Self-Fulfilling Credit Scores

Victor Duarte
UIUC*

Julia Fonseca
UIUC and NBER†

July 2026

Abstract

A decline in credit scores can worsen the borrower's financial position, making default a self-fulfilling prophecy. We estimate the causal effect of credit scores on default using a rule that counts multiple inquiries within 14 days as one, moving scores without changing credit reports. Regression discontinuity estimates around the 14-day cut-off show that an additional counted inquiry lowers scores by five points on average. Default does not respond among consumers with clean records, but rises by 3.2 percentage points over two years among those with prior derogatories. For these consumers, at least 13 percent of the relationship between scores and default is self-fulfilling. We show that scorers would want to limit self-fulfilling defaults if their objective is forecast accuracy, but not if it is separating defaulters from non-defaulters, because that objective rewards the score for the defaults it causes.

*Gies College of Business, 1206 South 6th Street, Champaign IL, 61820. vduarte@illinois.edu

†Gies College of Business, 1206 South 6th Street, Champaign IL, 61820. juliaf@illinois.edu

1 Introduction

Consumers with the lowest credit scores are over 60 times more likely to become seriously delinquent than those with the highest scores.[1] Because credit scores predict future default, they are widely used by lenders when deciding who obtains a loan and on what terms. A decline in credit scores may therefore tighten access to credit and worsen the borrower's financial position, so that default becomes a self-fulfilling prophecy. If so, some of the large individual and societal costs of default are potentially avoidable consequences of credit scoring itself.

Estimating the causal effect of credit scores on default is challenging because it requires a change in scores that is orthogonal to other drivers of default. Ordinary movements in credit scores clearly do not meet this standard, as scores are designed to fall precisely when a borrower becomes more likely to default. Even plausibly exogenous changes to a borrower's credit report, such as the removal of bankruptcy flags and other derogatory information, do not meet it. Credit reports are used directly by lenders, employers, landlords, and insurers, and thus changing the content of a report can directly affect the borrower through any of these uses, and not solely through credit scores.

This paper uses a regression discontinuity (RD) design to estimate the causal effect of credit scores on default, exploiting a rule that changes credit scores but not credit reports. To avoid penalizing consumers shopping for the best rate on a loan, credit scoring models count inquiries for major credit products, such as mortgages and auto loans, made within a short window of time as a single inquiry. The shopping window is 14 days in many credit scoring models—including the VantageScore model whose scores we observe—and our empirical strategy compares consumers whose second inquiry falls just inside or just outside the window. The credit report lists both inquiries in either case, but the second inquiry lowers the score when it falls outside the window, and thus the comparison fully

[1] Approximately 62 percent of consumers with credit scores below 580 become seriously delinquent, defined as more than 90 days past due on a debt payment, compared with less than 1 percent of consumers with scores above 800 (Experian, 2026a,b).

isolates the effect of the score from that of the underlying information. Using data from the Gies Consumer and Small Business Credit Panel (GCCP), a state-of-the-art credit panel covering between two and three million consumers per year, we observe roughly 700,000 pairs of major inquiries near the 14-day cut-off between 2021 and 2025.

The additional counted inquiry lowers credit scores by 5.1 points on average. As we would expect, this small decline in scores has no effect on new serious delinquency one or two years ahead among the most creditworthy consumers. However, for consumers in the major derogatory scorecard—those with a derogatory such as a collection, charge-off, or bankruptcy in their record—the additional inquiry lowers the score by 6.3 points and raises the probability of default by 1.1 percentage points after one year and by 3.2 percentage points after two years, with the latter effect being statistically significant.

We then measure what share of the relationship between credit scores and default is self-fulfilling for these consumers by scaling the causal effect of the score on default by the calibration slope, the cross-sectional relationship between scores and default. Our estimates imply that at least 13 percent of the predictive relationship between scores and default is self-fulfilling for consumers in the major derogatory scorecard. The data thus rule out the view that scores merely measure default risk for this group, and cannot rule out that the entire predictive relationship is self-fulfilling.

We find no evidence of manipulation of inquiry timing around the 14-day cut-off, and our placebo and falsification tests support the validity of the RD design. Consumers on either side of the boundary are indistinguishable on a wide range of predetermined characteristics, including their credit scores and delinquency histories before the inquiry. We also see no discontinuities in contemporaneous measures of credit-seeking behavior, such as inquiry types, newly opened accounts per inquiry, and the number of lenders from which the inquiries originate. Placebo cut-offs at 7 and 21 days—which mirror differences in the timing of applications but eliminate the difference in inquiry de-duplication—show no discontinuities, with confidence intervals that exclude our baseline estimates. Inquiry types that credit scoring models do not de-duplicate, such as retail cards and utilities,

also do not exhibit discontinuities at the cut-off.

We then ask whether a credit scorer would want to reduce self-fulfilling defaults by dampening adverse score movements. In a model of credit scoring with causal feedback, we show that the answer hinges on the scorer's objective function. A scorer evaluated on the accuracy of its default probabilities can dampen adverse revisions in a way that both lowers default and improves accuracy—a strict Pareto improvement. However, a scorer evaluated on how well the score separates defaulters from non-defaulters cannot, because the score is rewarded for correctly ranking the very defaults it causes. In practice, credit scoring models are typically validated on separation metrics, suggesting that credit scoring companies have no incentive to limit the self-fulfilling defaults we document.

Our study contributes to a largely theoretical literature on the feedback effects of credit ratings, in which a rating affects the outcome it is designed to predict. In Manso (2013), a downgrade raises the borrower's financing costs and with them the probability of default, so a rating agency concerned with accuracy should rate leniently to account for the effects of its own ratings. Related work shows that ratings can select among equilibria (Boot, Milbourn, and Schmeits, 2006), that rating agencies may optimally inflate ratings because of this feedback (Goldstein and Huang, 2020), and that corporate downgrades causally reduce investment and financing (Almeida, Cunha, Ferreira, and Restrepo, 2017). In consumer credit, Chatterjee, Corbae, Dempsey, and Ríos-Rull (2023) develop an equilibrium model in which the credit score is self-confirming, and Perdomo, Zrnic, Mendler-Dünner, and Hardt (2020) study prediction models whose deployment changes the outcomes they predict. The closest empirical evidence comes from Purnanandam and Wirth (2025), who estimate the effect of a pandemic-era suspension of delinquency reporting on subsequent default among borrowers of an online lending platform. We make two contributions to this literature, one empirical and one theoretical. First, we estimate the causal effect of credit scores on default, holding the contents of credit reports constant. We thus provide the first causal evidence of the score-to-default feedback effect that models in this literature often assume. Second, we show that the case for lenient ratings depends on

the scorer's objective function. Dampening adverse score movements improves forecast accuracy under proper scoring rules but not under the separation metrics used to validate credit scoring models in practice (Thomas, Crook, and Edelman, 2017; Lessmann, Baesens, Seow, and Thomas, 2015), because separation metrics reward the score for the defaults it causes.

We also relate to research on the effects of credit-report information on consumer outcomes. Prior work has studied the removal of information about past bankruptcies, debt collections, delinquencies, and unpaid debts from credit reports, and its effects on credit access and financial health (e.g., Musto, 2004; Dobbie, Keys, and Mahoney, 2017; Gross, Notowidigdo, and Wang, 2020; Duarte, Fonseca, Kohli, and Reif, 2025), labor market outcomes (Bos, Breza, and Liberman, 2018; Dobbie, Goldsmith-Pinkham, Mahoney, and Song, 2020; Herkenhoff, Phillips, and Cohen-Cole, 2021), and aggregate credit and welfare (Liberman, Neilson, Opazo, and Zimmerman, 2018; Jansen, Nagel, Yannelis, and Zhang, 2025). These studies measure the total effect of removing information, which is the relevant object for credit reporting policy and combines the effect of credit scores with those of all direct uses of the credit report, including by lenders, employers, landlords, and insurers. In contrast, our research question—whether credit scores cause default—requires a change in credit scores holding credit-report information fixed, and ours is the only instrument that we know of that provides such variation.

Finally, we relate to the literature on machine learning and credit scoring models. Prior studies have used machine learning to develop credit scoring models (e.g., Khandani, Kim, and Lo, 2010; Albanesi and Vamossy, 2019; Meursault, Moulton, Santucci, and Schor, 2022; Agarwal, Alok, Ghosh, and Gupta, 2023; Chioda, Gertler, Higgins, and Medina, 2024), to measure the accuracy of scores across borrowers (Blattner and Nelson, 2021; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2022), and to evaluate the predictive content of credit-report information (Duarte, Fonseca, Kohli, and Reif, 2025). Like Fuster, Goldsmith-Pinkham, Ramadorai, and Walther (2022), we study the economic consequences of the scoring technology rather than its predictive performance. We show

that the predictive performance of credit scores partly reflects the score's own effect on default, implying that standard validation metrics overstate the information credit scoring models extract.

The remainder of the paper proceeds as follows. Section 2 provides background on the de-duplication of credit inquiries and describes the data. Section 3 presents the regression discontinuity design and the empirical results. Section 4 develops a model of credit scoring with causal feedback. Section 5 concludes.

2 Background and Data

2.1 Background: inquiry de-duplication

When a consumer applies for credit, the lender's request for the credit report is recorded as a hard inquiry, and recent inquiries enter scoring models as a signal of new credit demand. Credit scoring models de-duplicate inquiries that arrive close together, so as not to treat consumers who are shopping around for a favorable rate as if they are seeking multiple loans (e.g., VantageScore, 2025). We focus on the de-duplication rule of the VantageScore model, which provides the credit scores we observe in our data. The model counts multiple major credit inquiries—such as mortgage, auto, installment, and bankcard—within a 14-day window as one inquiry regardless of the type of major credit product, while retail, collection, and utility inquiries are not de-duped (VantageScore, 2022). FICO also de-duplicate inquiries, with older models following a similar 14-day rule as the VantageScore and newer models de-duplicating over a 45-day window (FICO, 2025; Experian, 2024).

A consumer who applies for two loans 13 days apart and one who applies 15 days apart present the same information to lenders and other users of credit reports—two inquiries on the report—but the scoring model counts one inquiry for the first consumer and two for the second. Crossing the 14-day cutoff thus changes an input to the credit score but not the contents of the credit report. The public documentation is unclear about how inquiries exactly 14 days apart are treated, and whether their de-duplication depends

on the difference in calendar dates or on the elapsed time between the inquiries. As we discuss in Section 3, this ambiguity motivates excluding inquiries with a 14-day gap from our estimation sample, following Barreca, Lindo, and Waddell (2016).

2.2 Data

Our data come from the Gies Consumer and Small Business Credit Panel (GCCP), a panel of anonymized credit records from one of the three major credit bureaus. The GCCP features a one-percent random sample of individuals with a credit report, with annual snapshots as of the end of the first quarter of each year from 2004 to 2025. Sampling is based on the last two digits of Social Security numbers, which accounts for natural flows into and out of the panel and keeps the sample representative of the population with a credit record over time. Each record includes all reported debt obligations, or tradelines, with balances, limits, and payment history; debts in collections; public records such as bankruptcies; the consumer's credit score from the VantageScore model; and demographic variables such as age, sex, a bureau-provided income estimate, and zip code. We focus on inquiries from the post-pandemic 2021–2025 snapshots, which cover between 2.9 and 3.0 million consumers per year.

Central to our design, the GCCP records individual hard inquiries with their exact dates, the type of credit sought, and the kind of business that pulled the report. In each year, we observe between 2.6 and 3.3 million inquiries for roughly one million consumers. We classify each inquiry by whether the de-duplication rule applies to it: mortgage, auto, installment, and bankcard inquiries are the "major" types the rule covers, while retail, utility, and collection inquiries are the types its documentation excludes. Major types account for 87 to 95 percent of inquiries, and the non-de-duplicated types are reserved for a placebo check (Section 3).

We construct the estimation sample as follows. For each consumer and year, we group major inquiries according to the de-duplication rule: the first inquiry opens a window, and subsequent inquiries within 14 days of the initial inquiry are de-duplicated. The first

inquiry outside the initial 14-day window starts a new window. The running variable r_i is the number of days between a consumer's initial inquiry and the next major inquiry, which can fall inside or outside the 14-day window. We take each consumer's earliest spell with r_i between 6 and 22 days in a given year, yielding one observation per consumer-year, and require a valid credit score in both the year of the inquiry and the year before so that we can measure a change in scores. The resulting sample has 700,962 observations for major inquiries and 12,807 for the never-de-duplicated placebo types.

Our main outcome is new serious delinquency, measured as an indicator for any account reported 90 or more days past due during the 12 months ending at horizon $t + h$, for h equal to one or two years. This is the outcome definition and window that scoring models themselves are built to predict. Table 1 compares our estimation sample to the full GCCP. Consumers with multiple major inquiries in a short window are active credit seekers, and they are younger and riskier than the average consumer with a credit record. They carry lower credit scores (654 versus 702 on average), higher revolving utilization (41 versus 28 percent), and more newly opened accounts, and their incidence of new serious delinquency in the previous 12 months is twice that of the full panel (18.8 versus 9.4 percent). Our estimates are therefore local to this population, but it is a large one, accounting for about one in twenty consumers with a credit score in a given year, or roughly 14 million people nationwide.

Finally, we assign each consumer to one of three scorecards based on $t - 1$ information, mirroring the industry practice of scoring distinct segments with separate models. While the model in Section 4 has only two scorecards, for higher- and lower-risk consumers, we adopt the common practice of further separating consumers with thin files or short credit histories, as such files carry the noisiest scores and default predictions (Blattner and Nelson, 2021; Duarte, Fonseca, Kohli, and Reif, 2025), and we thus expect to be underpowered for this group. Our thin-young scorecard features consumers with fewer than three tradelines or a file younger than 36 months. Among the remainder, consumers with a prior major derogatory—a collection, a bankruptcy or other public record, or a

charge-off on file—form the second scorecard, and consumers with clean files form the third. Appendix Table A.1 reports summary statistics by scorecard. Consumers in the major derogatory scorecard show the most financial distress, with a 32.8 percent incidence of new serious delinquency in the previous 12 months, against 3.5 percent for clean files and 24.2 percent for thin and young files, whose credit histories are shortest and scores lowest.

3 Regression Discontinuity Design and Results

3.1 Empirical strategy

Our research design compares consumers whose second credit inquiry falls just outside the 14-day de-duplication window to consumers whose second inquiry falls just inside it. The running variable is r_i, the number of days between consumer i's initial inquiry and the next inquiry, as described in Section 2.2. When the gap is 13 days or fewer, the scoring model counts the two inquiries as one; when it is 15 days or more, it counts them separately. As we discuss in Section 2.1, a gap of exactly 14 days is ambiguous under the rule's public description, which does not specify whether the window is inclusive and whether it depends on the difference in calendar dates or on the elapsed time between the inquiries. We thus exclude 14-day gaps from the estimation sample, following Barreca, Lindo, and Waddell (2016).

We estimate the following specification:

$$Y_i \;=\; \alpha \;+\; \beta\,\mathrm{ABOVE}_i \;+\; \gamma\,(r_i - 14) \;+\; \delta\,\mathrm{ABOVE}_i \times (r_i - 14) \;+\; \varepsilon_i, \tag{1}$$

where Y_i is the change in the consumer's credit score from $t-1$ to t or a delinquency outcome at horizon $t+h$, and ABOVE_i indicates $r_i > 14$. We estimate Equation (1) with a triangular kernel on observations within h days of the cutoff, excluding day 14. We demean outcomes by sample year to absorb cohort effects. The parameter β captures the effect of having the second inquiry counted separately.

