Open Banking and Customer Data Sharing: Implications for FinTech Borrowers*

Rachel J. Nam†

October 20, 2023

Abstract

Open banking allows loan applicants to easily share payment data with prospective lenders during loan applications. In theory, this could broaden credit access by reducing information asymmetry but may also lead to price discrimination that exploits individuals' preferences and behavioral traits. This paper studies the impact of open banking on prospective borrowers and lends empirical support to the sizable benefits of data-sharing driven by improved inferences about borrower credit quality. Using loan application data from a leading German FinTech lender in consumer credit, I show that applicants with observably higher credit risk (with lower credit scores) are more likely to share data. By exploiting the variation of data sharing choices from observably similar applicants, I document that data sharing increases loan approval rates, reduces interest rates, and is associated with lower ex post default rates. These findings suggest that open banking can enhance credit allocation efficiency and reduce adverse selection. *JEL:* D12, G21, G28, G50)

Keywords: Open banking, FinTech, Marketplace lending, P2P lending, Big Data, Customer data sharing, Data access, Data portability

*I am deeply indebted to my thesis advisor, Loriana Pelizzon, for her invaluable guidance and support and am grateful to Matthias Efing, Luca Gemmi, Will Gornall, Jacob Hartwig, Mark Jansen, Xiang Li, Felix Martini, Christine Parlour, and Uday Rajan for thoughtful discussions. I also thank Ashwini Agrawal, Tania Babina, Konstantin Egorov, Andrew Ellul, Aoife Fitzpatrick, Michael Haliassos, Zhiguo He, Jing Huang, Vasso Ioannidou, Christine Laudenbach, Igor Makarov, Christian Muecke, Antoinette Schoar, Daniel Streitz, Huan Tang, and the organizers and participants of the following conferences and seminars for valuable comments: NBER Innovative Data in Household Finance 2022; Bank of England Annual Research Conference 2023; Swiss Winter Conference on Financial Intermediation 2023; Future of Financial Information 2023; SFS Cavalcade NA 2023; FIRS 2023; SAFE Household Finance Workshop 2023; Bank of Canada Payments and Securities Settlement Workshop 2023; German Finance Association Annual Conference 2023; Federal Reserve Bank of Philadelphia Conference on New Perspectives on Consumer Behavior 2023. I also thank Auxmoney for providing the data and the Leibniz Institute for Financial Research SAFE for a stimulating research environment and generous funding.

This paper was awarded the *JFI/FIRS Ph.D. Student Paper Award* from the FIRS Annual Conference 2023, *Best Ph.D. Paper Award* from the Future of Financial Information Conference 2023 (sponsored by BlackRock), *Best Paper Award in Digital Finance* from the DGF German Finance Association 2023, and was selected as a *finalist* for the ECB Young Economist Prize 2023.

†House of Finance, Goethe University Frankfurt & Leibniz Institute for Financial Research SAFE. e.V. Theodor-W.-Adorno-Platz 3, 60323 Frankfurt am Main, Germany. Email: rachel.nam@outlook.com

With the rapid pace of digital transformation and technological advancement, consumer financial activities generate large, diverse (structured and unstructured), high-dimensional, and complex sets of data, referred to as *Big Data* (Goldstein, Spatt, and Ye 2021). In financial markets, a notable example is payment data, which can offer insights into borrowers' cash flow, spending habits, and other financial behaviors that are not typically captured in traditional credit reports. Prospective borrowers, recognizing the potential for such data in reducing information asymmetry with a new lender, might be inclined to share it when seeking to switch providers or applying for loans from different lenders. However, consumers often face considerable friction in data sharing. Banks might be reluctant to facilitate such data transfers due to competitive concerns or for security reasons. Furthermore, the absence of infrastructure can add to the complexity, making the process cumbersome for consumers. This friction can introduce market inefficiencies such as credit rationing and reduced borrower choice due to elevated search and switch costs (Jaffee and Russell 1976; Stiglitz and Weiss 1981). It can also reinforce data monopoly, thereby consolidating the market power of incumbents (Lambrecht and Tucker 2015; de Ridder 2019; Kirpalani and Philippon 2020; Fracassi and Magnuson 2021; Eeckhout and Veldkamp 2022).

