.178	447.139
Wald Chi2 Prob.	0.000	0.000	0.000	0.000	0.000
Observations	126,015	126,015	126,015	126,015	126,015

Note: *p<0.1; **p<0.05; ***p<0.01

and Crook (2012) state adjusted values between 10.5 and 11.1% in an linear regression on a comprehensive set of debtor, contract, spatial and macroeconomic characteristics. Gürtler and Hibbeln (2013) state values between 4.4 and 18.9%.

Characteristics acquired by the collection agency

The results including the mean collection rate in a county or city ($CR\_C\_CODE\_ind$) are stated in the second column. The coefficient has a significantly positive relation to the collection rate of the individual debtor in all three samples. This is in line with the assumption that regional differences explain some part of the variation in the collection rate. The coefficient is higher in sample A and C with an increase of 4.5 and 2.7 percentage points compared to 0.8 percentage points in sample B. This is also reflected in the increase in the adjusted R^2 by 0.7 and 0.3 percentage points in sample A and C compared to 0.1 percentage points in sample B. We assess the drivers behind the regional differences in more detail in Section 4.3.

The credit bureau score ($SCORE$) is added in the third column. In Sample A, the debtors with a score of 'A' have an about three percentage points higher collection rate compared to missing scores. The worst score level has an about 32 percentage points lower collection rate. In sample C, the values range from a 12 percentage points higher collection rate for the score level 'A' to a more than 20 percentage points lower collection rate for the worst score levels. Besides the magnitude, it is interesting that the coefficients are monotonous over all levels in sample A and all levels except the worst two in sample C. This points towards a noteworthy level of accuracy. The coefficients have a change in sign in sample C but there is none in sample A. This seemingly reflects the fact that the score is obtained in most cases in sample C but more selectively in sample A. The fact that the score was obtained in sample A therefore likely conveys information of a collection process that was difficult before obtaining the score.[13] Given the magnitude of the coefficients, the credit bureau score has a strong relation to the collection rate in sample A and C. This is emphasized by the increase in the adjusted R^2. The values increase by 13.3 percentage points in sample A and 11.2 percentage points in sample C which is very sizable, especially, given the overall level of the R^2.

The adjusted R^2 further increases after adding the dummy for debtors that become insolvent during the workout process ($INSOLV\_PROC$; 2.2, 2.1 and 0.5 percentage points in sample A, B and C). The effect is particularly large considering the size of the coefficients. The collection rate of debtors in insolvency after the beginning of the workout process is significantly lower by 48.9, 56.1 and 25.7 percentage points.

[13] We discuss this aspect in Section 4.4.

The model in the fifth column includes the success on other accounts. The mean collection rate on other accounts (CR_other) has a significantly positive coefficient over all three samples. Given an increase of one standard deviation, the collection rate increases by 16.3, 3.9 and 9 percentage points. The coefficient for single encounters ($D\_NO\_MULTIPLE$) with the collection agency is inconsistent over the three samples. In sample A and B, it is negative. In sample C, the coefficient is positive. The adjusted R^2 increases by 14.1, 2.8 and 6.4 percentage points in sample A, B and C.

Considering the variables that are included in the first column, the key variables keep the initial sign an significance (EXP, $INSOLV\_ACQU$, TEL, L_ORIG_ACQU, $L\_ORIG\_ACQU\_NA$ and $END\_missing$). Some of the variables that are inconsistent in the initial set of variables (AGE, AGE_NA, $MALE$ and $FIRM$) display changes in sign or significance.

Comparing the initial adjusted R^2 values with the values including the characteristics gathered after the time of transfer, the difference is quite noteworthy. The overall increases amount to 30.3, 5 and 18.4 percentage points for the three samples A, B and C indicating that a large part of the quality of predictions relies on information that is gathered by the collection agency.

4.3 Spatial Area of Residence

We further assess which factors drive the regional differences in the collection rate. We therefore build a regression model explaining the mean collection rate in a county or city (CR_C_CODE) using the unemployment rate ($UNEMPL$), the population density ($INHAB\_DENS$), the distance between the collection agency and the debtor ($DISTANCE$), the number of cases per capita ($C\_CASES\_PER\_CAP$) and the total number of cases (CR_C_COUNT) as explanatory variables.

We include the unemployment rate to control for regional differences in the economic strength. The rationale is that when a region is economically weaker, the capacities to repay debt are lower. We therefore expect a negative relation to the collection rate.[14]

We also include the population density. There is a stylized assumption among practitioners that legal enforcement works better in less populated areas as the number of bailiffs is higher per capita and bailiffs have a deeper insight in the living conditions of debtors. We therefore expect a negative relation to the collection rate. The distance to the debtor is included in order to control for closeness to the collection agency. The rationale is that the collection agency might have a better understanding of the debtors that reside geographically closer. In this case, the distance should have a negative coefficient. In a

[14] We note that the effect of higher unemployment rates on the payment capability is more immediate for individuals. We estimate the results in Table 6 only including corporate debtors. The effect of the unemployment rate disappears in sample A and B but remains in sample C.

Table 6: Regression results - mean collections rate of counties and cities explained by the spatial characteristics

	CR_C_CODE					
	Sample A		Sample B		Sample C	
	(1)	(2)	(3)	(4)	(5)	(6)
UNEMPL	−0.010***	−0.010***	−0.008***	−0.008***	−0.010***	−0.008***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)
INHAB_DENS		−0.00002***		−0.00001		−0.00000
		(0.00001)		(0.00001)		(0.00000)
DISTANCE		0.00001		−0.00001		0.0001***
		(0.00003)		(0.00004)		(0.00002)
C_CASES_PER_CAP		6.145**		107.892**		−1.993
		(3.083)		(53.335)		(3.951)
CR_C_COUNT_ind		0.00001		0.0001		0.00000
		(0.00001)		(0.0001)		(0.00001)
Constant	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.160	0.219	0.044	0.055	0.240	0.255
F Statistic	77.249	23.515	19.361	5.652	127.890	28.458
F Statistic Prob.	0.000	0.000	0.000	0.000	0.000	0.000
Observations	402	402	402	402	402	402

Note: *p<0.1; **p<0.05; ***p<0.01

similar way, we include the cases per capita and the total number of cases in order to see whether the collection rate is higher in regions that appear in the collection portfolio more often.

The results are presented in Table 6. For each sample, we estimate one regression only including the unemployment rate and a second regression including all spatial explanatory variables. The regression models are calculated for all 402 counties and cities that appear in the unemployment data. As the unemployment rate changes over time, we use a mean over all observations in the respective region.

The results for the unemployment rate show a negative significant effect over all samples and regression settings. The results for the other variables are more impaired. There is a significant negative coefficient for the population density in sample A that is in line with the expected relation but none in the other samples. The cases per capita have a positive significant influence in sample A and B but not in sample C. The distance has a positive and significant coefficient in sample C, which is not in line with a lower distance standing for a better monitoring.

It is interesting to notice that the adjusted R^2 drops by less than a quarter in all samples when excluding all other variables except for the unemployment rate indicating that the regional differences appear to be mainly driven by economic factors. Other factors might be influential as well but are less consistent over the samples.

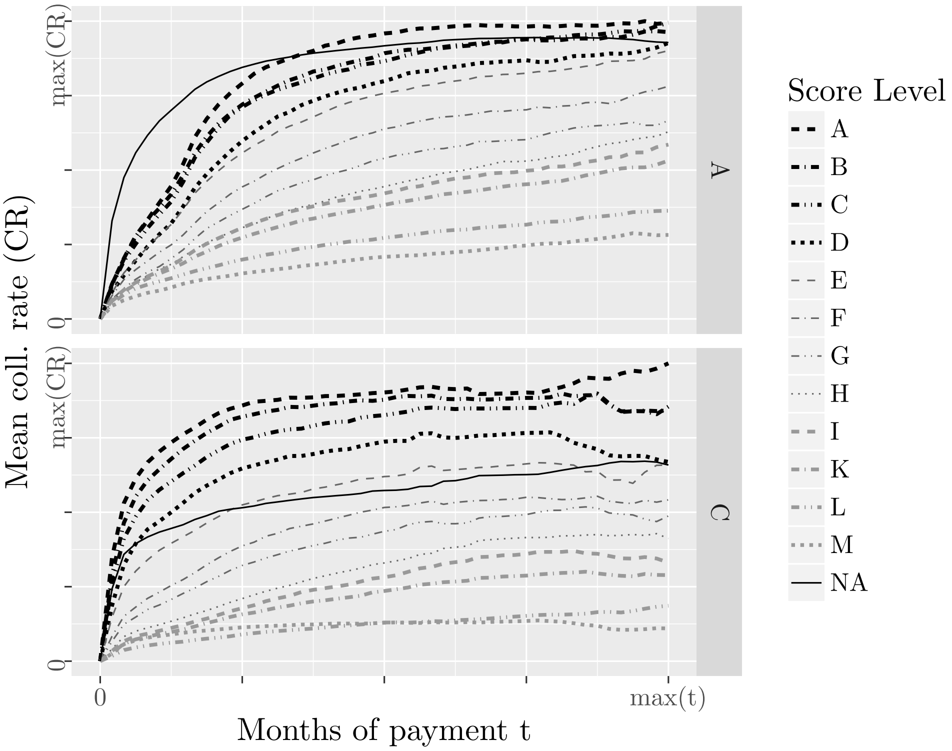

Figure 2: Mean collection rate for varying numbers of payment months by credit bureau score

4.4 Credit bureau score

As mentioned in Section 4.2, the fact that the credit bureau score is obtained might already contain some information of the pace of the collection process that is ex-post knowledge and is not in fact available in a prediction context. We therefore additionally calculate the change in the adjusted R^2 for removing the credit bureau score from a model estimated only on accounts with an obtained credit bureau score. These results are included in Table 15 in the Appendix. The adjusted R^2 is overall lower on this subset in sample (38.3%) and higher on this subset in sample C (37.8%). In sample A, the change in the adjusted R^2 is much lower compared to the values stated in Section 4.2 indicating that the missingness of the credit score contains some information of the pace of the collection process. It is, however, still considerable. In sample C, the change in the adjusted R^2 remains in a similar magnitude compared to Section 4.2 which is in line with more unconditionally obtained credit scores.