We measure the effect of the credit score decline on default as the Wald ratio:

$$\tau_H \;=\; \frac{\beta^D}{-\,\beta^S}, \tag{2}$$

where β^S is the effect of the additional counted inquiry on credit scores (the first stage) and β^D effect on delinquency (the reduced form), each the estimate of β from Equation (1). Because β^S is negative, τ_H measures the change in the probability of default, in percentage points, per one-point decline in the credit score, so a positive τ_H means that a decline in the score raises default.

The self-fulfilling share of defaults expresses this causal effect per unit of the default risk the score itself predicts,

$$\phi \;=\; \frac{\tau_H}{m_H}, \tag{3}$$

where the calibration slope m_H is the predicted default risk that one score point carries at the in-bandwidth average score $\bar{S}$, measured as the slope of the fitted score-to-default curve. In the notation of the model in Section 4, $m_H = -\mu'_H(\bar{S})$, and ϕ is the causal feedback parameter the model takes as its empirical primitive.

The identifying assumption behind the RD is that assignment to either side of the cut-off is as good as random. We support this assumption with falsification tests on predetermined outcomes, with covariate smoothness tests, and by examining the distribution of the running variable around the cutoff. We interpret the effect of the additional counted inquiry on future default as the causal effect of the decline in scores, which requires the exclusion restriction that the counted inquiry affects default only through credit scores. This assumption is natural in our setting, as the de-duplication affects the credit score while leaving the underlying information unchanged, with the same two hard inquiries visible to lenders and other users of credit reports. We address concerns that a difference of a few days between the two inquiries can directly affect outcomes with placebo checks at 7 and 21 days, mirroring the difference between treatment and control consumers in the timing of applications but eliminating the difference in credit scores. These placebo checks

show no discontinuities at the cut-off and are powered to reject our baseline estimates.

Standard errors are clustered at the consumer level. Since the running variable is discrete, we follow Kolesár and Rothe (2018) and report fixed-length confidence intervals (Armstrong and Kolesár, 2018) that account for the worst-case bias of the local linear estimator over a Hölder smoothness class, with a side-specific curvature bound as in Armstrong and Kolesár (2020). These intervals are honest in the sense of Li (1989), meaning that they achieve correct coverage uniformly over all conditional expectation functions in the class. For the Wald ratio of Equation (2), we report the bias-aware confidence intervals of Noack and Rothe (2024), which test candidate values of the ratio directly rather than relying on delta-method approximations. Because they guard against the worst-case bias allowed by the smoothness class, these confidence intervals are wider than the robust bias-corrected intervals typically used in RD designs with continuous running variables (Calonico, Cattaneo, and Titiunik, 2014), so our inference is conservative relative to standard practice. The optimal bandwidth is the one that minimizes the length of the honest confidence interval (Armstrong and Kolesár, 2020) and is allowed to vary by outcome.

3.2 Results

We start by estimating the effect of the additional counted inquiry on credit scores. Panel (a) of Figure 1 shows the effect in the full sample, where the marginal counted inquiry lowers credit scores by 5.1 points (column 1 of Table 2). The magnitude matches public documentation regarding the impact of inquiries on credit scores, which puts the effect of a single additional inquiry at five to ten points (Experian, 2024). Table 2 shows that the decline in scores is present in every scorecard: 6.3 points for borrowers with prior major derogatories (also shown in panel (b) of Figure 1), 6.7 for clean files, and a weaker 3.0 for thin and young files. As we expected, we are underpowered in the thin-young scorecard, where the first stage F-statistic is 5.8, compared to 22.7 and 46.3 in the major derogatory and clean file scorecards, respectively.

The modest decline in credit scores does not increase default rates for the average con-

sumer. At the one-year horizon, the effect on new serious delinquency is 0.10 percentage points (column 1 of Appendix Table A.2), and at the two-year horizon it is 0.16 percentage points (column 1 of Table 3). Panel (a) of Figure 2 makes this point visually, showing no evidence of a discontinuity in $t + 2$ default for the full sample.

However, the average effect masks considerable heterogeneity, with consumers with prior histories of serious default being more responsive to the credit-score shock. For the major derogatory scorecard, the two-year effect on default is 3.19 percentage points (column 2 of Table 3)—building from a statistically insignificant increase of 1.07 percentage points after one year (Table A.2)—and is visible as a discontinuity in panel (b) of Figure 2. We find no effect in the other two scorecards (column 3–4 of Table 3), though the weaker score response in the thin-young segment limits what our design can say about these borrowers.

Next, we estimate what share of predicted defaults are self-fulfilling in the major derogatory scorecard, according to Equation (3). We collect the inputs into this share in Table 4. For the major derogatory scorecard, the Wald ratio of Equation (2) is $\tau_H = 0.50$ percentage points of additional default per point of score decline, and the calibration slope at the in-bandwidth average score of $\bar{S} = 610$ is $m_H = 0.465$ percentage points of predicted default per point (Appendix Figure A.9). The resulting point estimate for the self-fulfilling share is thus $\phi = 1.08$, implying that, at the margin we identify, roughly all of the score's predictive slope for borrowers in the major derogatory scorecard is self-fulfilling (Table 4).

Our confidence intervals for the self-fulfilling share are wide, and we are thus underpowered to pin down the exact magnitude of this effect. We report the confidence interval from a clustered bootstrap that resamples consumers and re-estimates every input to the ratio in each of 1,000 draws, which suggests that the self-fulfilling share is at least 32 percent (Table 4). This confidence interval reflects only sampling uncertainty, and thus we also report a Noack and Rothe (2024) bias-aware confidence interval. This worst-case bias interval also excludes zero, so even the most conservative reading of our estimates implies that at least 13 percent (or roughly one eighth) of the score's predictive content for these borrowers reflects defaults caused by the score itself. The results in Section 4 depend only

on the share being bounded away from zero and not on its magnitude, and thus these findings provide the empirical primitive that our model takes as input.

3.3 Robustness

We evaluate the no-manipulation assumption by testing whether predetermined outcomes are continuous at the cutoff. Figure A.1 shows no statistically significant discontinuities in income, age, or sex, measured in $t-1$. Figure A.2 goes further by assessing contemporaneous (time t) measures of credit-seeking behavior. We find no discontinuities in whether the inquiries that define the running variable are of the same major type (e.g., for a mortgage loan, an auto loan, or a bankcard), indicating that rate-shopping versus seeking-multiple-loans behavior does not change at the cut-off. We also show no discontinuities in the number of accounts opened per inquiry, which is a function of both application success and credit demand. Finally, we proxy for the number of lenders from which inquiries originate as unique combinations of industry and subscriber codes, and find no discontinuities in that measure.

We also examine the distribution of the running variable. Because the running variable is discrete, the standard McCrary (2008) test, which estimates the density of a continuous running variable on each side of the cutoff, does not apply, and we instead examine the distribution of day gaps directly (Frandsen, 2017) in Appendix Figure A.3. The distribution bunches at weekly multiples of 7, 14, 21, and 28 days, reflecting the day-of-week rhythm of credit applications, and the excess mass at day 14 is statistically indistinguishable from the excess mass at the weekly days the scoring rule attaches no meaning to (p-value = 0.38). The standard response to bunching of this kind is a donut specification that drops observations at the cutoff (Barreca, Lindo, and Waddell, 2016), which we already do, since inquiries with a 14-day gap are excluded because their treatment under the de-duplication rule is ambiguous (Section 2.1).

We also conduct a series of falsification and placebo tests. Appendix Figures A.4 and A.5 replicate our RD specification using predetermined outcomes, including credit scores, bal-

ances, delinquencies, collections, bankruptcies, and past inquiries, all measured at $t-1$. As expected, we find no evidence of discontinuities. Appendix Figures A.6 and A.7 reproduce our RD at placebo cut-offs of 7 and 21 days, respectively. In these specifications, the difference in the timing of applications mirrors that of our design, but there are no differences in inquiry de-duplication across treatment and control groups, thus addressing concerns that a difference of a few days between inquiries can directly affect outcomes. Honest confidence intervals across both placebo cut-offs exclude our statistically significant estimates, including $t+2$ delinquency for the major derogatory scorecard (column 4, Appendix Tables A.5 and A.6). Finally, Appendix Figure A.8 repeats our analysis for retail, utility, and collection inquiries, which the scoring model does not de-duplicate, and likewise finds no effects on credit scores or subsequent delinquency, with point estimates of the opposite sign as our baseline but with wide confidence intervals. Together, these results show that the 14-day cutoff is not associated with breaks in predetermined outcomes, at cut-offs not associated with de-duplication, or in inquiry types the rule does not de-duplicate, supporting the internal validity of our RD design.

4 A Model of Credit Scoring

Our estimates imply that, for borrowers with prior derogatories, a lower credit score causes part of the default it predicts. This section develops a parsimonious model of what this causal feedback implies for the credit scorer. The model treats the causal effect of scores on default as an empirical primitive—the parameter ϕ we estimate in Section 3—and asks whether a scorer who could dampen adverse score movements would want to.

We show that the answer hinges entirely on the scorer's objective function. If the scorer evaluates the score as a numerical default-probability forecast, dampening adverse revisions can be a strict Pareto improvement, lowering default and improving measured forecast accuracy. If the scorer instead evaluates how well the score orders borrowers by realized default, no such improvement exists. Because the score causes defaults among borrowers it already ranks as risky, the score is rewarded for correctly ranking defaults it

helped create.

4.1 Environment

There are two periods. In period 1, borrower i has credit-report state x_i, which summarizes the information available to scoring models, such as payment history, balances, utilization, account age, recent inquiries, and public records. Default $Y_i \in \{0, 1\}$ is realized in period 2. Borrowers belong to one of two observed types, $T_i \in \{H, L\}$. The credit scorer observes (x_i, T_i) but is naive about the causal effect of scores on default. It learns equilibrium default probabilities under the status quo score and does not internalize that the score partly generates those defaults.

The naive credit score is $s_i^N = s^N(x_i, T_i) \in \mathcal{S}_T$, where higher scores indicate lower predicted default risk. The score is associated with a type-specific probability scale $\mu_T : \mathcal{S}_T \to (0, 1)$, continuously differentiable and strictly decreasing, so that $\mu_T(s)$ is the default probability forecast encoded by score s for type T. We write $p_i^N = \mu_{T_i}(s_i^N)$ for the status quo forecast.

Let $a_i \geq 0$ denote an adverse revision to the naive score, measured on the probability scale: a_i is the increase in the default-probability forecast implied by the naive rule's downward revision of borrower i's score, with $a_i = 0$ for borrowers whose score is not revised downward. Adverse revisions occur in the affected segment,

$$\mathbb{E}_H[a_i] > 0, \tag{4}$$

where $\mathbb{E}_H[\cdot]$ denotes expectation conditional on $T_i = H$.

We consider whether the scorer has an incentive to smooth adverse revisions to credit scores. A smoothing rule with parameter $\lambda_T \in [0, 1]$ replaces the naive forecast with

$$p_i(\lambda_T) = p_i^N - \lambda_T a_i, \tag{5}$$

reported as the credit score $s_i(\lambda_T) = \mu_{T_i}^{-1}\big(p_i(\lambda_T)\big) \geq s_i^N$. Setting $\lambda_T = 0$ reproduces the naive score, while $\lambda_T = 1$ fully removes the adverse revision. To match our empirical design, which identifies the effect of adverse score movements and has no corresponding positive score shock, we assume that score improvements have no effect on defaults and are not dampened.

Under the status quo rule, the naive score is calibrated to the equilibrium conditional mean:

$$\mathbb{P}(Y_i = 1 \mid x_i, T_i, \lambda_T = 0) = p_i^N. \tag{6}$$

Calibration does not mean the naive score is structurally correct, only that it is correct for the default distribution generated by its own deployment.

Motivated by our empirical findings, we assume scores have no causal effect on default for type L. For type H, credit score declines raise future default, so smoothing those declines lowers default relative to the no-smoothing counterfactual.

Let $\phi > 0$ denote the causal feedback parameter, the reduction in the true default probability per unit reduction in the forecast induced by smoothing. For an H borrower, under smoothing $\lambda \equiv \lambda_H$,

$$q_i(\lambda) \equiv \mathbb{P}(Y_i = 1 \mid x_i, T_i = H, \lambda) = p_i^N - \phi\lambda a_i, \tag{7}$$

while for an L borrower $\mathbb{P}(Y_i = 1 \mid x_i, T_i = L, \lambda_L) = p_i^N$. We assume throughout that $p_i(\lambda_T)$ remains in the range of μ_{T_i} and that $q_i(\lambda_T) \in [0, 1]$ for all $\lambda_T \in [0, 1]$. Because borrower type is observed and there is no causal feedback for L borrowers, the scorer sets $\lambda_L = 0$, and we focus on $\lambda \equiv \lambda_H$. The expected default rate of H borrowers is $D_H(\lambda) = \mathbb{E}_H[q_i(\lambda)]$. Borrowers incur a utility cost when they default, so borrower welfare is ordered by $D_H(\lambda)$.

The credit scorer's objective. The scorer evaluates a scoring rule through an objective $\mathcal{O}(\lambda) = \mathcal{O}\big(\{s_i(\lambda)\}, \{q_i(\lambda)\}\big)$. A smoothing rule is a Pareto improvement over the

naive rule if it weakly improves the scorer's objective and strictly lowers expected borrower default costs. For objectives written as losses, lower values are better; for objectives written as performance metrics, higher values are better. Our results hinge on the scorer's objective, and we consider two classes of objectives in turn.

4.2 Case 1: The objective is probability-forecast accuracy

The first class of objectives evaluates the numerical probability forecast encoded by the score. Metrics in this class depend on the levels of the predicted probabilities and not only on how they rank borrowers, rewarding a score whose borrowers default at rates compatible with its predictions.

Definition 1 (Probability-forecast loss)**.** *A binary probability-forecast loss is a function* $\ell : (0,1) \times \{0,1\} \to \mathbb{R}$*. If a borrower defaults with probability* q *and the score reports default probability* p*, the expected loss is*

$$L(p,q) = q\ell(p,1) + (1-q)\ell(p,0). \tag{8}$$

The loss is standard *on a set* $\mathcal{P} \subset (0,1)$ *if* L *is continuously differentiable in a neighborhood of* $\{(p,p) : p \in \mathcal{P}\}$ *and satisfies two conditions. First, it is locally proper:*

$$L_p(p,p) = 0 \quad \textit{for all } p \in \mathcal{P}. \tag{9}$$

Second, defaults are the more costly forecast error:

$$L_q(p,p) = \ell(p,1) - \ell(p,0) > 0 \quad \textit{for all } p \in \mathcal{P}. \tag{10}$$

Lower expected loss represents better measured probability-forecast accuracy.

Local propriety says that, holding the outcome distribution fixed, reporting the true default probability is locally optimal. It holds for any proper scoring rule that is differentiable in the forecast. Condition (10) says that a realized default is the worse outcome

for measured accuracy. At any forecast in $\mathcal{P}$, the loss recorded when a borrower defaults exceeds the loss recorded when that borrower repays, so fewer defaults improve the measured accuracy of a given forecast. The two conditions are linked through the generalized entropy of the loss, $G(p) = L(p, p)$: local propriety gives $G'(p) = L_q(p, p)$, so condition (10) requires the entropy to be strictly increasing on $\mathcal{P}$.

For a probability-forecast loss, the credit scorer's objective is $\mathcal{L}_H(\lambda) = \mathbb{E}_H[L(p_i(\lambda), q_i(\lambda))]$, where lower values are better. Our first theorem shows that if the self-fulfilling share is greater than zero ($\phi > 0$), a strict Pareto improvement exists under this objective.