Against this backdrop, countries worldwide are adopting open banking, which provides consumers with enhanced control over data sharing. As of October 2021, 80 countries have taken government-led initiatives to promote open banking.[1] In credit markets, this allows potential borrowers to share their transaction data during loan applications easily and securely.

Exploiting the variation in this optional data disclosure and the consequent access to highly detailed consumer financial data, this paper investigates the following questions: which factors influence data sharing decisions, and does such data sharing ultimately benefit borrowers? The answers to these questions are not obvious ex ante because the benefits and costs of data sharing that influence sharing decisions can be heterogeneous across agents, and the subsequent impact will depend on the main use of the shared data. Some users will choose to share data due to the perceived benefits, while others may refrain when costs outweigh (i.e., privacy concerns and the risks associated with revealing negative information). Importantly, willingness to share data can also depend on the type of data being shared (Tang 2019b; Lin 2022). In theory, sharing highly detailed transaction data could mitigate asymmetric information and adverse selection, allowing technology-enabled entities like FinTech lenders to im-

[1] See Babina, Buchak, and Gornall (2022) for an overview of the status of open banking worldwide.

prove credit assessments using advanced algorithms, thereby expanding credit access. On the other hand, it can also open the door to first-degree price discrimination, allowing lenders to extract consumer surplus by capitalizing on individuals' preferences and behaviors. For instance, lenders may use detailed payment data to infer price sensitivity or level of search efforts and offer different interest rates to borrowers with a similar level of credit risk. This practice aligns with existing theoretical models that underscore the potential use of data for digital price discrimination (Chen and Iyer 2002; Taylor 2004; Acquisti, Brandimarte, and Loewenstein 2015; He, Huang, and Zhou 2020; Babina, Buchak, and Gornall 2022).[2] Therefore, who decides to share and whether or not borrowers benefit from doing so is an empirical question.

To the best of my knowledge, this is the first paper to provide empirical evidence addressing contrasting theories on the implications of voluntary disclosure of consumer financial data within the context of open banking in consumer credit markets. Using a rich set of granular loan application data from the largest German FinTech lender in consumer credit, this study examines the potential drivers behind data sharing and its subsequent effects on loan application outcomes. This study exploits a unique empirical setting in which loan applicants are presented with an option to share transaction details from their bank accounts during the loan application process. Through this analysis, I aim to shed light on how detailed payment data shared through open banking is used within credit markets and contribute to the broader policy discussion on potential benefits (and costs) of optional data disclosure for borrowers.

The first question I investigate involves the determinants of data sharing decisions, with a particular focus on observable credit risk as implied by credit scores.[3] The analyses show that applicants with observably higher credit risk are more likely to share data. Applicants with the lowest credit scores are 2.1 percentage points more likely to opt in than those with the highest credit scores, who share data at an average rate of 6.6%. The likelihood of data sharing monotonically decreases as the credit score increases. The results are robust to controlling for other factors, such as age, that might be driving the data sharing decision and are simultaneously correlated with credit score. At first glance, these findings might appear counterintuitive, given the

[2] The EU Consumer Credit Directive/2021 contains an explicit anti-discrimination provision on the basis of nationality, place of residence, sex, race, among other identifiers. First-degree price discrimination, often referred to as personalized pricing, is not directly addressed unless it explicitly discriminates on the protected attributes.

[3] *Schufa* scores are consumer credit ratings generated by Schufa Holding AG, a German credit bureau. It is similar to the FICO score in the US, but Schufa scores use a discrete scale ranging from A (best) to M (worst).

conventional understanding of adverse selection in financial markets. Traditionally, the prevailing theory suggests that individuals with lower credit risk, who are more likely to have better credit scores, would be more willing to disclose their data. Thus, these empirical findings show that a borrower's decision to reveal data is more nuanced.