One further concern that we want to address is whether the credit bureau score may only work as good in predicting the collection rate due to reactions of the collection agency after obtaining the credit bureau score. While we cannot entirely rule this out, we want to stress that the credit bureau score is already a good predictor of recoveries in early stages of the collection process, where few or no actual workout actions have been conducted. We therefore plot the mean collection rate over time for each level of the credit bureau score. The resulting plot is presented in Figure 2 for sample A and C. It is noteworthy that the credit bureau score almost perfectly separates the mean collection rate by score levels. Only for some levels at the longest collection periods, where there are less accounts, and for the score of 'L' and 'M' in sample C, the lines touch or intersect. This applies to all points from end-to-end. Given this information, it appears that the credit bureau score is already a good predictor of the collection rate at the early stages of the collection process.

5 Post-Hoc Analysis: Transfer Policy and Undisclosed Information

In Section 4, we find the exposure size and the age of the account to be negatively related to the collection rate. Insolvent debtors have lower collection rates while debtors with available contact information have higher collection rates. The credit bureau score and the collection rate on other accounts of the same debtor are strong predictors of the collection success. High collection rates in a region are predictive of the collection success but only explain a small part of the variation in the data. The relation of the age, the gender, the legal status and the existence of other accounts are less consistent but might influence the collection rate in some of the samples.

This section presents a more in-depth analysis of two more aspects. (1) Considering the situation of LGD/RR prediction or predictions of recoveries on distressed trade credit in in-house collection, the in-house department makes predictions on a more or less homogenous portfolio of claims. This is not necessarily the case in third-party debt collection considering that the third-party collector is receiving recoveries from claims that originated from different original creditors that likely have a different quality of debtors and likely have different in-house collection processes. (2) Further, the information disclosed to the collection agency is dependent on the willingness of the original creditor to disclose (besides legal restrictions).

We conduct a more detailed analysis in the exposure size and the telephone contact details as indicators of the transfer policy and undisclosed information.

5.1 Transfer Policy

Our argument for the transfer policy is motivated by Walter et al. (2017) who discuss, whether their results of a positive relation between exposures and the collection rate is due to large exposures being more profitable and therefore being worked out more thoroughly by the debt collector. We follow this argument but alter it insofar as this could in a first place apply to the in-house collection process resulting in a preselection of large exposures that are handed over to the third-party collection. We test another hypothesis, whether the negative relation results from small exposures being covered faster in order to avoid lengthy complications with the collection agency that is also discussed in Walter et al. (2017).

Preselection in in-house collection

To assess whether the data reveals a preselection in in-house collection, we make use of the age of the account and the credit bureau score. Assuming that large exposures are worked out more thoroughly, the age of the account (L_ORIG_ACQU) should positively correlate with the exposure size. The correlation is at about 20% in sample A and at about 30% in sample C. The correlation is close to zero for sample B. This supports the hypothesis that the initial holders of the claims invest more effort to collect large claims in in-house collection in sample A and C.

One could further argue that, if the in-house collection department spends more effort in recovering large exposures, the transferred large exposures might be of worse quality compared to the smaller cases. This is assessed in Figure 3. The figure displays the proportion of accounts belonging to the four exposure quartiles of the exposures in sample A and C (EXP_1 to EXP_4) over the levels of the credit bureau score ($SCORE$) as a measure of debtor quality. The proportions for sample A are plotted on the left, the proportions for sample C are plotted on the right panel. In sample A, the proportion of the exposure quartiles appears relatively stable. However, in sample C, the highest exposure quartile is considerably more frequent for bad levels of the credit bureau score. This seems to be particularly apparent for score values above 'E'. The proportion of the highest exposure quartile is about 34% for debtors with a score of 'L' compared to about 24% for score values of 'A'.[15]

We also test the relation between the exposure on the one hand and the age of the account and credit bureau score on the other hand. The respective linear regression results are presented in Table 7. The age of the account has a significantly positive relation to the exposure in sample A and C but no significant coefficient in sample B. The credit bureau score has some significant coefficients in sample A but there is no obvious monotonous relation to the exposure size. Most score levels are insignificant. This is different for sample C. All score values are significant and there is considerable variation between the levels. The score levels from 'A' to 'C' have particularly low coefficients, while the score

[15] The credit bureau score further appears to be obtained less often for small exposures, which is reasonable given that obtaining the score is costly.

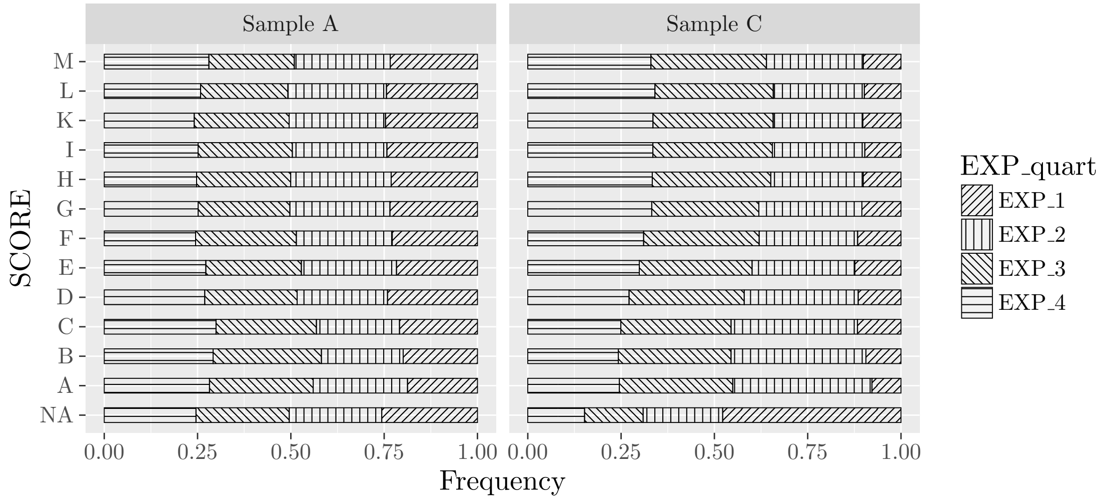

Figure 3: Barplot of the frequency of exposure quartiles by credit bureau score

values from 'F' to 'M' have particularly high coefficients. The adjusted R^2 is at around six percent in sample A, near zero in sample B and at around ten percent in sample C.

The bivariate and the multivariate assessment of the relation between the exposure size and the age of the account and credit bureau score seem to suggest that the initial holders treat large exposures differently in sample A and C, but we do not find the same relation in sample B. In sample A and C, larger exposures are kept longer in in-house collection. In sample C, large exposures are in addition more likely to have a lower credit score indicating that there is a preselection of transferred large exposures besides a longer effort for larger exposures in in-house collection.

Earlier payment for small exposures

Concerning the earlier payment for small exposures, we fit a linear regression model for the time period until the first payment ($FIRST\_PAYM$) including all variables in the full model but replacing the exposure with dummies for exposure sizes being in one of four exposure size quartiles (EXP_1 to EXP_4). We include all accounts that have at least one in-going payment. The results are presented in Table 8.

The linear regression results are in line with the assumed earlier payment for smaller exposures. The difference is, however, small in sample A and B. Compared to accounts in the first exposure size quartile in sample A, payments are received about a third of a month later (0.289, 0.361, 0.416). This relation is slightly higher in sample B but the difference is still lower than one month (0.223, 0.339, 0.605). Small accounts pay considerably faster for sample C (0.359, 1.182, 2.291). It appears that debtors tend to settle small accounts faster here, which supports the notion that debtors attempt to avoid complications. In

Table 7: Regression results - exposure explained by the credit bureau score and the age of the account

	EXP					
			Sample			
	(A)		(B)		(C)	
D_SCORE_A	-12.256	(17.841)			48.849^{***}	(11.806)
D_SCORE_B	-1.931	(16.465)			58.979^{***}	(12.856)
D_SCORE_C	-13.527	(16.047)			59.146^{***}	(13.167)
D_SCORE_D	-0.693	(17.373)			119.298^{***}	(14.288)
D_SCORE_E	-1.955	(12.172)			139.510^{***}	(11.086)
D_SCORE_F	-40.957^{***}	(12.844)			161.135^{***}	(12.967)
D_SCORE_G	-20.367	(13.081)			179.744^{***}	(12.954)
D_SCORE_H	-19.535	(12.513)			173.522^{***}	(12.458)
D_SCORE_I	-18.693^{*}	(10.424)			185.132^{***}	(10.661)
D_SCORE_K	-41.147^{***}	(9.083)			173.800^{***}	(9.047)
D_SCORE_L	-2.687	(8.157)			194.822^{***}	(7.781)
D_SCORE_M	8.505	(9.461)			165.493^{***}	(7.413)
L_ORIG_ACQU	1.666^{***}	(0.019)	0.055	(0.046)	1.262^{***}	(0.012)
L_ORIG_ACQU_NA	$3,126.784^{***}$	(54.923)	33.276^{***}	(4.794)	115.309^{***}	(5.719)
Constant	-40.066^{***}	(3.280)	154.013^{***}	(6.754)	23.704^{***}	(4.059)
Base *SCORE*	NA		—		NA	
Adjusted R^2	0.058		0.003		0.101	
F Statistic	804.086		24.812		1,015.722	
F Statistic Prob.	0.000		0.000		0.000	
Observations	182,880		16,623		126,015	
Note:						*p<0.1; **p<0.05; ***p<0.01

order to assess how this influences the collection rate, we fit a fractional response model for the collection rate regressed over the same variables as in Table 8. The results are presented in Table 9.

For sample A and B, in terms of the collection rate (Table 9) the second quartile does not differ significantly from the first quartile. For the third quartile, the coefficient is even positive for sample A and small for sample B. The key difference appears to be the one to the highest quartile. This does not seem to be in line with the monotonous relation between the time until the first payment and the exposure. In sample C, the difference between the quartiles is more monotonous. The second and third quartile differ noteworthy from the first but the main difference seems to be the one to the fourth quartile (more than 18 percentage points lower collection rate). While this is generally in line with the assumption that small exposures are paid faster and therefore have a higher collection rate, the central difference appears to be the difference to the highest exposures.