Theorem 1 (Probability-forecast loss)**.** *Suppose the population of H borrowers is finite and the credit scorer evaluates scores using a probability-forecast loss that is standard on a set $\mathcal{P}$ in the sense of Definition 1. Suppose $\phi > 0$ and*

$$p_i^N \in \mathcal{P} \quad \text{for every } H \text{ borrower with } a_i > 0. \tag{11}$$

Then, with $\lambda_L = 0$, there exists $\bar{\lambda} \in (0, 1]$ such that every $\lambda \in (0, \bar{\lambda}]$ satisfies

$$D_H(\lambda) < D_H(0) \quad \text{and} \quad \mathcal{L}_H(\lambda) < \mathcal{L}_H(0). \tag{12}$$

Under a probability-forecast loss, a small amount of smoothing is therefore a Pareto improvement, strictly lowering default for H borrowers and strictly improving the credit scorer's measured forecast accuracy.

The proof, in Appendix B.1, is an envelope argument. At a calibrated score, the forecast is locally optimal, so the accuracy cost of reporting a slightly more optimistic forecast is second order. The gain is first order, because smoothing impacts defaults. With $\phi > 0$, the true default probability of smoothed borrowers falls, and when defaults are the more costly forecast error, fewer defaults mean lower measured loss.

4.3 Case 2: The objective is rank-separation

The second class of objectives evaluates how well the score orders borrowers by realized default status. These objectives do not require the numerical score to be a correct default probability, but rather reward the separation between the score ranks of realized defaulters and non-defaulters. Examples of rank-separation metrics include the area under the Receiver Operating Characteristic curve (AUC) and Kolmogorov-Smirnov (KS).

Throughout, the metric is computed within the affected segment H. For a smoothing rule λ, let

$$F_1^\lambda(s) = \mathbb{P}(s_i^N \leq s \mid Y_i = 1, T_i = H; \lambda), \qquad F_0^\lambda(s) = \mathbb{P}(s_i^N \leq s \mid Y_i = 0, T_i = H; \lambda) \quad (13)$$

be the score-rank distributions among realized defaulters and non-defaulters in the H population. A better ranking has defaulters shifted to lower scores and non-defaulters shifted to higher scores.

Definition 2 (Rank-separation metric)**.** *A metric $\mathfrak{R}(F_1, F_0)$ is a rank-separation metric if it depends on scores only through their ranks—it is invariant to strictly increasing transformations of the score applied to both distributions—and it is weakly increasing in score-rank separation: whenever*

$$\widetilde{F}_1(s) \geq F_1(s) \quad \textit{and} \quad \widetilde{F}_0(s) \leq F_0(s) \quad \textit{for all } s, \qquad (14)$$

then $\mathfrak{R}(\widetilde{F}_1, \widetilde{F}_0) \geq \mathfrak{R}(F_1, F_0)$.

The first input into our second theoretical result is that small smoothing preserves the ranking. Because smoothing is proportional to each borrower's adverse revision a_i, the reported score $s_i(\lambda)$ need not be a monotone transformation of s_i^N at every λ: a borrower with a large adverse revision can overtake one with a small revision once λ is large enough. Lemma 1 shows that this cannot happen below an explicit threshold.

Lemma 1 (Rank preservation under small smoothing)**.** *Suppose the population of H*

borrowers is finite with distinct naive scores. Define

$$\bar{\lambda}_{\text{rank}} \;=\; \min\left\{ \frac{p_i^N - p_j^N}{a_i - a_j} \;:\; p_i^N > p_j^N \;\; \textit{and}\; a_i > a_j \right\}, \tag{15}$$

with $\bar{\lambda}_{\text{rank}} = +\infty$ *when no such pair of borrowers exists. Then* $\bar{\lambda}_{\text{rank}} > 0$*, and for every* $\lambda \in [0,1]$ *with* $\lambda < \bar{\lambda}_{\text{rank}}$ *the smoothed score is a strictly increasing transformation of the naive score:* $s_i(\lambda) = \varphi_\lambda(s_i^N)$ *for some strictly increasing* φ_λ*, so the ranking of borrowers by score is unchanged.*

The proof is in Appendix B.2. Smoothing moves a pair of borrowers together only when the difference in adverse revisions outruns the difference in naive forecasts, and $\bar{\lambda}_{\text{rank}}$ is the smoothing weight at which the closest such pair first ties. Below it, no pair ties or reverses.

The second input into our second theoretical result is the assumption that the causal feedback from scores is rank-aligned, formalized as Assumption 1.

Assumption 1 (Rank-aligned causal feedback)**.** *Let* F_C *be the score distribution of the marginal score-induced defaults:*

$$F_C(s) = \frac{\mathbb{E}_H[a_i \mathbf{1}\{s_i^N \le s\}]}{\mathbb{E}_H[a_i]}. \tag{16}$$

The marginal score-induced defaults are concentrated at lower scores than the realized default and non-default distributions under the naive rule:

$$F_C(s) \ge F_1^0(s) \quad \textit{and} \quad F_C(s) \ge F_0^0(s) \quad \textit{for all } s. \tag{17}$$

Assumption 1 says that the defaults the score creates arise among borrowers the score already ranks as high risk. Under the status quo calibration, realized defaulters are already concentrated at low scores, so (17) requires adverse revisions to be concentrated at low scores even more heavily than defaults themselves. Whether feedback is rank-

aligned is an empirical question, and we show evidence that this assumption holds in our setting in Appendix Figure A.10. The figure plots the risk-rank distribution of adverse score revisions in the major derogatory scorecard (our empirical proxy for H) against the distributions of realized defaulters and non-defaulters. The revision-weighted distribution lies to the right of both at every point, so the score-induced marginal defaults sit among borrowers the score already ranks riskiest, as required by Assumption 1.

Our second theorem shows that if the self-fulfilling share is greater than zero ($\phi > 0$) and the causal effect on default is rank-aligned, smoothing cannot improve the rank-separation objective.

Theorem 2 (Rank-separation metric)**.** *Suppose the population of H borrowers is finite with distinct naive scores, $\phi > 0$, and Assumption 1 holds. Let $\bar{\lambda}_{\text{rank}} > 0$ be the rank-preservation threshold of Lemma 1, and let the credit scorer's objective be a rank-separation metric $\mathfrak{R}(F_1^\lambda, F_0^\lambda)$ in the sense of Definition 2, computed within the H segment. Then, with $\lambda_L = 0$, every $\lambda \in (0, 1]$ with $\lambda < \bar{\lambda}_{\text{rank}}$ and $D_H(\lambda) > 0$ satisfies*

$$D_H(\lambda) < D_H(0) \quad \textit{but} \quad \mathfrak{R}(F_1^\lambda, F_0^\lambda) \leq \mathfrak{R}(F_1^0, F_0^0)\,: \tag{18}$$

smoothing strictly lowers default but cannot improve the rank-separation objective.

The proof, in Appendix B.3, works through the realized default distribution. Smoothing removes default mass from each borrower in proportion to the adverse revision a_i, so the removed mass is distributed across scores according to F_C. Under Assumption 1, that mass sits at the bottom of the ranking, so its removal shifts the remaining defaulters toward safer scores and the non-defaulters toward riskier scores, weakly compressing rank separation. In other words, rank-separation metrics reward the score for the defaults it causes.

Note that Theorem 2 still allows a Pareto improvement if smoothing leaves the rank-separation objective exactly unchanged. The theorem cannot rule this case out for the general class of rank-separation metrics because smoothing compresses separation at spe-

cific scores, and a metric can be exactly unchanged if it puts no weight on those scores. For instance, KS measures only the largest gap between the two distributions, so it does not move when compression occurs at other scores. Corollary 1 rules out a Pareto improvement for commonly used rank-separation metrics, giving conditions under which AUC, KS, and top-group recall and lift strictly decline with smoothing.

Corollary 1 (Common separation metrics)**.** *Under the conditions of Theorem 2, each of AUC, KS, recall-at-top-p, and top-p lift strictly falls for any* $\lambda \in (0, 1]$ *with* $\lambda < \bar{\lambda}_{\text{rank}}$ *and* $D_H(\lambda) > 0$*, provided the corresponding strictness condition holds: for AUC,* $F_C > F_1^0$ *on a set of scores with positive non-defaulter mass under the naive rule, or* $F_C > F_0^0$ *on a set with positive defaulter mass under the naive rule; for KS,* $\sup_s[F_1^0(s) - F_0^0(s)] > 0$*, which holds unless all H borrowers share a single naive score; for recall and lift,* $F_C(s_p) > F_1^0(s_p)$ *at the top-group cutoff* s_p*. In each case, smoothing strictly lowers default and strictly lowers the metric, so no positive smoothing rule below the rank-preservation threshold is a Pareto improvement under these objectives.*

Appendix B.4 shows that AUC, KS, recall-at-top-p, and top-p lift are rank-separation metrics in the sense of Definition 2, and Appendix B.5 proves the corollary.

4.4 Discussion

In the empirics of Section 3, we estimate the self-fulfilling share ϕ for borrowers in the major derogatory scorecard—for whom adverse score shocks cause default—and bound this share away from zero with 95 percent confidence. We also provide evidence that this effect is concentrated among borrowers that the scoring model already ranks as high risk (Assumption 1) in Appendix Figure A.10. Under these empirically supported assumptions, our theoretical results say that whether this causal feedback is a defect the credit scorer would want to correct depends on the objective. A scorer targeting forecast accuracy would want to smooth adverse revisions when defaults are the more costly forecast error (Theorem 1). In contrast, a scorer targeting rank separation would not want to smooth, because the causal effect is concentrated on borrowers that the score thinks will

default, and thus the score is rewarded for the defaults it causes (Theorem 2).

In practice, credit scoring models are typically evaluated on separation metrics (Thomas, Crook, and Edelman, 2017; Lessmann, Baesens, Seow, and Thomas, 2015). Credit scoring companies themselves describe models as enabling lenders "to rank consumers by their potential risk" and note that scores are "not designed for predicting absolute default rates" (VantageScore, 2022). Our results thus suggest that scorers have no incentive to reduce self-fulfilling defaults.

This conclusion relies on Theorem 2, which covers only smoothing parameters that are small enough to satisfy the rank-preservation threshold of Lemma 1. We view small smoothing as the empirically relevant case, as a scorer considering implementing a dampening rule would likely begin with modest departures from its current model. However, the data allow us to evaluate large interventions directly. We do so by tracing the two most common rank-separation metrics, AUC and KS, along the entire smoothing path implied by our estimates. For every consumer in the major derogatory scorecard we form the smoothed forecast $p_i(\lambda) = r_i^N - \lambda a_i$ and the counterfactual default probability $q_i(\lambda) = r_i^N - \hat{\phi}\lambda a_i$, where r_i^N, the empirical counterpart of the naive forecast p_i^N, and a_i are constructed from the calibration curve as in Appendix Figure A.10 and $\hat{\phi}$ is the self-fulfilling share estimated in Section 3. We then compute AUC and KS at each possible value of the smoothing weight.

Appendix Figure A.11 reports results of this exercise. Panels (a) and (b) show AUC and KS, respectively, at the point estimate of $\phi = 1.08$ (Table 4), both within the major derogatory scorecard and pooled across all scorecards. Panels (c) and (d) repeat the within-scorecard paths with ϕ at the bounds of the honest confidence interval. Across all versions of this exercise, every path lies strictly below its no-smoothing value at every positive value of the smoothing weight λ. In the data, our estimates thus imply that any amount of smoothing strictly lowers common rank-separation metrics, supporting the conclusion that scorers would have no incentive to curb the self-fulfilling defaults we document.

5 Conclusion

This paper asks whether credit scores cause some of the defaults they predict. Our regression discontinuity design exploits the deduplication of credit inquiries, which moves credit scores while leaving credit reports unchanged. Among consumers with prior derogatories, scores fall by over six points and new serious delinquency rises by 3.2 percentage points after two years. Our estimates imply that at least 13 percent of the relationship between scores and default is self-fulfilling for these consumers.

In a model of credit scoring with causal feedback, we then show that a scorer evaluated on forecast accuracy would want to dampen adverse score movements, lowering default and improving measured accuracy at the same time. However, a scorer evaluated on separating defaulters from non-defaulters would not, because separation metrics reward the score for the defaults it causes. In practice, credit scoring models are typically validated on separation metrics, suggesting that scorers have no incentive to curb the self-fulfilling defaults we document.

References

Agarwal, S., S. Alok, P. Ghosh, and S. Gupta (2023). Financial inclusion and alternate credit scoring: Role of big data and machine learning in fintech. *Journal of Money, Credit and Banking*. Forthcoming.

Albanesi, S. and D. F. Vamossy (2019). Predicting consumer default: A deep learning approach. NBER Working Paper 26165, National Bureau of Economic Research.

Almeida, H., I. Cunha, M. A. Ferreira, and F. Restrepo (2017). The real effects of credit ratings: The sovereign ceiling channel. *Journal of Finance 72*(1), 249–290.

Armstrong, T. B. and M. Kolesár (2018). Optimal inference in a class of regression models. *Econometrica 86*(2), 655–683.

Armstrong, T. B. and M. Kolesár (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics 11*(1), 1–39.

Barreca, A. I., J. M. Lindo, and G. R. Waddell (2016). Heaping-induced bias in regression-discontinuity designs. *Economic Inquiry 54*(1), 268–293.

Blattner, L. and S. Nelson (2021). How costly is noise? Data and disparities in consumer credit. Working paper.

Boot, A. W. A., T. T. Milbourn, and A. Schmeits (2006). Credit ratings as coordination mechanisms. *Review of Financial Studies 19*(1), 81–118.

Bos, M., E. Breza, and A. Liberman (2018, 01). The Labor Market Effects of Credit Market Information. *The Review of Financial Studies 31*(6), 2005–2037.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Chatterjee, S., D. Corbae, K. Dempsey, and J.-V. Ríos-Rull (2023). A quantitative theory of the credit score. *Econometrica 91*(5), 1803–1840.

Chioda, L., P. Gertler, S. Higgins, and P. C. Medina (2024, November). Fintech lending to borrowers with no credit history. Working Paper 33208, National Bureau of Economic Research.

Dobbie, W., P. Goldsmith-Pinkham, N. Mahoney, and J. Song (2020). Bad credit, no problem? credit and labor market consequences of bad credit reports. *The Journal of Finance 75*(5), 2377–2419.

Dobbie, W., B. J. Keys, and N. Mahoney (2017). Credit market consequences of credit flag removals. Working Paper 2991711, Princeton University.

Duarte, V., J. Fonseca, D. Kohli, and J. Reif (2025). The effects of deleting medical debt from consumer credit reports. NBER Working Paper 33644, National Bureau of Economic Research.

Experian (2024). Do multiple loan inquiries affect your credit score? Accessed July 2026.

Experian (2026a). 570 credit score: Is it good or bad? Accessed July 2026.

Experian (2026b). 800 credit score: Is it good or bad? Accessed July 2026.

FICO (2025). Does checking your credit score lower it? Accessed July 2026.

Frandsen, B. R. (2017). Party bias in union representation elections: Testing for manipulation in the regression discontinuity design when the running variable is discrete. In *Regression Discontinuity Designs: Theory and Applications*, Volume 38 of *Advances in Econometrics*, pp. 281–315. Emerald Publishing.

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance 77*(1), 5–47.

Goldstein, I. and C. Huang (2020). Credit rating inflation and firms' investments. *Journal of Finance 75*(6), 2929–2972.

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

Herkenhoff, K., G. Phillips, and E. Cohen-Cole (2021). The impact of consumer credit access on self-employment and entrepreneurship. *Journal of Financial Economics 141*(1), 345–371.

Jansen, M., F. Nagel, C. Yannelis, and A. L. Zhang (2025). Data and welfare in credit markets. *Journal of Financial Economics 174*.