Individuals might strategically share data to distinguish themselves from others with comparable scores. This sharing serves as an attempt to convey to lenders that there are additional facets of their financial behavior and circumstances not captured by the score alone. This is particularly the case if the revealed credit risk does not match the underlying fundamental risk. Importantly, credit scores, while commonly used, may not always provide an accurate representation of a borrower's credit risk. Such imprecision can arise from the limited availability of relevant information, particularly for borrowers with lower credit scores (Albanesi and Vamossy 2019; Gambacorta et al. 2019; Jansen, Nguyen, and Shams 2023). Using ex post loan payment data and mean squared error (MSE), I confirm that this imprecision is especially pronounced for those with lower scores. In this context, they have more to gain from disclosing private information because that can help correct the larger mismatch between the revealed and fundamental credit risk. Beyond this empirical observation, several other factors may drive data sharing decisions. Sharing data entails costs—from the intrinsic value of privacy to the risk of revealing negative information—which can be heterogeneous across individuals (Lin 2022). For those with low scores and limited outside options, the relative cost might be lower, making them more inclined to disclose. Furthermore, variations in financial literacy could influence disclosure decisions, depending on individuals' understanding of data use.

Given the decision to share data, the natural next question concerns direct consequences. Investigating how data sharing affects loan application outcomes allows me to address different theories surrounding the effects of detailed consumer financial data on lending: improved credit assessments through reduced information asymmetry vs. price discrimination by exploiting consumer preferences. To test this, I quantify the impact of data sharing on loan approvals and interest rates by leveraging variations in data sharing decisions among observably similar applicants using hybrid matching.

The analyses show that data sharing increases the probability of loan approval by up to 11.7 percentage points and leads to lower borrowing costs, with a maximum reduction in the interest rate of 2.2 percentage points. The results are economically sizable and statistically significant given the average loan approval rate is 35.2%, with an average interest rate of 8.9% for each respective group. Data sharing benefits applicants from all credit score groups, but the impact is particularly pronounced for applicants with lower credit scores on the extensive margin (i.e., who enjoy a greater increase in the probability of getting a loan). This result can be explained by the fact that applicants with high credit scores have ex ante a sufficiently high probability of obtaining a loan; thus, data sharing decisions affect loan approval decisions to a lesser degree. However, for applicants on the margin with borderline credit profiles, even a slight improvement in perceived creditworthiness from this supplementary information can significantly increase their chances of loan approval.

The effects of data sharing on interest rates also show heterogeneity, with high-score applicants benefiting from larger reductions in interest rates. This result may appear at odds with the earlier observation that low-score applicants face greater credit score imprecision. Given this imprecision, one may anticipate that low-score applicants would also see more pronounced benefits from data sharing on the intensive margin. Therefore, there is a need to further explore the underlying sources of this heterogeneity. Following the literature that data helps reduce uncertainty (Farboodi and Veldkamp 2020), I investigate the following two channels through which data can affect loan prices: 1) data reveal information, thus changing the lender's prior about the borrower type, and 2) data reduce uncertainty. To test this, I examine the differences in platform-provided scores (internal scores that incorporate information from the shared data) from observably similar applicants. Assuming that observably similar individuals would receive similar platform scores, any improvement in the scores can be at least partially attributed to improvement in the lender's prior as a result of data sharing. If the degree of improvement varies across credit score groups, this suggests heterogeneity in information content and quality. I document that the lender's prior improves by a larger margin, and default predictions become less uncertain with data for high-score applicants. This implies that data shared by high-score applicants contain more positive information and are of higher quality, which could underlie the heterogeneous effects of data sharing on interest rates.

It is important to note that the main analyses employ hybrid matching and use a sample that includes only one application per borrower (when an applicant has made multiple applications, only the initial one is included). However, matching methods are limited in their ability to account for unobserved characteristics that might simultaneously influence both the decision to share data and the outcome variable, which could bias the results. A standard way to address this empirically is by using individual fixed effects. To this end, I examine a subset of applicants who filed multiple applications and whose initial applications were without data sharing and subsequent ones with sharing. This allows me to control for unobserved attributes through individual fixed effects and isolate the effect of data sharing from potential confounders. Additionally, I employ a Rosenbaum sensitivity analysis (Rosenbaum 2002) to test the sensitivity of causal inferences by quantifying how severe unmeasured confounding variables would have to be between the treated and control units to nullify the treatment effect. These robustness tests yield results that are both quantitatively and qualitatively consistent.