Table 8: Regression results - time until the first payment explained by the exposure quartile

	FIRST_PAYM					
	Sample					
	(A)		**(B)**		**(C)**	
EXP_2	0.289***	(0.030)	0.223***	(0.058)	0.359***	(0.059)
EXP_3	0.361***	(0.030)	0.339***	(0.058)	1.182***	(0.063)
EXP_4	0.416***	(0.032)	0.605***	(0.062)	2.291***	(0.073)
Controls	Yes		Yes		Yes	
Base EXP	EXP_1		EXP_1		EXP_1	
Adjusted R^2	0.141		0.060		0.127	
F Statistic	729.761		53.843		228.740	
F Statistic Prob.	0.000		0.000		0.000	
Note:					*p<0.1; **p<0.05; ***p<0.01	

Table 9: Regression results - collection rate explained by the exposure quartile

	CR					
	Sample					
	(A)		**(B)**		**(C)**	
EXP_2	0.002	(0.003)	−0.006	(0.005)	−0.017***	(0.003)
EXP_3	0.007***	(0.003)	−0.008*	(0.005)	−0.045***	(0.003)
EXP_4	−0.008***	(0.003)	−0.040***	(0.005)	−0.185***	(0.004)
Controls	Yes		Yes		Yes	
Base EXP	EXP_1		EXP_1		EXP_1	
Adjusted R^2	0.431		0.158		0.319	
Wald Chi2	1403.375		51.454		370.374	
Wald Chi2 Prob.	0.000		0.000		0.000	
Observations	182,880		16,623		126,015	
Note:					*p<0.1; **p<0.05; ***p<0.01	

5.2 Undisclosed Information

The availability of telephone contact information has a strong positive relation to the collection rate in Section 4. In terms of undisclosed debtor information available to the original creditor but not to the collection agency, we propose that the availability of contact information is informative of a higher debtor quality. We further test another important explanation as mentioned by Thomas et al. (2012) that this is mainly due to the practical use value of telephone contact details for getting in contact with the debtor more easily.

Signal of debtor quality

In terms of undisclosed information, if the original creditor had difficulties to get in contact with the debtor via telephone, this should be predictive of lower collection rates in third-party collection as well. To assess whether debtors with and without contact details are

Table 10: Regression results - telephone contact details explained by the credit bureau score

	TEL	
	Full Samples	
	(A)	**(C)**
D_SCORE_A	0.120** (0.053)	0.942*** (0.031)
D_SCORE_B	0.042 (0.048)	0.860*** (0.034)
D_SCORE_C	−0.037 (0.046)	0.648*** (0.034)
D_SCORE_D	−0.366*** (0.048)	0.422*** (0.036)
D_SCORE_E	−0.538*** (0.033)	0.272*** (0.028)
D_SCORE_F	−0.518*** (0.035)	0.087*** (0.033)
D_SCORE_G	−0.766*** (0.035)	0.006 (0.033)
D_SCORE_H	−0.928*** (0.034)	−0.219*** (0.033)
D_SCORE_I	−0.963*** (0.028)	−0.307*** (0.029)
D_SCORE_K	−0.950*** (0.025)	−0.439*** (0.025)
D_SCORE_L	−1.145*** (0.023)	−0.577*** (0.022)
D_SCORE_M	−1.230*** (0.026)	−0.606*** (0.021)
Constant	0.739*** (0.006)	−0.364*** (0.009)
Base SCORE	NA	NA
Wald Chi2	7524.7	4992
Wald Chi2 Prob.	0.000	0.000
Observations	182,880	126,015
Note:		*p<0.1; **p<0.05; ***p<0.01

different, we plot the proportion of debtors with telephone details for the levels of the credit bureau score in Figure 4. It is interesting to notice that the number of cases with missing telephone details increases for worse score values in both samples A and C. In sample A, the debtors with the best score levels have a similar level of telephone details with the missing credit bureau scores. This is in line with the earlier finding that the score is only obtained in worse cases in this sample. In sample C, the proportion of missing telephone details for missing credit bureau scores is on a similar level as for intermediate levels of the score.

We test this more formally by assessing whether the credit bureau score can explain the availability of telephone contact information in a logistic regression (Table 10). These multivariate results are in line with the results from Figure 4. Bad credit bureau scores are linked to a lower probability of having telephone contact information.

Accelerated collection process

We continue with considering the alternative line of argument. If available contact details enable an easier contact with the debtor, one would expect that payments from accounts with contact details are received earlier. In order to test this assumption, we fit a linear regression explaining the time on books before the first payment ($FIRST\_PAYM$). The

Table 11: Regression results - time until the first payment explained by the telephone contact dummy

	FIRST_PAYM					
	Full Samples					
	(A)		(B)		(C)	
TEL	0.030	(0.024)	-0.114^{**}	(0.055)	0.175^{***}	(0.042)
Controls	Yes		Yes		Yes	
Adjusted R^2	0.139		0.056		0.122	
F Statistic	776.773		57.618		235.179	
F Statistic Prob.	0.000		0.000		0.000	
Observations	124,514		12,413		43,943	
	Most Recent Accounts (6 months)					
	(A)		(B)		(C)	
TEL	0.104^{***}	(0.028)	0.142^{***}	(0.041)	0.071^{***}	(0.026)
Controls	Yes		Yes		Yes	
Adjusted R^2	0.041		0.052		0.035	
F Statistic	21.087		22.067		17.622	
F Statistic Prob.	0.000		0.000		0.000	
Observations	5,620		4,228		5,560	

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

coefficients for the telephone contact details are presented in Table 11. In the upper panel of the table, we present the results for the full samples A, B and C. In the lower panel, we include only accounts that were transferred within the last six months previous to the sample extraction and which were not in contact with the collection agency in the past. This way we attempt to minimize the cases where the telephone contact details were recorded during the ongoing collection process or a previous one.[16]

Given this rationale, one would expect that accounts with telephone contact details receive payments earlier compared to the accounts with no telephone details. This is contrary to the results for sample A and C in Table 11. In sample C, available telephone contacts have a positive relation to the time until the first payment in the full and in the short-term sample. In sample A, the relation is significantly positive in the short-term sample but insignificant in the full sample. For sample B, there is a positive relation over the first six months but a negative relation for the full period. The result for sample B on the full sample is not in line with rejecting the contact details as a mean of better communication. However, the contact details in sample B appear to be obtained later making these results less reliable (refer to Figure 5 and Table 16 in the Appendix). For

[16] Except for sample B, there is no visual relation between the time on books and the availability of telephone contact details over the first six months of the collection process (refer to Figure 5 in the Appendix). We test this more formally in Table 16 in the Appendix. There is no significant positive relation of the time on books on the telephone contact information compared to the first month for sample A and C but for sample B.

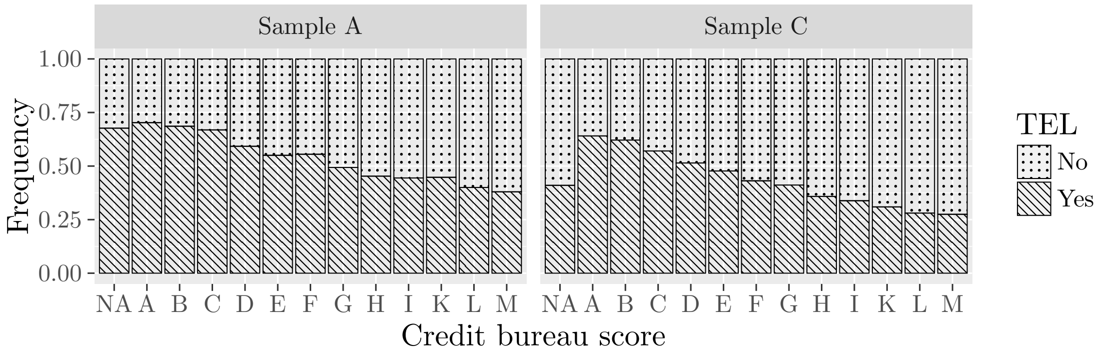

Figure 4: Barplot of the frequency of telephone contact details by credit bureau score

sample A and sample C the results are opposed to the assumption of telephone contact details serving to settle the workout process more easily.

6 Robustness Checks

Repayment horizon

In Section 3.1 we calculate the collection rate over a uniform repayment horizon. Among others, this could particularly affect the results for the exposure size, given that the length of the workout process might differ for different exposure sizes which is likely given the results in Section 5.1. We therefore replicate the results from Section 4.2 using different payment horizons. The repayment horizons are displayed in Figure 6 in the Appendix (The vertical lines marked with the number (3) in the upper and lower panel indicate the baseline payment horizon for sample A and C. The line with the number (2) indicates the baseline payment horizon for sample B in the panel in the center.). The results for these payment horizons are stated in the Tables 17, 18 and 19 in the Appendix. The coefficients of the exposure become slightly smaller but remain their direction and significance. The results for the other characteristics, as well, remain qualitatively unchanged.

Right-skewed exposures

As outlined in Section 3.1, the distribution of the exposure size displays a decisive right-skewness. In order to ensure that the results are not driven by large exposure values imposing a heavy weight on estimations, we present the results from section 4.2 estimated with the log of the exposures in Table 20 in the Appendix. The results are qualitatively unchanged.

Multiple accounts

In the three samples, some debtors are linked to multiple accounts. These cases are in the data for several reasons. First, there might be debtors that miss payments repeatedly due to obliviousness. Second, there might be cases that are in a continued distress resulting in multiple claims being handed over to the collection agency. We therefore estimate the results from Section 4.2 only including accounts of one-time debtors. These results are presented in Table 21 in the Appendix. The results remain qualitatively unchanged.

Individual and corporate debtors

Our samples contain both individual and corporate debtors. As one could argue that these groups might be driven by different factors, we estimate the regression models from Section 4.2 for both groups separately. The results are presented in Table 22 in the Appendix. For the individual debtors that account for the majority of cases, all results remain qualitatively unchanged. For the corporate debtors, most of the results remain qualitatively unchanged besides four minor changes in the significance (individual: age of the account in sample C, corporate: exposure in sample A and B and mean regional collection rate in sample B).

7 Conclusion

Delegating the collection of distressed receivables to debt collection agencies is common practice in many industries. Surprisingly, very little is known considering the successful management of receivables in third-party debt collection. This study aims at providing unique insights on how collection agencies can predict collection rates. These results are both valuable to collection agencies as well as banks and suppliers offering bank loans, consumer finance and trade credit.