Khandani, A. E., A. J. Kim, and A. W. Lo (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance 34*(11), 2767–2787.

Kolesár, M. and C. Rothe (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review 108*(8), 2277–2304.

Lessmann, S., B. Baesens, H.-V. Seow, and L. C. Thomas (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research 247*(1), 124–136.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics 17*(3), 1001–1008.

Liberman, A., C. Neilson, L. Opazo, and S. Zimmerman (2018). The equilibrium effects of asymmetric information: Evidence from consumer credit markets. NBER Working Paper 25097, National Bureau of Economic Research.

Manso, G. (2013). Feedback effects of credit ratings. *Journal of Financial Economics 109*(2), 535–548.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

Meursault, V., D. Moulton, L. Santucci, and N. Schor (2022). The time is now: Advancing fairness in lending through machine learning. Working paper.

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

Noack, C. and C. Rothe (2024). Bias-aware inference in fuzzy regression discontinuity designs. *Econometrica 92*(3), 687–711.

Perdomo, J., T. Zrnic, C. Mendler-Dünner, and M. Hardt (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 7599–7609.

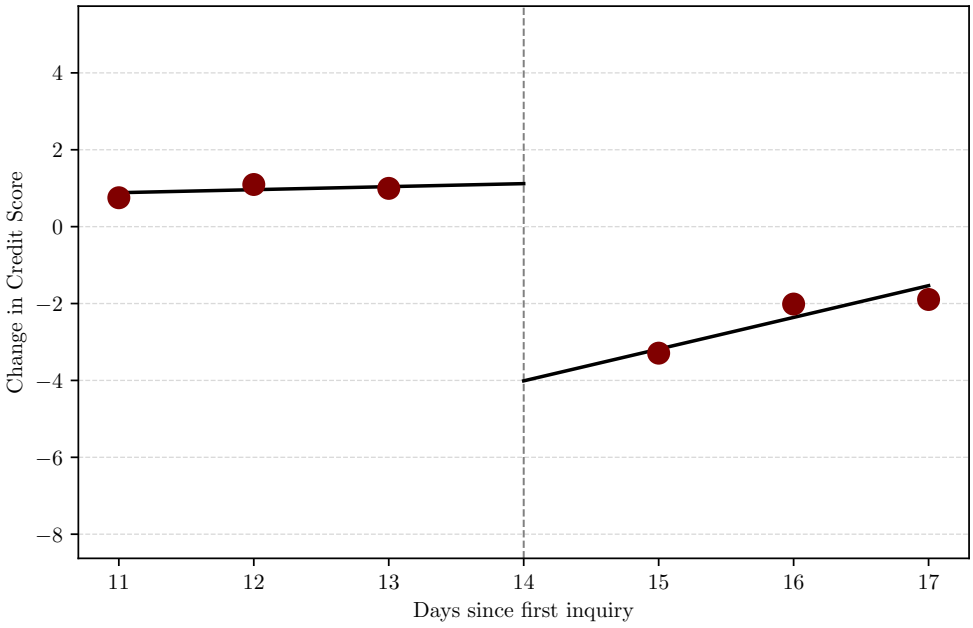
Purnanandam, A. and A. Wirth (2025). Can credit rating affect credit risk? Causal evidence from an online lending marketplace. Working paper, University of Michigan.

Thomas, L. C., J. N. Crook, and D. B. Edelman (2017). *Credit Scoring and Its Applications* (Second ed.). Philadelphia: SIAM.

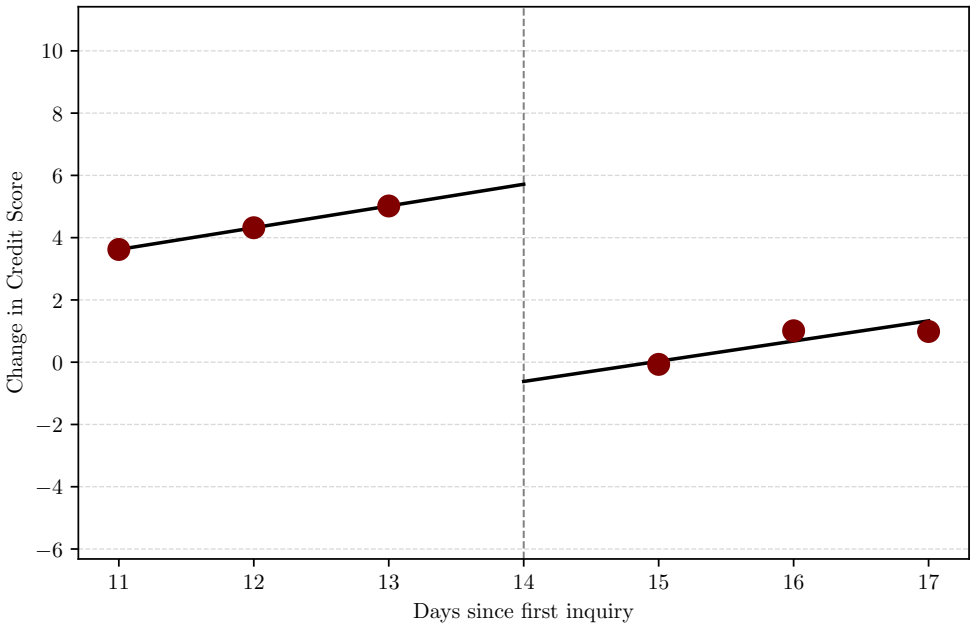
VantageScore (2022, September). Vantagescore 4.0 user guide. Abridged edition, revised September 2022.

VantageScore (2025). Lender FAQs. Accessed July 2026.

Figure 1: Effect of an Additional Counted Inquiry on Credit Scores



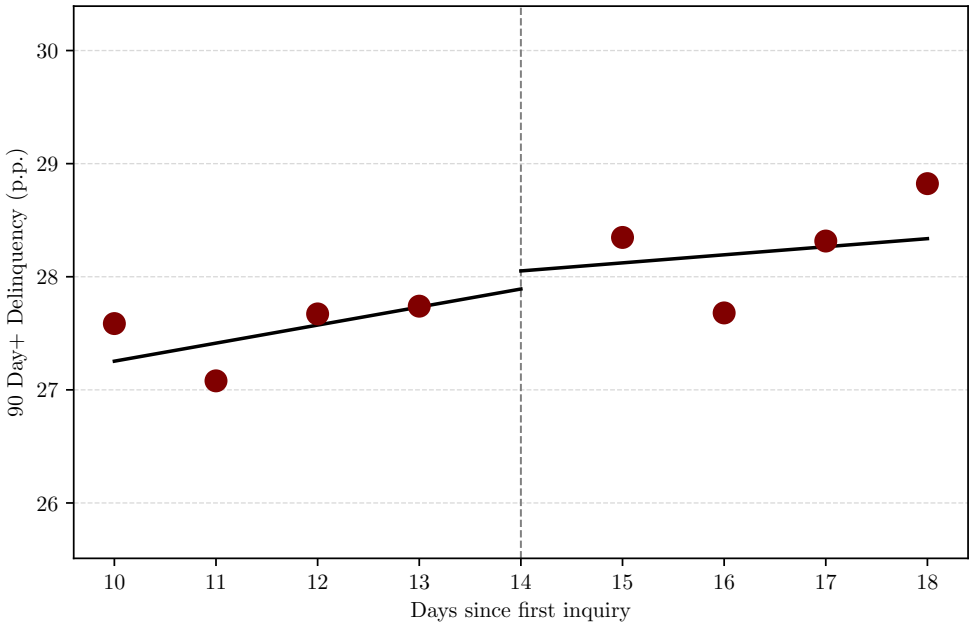
(a) Full sample



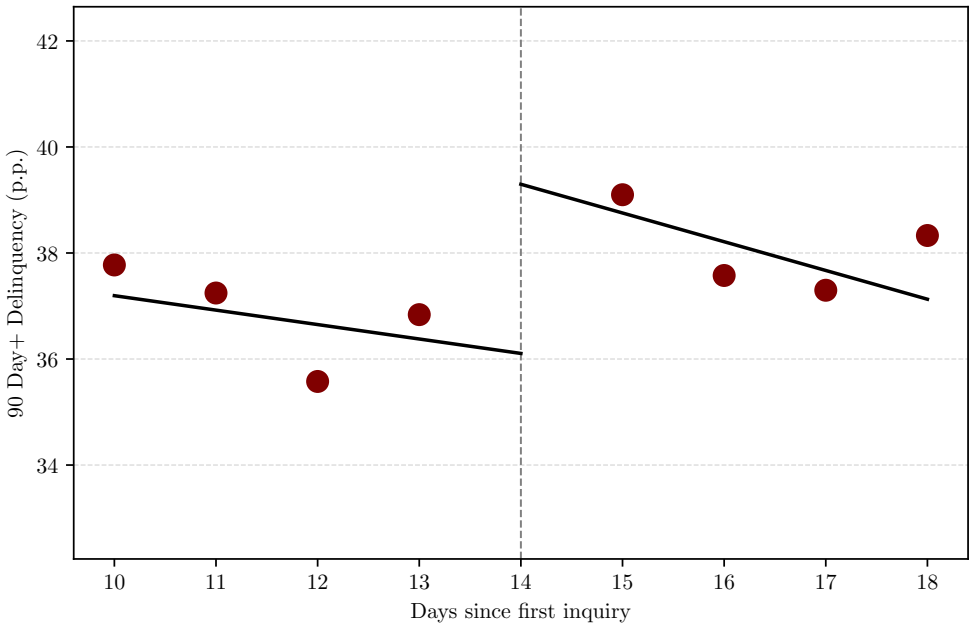
(b) Prior major derogatories

Notes: This figure shows the relationship between the running variable and the change in credit scores from $t-1$ to t. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. Panel (a) shows the full sample and Panel (b) the major derogatory scorecard. The corresponding RD estimates from Equation (1) are reported in Table 2.

Figure 2: Effect of an Additional Counted Inquiry on $t + 2$ Delinquency



(a) Full sample



(b) Prior major derogatories

Notes: This figure shows the relationship between the running variable and the share of consumers with a new account 90 or more days past due in the 12 months ending at $t + 2$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. Panel (a) shows the full sample and Panel (b) the major derogatory scorecard. The corresponding RD estimates from Equation (1) are reported in Table 3.

Table 1: Summary Statistics

	(1)	(2)	(3)	(4)	(5)	(6)
	GCCP			Active Credit Seekers		
	Mean	St. Dev.	Median	Mean	St. Dev.	Median
A. Demographics						
Income ($1,000)	51.59	32.92	41.00	52.71	34.30	41.00
Age (years)	50.87	19.65	50.00	42.17	14.35	40.00
Female (%)	49.82	50.00	0.00	47.70	49.95	0.00
B. Access to Credit						
Credit Score	701.90	99.48	714.00	653.55	100.30	647.00
Total Credit Balance ($1,000)	97.89	212.75	19.66	119.25	246.18	30.24
Revolving Utilization (%)	27.74	32.74	13.00	41.31	36.42	33.00
Number of Accounts	10.78	9.75	8.00	15.00	12.52	12.00
Age of Oldest Account (months)	211.61	144.26	189.00	154.11	108.14	135.00
Accounts Opened in Last 6 Months	0.41	0.83	0.00	0.96	1.36	1.00
C. Financial Distress						
New Serious Delinquency (%)	9.38	29.15	0.00	18.78	39.06	0.00
Number of Accounts 90+ Days Past Due	0.19	0.90	0.00	0.41	1.32	0.00
Balance 90+ Days Past Due ($1,000)	0.85	12.38	0.00	1.29	13.30	0.00
Number of Collections	0.55	1.75	0.00	1.23	2.53	0.00
Bankruptcy in Last 7 Years (%)	2.12	14.39	0.00	5.47	22.74	0.00
Observations	13,527,988			700,962		

Notes: This table presents summary statistics for the full GCCP and for our estimation sample of active credit seekers, consumers with two major inquiries within 6 to 22 days (Section 2.2). All variables are measured at $t-1$, the year before the inquiries.

Table 2: Effect of an Additional Counted Inquiry on Credit Scores

	Full sample	Major derogatory	Clean file	Thin-young
ABOVE	-5.13***	-6.33***	-6.72***	-2.96**
	[-7.03, -3.23]	[-9.32, -3.34]	[-9.12, -4.31]	[-5.78, -0.14]
Control Mean	654	614	740	601
Optimal Bandwidth	± 3	± 3	± 3	± 3
F-statistic	55.9	22.7	46.3	5.8
Observations				
Total	700,962	175,250	249,655	276,057
In-Bandwidth	221,893	55,988	78,469	87,436

Notes: This table presents the estimated coefficient β and honest 95 percent confidence intervals from Equation (1), where the outcome is the change in credit scores from $t-1$ to t, in points. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. Columns split the sample by the scorecards defined in Section 2. We report honest confidence intervals in brackets, which account for the worst-case bias of the local linear estimator (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean score at $t-1$ on the control side of the bandwidth. The reported optimal bandwidth minimizes the length of the honest confidence interval. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table 3: Effect of an Additional Counted Inquiry on $t+2$ Delinquency

	Full sample	Major derogatory	Clean file	Thin-young
ABOVE	0.16	3.19**	-0.39	-1.26
	[-1.68, 2.00]	[0.37, 6.01]	[-2.37, 1.58]	[-3.73, 1.21]
Control Mean	27.52	36.88	11.09	36.34
Optimal Bandwidth	± 4	± 4	± 4	± 4
Observations				
Total	451,475	113,490	160,065	177,920
In-Bandwidth	190,234	48,358	66,664	75,212

Notes: This table presents the estimated coefficient β and honest 95 percent confidence intervals from Equation (1), where the outcome is an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. Columns split the sample by the scorecards defined in Section 2. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean of the dependent variable on the control side of the bandwidth. The reported optimal bandwidth minimizes the length of the honest confidence interval. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table 4: The Self-Fulfilling Share ϕ: Inputs and Estimates

	Estimate	95% CI
Reduced form on default, $t+2$ (p.p.)	3.19	[0.37, 6.01]
First stage on credit score (points)	-6.33	[-9.32, -3.34]
τ_H (p.p. per point)	0.50	[0.06, 1.99]
m_H (p.p. per point)	0.465	[0.434, 0.490]
Self-fulfilling share ϕ	1.08	[0.32, 2.25]
		(0.13, 4.29)

This table collects the inputs to the self-fulfilling share $\phi = \tau_H / m_H$ of Equation (3) for the major derogatory scorecard. The reduced form, the first stage, and the Wald ratio τ_H of Equation (2) carry honest 95 percent confidence intervals (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). For τ_H, we report the bias-aware interval of Noack and Rothe (2024). The calibration slope m_H, evaluated at the in-bandwidth average score of 610 (Appendix Figure A.9), carries a 95 percent confidence interval from a clustered bootstrap with 1,000 draws. For the self-fulfilling share ϕ, the interval in brackets is the clustered bootstrap and the interval in parentheses is the Noack and Rothe (2024) bias-aware interval.