We contribute to the literature on collection agencies in at least three important ways:
(1) Considering the information that is initially provided by the original creditor, the exposure size and the age of the account negatively relate to collection rates. The collection rate is higher for accounts with available telephone contact information. The results are in line with usual findings of the LGD/RR and collection rate literature. The initial information accounts for an adjusted R^2 of 10.7 to 14.3%. This is a relatively low extent but, as well, similar to work on the LGD/RR.
(2) We further consider information that is gathered by the collection agency from external sources or over the relationship to the debtor. The quality of the area of residence has a positive relation to the collection rate. Worse score levels relate to lower collection rates. Collection rates on other accounts are predictive of the collection success. The quality of

the residence area can be linked mainly to regional economic conditions. Especially the credit bureau score and the collections on other accounts considerably improve predictions. The credit bureau score is very accurate in distinguishing different levels of debtor quality. The information gathered by the collection agency substantially increases the quality of predictions with an overall adjusted R^2 of 15.7 to 43.1%.
(3) In a third step, we examine how the transfer policy in in-house collection and undisclosed information affect collection rates. Our analyzes show that there is a preselection of (supposedly more profitable) large exposures in a way that older cases and cases with a low credit quality are more likely to be handed over. This needs to be taken into account in making predictions. We further assess the information content of the telephone contact details and find it to contain information of the debtor quality in in-house collection besides its practical use in the third-party collection process.

In addition to the previously mentioned literature, our results have direct implications for the trade credit literature by identifying driving forces behind the trade credit recovery risk. In a more general reference to the trade credit literature, we provide analyze the motivation for commissioning a debt collection as a specialized external contractor. Given that in the segment of large exposures more difficult cases are transferred and that gathered information is a central factor in the quality of predictions, we provide evidence that collection agencies might be competitive due to economics of scale and specialization gains (as mentioned in Mian and Smith, 1992 and Mian and Smith, 1994). Our results are further relevant for the LGD/RR literature. The quality of the residence area and the strong impact of the credit bureau score are likely to extend to recoveries on bank loans and consumer finance. By linking the quality of the residence area to economic factors, we further contribute to the discussion on the impact of macroeconomic conditions on the LGD/RR.

The aspect that is central to our results is the information that is gathered in addition to the initially disclosed characteristics. The collection success on other accounts is obtained over repeated contact with the debtor. The score is obtained (at cost) from a credit bureau. The collection agency obtains information about the original creditors' policy to hand-over distressed accounts over a longstanding relationship. There is likely to be more debtor information available at the original creditor that could be integrated for a mutual benefit. While the quality of LGD/RR prediction relies on comprehensive credit application data, most of the information used in third-party collection prediction needs to be obtained from the original creditor or additional sources first. Besides the evident role of collecting recoveries, debt collection agencies fulfill a role as a collector of information.

Appendix

Table 12: Correlation table of the independent variables in Sample A

	EXP	AGE	AGE_NA	MALE	INSOLV_ACQU	FIRM	L_ORIG_ACQU	L_ORIG_ACQU_NA	END_missing	TEL	CR_C_CODE_ind	D_NO_MULTIPLE	CR_other	INSOLV_PROC
EXP	1.00	0.01	0.07	0.01	0.01	0.07	0.20	0.13	0.20	-0.00	-0.01	0.01	-0.03	0.02
AGE	0.01	1.00	-0.00	0.02	0.00	0.00	-0.09	0.00	-0.07	0.11	0.05	-0.04	0.06	-0.03
AGE_NA	0.07	-0.00	1.00	-0.25	0.00	0.87	-0.04	0.01	-0.00	-0.06	-0.03	0.04	-0.03	0.01
MALE	0.01	0.02	-0.25	1.00	-0.01	-0.30	0.02	0.01	0.02	0.03	0.02	-0.03	0.01	-0.01
INSOLV_ACQU	0.01	0.00	0.00	-0.01	1.00	0.00	0.02	0.01	0.01	-0.01	-0.01	0.08	-0.05	-0.02
FIRM	0.07	0.00	0.87	-0.30	0.00	1.00	-0.04	-0.00	-0.01	-0.03	-0.03	0.01	-0.04	0.01
L_ORIG_ACQU	0.20	-0.09	-0.04	0.02	0.02	-0.04	1.00	-0.00	0.26	-0.05	-0.05	0.05	-0.12	0.02
L_ORIG_ACQU_NA	0.13	0.00	0.01	0.01	0.01	-0.00	-0.00	1.00	0.17	-0.01	-0.00	0.03	-0.01	0.00
END_missing	0.20	-0.07	-0.00	0.02	0.01	-0.01	0.26	0.17	1.00	-0.06	-0.04	0.12	-0.08	0.02
TEL	-0.00	0.11	-0.06	0.03	-0.01	-0.03	-0.05	-0.01	-0.06	1.00	0.05	-0.13	0.20	-0.03
CR_C_CODE_ind	-0.01	0.05	-0.03	0.02	-0.01	-0.03	-0.05	-0.00	-0.04	0.05	1.00	-0.04	0.09	-0.02
D_NO_MULTIPLE	0.01	-0.04	0.04	-0.03	0.08	0.01	0.05	0.03	0.12	-0.13	-0.04	1.00	0.00	0.06
CR_other	-0.03	0.06	-0.03	0.01	-0.05	-0.04	-0.12	-0.01	-0.08	0.20	0.09	0.00	1.00	-0.07
INSOLV_PROC	0.02	-0.03	0.01	-0.01	-0.02	0.01	0.02	0.00	0.02	-0.03	-0.02	0.06	-0.07	1.00
D_SCORE_NA	-0.02	0.06	0.15	-0.04	0.00	0.13	-0.13	-0.02	-0.10	0.18	0.05	0.01	0.29	-0.00
D_SCORE_A	0.00	0.02	-0.02	-0.04	-0.01	-0.02	0.01	0.00	-0.00	0.02	-0.00	-0.04	0.02	-0.00
D_SCORE_B	0.00	0.01	-0.03	-0.03	-0.01	-0.02	0.01	-0.00	-0.01	0.01	0.01	-0.04	0.01	-0.01
D_SCORE_C	0.00	0.03	-0.03	-0.01	-0.01	-0.02	0.01	-0.00	-0.00	0.01	0.01	-0.04	0.01	-0.01
D_SCORE_D	0.00	0.00	-0.02	-0.01	-0.01	-0.02	0.01	0.00	0.00	-0.01	-0.00	-0.02	-0.00	-0.01
D_SCORE_E	0.01	0.01	-0.04	0.01	-0.01	-0.03	0.02	0.01	0.01	-0.02	-0.00	-0.02	-0.01	-0.01
D_SCORE_F	-0.00	-0.00	-0.03	0.00	-0.01	-0.03	0.03	0.00	0.01	-0.02	-0.00	-0.01	-0.04	-0.00
D_SCORE_G	0.00	-0.02	-0.03	-0.00	-0.01	-0.03	0.03	0.00	0.02	-0.04	-0.01	-0.00	-0.06	0.00
D_SCORE_H	0.00	-0.01	-0.03	-0.02	-0.01	-0.03	0.03	0.01	0.02	-0.05	-0.02	-0.00	-0.09	0.00
D_SCORE_I	0.00	-0.01	-0.04	-0.02	-0.01	-0.04	0.04	0.00	0.02	-0.06	-0.01	0.00	-0.10	0.00
D_SCORE_K	0.00	-0.03	-0.05	0.01	-0.02	-0.04	0.05	0.02	0.04	-0.07	-0.03	0.01	-0.13	0.00
D_SCORE_L	0.01	-0.04	-0.06	0.07	-0.02	-0.05	0.06	0.01	0.07	-0.10	-0.03	0.03	-0.17	0.00
D_SCORE_M	0.02	-0.06	-0.05	0.06	0.10	-0.04	0.06	0.01	0.07	-0.10	-0.03	0.03	-0.16	0.02

Table 13: Correlation table of the independent variables in Sample B

	EXP	AGE	AGE_NA	MALE	INSOLV_ACQU	FIRM	L_ORIG_ACQU	L_ORIG_ACQU_NA	TEL	CR_C_CODE_ind	D_NO_MULTIPLE	CR_other	INSOLV_PROC
EXP	1.00	0.04	0.06	-0.11	0.02	0.35	0.01	0.05	0.02	0.00	0.02	-0.07	0.03
AGE	0.04	1.00	-0.00	-0.01	-0.00	-0.00	-0.01	-0.09	0.04	0.01	-0.00	-0.02	-0.03
AGE_NA	0.06	-0.00	1.00	-0.08	-0.01	0.23	-0.03	-0.08	-0.60	0.00	0.05	-0.03	-0.05
MALE	-0.11	-0.01	-0.08	1.00	-0.01	-0.38	0.01	-0.02	0.02	0.01	0.01	0.02	-0.02
INSOLV_ACQU	0.02	-0.00	-0.01	-0.01	1.00	0.02	0.01	0.02	-0.04	-0.01	0.03	-0.03	-0.02
FIRM	0.35	-0.00	0.23	-0.38	0.02	1.00	-0.00	-0.05	-0.03	-0.02	-0.04	-0.07	0.06
L_ORIG_ACQU	0.01	-0.01	-0.03	0.01	0.01	-0.00	1.00	-0.00	-0.01	-0.03	0.04	-0.06	0.01
L_ORIG_ACQU_NA	0.05	-0.09	-0.08	-0.02	0.02	-0.05	-0.00	1.00	0.01	-0.01	-0.00	-0.01	0.04
TEL	0.02	0.04	-0.60	0.02	-0.04	-0.03	-0.01	0.01	1.00	0.02	-0.07	0.05	-0.03
CR_C_CODE_ind	0.00	0.01	0.00	0.01	-0.01	-0.02	-0.03	-0.01	0.02	1.00	0.00	0.00	-0.01
D_NO_MULTIPLE	0.02	-0.00	0.05	0.01	0.03	-0.04	0.04	-0.00	-0.07	0.00	1.00	0.00	0.03
CR_other	-0.07	-0.02	-0.03	0.02	-0.03	-0.07	-0.06	-0.01	0.05	0.00	0.00	1.00	-0.01
INSOLV_PROC	0.03	-0.03	-0.05	-0.02	-0.02	0.06	0.01	0.04	-0.03	-0.01	0.03	-0.01	1.00

Table 14: Correlation table of the independent variables in Sample C

	EXP	AGE	AGE_NA	MALE	INSOLV_ACQU	FIRM	L_ORIG_ACQU	L_ORIG_ACQU_NA	END_missing	TEL	CR_C_CODE_ind	D_NO_MULTIPLE	CR_other	INSOLV_PROC
EXP	1.00	-0.04	0.05	0.01	-0.01	0.03	0.30	0.07	0.13	-0.02	-0.03	-0.02	-0.05	0.01
AGE	-0.04	1.00	-0.00	0.02	0.02	-0.00	-0.01	-0.06	0.01	0.04	0.02	0.02	0.02	-0.01
AGE_NA	0.05	-0.00	1.00	-0.28	-0.01	0.86	-0.00	-0.04	0.03	-0.04	-0.01	-0.11	-0.03	0.00
MALE	0.01	0.02	-0.28	1.00	-0.01	-0.34	0.03	-0.02	0.02	0.03	0.01	0.02	0.02	-0.00
INSOLV_ACQU	-0.01	0.02	-0.01	-0.01	1.00	-0.01	-0.00	0.03	0.01	-0.01	-0.00	0.05	-0.01	-0.04
FIRM	0.03	-0.00	0.86	-0.34	-0.01	1.00	-0.00	-0.03	0.04	-0.01	-0.01	-0.14	-0.03	0.01
L_ORIG_ACQU	0.30	-0.01	-0.00	0.03	-0.00	-0.00	1.00	-0.00	-0.17	-0.07	0.00	0.02	-0.01	-0.00
L_ORIG_ACQU_NA	0.07	-0.06	-0.04	-0.02	0.03	-0.03	-0.00	1.00	0.20	-0.06	-0.03	-0.01	-0.00	-0.01
END_missing	0.13	0.01	0.03	0.02	0.01	0.04	-0.17	0.20	1.00	0.11	-0.01	-0.07	-0.03	0.02
TEL	-0.02	0.04	-0.04	0.03	-0.01	-0.01	-0.07	-0.06	0.11	1.00	0.04	-0.00	0.17	-0.01
CR_C_CODE_ind	-0.03	0.02	-0.01	0.01	-0.00	-0.01	0.00	-0.03	-0.01	0.04	1.00	0.03	0.04	-0.00
D_NO_MULTIPLE	-0.02	0.02	-0.11	0.02	0.05	-0.14	0.02	-0.01	-0.07	-0.00	0.03	1.00	0.00	0.03
CR_other	-0.05	0.02	-0.03	0.02	-0.01	-0.03	-0.01	-0.00	-0.03	0.17	0.04	0.00	1.00	-0.01
INSOLV_PROC	0.01	-0.01	0.00	-0.00	-0.04	0.01	-0.00	-0.01	0.02	-0.01	-0.00	0.03	-0.01	1.00
D_SCORE_NA	-0.11	0.08	0.35	-0.15	0.04	0.31	-0.04	-0.12	-0.14	0.02	0.01	0.07	0.04	-0.00
D_SCORE_A	-0.01	0.01	-0.06	-0.04	-0.04	-0.05	0.00	0.02	0.02	0.10	0.02	0.03	0.08	-0.01
D_SCORE_B	-0.00	-0.01	-0.05	0.00	-0.04	-0.05	0.01	0.02	0.02	0.08	0.02	0.03	0.07	-0.01
D_SCORE_C	-0.00	0.00	-0.05	0.03	-0.04	-0.04	0.01	0.01	0.01	0.06	0.02	0.02	0.06	-0.01
D_SCORE_D	0.01	-0.00	-0.05	0.02	-0.03	-0.04	0.01	0.01	0.01	0.04	0.00	0.02	0.04	-0.01
D_SCORE_E	0.02	-0.00	-0.06	0.06	-0.04	-0.05	0.01	0.02	0.02	0.04	0.00	0.01	0.03	-0.01
D_SCORE_F	0.02	-0.01	-0.05	0.01	-0.03	-0.04	0.01	0.02	0.02	0.01	-0.01	0.00	0.02	-0.00
D_SCORE_G	0.02	-0.01	-0.04	0.02	-0.03	-0.04	0.00	0.02	0.02	0.01	-0.00	-0.01	0.01	0.01
D_SCORE_H	0.02	-0.02	-0.05	-0.03	-0.03	-0.05	0.01	0.03	0.02	-0.01	-0.00	-0.02	-0.01	0.01
D_SCORE_I	0.03	-0.01	-0.06	-0.03	-0.04	-0.06	0.01	0.03	0.03	-0.03	-0.01	-0.01	-0.02	0.01
D_SCORE_K	0.03	-0.01	-0.08	0.01	-0.05	-0.07	0.01	0.03	0.03	-0.05	-0.01	-0.02	-0.04	0.01
D_SCORE_L	0.05	-0.03	-0.10	0.09	-0.07	-0.09	0.01	0.02	0.04	-0.08	-0.02	-0.05	-0.08	0.00
D_SCORE_M	0.03	-0.05	-0.10	0.11	0.24	-0.09	0.00	0.04	0.06	-0.09	-0.02	-0.09	-0.10	0.00

Table 15: Change in the adjusted R^2 for removing individual characteristics

	R-square/ΔR-square		
	Sample		
	A	**B**	**C**
Baseline case			
Full model	0.4313	0.1569	0.3269
Non-missing score	0.3834		0.3779
Excluded variable (Δ)			
EXP	-0.0003	-0.0041	-0.0290
AGE	-0.0001	-0.0021	-0.0008
MALE	-0.0001	-0.0001	-0.0015
INSOLV_ACQU	-0.0051	-0.0109	-0.0016
FIRM	0.0000	-0.0003	-0.0000
L_ORIG_ACQU	-0.0097	-0.0354	-0.0001
END_missing	-0.0025		-0.0001
TEL	-0.0083	-0.0322	-0.0373
CR_C_CODE_ind	-0.0016	-0.0002	-0.0014
CR_other	-0.1416	-0.0281	-0.0638
INSOLV_PROC	-0.0139	-0.0200	-0.0053
SCORE	-0.0690		-0.0833
Non-missing SCORE (Δ)			
SCORE	-0.0310		-0.1357

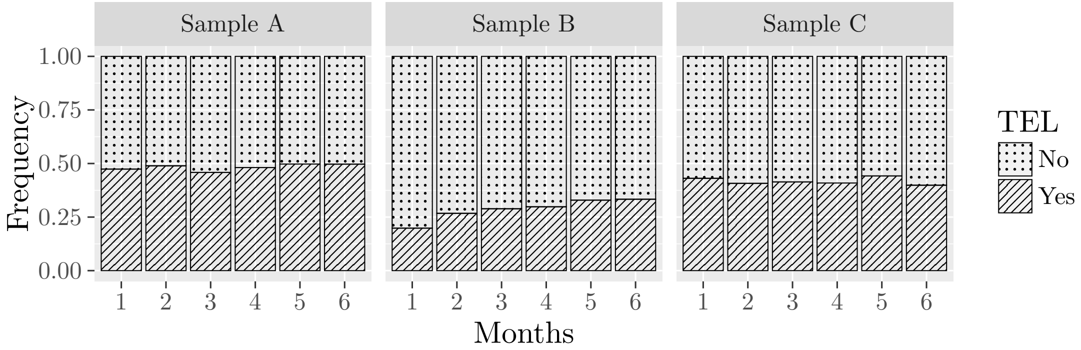

Figure 5: Barplot of the available telephone contact details by time period in the collection process

Table 16: Regression results - telephone contact dummy explained by time period in the collection process

	TEL		
		Sample	
	(A)	**(B)**	**(C)**
Month_2	0.061	0.388***	−0.097**
	(0.050)	(0.085)	(0.048)
Month_3	−0.063	0.495***	−0.069
	(0.061)	(0.086)	(0.050)
Month_4	0.029	0.540***	−0.090**
	(0.050)	(0.083)	(0.045)
Month_5	0.095	0.683***	0.045
	(0.058)	(0.089)	(0.057)
Month_6	0.093*	0.703***	−0.133***
	(0.056)	(0.082)	(0.051)
Constant	−0.104***	−1.396***	−0.278***
	(0.032)	(0.058)	(0.033)
Omitted Month	Month_1	Month_1	Month_1
Observations	14,367	7,999	19,167
Wald Chi2	8.323	99.883	13.922
Wald Chi2 Prob.	0.139	0.000	0.016
Note:			*p<0.1; **p<0.05; ***p<0.01