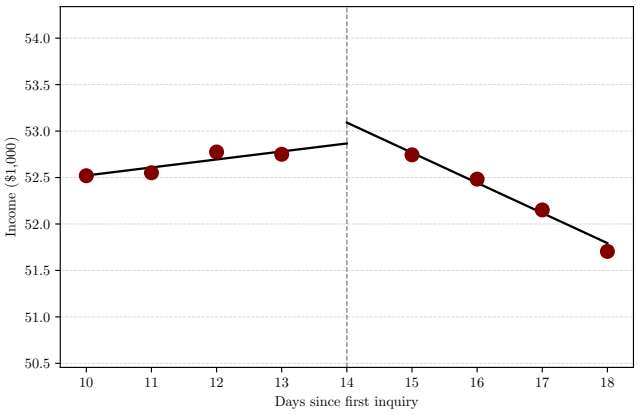
Online Appendix

"Self-Fulfilling Credit Scores"

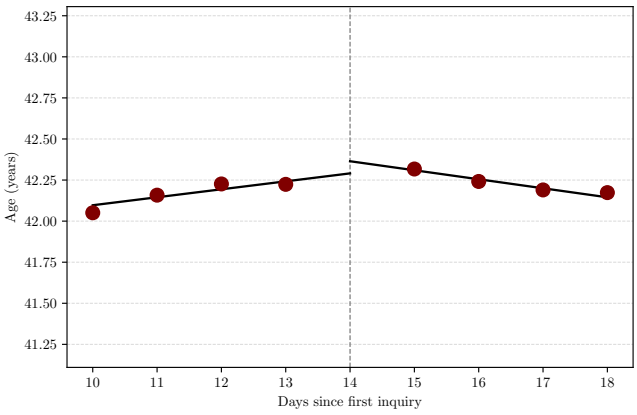
Victor Duarte and Julia Fonseca

A Additional Figures and Tables

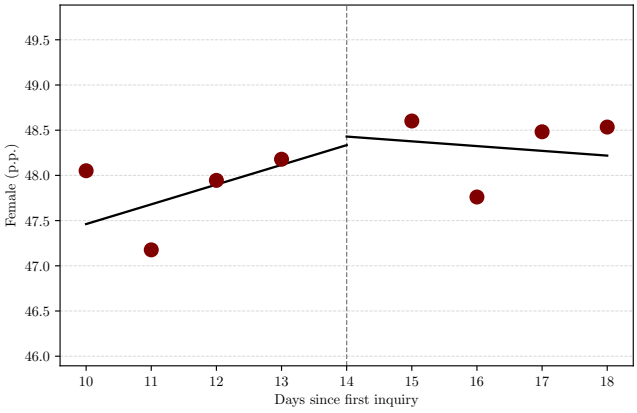
Figure A.1: Covariate Smoothness: Demographics



(a) Income ($1,000)



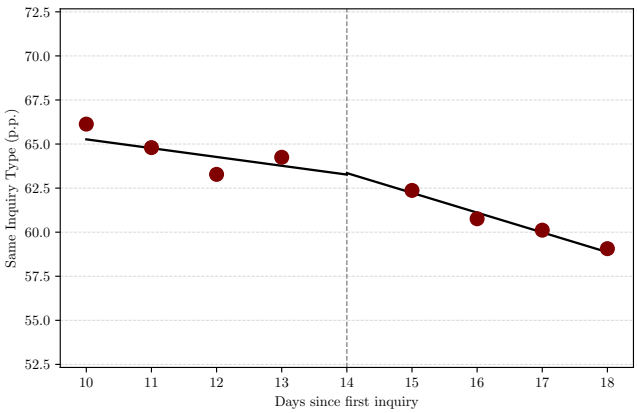
(b) Age (years)



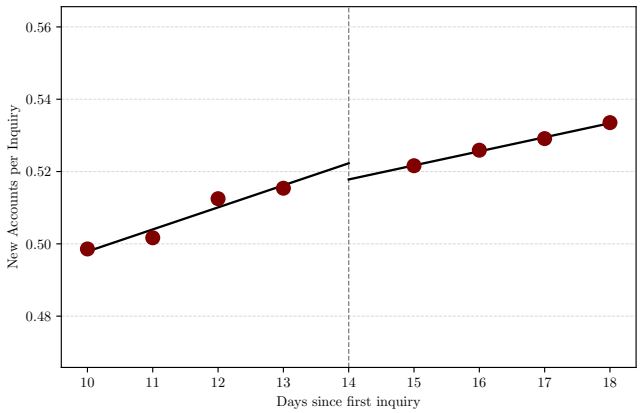
(c) Female (p.p.)

Notes: This figure shows the relationship between the running variable and three demographic variables, measured in $t-1$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. The corresponding RD estimates from Equation (1) are reported in Panel A of Table A.3.

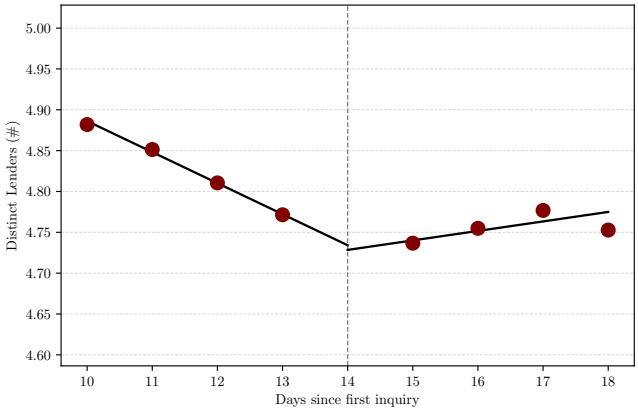
Figure A.2: Covariate Smoothness: Credit-Seeking Behavior



(a) Inquiries are of same type (p.p.)



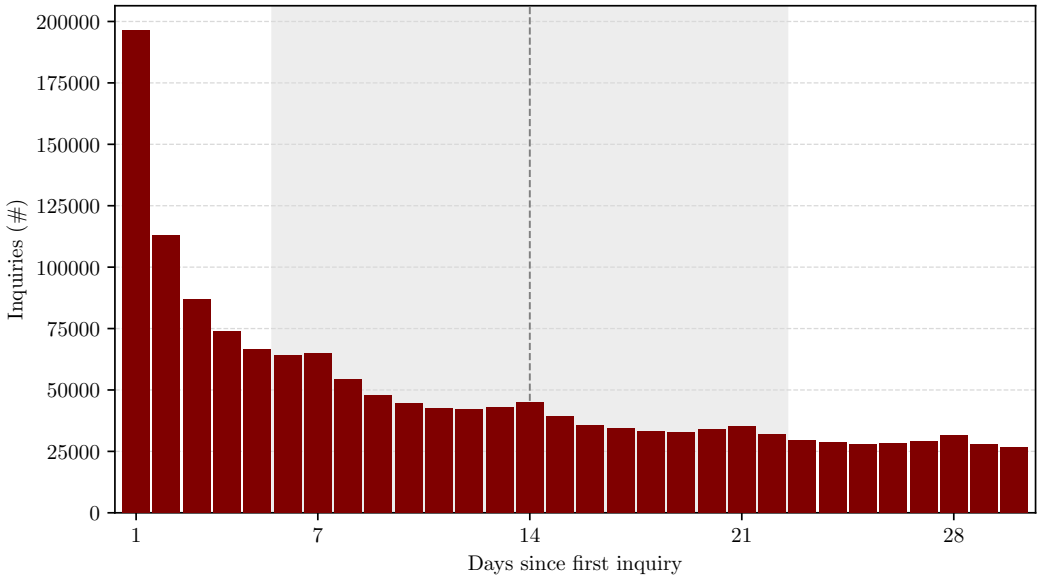
(b) New accounts per inquiry



(c) Distinct lenders (#)

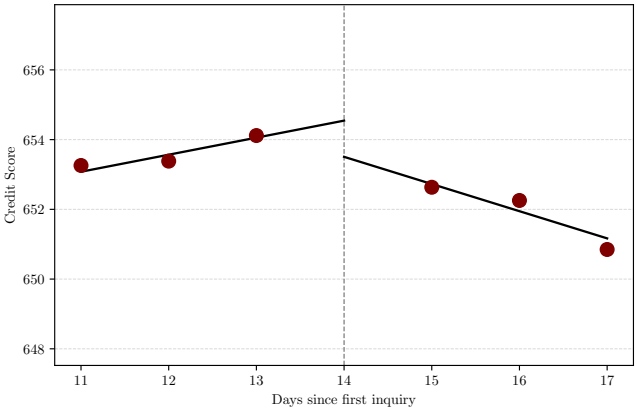
Notes: This figure shows the relationship between the running variable and three measures of credit-seeking behavior, measured at time t. Panel (a) plots a binary variable that takes the value of one if the inquiries that define the running variable are of the same major type (e.g. mortgage, auto, or bankcard). The product category is measured from the inquiry's type code, supplemented by the lender's industry code when the type code carries no product information. Panel (b) plots the number of new accounts opened per inquiry, and panel (c) plots the number of distinct lenders from which the consumer's inquiries originate, measured as unique combinations of industry and subscriber codes. The corresponding RD estimates from Equation (1) are reported in Panel B of Table A.3.

Figure A.3: Density of the Running Variable

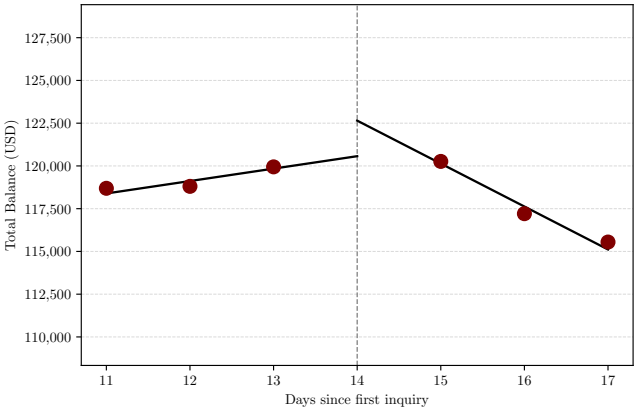


Notes: This figure shows the distribution of the running variable, the number of days between each initial inquiry and the next major inquiry. The dashed vertical line marks the 14-day de-duplication boundary and the shaded band marks the $[6, 22]$ window from which the estimation sample is drawn. The distribution bunches at weekly multiples—7, 14, 21, and 28 days—reflecting the day-of-week rhythm of credit applications. The excess mass at day 14 is statistically indistinguishable from the excess mass at the other weekly bunching days (p-value = 0.38).

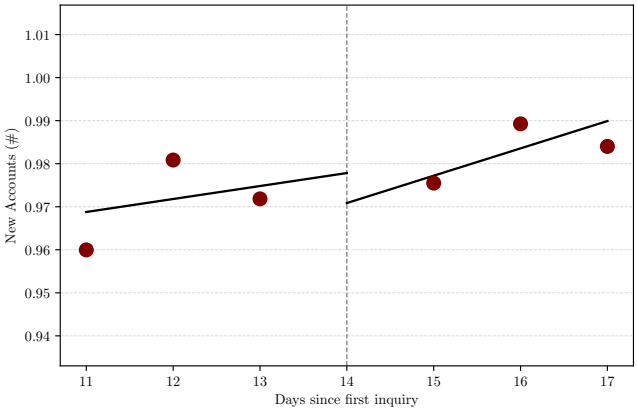
Figure A.4: Falsification Tests: Access to Credit



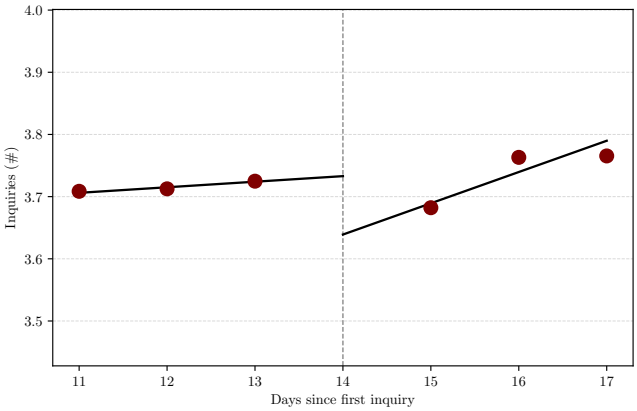
(a) Credit score



(b) Total Balance ($)



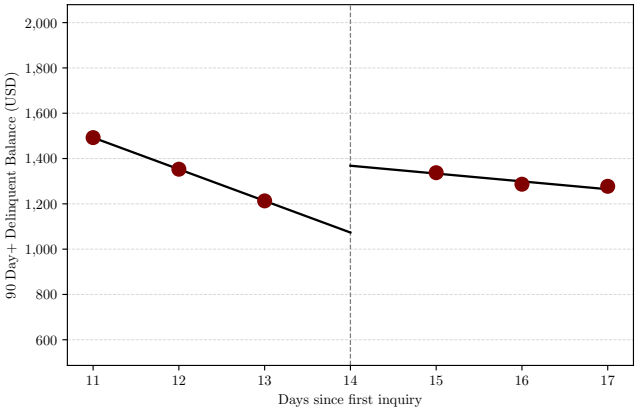
(c) New Accounts (#)



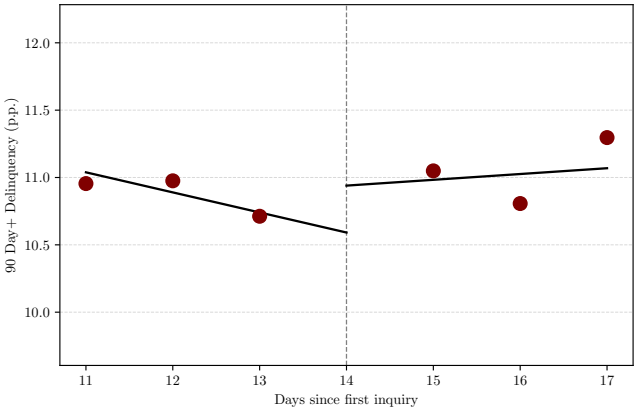
(d) Inquiries (#)

Notes: This figure shows the relationship between the running variable at t and four measures of credit access measured at $t-1$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. The corresponding RD estimates from Equation (1) are reported in Panel A of Table A.4.

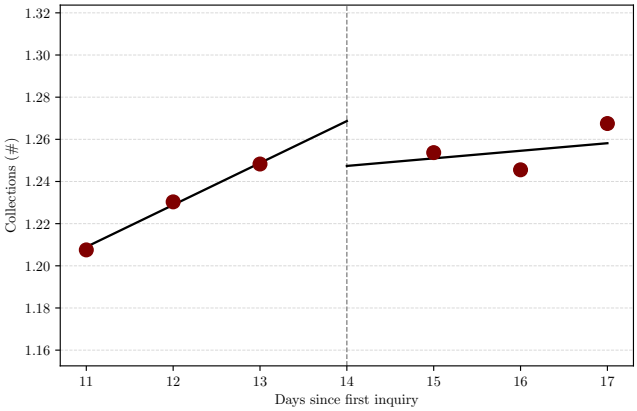
Figure A.5: Falsification Tests: Financial Distress



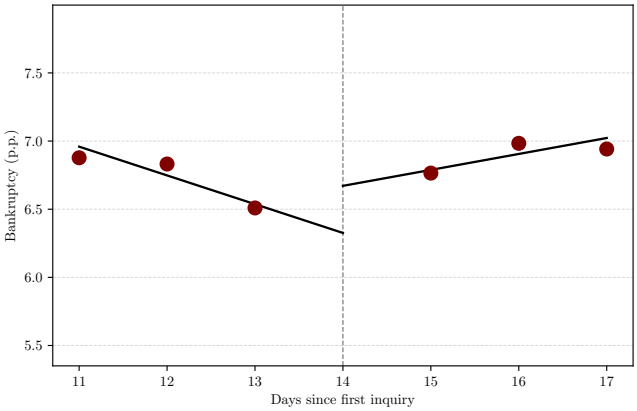
(a) Balance 90 Day+ Delinquent ($)



(b) 90 Day+ Delinquency (p.p.)