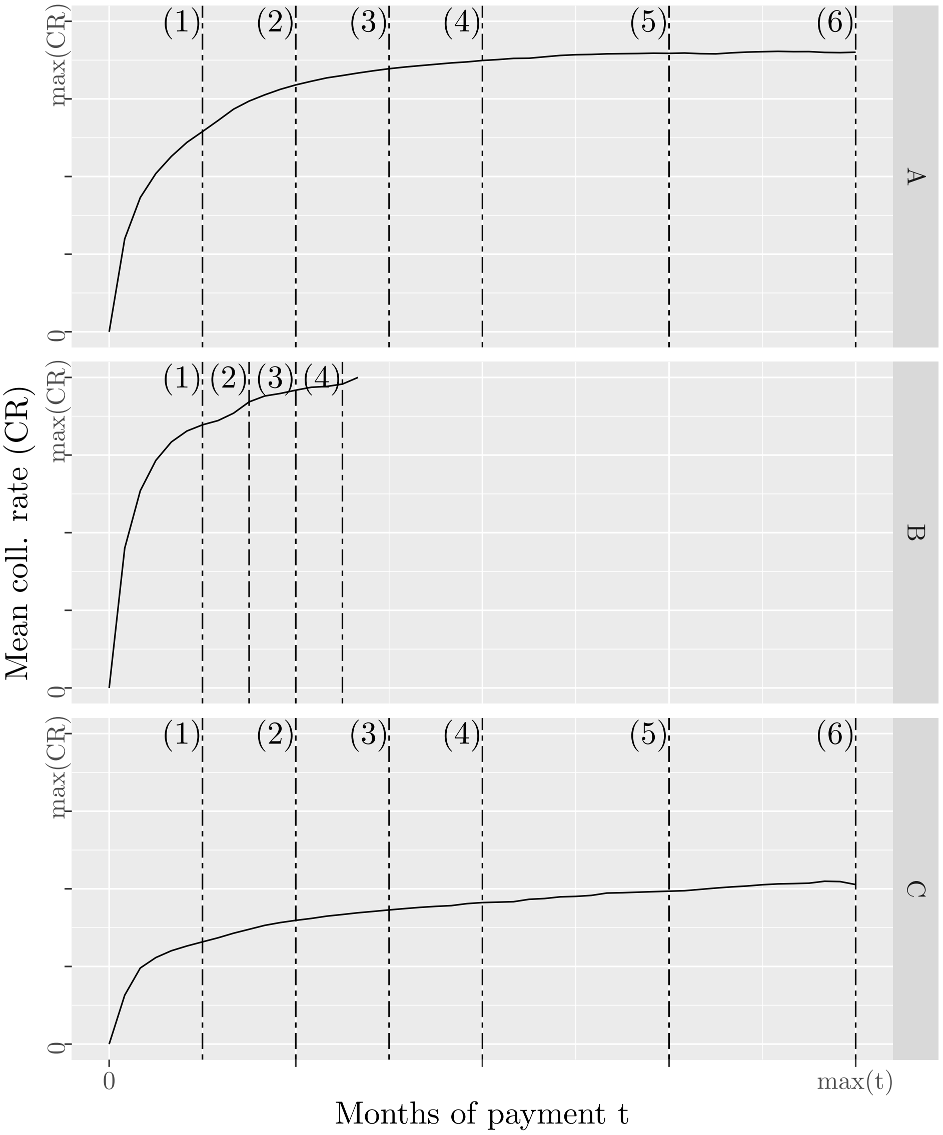

Figure 6: Overview robustness-checks workout period

Table 17: Regression results for varying workout periods in Sample A

	Dependent variable											
	(1)		(2)		(3)		(4)		(5)		(6)	
EXP	−0.031***	(0.004)	−0.024***	(0.003)	−0.017***	(0.002)	−0.015***	(0.002)	−0.015***	(0.002)	−0.016***	(0.004)
AGE	0.010***	(0.001)	0.004***	(0.001)	0.001	(0.001)	−0.002***	(0.001)	−0.005***	(0.001)	−0.005***	(0.002)
AGE_NA	0.003	(0.011)	−0.029***	(0.008)	−0.035***	(0.007)	−0.029***	(0.006)	−0.022***	(0.007)	−0.014	(0.012)
MALE	0.030***	(0.003)	0.017***	(0.002)	0.010***	(0.002)	0.008***	(0.002)	0.007***	(0.002)	0.008**	(0.003)
FIRM	0.065***	(0.012)	0.012	(0.009)	−0.002	(0.008)	−0.017**	(0.007)	−0.029***	(0.008)	−0.022	(0.014)
INSOLV_ACQU	−0.278***	(0.011)	−0.243***	(0.008)	−0.206***	(0.007)	−0.171***	(0.006)	−0.139***	(0.006)	−0.111***	(0.011)
INSOLV_PROC	−0.429***	(0.011)	−0.367***	(0.008)	−0.325***	(0.007)	−0.284***	(0.007)	−0.208***	(0.006)	−0.180***	(0.010)
TEL	0.085***	(0.003)	0.079***	(0.002)	0.083***	(0.002)	0.075***	(0.002)	0.060***	(0.002)	0.066***	(0.003)
D_SCORE_A	−0.302***	(0.014)	−0.105***	(0.011)	−0.034***	(0.010)	−0.001	(0.010)	0.011	(0.009)	0.045***	(0.014)
D_SCORE_B	−0.289***	(0.013)	−0.135***	(0.009)	−0.075***	(0.009)	−0.043***	(0.008)	−0.021***	(0.008)	0.019	(0.013)
D_SCORE_C	−0.323***	(0.013)	−0.140***	(0.009)	−0.089***	(0.008)	−0.056***	(0.008)	−0.011	(0.008)	0.027*	(0.014)
D_SCORE_D	−0.339***	(0.014)	−0.186***	(0.010)	−0.111***	(0.009)	−0.074***	(0.008)	−0.038***	(0.008)	0.011	(0.014)
D_SCORE_E	−0.334***	(0.010)	−0.196***	(0.007)	−0.129***	(0.006)	−0.084***	(0.006)	−0.049***	(0.005)	−0.006	(0.009)
D_SCORE_F	−0.385***	(0.011)	−0.242***	(0.008)	−0.170***	(0.006)	−0.117***	(0.006)	−0.081***	(0.006)	−0.044***	(0.009)
D_SCORE_G	−0.414***	(0.013)	−0.265***	(0.008)	−0.196***	(0.006)	−0.145***	(0.006)	−0.103***	(0.005)	−0.075***	(0.009)
D_SCORE_H	−0.409***	(0.012)	−0.291***	(0.008)	−0.230***	(0.006)	−0.180***	(0.006)	−0.119***	(0.005)	−0.073***	(0.009)
D_SCORE_I	−0.428***	(0.010)	−0.300***	(0.007)	−0.238***	(0.005)	−0.196***	(0.005)	−0.136***	(0.004)	−0.084***	(0.007)
D_SCORE_K	−0.432***	(0.009)	−0.314***	(0.006)	−0.254***	(0.005)	−0.207***	(0.004)	−0.150***	(0.004)	−0.104***	(0.007)
D_SCORE_L	−0.465***	(0.009)	−0.370***	(0.006)	−0.306***	(0.005)	−0.255***	(0.004)	−0.194***	(0.004)	−0.155***	(0.006)
D_SCORE_M	−0.493***	(0.012)	−0.389***	(0.008)	−0.321***	(0.006)	−0.270***	(0.005)	−0.208***	(0.005)	−0.175***	(0.008)
L_ORIG_ACQU	−0.153***	(0.002)	−0.077***	(0.002)	−0.052***	(0.001)	−0.038***	(0.001)	−0.021***	(0.001)	−0.019***	(0.002)
L_ORIG_ACQU_NA	−0.152**	(0.064)	−0.114***	(0.039)	−0.082***	(0.029)	−0.063**	(0.026)	−0.034	(0.025)	−0.032	(0.042)
END_missing	−0.232***	(0.012)	−0.158***	(0.007)	−0.130***	(0.006)	−0.095***	(0.005)	−0.073***	(0.005)	−0.064***	(0.009)
D_NO_MULTIPLE	−0.021***	(0.003)	−0.119***	(0.002)	−0.127***	(0.002)	−0.130***	(0.002)	−0.125***	(0.002)	−0.145***	(0.003)
CR_other	0.179***	(0.002)	0.179***	(0.001)	0.163***	(0.001)	0.141***	(0.001)	0.111***	(0.001)	0.108***	(0.002)
CR_C_CODE_ind	0.019***	(0.001)	0.020***	(0.001)	0.018***	(0.001)	0.014***	(0.001)	0.011***	(0.001)	0.011***	(0.002)
Baseline SCORE	NA		NA		NA		NA		NA		NA	
R2 adj.	0.321		0.409		0.431		0.436		0.428		0.383	
Wald Chi2	1696.679		1355.468		1437.846		1059.841		502.483		125.33	
Wald Chi2 Prob.	0.000		0.000		0.000		0.000		0.000		0.000	
Observations	182,880		182,880		182,880		152,087		90,862		27,438	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 18: Regression results for varying workout periods in Sample B

	Dependent variable							
	(1)		**(2)**		**(3)**		**(4)**	
EXP	−0.025***	(0.006)	−0.016***	(0.004)	−0.007***	(0.002)	−0.004*	(0.002)
AGE	0.001	(0.002)	−0.001	(0.002)	−0.001	(0.002)	−0.005**	(0.002)
AGE_NA	0.029***	(0.006)	0.022***	(0.005)	0.019***	(0.004)	0.018***	(0.005)
MALE	−0.009*	(0.005)	−0.007*	(0.004)	−0.007**	(0.003)	−0.002	(0.004)
FIRM	0.049***	(0.013)	0.015	(0.010)	−0.006	(0.007)	−0.008	(0.010)
INSOLV_ACQU	−0.179***	(0.017)	−0.154***	(0.012)	−0.114***	(0.010)	−0.074***	(0.013)
INSOLV_PROC	−0.322***	(0.024)	−0.256***	(0.018)	−0.174***	(0.014)	−0.118***	(0.016)
TEL	0.124***	(0.006)	0.106***	(0.005)	0.082***	(0.004)	0.065***	(0.006)
L_ORIG_ACQU	−0.053***	(0.006)	−0.034***	(0.004)	−0.021***	(0.003)	−0.011***	(0.003)
L_ORIG_ACQU_NA	−0.092***	(0.005)	−0.063***	(0.004)	−0.033***	(0.003)	−0.015***	(0.004)
D_NO_MULTIPLE	−0.157***	(0.011)	−0.135***	(0.009)	−0.114***	(0.010)	−0.101***	(0.012)
CR_other	0.047***	(0.003)	0.039***	(0.002)	0.024***	(0.002)	0.017***	(0.002)
CR_C_CODE_ind	0.003	(0.002)	0.004**	(0.002)	0.003*	(0.001)	0.001	(0.002)
R2 adj.	0.15		0.157		0.166		0.16	
Wald Chi2	60.899		59.62		29.493		3.328	
Wald Chi2 Prob.	0.000		0.000		0.000		0.068	
Observations	16,623		16,623		10,568		3,452	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 19: Regression results for varying workout periods in Sample C

	Dependent variable											
	(1)		**(2)**		**(3)**		**(4)**		**(5)**		**(6)**	
EXP	−0.154***	(0.003)	−0.152***	(0.003)	−0.144***	(0.003)	−0.156***	(0.004)	−0.139***	(0.006)	−0.134***	(0.011)
AGE	−0.001**	(0.001)	−0.005***	(0.001)	−0.008***	(0.001)	−0.016***	(0.001)	−0.025***	(0.002)	−0.022***	(0.004)
AGE_NA	0.009*	(0.005)	0.006	(0.006)	0.001	(0.007)	−0.013	(0.009)	−0.044***	(0.014)	−0.058***	(0.022)
MALE	0.025***	(0.002)	0.031***	(0.002)	0.032***	(0.002)	0.038***	(0.003)	0.037***	(0.005)	0.032***	(0.008)
FIRM	0.008	(0.005)	−0.001	(0.007)	−0.014*	(0.008)	−0.028***	(0.011)	−0.057***	(0.016)	−0.065**	(0.026)
INSOLV_ACQU	−0.033***	(0.003)	−0.046***	(0.004)	−0.054***	(0.005)	−0.077***	(0.007)	−0.111***	(0.011)	−0.146***	(0.019)
INSOLV_PROC	−0.086***	(0.005)	−0.132***	(0.006)	−0.160***	(0.006)	−0.197***	(0.008)	−0.233***	(0.012)	−0.211***	(0.019)
TEL	0.067***	(0.002)	0.113***	(0.002)	0.146***	(0.002)	0.184***	(0.003)	0.200***	(0.004)	0.178***	(0.008)
D_SCORE_A	0.059***	(0.003)	0.097***	(0.004)	0.120***	(0.005)	0.142***	(0.007)	0.117***	(0.013)	0.183***	(0.042)
D_SCORE_B	0.047***	(0.003)	0.083***	(0.004)	0.105***	(0.005)	0.124***	(0.008)	0.107***	(0.014)	0.075*	(0.041)
D_SCORE_C	0.032***	(0.003)	0.062***	(0.004)	0.081***	(0.005)	0.105***	(0.008)	0.097***	(0.014)	0.082**	(0.039)
D_SCORE_D	0.005	(0.004)	0.030***	(0.005)	0.050***	(0.006)	0.065***	(0.008)	0.058***	(0.014)	0.020	(0.035)
D_SCORE_E	−0.026***	(0.003)	−0.014***	(0.004)	−0.001	(0.004)	0.004	(0.006)	−0.013	(0.011)	−0.032	(0.025)
D_SCORE_F	−0.078***	(0.004)	−0.062***	(0.005)	−0.047***	(0.005)	−0.041***	(0.007)	−0.058***	(0.012)	−0.086***	(0.023)
D_SCORE_G	−0.095***	(0.005)	−0.085***	(0.005)	−0.074***	(0.006)	−0.076***	(0.008)	−0.093***	(0.012)	−0.122***	(0.025)
D_SCORE_H	−0.129***	(0.005)	−0.133***	(0.005)	−0.133***	(0.006)	−0.143***	(0.008)	−0.140***	(0.012)	−0.137***	(0.023)
D_SCORE_I	−0.144***	(0.005)	−0.152***	(0.005)	−0.153***	(0.005)	−0.173***	(0.007)	−0.179***	(0.010)	−0.208***	(0.021)
D_SCORE_K	−0.158***	(0.004)	−0.171***	(0.004)	−0.180***	(0.005)	−0.204***	(0.006)	−0.240***	(0.009)	−0.241***	(0.017)
D_SCORE_L	−0.200***	(0.004)	−0.247***	(0.005)	−0.270***	(0.005)	−0.322***	(0.006)	−0.404***	(0.009)	−0.365***	(0.016)
D_SCORE_M	−0.163***	(0.004)	−0.207***	(0.004)	−0.237***	(0.004)	−0.304***	(0.006)	−0.398***	(0.009)	−0.447***	(0.020)
L_ORIG_ACQU	0.0003	(0.001)	−0.001	(0.001)	−0.003**	(0.001)	−0.007***	(0.002)	−0.020***	(0.004)	−0.022***	(0.007)
L_ORIG_ACQU_NA	−0.017***	(0.002)	−0.011***	(0.002)	−0.007***	(0.003)	−0.010***	(0.004)	−0.044	(0.092)	−0.015	(0.207)
END_missing	−0.002	(0.002)	−0.009***	(0.002)	−0.014***	(0.003)	−0.036***	(0.005)	−0.032	(0.036)	0.038	(0.063)
D_NO_MULTIPLE	0.116***	(0.002)	0.143***	(0.003)	0.152***	(0.003)	0.164***	(0.004)	0.155***	(0.006)	0.124***	(0.010)
CR_other	0.047***	(0.001)	0.072***	(0.001)	0.090***	(0.001)	0.112***	(0.002)	0.131***	(0.003)	0.111***	(0.005)
CR_C_CODE_ind	0.009***	(0.001)	0.012***	(0.001)	0.014***	(0.001)	0.018***	(0.001)	0.018***	(0.002)	0.011***	(0.004)
Baseline SCORE	NA		NA		NA		NA		NA		NA	
R2 adj.	0.267		0.313		0.327		0.332		0.313		0.293	
Wald Chi2	333.07		388.638		447.139		324.038		194.927		147.865	
Wald Chi2 Prob.	0.000		0.000		0.000		0.000		0.000		0.000	
Observations	126,015		126,015		126,015		98,192		53,007		14,745	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 20: Regression results - log exposures

	CR					
			Sample			
	(A)		(B)		(C)	
log(EXP)	−0.006***	(0.001)	−0.015***	(0.002)	−0.048***	(0.001)
AGE	0.001	(0.001)	−0.0004	(0.002)	−0.005***	(0.001)
AGE_NA	−0.035***	(0.007)	0.022***	(0.005)	0.003	(0.005)
MALE	0.010***	(0.002)	−0.007*	(0.004)	0.024***	(0.002)
FIRM	−0.002	(0.008)	0.015*	(0.009)	0.006	(0.005)
INSOLV_ACQU	−0.208***	(0.007)	−0.159***	(0.013)	−0.032***	(0.004)
INSOLV_PROC	−0.327***	(0.007)	−0.263***	(0.018)	−0.111***	(0.004)
TEL	0.083***	(0.002)	0.110***	(0.005)	0.107***	(0.002)
D_SCORE_A	−0.034***	(0.010)			0.104***	(0.004)
D_SCORE_B	−0.075***	(0.009)			0.091***	(0.004)
D_SCORE_C	−0.089***	(0.009)			0.074***	(0.004)
D_SCORE_D	−0.112***	(0.009)			0.052***	(0.004)
D_SCORE_E	−0.129***	(0.006)			0.016***	(0.003)
D_SCORE_F	−0.171***	(0.006)			−0.014***	(0.004)
D_SCORE_G	−0.197***	(0.007)			−0.033***	(0.004)
D_SCORE_H	−0.231***	(0.006)			−0.073***	(0.004)
D_SCORE_I	−0.239***	(0.005)			−0.087***	(0.004)
D_SCORE_K	−0.255***	(0.005)			−0.106***	(0.003)
D_SCORE_L	−0.308***	(0.005)			−0.170***	(0.003)
D_SCORE_M	−0.323***	(0.006)			−0.148***	(0.003)
L_ORIG_ACQU	−0.053***	(0.001)	−0.035***	(0.004)	−0.007***	(0.001)
L_ORIG_ACQU_NA	−0.102***	(0.030)	−0.064***	(0.004)	−0.004*	(0.002)
END_missing	−0.133***	(0.006)			0.003	(0.002)
D_NO_MULTIPLE	−0.128***	(0.002)	−0.137***	(0.010)	0.106***	(0.002)
CR_other	0.164***	(0.001)	0.040***	(0.002)	0.063***	(0.001)
CR_C_CODE_ind	0.018***	(0.001)	0.004**	(0.002)	0.010***	(0.001)
Baseline SCORE	NA		—		NA	
R2 adj.	0.431		0.157		0.318	
Wald Chi2	1376.786		46.768		429.652	
Wald Chi2 Prob.	0.000		0.000		0.000	
Observations	182,880		16,623		126,015	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 21: Regression results - only single accounts

	CR					
			Sample			
	(A)		(B)		(C)	
EXP	−0.027***	(0.006)	−0.034***	(0.010)	−0.261***	(0.006)
AGE	−0.001	(0.002)	−0.0001	(0.004)	−0.017***	(0.002)
AGE_NA	−0.063***	(0.012)	0.051***	(0.011)	0.005	(0.012)
MALE	0.015***	(0.004)	−0.016*	(0.009)	0.058***	(0.004)
FIRM	0.046***	(0.015)	0.042*	(0.023)	0.002	(0.015)
INSOLV_ACQU	−0.318***	(0.012)	−0.338***	(0.027)	−0.098***	(0.009)
INSOLV_PROC	−0.423***	(0.012)	−0.564***	(0.038)	−0.270***	(0.012)
TEL	0.159***	(0.004)	0.243***	(0.012)	0.264***	(0.004)
D_SCORE_A	−0.066**	(0.027)			0.213***	(0.009)
D_SCORE_B	−0.206***	(0.022)			0.184***	(0.010)
D_SCORE_C	−0.225***	(0.022)			0.142***	(0.010)
D_SCORE_D	−0.287***	(0.021)			0.087***	(0.010)
D_SCORE_E	−0.317***	(0.014)			−0.007	(0.008)
D_SCORE_F	−0.390***	(0.016)			−0.094***	(0.010)
D_SCORE_H	−0.541***	(0.017)			−0.261***	(0.011)
D_SCORE_I	−0.565***	(0.015)			−0.301***	(0.010)
D_SCORE_K	−0.581***	(0.013)			−0.343***	(0.009)
D_SCORE_L	−0.688***	(0.013)			−0.513***	(0.009)
D_SCORE_M	−0.714***	(0.017)			−0.450***	(0.009)
L_ORIG_ACQU	−0.092***	(0.004)	−0.078***	(0.009)	−0.002	(0.003)
L_ORIG_ACQU_NA	−0.138**	(0.054)	−0.146***	(0.008)	−0.007	(0.005)
END_missing	−0.195***	(0.010)			−0.025***	(0.005)
CR_C_CODE_ind	0.033***	(0.002)	0.008*	(0.004)	0.024***	(0.002)
Baseline SCORE	NA		—		N	
R2 adj.	0.33		0.127		0.264	
Wald Chi2	243.914		35.105		132.72	
Wald Chi2 Prob.	0.000		0.000		0.000	
Observations	69,292		14,814		91,748	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 22: Regression results - individual and corporate debtors