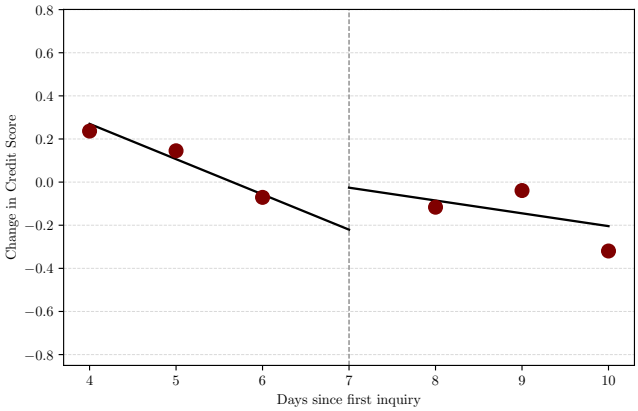
(c) Accounts in Collection (#)



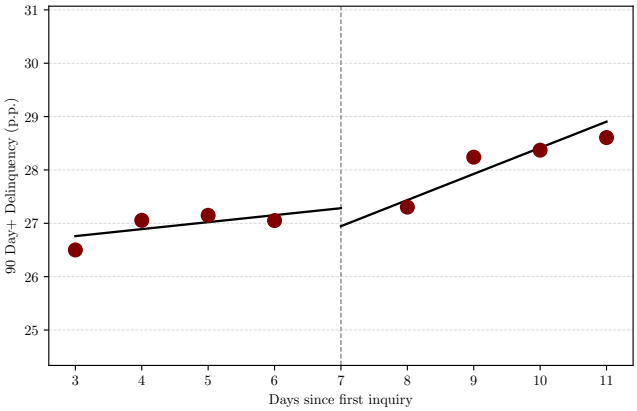
(d) Bankruptcies (p.p.)

Notes: This figure shows the relationship between the running variable at t and four measures of financial distress measured at $t-1$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. The corresponding RD estimates from Equation (1) are reported in Panel B of Table A.4.

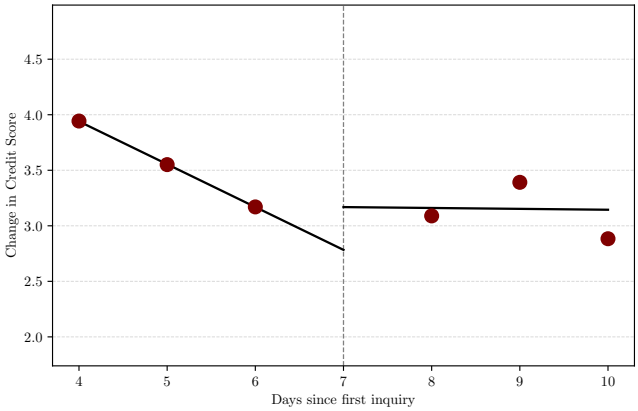
Figure A.6: Placebo Test: 7-Day Cutoff



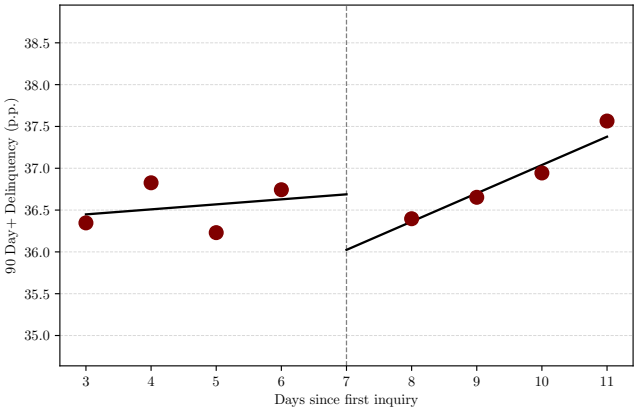
(a) Full Sample: Credit Score



(b) Full Sample: Delinquency at $t+2$



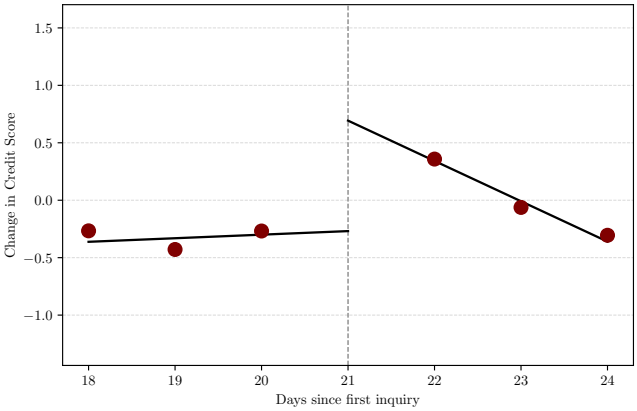
(c) Major Derogatory: Credit Score



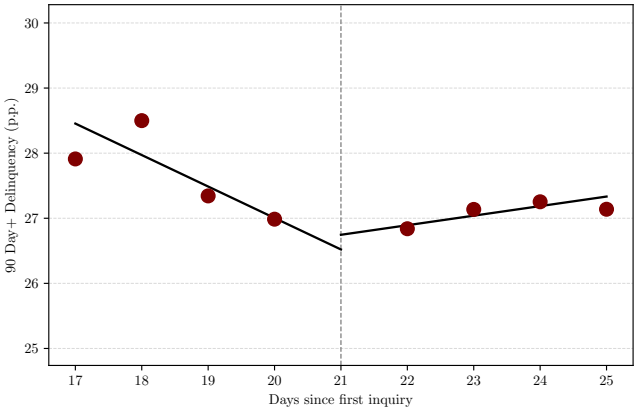
(d) Major Derogatory: Delinquency at $t+2$

Notes: This figure applies the research design to a placebo cutoff at 7 days, using major inquiries with day gaps between 3 and 11. The de-duplication rule implies no discontinuity at 7 days: on both sides of the placebo cutoff, the second inquiry falls within the initial 14-day window and is de-duplicated. Panels (a) and (c) show the change in credit scores from $t-1$ to t for the full placebo sample and the major derogatory scorecard. Panels (b) and (d) plot an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. The corresponding RD estimates from Equation (1) are reported in Table A.5.

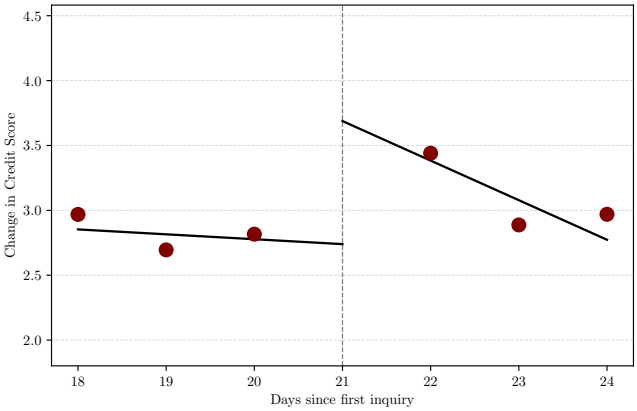
Figure A.7: Placebo Test: 21-Day Cutoff



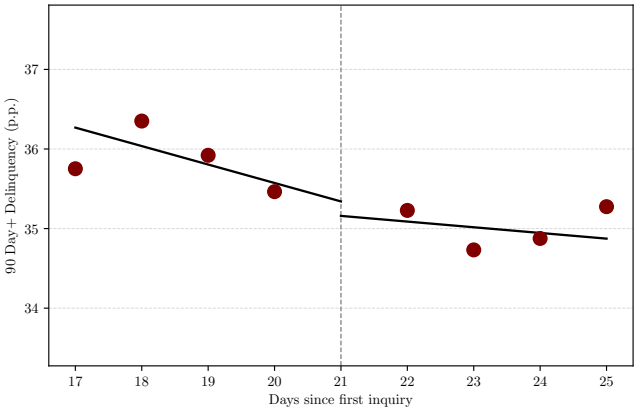
(a) Full Sample: Credit Score



(b) Full Sample: Delinquency at $t+2$



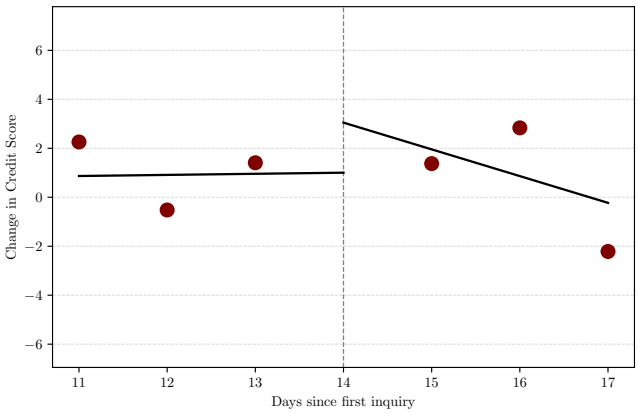
(c) Major Derogatory: Credit Score



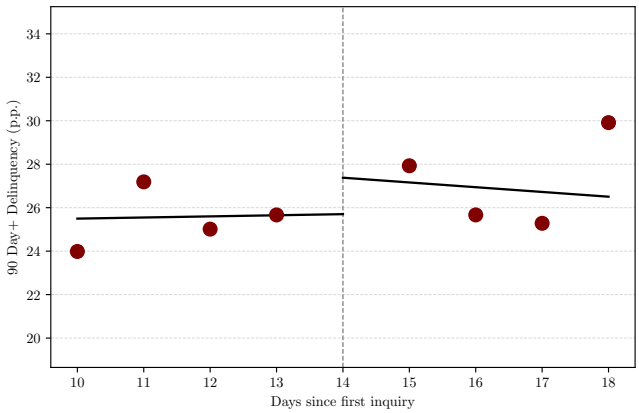
(d) Major Derogatory: Delinquency at $t+2$

Notes: This figure applies the research design to a placebo cutoff at 21 days, using major inquiries with day gaps between 17 and 25. The de-duplication rule implies no discontinuity at 21 days: on both sides of the placebo cutoff, the second inquiry falls outside the initial 14-day window and is counted separately. Panels (a) and (c) show the change in credit scores from $t-1$ to t for the full placebo sample and the major derogatory scorecard. Panels (b) and (d) plot an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. The corresponding RD estimates from Equation (1) are reported in Table A.6.

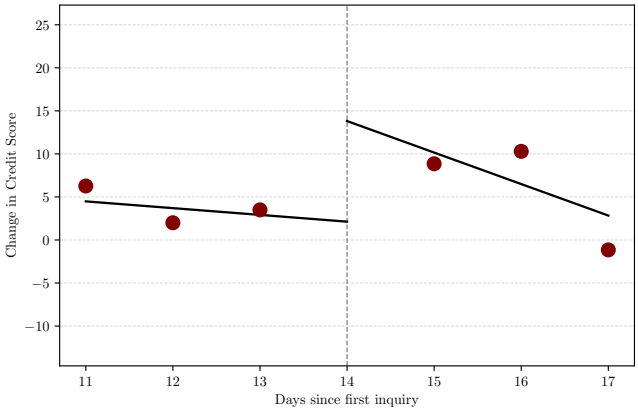
Figure A.8: Placebo Test: Never-De-Duplicated Inquiry Types



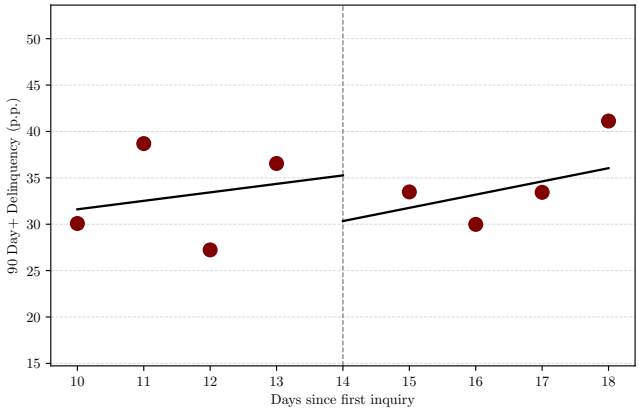
(a) Full Sample: Credit Score



(b) Full Sample: Delinquency at $t+2$



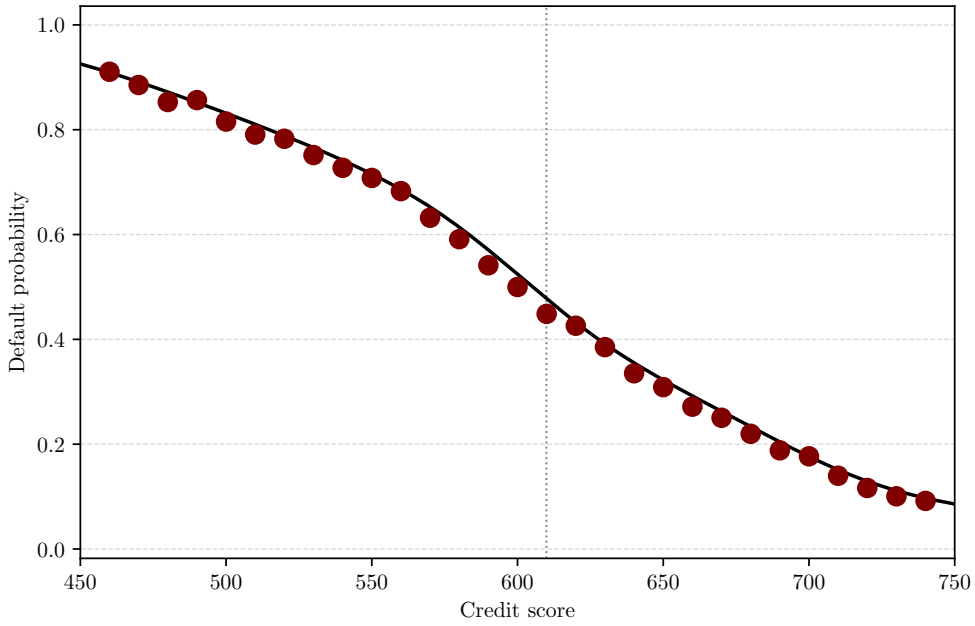
(c) Major Derogatory: Credit Score



(d) Major Derogatory: Delinquency at $t+2$

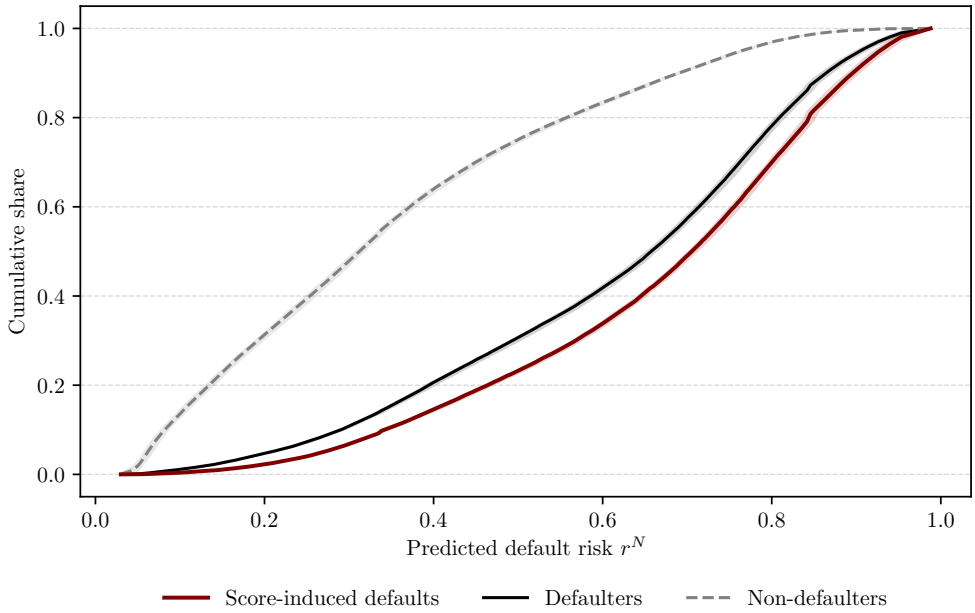
Notes: This figure applies the research design to retail, utility, and collection inquiries, which the credit scoring model does not de-duplicate, and so the rule implies no discontinuity at 14 days. Panels (a) and (c) show the change in credit scores from $t-1$ to t for the full placebo sample and the major derogatory scorecard. Panels (b) and (d) plot an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. The corresponding RD estimates from Equation (1) are reported in Table A.7.

Figure A.9: Score-to-Default Calibration for the Major Derogatory Scorecard



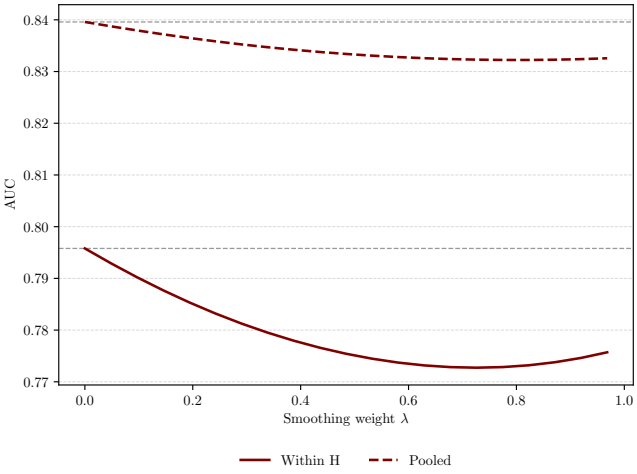
Notes: This figure shows the probability of any new serious delinquency by $t+2$ against the credit score at t for the major derogatory scorecard. Dots are empirical means in ten-point score bins; the line is a natural cubic spline logit fit. The dotted vertical line marks the in-bandwidth average score of 610, at which the calibration slope is $m_H = 0.46$ percentage points per point, which we use to construct the self-fulfilling share ϕ in Section 3.2.

Figure A.10: Rank-Aligned Causal Feedback

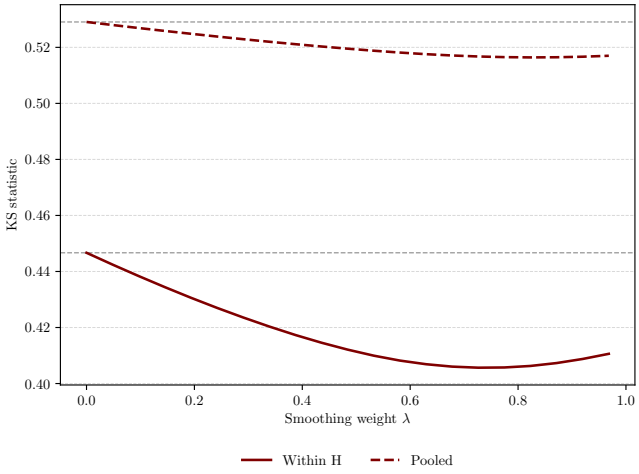


Notes: This figure provides empirical support for Assumption 1. For the major derogatory scorecard, it plots cumulative distributions over predicted default risk r^N of three groups: score-induced marginal defaults, weighting adversely revised consumers by a_i as in equation (16); realized defaulters, consumers with any new serious delinquency by $t+2$; and realized non-defaulters. Predicted risk is $r^N = g(s_{\text{post}})$, where g is the calibration curve of Figure A.9 and s_{post} is the credit score after the revision, and the adverse revision $a_i = \max\{g(s_{\text{post},i}) - g(s_{\text{pre},i}),\, 0\}$ is the increase in predicted risk implied by the score decline, the probability-scale revision of Section 4. Shaded regions are pointwise 95 percent confidence bands from a clustered bootstrap over consumers that re-estimates the calibration curve in each of 1,000 draws. The revision-weighted distribution lies weakly to the right of both outcome distributions at every point, so adverse revisions are concentrated among borrowers the score already ranks riskiest.

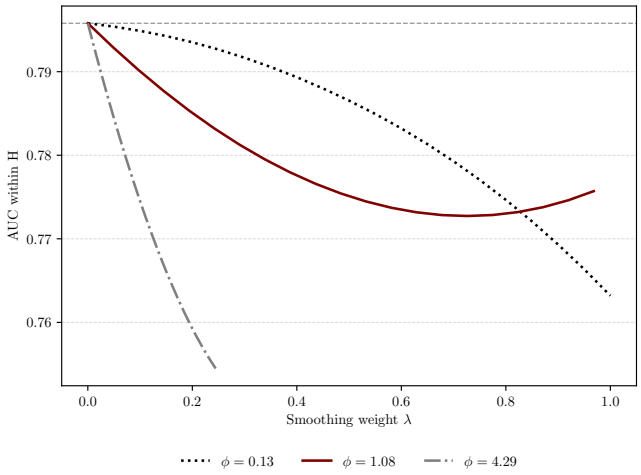
Figure A.11: Separation Metrics Along the Smoothing Path



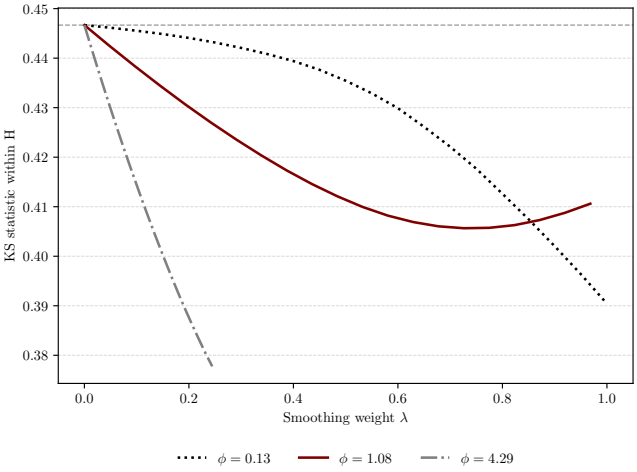
(a) AUC at the point estimate of ϕ



(b) KS at the point estimate of ϕ



(c) AUC across the ϕ confidence interval



(d) KS across the ϕ confidence interval

Notes: This figure traces two common rank-separation metrics along the counterfactual smoothing path of Section 4. For each consumer in the major derogatory scorecard, the smoothed forecast is $p_i(\lambda) = r_i^N - \lambda a_i$ and the counterfactual default probability is $q_i(\lambda) = r_i^N - \hat{\phi}\lambda a_i$, with r_i^N and a_i constructed as in Figure A.10 and $\hat{\phi}$ taken from Table 4. AUC and KS are computed in expectation over default outcomes realized with probability $q_i(\lambda)$, ranking consumers by the smoothed score. Pooled paths rank all consumers in the estimation sample together, holding the scores and default probabilities of consumers outside the major derogatory scorecard fixed. Panels (c) and (d) evaluate ϕ at the endpoints of the honest confidence interval (Table 4). For values of ϕ above one, each path ends at the largest λ for which all counterfactual default probabilities remain nonnegative. Horizontal dashed grey lines mark each metric's no-smoothing value.

Table A.1: Summary Statistics by Scorecard

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Major derogatory			Clean file			Thin-young		
	Mean	St. Dev.	Median	Mean	St. Dev.	Median	Mean	St. Dev.	Median
A. Demographics									
Income ($1,000)	50.87	24.30	44.00	75.62	41.51	63.00	33.15	13.96	31.00
Age (years)	44.11	12.76	43.00	46.12	14.56	44.00	37.35	13.73	35.00
Female (%)	54.37	49.81	100.00	43.93	49.63	0.00	46.85	49.90	0.00
B. Access to Credit									
Credit Score	613.73	77.98	612.00	739.89	73.69	753.00	600.74	78.96	597.00
Total Credit Balance ($1,000)	94.55	143.56	39.54	214.13	345.43	114.56	29.82	79.46	7.62
Revolving Utilization (%)	51.76	36.67	51.00	29.59	29.12	19.00	49.54	40.78	45.00
Number of Accounts	18.78	12.12	16.00	20.74	12.74	18.00	7.19	7.58	5.00
Age of Oldest Account (months)	171.76	84.60	158.00	214.68	112.28	197.00	86.38	74.89	68.00
Accounts Opened in Last 6 Months	1.15	1.53	1.00	0.89	1.25	0.00	0.90	1.32	0.00
C. Financial Distress									
New Serious Delinquency (%)	32.78	46.94	0.00	3.46	18.28	0.00	24.15	42.80	0.00
Number of Accounts 90+ Days Past Due	0.81	1.89	0.00	0.06	0.44	0.00	0.48	1.32	0.00
Balance 90+ Days Past Due ($1,000)	2.80	20.02	0.00	0.37	9.28	0.00	1.15	10.38	0.00
Number of Collections	2.24	2.89	1.00	0.00	0.00	0.00	1.71	2.94	1.00
Bankruptcy in Last 7 Years (%)	12.08	32.59	0.00	0.00	0.00	0.00	6.22	24.16	0.00
Observations	175,250			249,655			276,057		

Notes: This table presents summary statistics for the estimation sample by scorecard, as defined in Section 2. All variables are measured at $t-1$, the year before the inquiries.

Table A.2: Effect of an Additional Counted Inquiry on $t+1$ Delinquency

	Full sample	Major derogatory	Clean file	Thin-young
ABOVE	0.10	1.07	0.54	-0.81
	[-1.72, 1.92]	[-1.60, 3.73]	[-1.38, 2.47]	[-3.18, 1.56]
Control Mean	26.51	35.27	10.38	35.44
Optimal Bandwidth	± 4	± 4	± 4	± 4
Observations				
Total	580,088	145,890	205,593	228,605
In-Bandwidth	244,885	62,324	85,880	96,681

Notes: This table presents the estimated coefficient β and honest 95 percent confidence intervals from Equation (1), where the outcome is an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+1$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. Columns split the sample by the scorecards defined in Section 2. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean of the dependent variable on the control side of the bandwidth. The reported optimal bandwidth minimizes the length of the honest confidence interval. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table A.3: Covariate Smoothness

	Panel A: Demographics		
	Income ($1,000)	Age (years)	Female (p.p.)
ABOVE	0.22	0.07	0.09
	[-3.81, 4.26]	[-1.03, 1.17]	[-2.53, 2.72]
Control Mean	52.65	42.16	47.84
Bandwidth	± 4	± 4	± 4
Observations			
Total	700,854	699,694	646,592
In-Bandwidth	295,694	295,194	272,786
	Panel B: Credit-Seeking Behavior		
	New Accounts per Inquiry	Distinct Lenders (#)	Same Inquiry Type (p.p.)
ABOVE	-0.00	-0.01	0.08
	[-0.05, 0.04]	[-0.08, 0.07]	[-2.67, 2.84]
Control Mean	0.51	4.83	64.65
Bandwidth	± 4	± 4	± 4
Observations			
Total	697,752	700,962	582,041
In-Bandwidth	294,386	295,739	244,854

Notes: This table presents the estimated coefficient β from Equation (1) for the demographic variables shown in Figure A.1 (Panel A) and the measures of credit-seeking behavior shown in Figure A.2 (Panel B). The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean of the dependent variable on the control side of the bandwidth, which is fixed at four days. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table A.4: Falsification Tests

	Panel A: Access to Credit			
	Credit Score	Total Balance ($)	New Accounts (#)	Inquiries (#)
ABOVE	-0.57 [-8.73, 7.58]	1,679 [-13,207, 16,565]	-0.01 [-0.07, 0.05]	-0.06 [-0.20, 0.07]
Control Mean	653	118,672	0.97	3.70
Bandwidth	± 4	± 4	± 4	± 4
Observations				
Total	700,962	640,670	693,910	700,962
In-Bandwidth	295,739	270,137	292,750	295,739
	Panel B: Financial Distress			
	Balance 90 Day+ Delinquent ($)	90 Day+ Delinquency (p.p.)	Accounts in Collection (#)	Bankruptcies (p.p.)
ABOVE	105 [-364, 575]	0.24 [-2.79, 3.27]	-0.02 [-0.16, 0.13]	0.24 [-0.93, 1.42]
Control Mean	1,307	10.93	1.23	6.73
Bandwidth	± 4	± 4	± 4	± 4
Observations				
Total	673,580	673,580	700,960	700,962
In-Bandwidth	284,138	284,138	295,739	295,739

Notes: This table presents the estimated coefficient β from Equation (1) for the measures of credit access shown in Figure A.4 (Panel A) and the measures of financial distress shown in Figure A.5 (Panel B), all measured at $t-1$. The running variable is the number of days between a consumer's initial inquiry and the next major inquiry. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean of the dependent variable on the control side of the bandwidth, which is fixed at four days. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table A.5: Placebo Test: 7-Day Cutoff

	Full sample		Major derogatory	
	Credit Score	Delinquency	Credit Score	Delinquency
ABOVE	0.19	-0.34	0.39	-0.66
	[-1.80, 2.19]	[-2.14, 1.46]	[-2.47, 3.25]	[-3.15, 1.82]
Control Mean	655	26.90	613	36.53
Optimal Bandwidth	± 3	± 4	± 3	± 4
Observations				
Total	579,646	479,470	142,132	118,107
In-Bandwidth	366,121	424,136	90,438	104,295

Notes: This table presents the estimated coefficient β and honest 95 percent confidence intervals from Equation (1) at the placebo cutoff of 7 days shown in Figure A.6, using major inquiries with day gaps between 3 and 11. Credit Score is the change in credit scores from $t-1$ to t, in points. Delinquency is an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean score at $t-1$ in the Credit Score columns and the mean of the dependent variable in the Delinquency columns, on the control side of the bandwidth. The reported optimal bandwidth minimizes the length of the honest confidence interval. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table A.6: Placebo Test: 21-Day Cutoff

	Full sample		Major derogatory	
	Credit Score	Delinquency	Credit Score	Delinquency
ABOVE	0.96	0.23	0.95	-0.18
	[-1.29, 3.21]	[-1.73, 2.18]	[-2.30, 4.20]	[-2.95, 2.59]
Control Mean	644	27.69	609	35.87
Optimal Bandwidth	± 3	± 4	± 3	± 4
Observations				
Total	333,021	276,318	88,395	73,730
In-Bandwidth	219,409	242,113	58,305	64,732

Notes: This table presents the estimated coefficient β and honest 95 percent confidence intervals from Equation (1) at the placebo cutoff of 21 days shown in Figure A.7, using major inquiries with day gaps between 17 and 25. Credit Score is the change in credit scores from $t-1$ to t, in points. Delinquency is an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean score at $t-1$ in the Credit Score columns and the mean of the dependent variable in the Delinquency columns, on the control side of the bandwidth. The reported optimal bandwidth minimizes the length of the honest confidence interval. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

Table A.7: Placebo Test: Never-De-Duplicated Inquiry Types

	Full sample		Major derogatory	
	Credit Score	Delinquency	Credit Score	Delinquency
ABOVE	2.04	1.68	11.69	-4.92
	[-7.81, 11.90]	[-5.22, 8.57]	[-5.60, 28.97]	[-17.52, 7.68]
Control Mean	620	25.44	601	33.01
Optimal Bandwidth	± 3	± 4	± 3	± 4
Observations				
Total	12,807	11,739	4,026	3,688
In-Bandwidth	4,184	5,097	1,307	1,602

Notes: This table presents the estimated coefficient β and honest 95 percent confidence intervals from Equation (1) applied to retail, utility, and collection inquiries, which the credit scoring model does not de-duplicate, as shown in Figure A.8. Credit Score is the change in credit scores from $t-1$ to t, in points. Delinquency is an indicator equal to 100 if the consumer has a new account 90 or more days past due in the 12 months ending at $t+2$. We report honest confidence intervals in brackets (Armstrong and Kolesár, 2020; Kolesár and Rothe, 2018). Control Mean reports the mean score at $t-1$ in the Credit Score columns and the mean of the dependent variable in the Delinquency columns, on the control side of the bandwidth. The reported optimal bandwidth minimizes the length of the honest confidence interval. "In-Bandwidth" observations report the number of observations falling within this window, while "Total" observations report the full estimation sample. A */**/*** denotes statistical significance at the 10%, 5%, and 1% levels respectively, based on honest confidence intervals.