	CR											
	Individual						Corporate					
	Sample						Sample					
	(A)		(B)		(C)		(A)		(B)		(C)	
EXP	−0.025***	(0.002)	−0.020***	(0.002)	−0.155***	(0.003)	0.004	(0.008)	−0.002	(0.004)	−0.056***	(0.011)
AGE	0.001	(0.001)	−0.00005	(0.002)	−0.009***	(0.001)						
AGE_NA	−0.033***	(0.007)	0.028***	(0.005)	0.004	(0.007)						
MALE	0.010***	(0.002)	−0.007	(0.004)	0.034***	(0.002)						
INSOLV_ACQU	−0.195***	(0.007)	−0.149***	(0.013)	−0.046***	(0.005)	−0.684***	(0.074)	−0.177***	(0.038)	−0.331***	(0.040)
INSOLV_PROC	−0.316***	(0.007)	−0.250***	(0.020)	−0.164***	(0.007)	−0.563***	(0.040)	−0.201***	(0.033)	−0.168***	(0.023)
TEL	0.085***	(0.002)	0.120***	(0.006)	0.158***	(0.002)	0.068***	(0.010)	0.025***	(0.009)	0.053***	(0.007)
D_SCORE_A	−0.035***	(0.010)			0.128***	(0.005)						
D_SCORE_B	−0.076***	(0.009)			0.111***	(0.006)						
D_SCORE_C	−0.089***	(0.008)			0.087***	(0.006)						
D_SCORE_D	−0.111***	(0.009)			0.054***	(0.006)						
D_SCORE_E	−0.129***	(0.006)			0.001	(0.005)						
D_SCORE_F	−0.170***	(0.006)			−0.048***	(0.005)						
D_SCORE_G	−0.196***	(0.006)			−0.075***	(0.006)						
D_SCORE_H	−0.230***	(0.006)			−0.136***	(0.006)						
D_SCORE_I	−0.237***	(0.005)			−0.158***	(0.005)						
D_SCORE_K	−0.253***	(0.005)			−0.185***	(0.005)						
D_SCORE_L	−0.305***	(0.005)			−0.280***	(0.005)						
D_SCORE_M	−0.321***	(0.006)			−0.248***	(0.005)						
L_ORIG_ACQU	−0.051***	(0.001)	−0.035***	(0.005)	−0.002	(0.002)	−0.062***	(0.007)	−0.019***	(0.004)	−0.015***	(0.004)
L_ORIG_ACQU_NA	−0.072**	(0.030)	−0.064***	(0.004)	−0.006**	(0.003)	−0.104	(0.187)	−0.032***	(0.009)	−0.025**	(0.011)
END_missing	−0.123***	(0.006)			−0.013***	(0.003)	−0.180***	(0.039)			−0.025**	(0.011)
D_NO_MULTIPLE	−0.132***	(0.002)	−0.139***	(0.010)	0.146***	(0.003)	−0.053***	(0.011)	−0.091***	(0.019)	0.185***	(0.007)
CR_other	0.159***	(0.001)	0.034***	(0.002)	0.085***	(0.001)	0.265***	(0.007)	0.061***	(0.008)	0.131***	(0.005)
CR_C_CODE_ind	0.018***	(0.001)	0.004**	(0.002)	0.014***	(0.001)	0.024***	(0.005)	0.001	(0.004)	0.012***	(0.003)
Base SCORE	NA		—		NA		—		—		—	
Adjusted R^2	0.433		0.153		0.328		0.435		0.282		0.334	
Wald Chi2	1283.637		42.682		372.639		4.955		5.742		5.308	
Wald Chi2 Prob.	0.000		0.000		0.000		0.026		0.017		0.021	
Observations	174,406		15,469		118,505		8,474		1,154		7,510	

Note: *p<0.1; **p<0.05; ***p<0.01

References

Aktas, N., E. de Bodt, F. Lobez, and J.-C. Statnik (2011). The information content of trade credit. *Journal of Banking & Finance 36*(5), 1402–1403.

Bellotti, T. and J. Crook (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting 28*(1), 171–182.

Biais, B. and C. Gollier (1997). Trade credit and credit rationing. *Review of Financial Studies 10*(4), 903–937.

Bijak, K. and L. C. Thomas (2015). Modelling LGD for unsecured retail loans using Bayesian methods. *Journal of the Operational Research Society 66*(2), 342–352.

Boissay, F. and R. Gropp (2013). Payment defaults and interfirm liquidity provision. *Review of Finance 17*(6), 1853–1894.

Brennan, M. J., V. Miksimovic, and J. Zechner (1988). Vendor financing. *The Journal of Finance 43*(5), 1127.

Buelow, S. (2016). Update Branchenstudie Inkasso: Neue Zahlen, Daten und Fakten ueber das Forderungsmanagement. *Zeitschrift fuer das Forderungsmanagement (zfm)* (06/2016).

Burkart, M. and T. Ellingsen (2004). In-kind finance: A theory of trade credit. *American Economic Review 94*(3), 569–590.

Calabrese, R. (2014). Downturn loss given default: Mixture distribution estimation. *European Journal of Operational Research 237*(1), 271–277.

Calabrese, R. and M. Zenga (2010). Bank loan recovery rates: Measuring and nonparametric density estimation. *Journal of Banking & Finance 34*(5), 903–911.

Caselli, S., S. Gatti, and F. Querci (2008). The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services Research 34*(1), 1–34.

Cuñat, V. (2007). Trade credit: Suppliers as debt collectors and insurance providers. *Review of Financial Studies 20*(2), 491–527.

Davydenko, S. A. and J. R. Franks (2008). Do bankruptcy codes matter? A study of defaults in France, Germany, and the U.K. *The Journal of Finance 63*(2), 565–608.

De Almeida Filho, A. T., C. Mues, and L. C. Thomas (2010). Optimizing the collections process in consumer credit. *Production and Operations Management 19*(6), 698–708.

Deloof, M. and M. Jegers (1996). Trade credit, product quality, and intragroup trade: Some European evidence. *Financial Management 25*(3), 33.

Dermine, J. and C. N. de Carvalho (2006). Bank loan losses-given-default: A case study. *Journal of Banking & Finance 30*(4), 1219–1243.

Ernst & Young (2014). The impact of third-party debt collection on the national and state economies in 2013. Technical report.

Fedaseyeu, V. (2015). Debt collection agencies and the supply of consumer credit. *Working paper, Research Department, Federal Reserve Bank of Philadelphia* (23).

Fedaseyeu, V. and R. M. Hunt (2015). The economics of debt collection: Enforcement of consumer credit contracts. *Working paper, Research Department, Federal Reserve Bank of Philadelphia* (43).

Ferris, J. S. (1981). A transactions theory of trade credit use. *The Quarterly Journal of Economics 96*(2), 243.

Fonseca, J., K. Strair, and B. Zafar (2017). Access to credit and financial health: Evaluating the impact of debt collection. *Staff Report, Federal Reserve Bank of New York* (814).

Gürtler, M. and M. Hibbeln (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance 37*(7), 2354–2366.

Han, C. and Y. Jang (2013). Effects of debt collection practices on loss given default. *Journal of Banking & Finance 37*(1), 21–31.

Hoechstoetter, M., A. Nazemi, S. T. Rachev, and C. Bozic (2012). Recovery rate modelling of non-performing consumer credit using data mining algorithms. *RMI Working Paper - National University of Singapore* (12/09).

Hoyer, M. (2011). *Entwicklung eines Ratingsystems für Inkassoforderungen: Ein Prognosemodell für die Rückzahlung zahlungsgestörter Forderungen aus Handel, Industrie und Gewerbe*. Wiesbaden: Springer Gabler.

Ingermann, P.-H., F. Hesse, C. Bélorgey, and A. Pfingsten (2016). The recovery rate for retail and commercial customers in Germany: A look at collateral and its adjusted market values. *Business Research 9*(2), 179–228.

Leow, M. and C. Mues (2012). Predicting loss given default (LGD) for residential mortgage loans: A two-stage model and empirical evidence for UK bank data. *International Journal of Forecasting 28*(1), 183–195.

Leow, M., C. Mues, and L. Thomas (2014). The economy and loss given default: Evidence from two UK retail lending data sets. *Journal of the Operational Research Society 65*(3), 363–375.

Long, M. S., I. B. Malitz, and S. A. Ravid (1993). Trade credit, quality guarantees, and product marketability. *Financial Management 22*(4), 117–127.

Loterman, G., I. Brown, D. Martens, C. Mues, and B. Baesens (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting 28*(1), 161–170.

Makuch, W. M., J. L. Dodge, J. G. Ecker, D. C. Granfors, and G. J. Hahn (1992). Managing consumer credit delinquency in the US economy: A multi-billion dollar management science application. *Interfaces 22*(1), 90–109.

Matuszyk, A., C. Mues, and L. C. Thomas (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society 61*(3), 393–398.

Mian, S. L. and C. W. Smith (1992). Accounts receivable management policy: Theory and evidence. *The Journal of Finance 47*(1), 169–200.

Mian, S. L. and C. W. Smith (1994). Extending trade credit and financing receivables. *Journal of Applied Corporate Finance 7*(1), 75–84.

Ng, C. K., J. K. Smith, and R. L. Smith (1999). Evidence on the determinants of credit terms used in interfirm trade. *The Journal of Finance 54*(3), 1109–1129.

Papke, L. E. and J. M. Wooldridge (1996). Econometric methods for fractional response variables with an application to 401 (K) plan participation rates. *Journal of Applied Econometrics 11*(6), 619–632.

Petersen, M. A. and R. G. Rajan (1997). Trade credit: Theories and evidence. *Review of Financial Studies 10*(3), 661–691.

Qi, M. and X. Yang (2009). Loss given default of high loan-to-value residential mortgages. *Journal of Banking & Finance 33*(5), 788–799.

Schwartz, R. A. (1974). An economic model of trade credit. *The Journal of Financial and Quantitative Analysis 9*(4), 643.

Thomas, L. C., A. Matuszyk, and A. Moore (2012). Comparing debt characteristics and LGD models for different collections policies. *International Journal of Forecasting 28*(1), 196–203.

Tong, E. N., C. Mues, and L. Thomas (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting 29*(4), 548–562.

Walter, A., T. Beck, J. Grunert, and W. Neus (2017). What determines collection rates of debt collection agencies? *Financial Review 52*(2).

Zhang, J. and L. C. Thomas (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting 28*(1), 204–215.