B Model Appendix

B.1 Proof of Theorem 1

Proof. The effect on default follows directly from (7). For any $\lambda > 0$,

$$D_H(\lambda) - D_H(0) = -\phi\lambda\mathbb{E}_H[a_i] < 0, \tag{19}$$

where the strict inequality uses $\phi > 0$, $\lambda > 0$, and assumption (4).

For the credit scorer's loss, fix an H borrower and write $p = p_i^N$ and $a = a_i$. Along the smoothing path,

$$p(\lambda) = p - \lambda a, \qquad q(\lambda) = p - \phi\lambda a. \tag{20}$$

The derivative of expected loss at the status quo is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}L(p(\lambda), q(\lambda))\bigg|_{\lambda=0} &= L_p(p,p)(-a) + L_q(p,p)(-\phi a) \\ &= -\phi a\, L_q(p,p),\end{aligned} \tag{21}$$

where the second equality uses local propriety, $L_p(p,p) = 0$; the derivative exists because $p \in \mathcal{P}$ by (11), so L is continuously differentiable near (p,p) and $(p(\lambda), q(\lambda))$ lies in that neighborhood for λ small. If $a > 0$, Definition 1 implies $L_q(p,p) > 0$, so the derivative in (21) is strictly negative; if $a = 0$, the borrower's loss term is constant in λ. Averaging over the finite H population, $\mathcal{L}_H$ is right-differentiable at zero with

$$\mathcal{L}_H'(0) = -\phi\mathbb{E}_H\left[a_i L_q(p_i^N, p_i^N)\right] < 0, \tag{22}$$

where the strict inequality holds because every term $a_i L_q(p_i^N, p_i^N)$ is nonnegative and, by (4), positive for a positive fraction of H borrowers. Since $\mathcal{L}_H$ is right-differentiable at zero with $\mathcal{L}_H'(0) < 0$, there exists $\bar{\lambda} \in (0,1]$ such that $\mathcal{L}_H(\lambda) < \mathcal{L}_H(0)$ for all $\lambda \in (0, \bar{\lambda}]$. Since $\lambda_L = 0$ leaves both scores and outcomes unchanged for L borrowers, the type-specific rule is a Pareto improvement. □

Note that condition (11) is sufficient but not necessary. The proof uses it only to sign each term of $\mathcal{L}'_H(0) = -\phi\mathbb{E}_H[a_i L_q(p_i^N, p_i^N)]$. Provided L is continuously differentiable in a neighborhood of (p_i^N, p_i^N) for every H borrower with $a_i > 0$, the conclusion of Theorem 1 holds whenever the revision-weighted average $\mathbb{E}_H[a_i L_q(p_i^N, p_i^N)]$ is strictly positive, even if $L_q(p_i^N, p_i^N) \leq 0$ for some borrowers.

B.2 Proof of Lemma 1

Proof. Every ratio in (15) is strictly positive, because the pairs over which the minimum is taken have $p_i^N - p_j^N > 0$ and $a_i - a_j > 0$, and with finitely many borrowers the minimum runs over finitely many pairs; hence $\bar{\lambda}_{\text{rank}} > 0$. Distinct naive scores and μ_H strictly decreasing imply distinct naive forecasts, so every pair of borrowers is strictly ordered by the naive forecast. Fix $\lambda \in [0, 1]$ with $\lambda < \bar{\lambda}_{\text{rank}}$ and a pair of borrowers with $p_i^N > p_j^N$. If $a_i \leq a_j$, then

$$p_i(\lambda) - p_j(\lambda) = (p_i^N - p_j^N) - \lambda(a_i - a_j) \geq p_i^N - p_j^N > 0. \tag{23}$$

If $a_i > a_j$, then $\lambda(a_i - a_j) < \bar{\lambda}_{\text{rank}}(a_i - a_j) \leq p_i^N - p_j^N$ by the definition of $\bar{\lambda}_{\text{rank}}$, so again $p_i(\lambda) > p_j(\lambda)$. Smoothed forecasts therefore preserve the strict ordering of naive forecasts. Since $p_i^N = \mu_H(s_i^N)$ and $s_i(\lambda) = \mu_H^{-1}(p_i(\lambda))$ with μ_H strictly decreasing, smoothed scores are strictly increasing in naive scores across the finitely many support points, and any strictly increasing interpolation through the points $(s_i^N, s_i(\lambda))$ gives the required φ_λ. □

B.3 Proof of Theorem 2

Proof. Smoothing strictly lowers default exactly as in the proof of Theorem 1: $D_H(\lambda) - D_H(0) = -\phi\lambda\mathbb{E}_H[a_i] < 0$ by $\phi > 0$ and (4).

It remains to show that smoothing cannot raise the rank-separation objective. Because $\lambda < \bar{\lambda}_{\text{rank}}$, Lemma 1 applies: the smoothed score is a strictly increasing transformation of the naive score, and since the metric depends on scores only through their ranks, the credit scorer's objective under

smoothing coincides with $\mathfrak{R}(F_1^\lambda, F_0^\lambda)$ evaluated in the naive score coordinate. Let

$$Q_0 = \mathbb{E}_H[p_i^N], \qquad Q_\lambda = \mathbb{E}_H[q_i(\lambda)] = Q_0 - \phi\lambda\mathbb{E}_H[a_i] < Q_0 \tag{24}$$

be the total default probabilities under the naive and smoothed rules, and let $N_0 = 1 - Q_0$ and $N_\lambda = 1 - Q_\lambda$ denote the corresponding non-default probabilities; $Q_0, N_0 \in (0,1)$ because $p_i^N \in (0,1)$, $Q_\lambda = D_H(\lambda) > 0$ by hypothesis, and $N_\lambda \geq N_0 > 0$. Smoothing by λ removes default mass $\phi\lambda a_i \geq 0$ from each borrower and adds the same mass to the non-default state. The score distribution of the removed default mass is F_C from (16). Conditioning on realized default status, the score distribution among realized defaulters after smoothing is

$$F_1^\lambda(s) = \frac{Q_0 F_1^0(s) - \phi\lambda\mathbb{E}_H[a_i]F_C(s)}{Q_0 - \phi\lambda\mathbb{E}_H[a_i]}, \tag{25}$$

a legitimate distribution function because its numerator equals $\mathbb{E}_H[q_i(\lambda)\mathbf{1}\{s_i^N \leq s\}]$, which is nonnegative and nondecreasing in s since $q_i(\lambda) \in [0,1]$; the same argument with weights $1 - q_i(\lambda)$ applies to F_0^λ below. Subtracting $F_1^0(s)$ gives

$$F_1^\lambda(s) - F_1^0(s) = \frac{\phi\lambda\mathbb{E}_H[a_i]}{Q_0 - \phi\lambda\mathbb{E}_H[a_i]}\left[F_1^0(s) - F_C(s)\right] \leq 0, \tag{26}$$

where the inequality follows from Assumption 1. Thus smoothing shifts the score distribution of defaulters weakly upward toward safer scores.

The score distribution among realized non-defaulters after smoothing is

$$F_0^\lambda(s) = \frac{N_0 F_0^0(s) + \phi\lambda\mathbb{E}_H[a_i]F_C(s)}{N_0 + \phi\lambda\mathbb{E}_H[a_i]}. \tag{27}$$

Subtracting $F_0^0(s)$ gives

$$F_0^\lambda(s) - F_0^0(s) = \frac{\phi\lambda\mathbb{E}_H[a_i]}{N_0 + \phi\lambda\mathbb{E}_H[a_i]}\left[F_C(s) - F_0^0(s)\right] \geq 0, \tag{28}$$

again by Assumption 1. Thus smoothing shifts the score distribution of non-defaulters weakly downward toward riskier scores.

Combining these two dominance relations, the naive rule has weakly more score-rank separation than the smoothed rule:

$$F_1^0(s) \geq F_1^\lambda(s), \qquad F_0^0(s) \leq F_0^\lambda(s) \quad \text{for all } s. \tag{29}$$

Definition 2 therefore implies

$$\mathfrak{R}(F_1^0, F_0^0) \geq \mathfrak{R}(F_1^\lambda, F_0^\lambda). \tag{30}$$

□

B.4 Covered rank-separation metrics

This subsection shows that AUC and KS are rank-separation metrics in the sense of Definition 2, and that recall-at-top-p and top-p lift are covered once their cutoff is fixed by Lemma 2.

Write the tie-corrected AUC of a pair of score-rank distributions (F_1, F_0) as

$$\mathrm{AUC}(F_1, F_0) = \mathbb{P}(S_0 > S_1) + \tfrac{1}{2}\mathbb{P}(S_0 = S_1) = \int h_{F_1}(s)\,\mathrm{d}F_0(s), \qquad h_{F_1}(s) = \tfrac{1}{2}\left[F_1(s) + F_1(s^-)\right], \tag{31}$$

where $S_1 \sim F_1$ and $S_0 \sim F_0$ are independent. Both AUC and $\mathrm{KS}(F_1, F_0) = \sup_s[F_1(s) - F_0(s)]$ depend on scores only through ranks: a strictly increasing transformation applied to both distributions relabels the score axis without changing any of the probabilities in (31) or the set of pointwise differences over which the supremum is taken. For monotonicity, suppose $\widetilde{F}_1 \geq F_1$ and $\widetilde{F}_0 \leq F_0$ pointwise as in (14). Pointwise ordering of distribution functions is inherited by their left limits, so $h_{\widetilde{F}_1} \geq h_{F_1}$ pointwise, and $h_{\widetilde{F}_1}$ is bounded and nondecreasing. Then

$$\mathrm{AUC}(\widetilde{F}_1, \widetilde{F}_0) = \int h_{\widetilde{F}_1}\,\mathrm{d}\widetilde{F}_0 \geq \int h_{F_1}\,\mathrm{d}\widetilde{F}_0 \geq \int h_{F_1}\,\mathrm{d}F_0 = \mathrm{AUC}(F_1, F_0), \tag{32}$$

where the first inequality is pointwise dominance of the integrand and the second holds because $\widetilde{F}_0 \leq F_0$ means $\widetilde{F}_0$ first-order stochastically dominates F_0 and h_{F_1} is nondecreasing. For KS, $\widetilde{F}_1 - \widetilde{F}_0 \geq F_1 - F_0$ pointwise, so the supremum is weakly larger. Both metrics are therefore rank-separation metrics.

Lemma 2 (Smoothing preserves the score distribution). *For every $\lambda \in [0,1]$ and every s,*

$$Q_\lambda F_1^\lambda(s) + N_\lambda F_0^\lambda(s) = \mathbb{P}(s_i^N \le s \mid T_i = H), \tag{33}$$

where $Q_\lambda = D_H(\lambda)$ and $N_\lambda = 1 - Q_\lambda$. In particular, a cutoff defined as a fixed share of the H population is unchanged by smoothing.

Proof. The left side equals $\mathbb{E}_H[q_i(\lambda)\mathbf{1}\{s_i^N \le s\}] + \mathbb{E}_H[(1-q_i(\lambda))\mathbf{1}\{s_i^N \le s\}] = \mathbb{P}(s_i^N \le s \mid T_i = H)$: smoothing reallocates probability mass between the default and non-default states of each borrower at a fixed score. □

For a fixed population share p, define the top-group cutoff $s_p = \inf\{s : \mathbb{P}(s_i^N \le s \mid T_i = H) \ge p\}$, which is invariant to smoothing by Lemma 2, as is the realized top-group share, which replaces p below when an atom sits at the cutoff. The cutoff is defined in the naive score coordinate; for $\lambda < \bar{\lambda}_{\text{rank}}$, Lemma 1 implies that the scorer's top group on reported scores contains exactly the borrowers in the naive-coordinate bottom-p group, so the naive-coordinate expressions below equal the deployed metrics. Recall-at-top-p equals $F_1(s_p)$ and top-p lift equals $F_1(s_p)/p$. With s_p fixed, both are weakly increasing in F_1 pointwise and depend on scores only through ranks, so both satisfy the monotonicity in Definition 2.

B.5 Proof of Corollary 1

Proof. Membership of each metric in the class of Definition 2 is established in Appendix B.4, with the recall and lift cutoff fixed by Lemma 2. For recall and lift, the difference formula (26) evaluated at s_p shows $F_1^\lambda(s_p) < F_1^0(s_p)$ whenever $F_C(s_p) > F_1^0(s_p)$. For KS, write $g_\lambda = F_1^\lambda - F_0^\lambda$ and $d_\lambda = \phi\lambda\mathbb{E}_H[a_i]/N_\lambda \in (0,1)$. Assumption 1 gives $F_C - F_0^0 \ge F_1^0 - F_0^0 = g_0$ pointwise, so the difference formulas (26) and (28) imply

$$g_\lambda \le g_0 - d_\lambda(F_C - F_0^0) \le (1 - d_\lambda)g_0, \tag{34}$$

hence $\sup_s g_\lambda(s) \le (1-d_\lambda)\sup_s g_0(s) < \sup_s g_0(s)$ whenever the naive KS is positive; it is positive except in the single-score case because, under the status quo calibration with μ_H decreasing, $F_1^0 \ge F_0^0$ pointwise with strict inequality at some s. For AUC, writing the tie-corrected statistic as in (31) and using (25) and (27),

$$\mathrm{AUC}(0) - \mathrm{AUC}(\lambda) = \frac{\phi\lambda\mathbb{E}_H[a_i]}{N_\lambda}\int \Delta_0(s)\,\mathrm{d}F_1^0(s) + \frac{\phi\lambda\mathbb{E}_H[a_i]}{Q_\lambda}\int \Delta_1(s)\,\mathrm{d}F_0^\lambda(s), \tag{35}$$

where $\Delta_j(s) = \frac{1}{2}[(F_C - F_j^0)(s) + (F_C - F_j^0)(s^-)] \ge 0$ by Assumption 1. The second stated condition makes the first term strictly positive; the first stated condition makes the second term strictly positive, because $N_\lambda F_0^\lambda = N_0 F_0^0 + \phi\lambda\mathbb{E}_H[a_i]F_C$ dominates $N_0F_0^0$ as a measure, so every set with positive non-defaulter mass under the naive rule retains positive mass after smoothing. □

B.6 Precision-at-top-p

For $\lambda < \bar{\lambda}_{\mathrm{rank}}$, precision-at-top-$p$ equals $Q_\lambda F_1^\lambda(s_p)/p$, with the cutoff s_p fixed by Lemma 2 and top-group membership fixed by Lemma 1; it depends on the default rate as well as the rank distributions, so it is not a rank-separation metric in the sense of Definition 2. It nonetheless falls with smoothing below the threshold. By the numerator identity behind (25),

$$Q_\lambda F_1^\lambda(s_p) = Q_0 F_1^0(s_p) - \phi\lambda\mathbb{E}_H[a_i]F_C(s_p), \tag{36}$$

which is strictly decreasing in λ whenever $F_C(s_p) > 0$. Part of that decline reflects the mechanically lower default rate rather than a worse ranking.