Having established that data sharing benefits loan applicants with higher approval rates and lower interest rates, I turn to understanding the nature of the relationship between data sharing and borrower type. The decision to share data can be viewed through the lens of self-selection. By opting to share their financial data, borrowers may be revealing unobservable traits, possibly based on their own assessment of their creditworthiness. From a theoretical perspective, data sharing may be a tool for differentiation from an otherwise similar pool of applicants. If such self-selection is associated with ex post loan performance, it underscores the importance of understanding the motivations for data sharing, not just its immediate benefits. By integrating this theoretical lens with empirical data, I aim to shed light on the relationship between data sharing and latent borrower type.

To test this relationship, I use borrowers' loan payment status conditional on obtaining a loan and document that data sharing is associated with lower *ex-post* default rates among observably similar borrowers who would otherwise have been pooled in the same risk bracket without the additional data. These findings, therefore, confirm the existing theoretical literature which claims that under open banking, latent high types are more likely to opt in (He, Huang, and Zhou 2020; Babina, Buchak, and Gornall 2022). Given that data sharing affects loan prices, I also control for the interest rate in a separate regression to account for the causal impact of loan prices on defaults (due to moral hazard or inability to pay) and provide evidence that once the loan rate is accounted for, data sharing has little to no effect on ex post defaults. This finding suggests that the main use of the data by the lender is to assess credit risk and that the platform effectively prices the risk using the data provided.

This study's findings have far-reaching policy implications. The particularly pronounced positive effects of data sharing on loan approvals for those with lower credit scores and without tangible assets suggest that open banking can be particularly beneficial for asset-light borrowers with thin credit files. Importantly, this may give borrowers more choice and flexibility in selecting financial products and could help address hold-up challenges tied to information asymmetry or limited credit avenues (Fracassi and Magnuson 2021). Moreover, as the importance of consumer financial data grows, more institutions will pursue access. This growth is occurring alongside an increased focus on consumer privacy and governmental regulations regarding data such as that General Data Protection Regulation in Europe (GDRP). Therefore, customer consent will be an essential element in a data-driven economy, and the implications of this study may extend beyond open banking.[4]

The rest of the paper is organized as follows. Section 1 provides a literature review, and Section 2 describes the data and provides descriptive statistics and preliminary evidence of open banking. Section 3 details the empirical methodology, Section 4 reports the empirical results, and Section 5 presents robustness checks. In Section 6, I provide potential avenues for future research and conclude.

1 Related Literature

The existing literature on open banking and sharing consumer payment data is primarily theoretical. Theoretical models indicate that the effects of data portability on welfare may vary with consumers' affiliations with the type of lender (Parlour, Rajan, and Zhu 2022), and whether open banking results in large lender asymmetry favoring FinTech lenders over traditional banks (He, Huang, and Zhou 2020) even when consumers have the option to share data. Providing an empirical perspective, Babina, Buchak, and Gornall 2022 study the role of open banking in fostering innovation and underscore the dual nature of its effects on consumer welfare, depending on the mode of data utilization. Other studies, such as Goldstein, Huang, and Yang 2022 and Brunnermeier and Payne 2022, provide theoretical perspectives on banking competition, resource allocation, and borrower welfare within the open banking ecosystem.

Building on this largely theoretical foundation, this paper makes several contributions. First, I provide empirical evidence of open banking and customer-driven data sharing by leveraging rich loan application data. The granularity of these data allows

[4]As pointed out by Babina, Buchak, and Gornall 2022, open banking has some similarities with credit registries (Djankov, McLiesh, and Shleifer 2007; Hertzberg, Liberti, and Paravisini 2011), but it differs in several respects. Customer financial data often contain a richer set of information; transactions, income, spending, consumption behavior, and so on, and open banking gives the customer the option to sign up. While credit registers are centralized databases that only cover consumers with credit products above a certain threshold, open banking data is available for anyone with a payment account. Importantly, these data are updated in real time, providing a more up-to-date representation of a given consumer's financial state. Most importantly, open banking is a way for consumers to share their financial data with third-party providers for a range of purposes that may extend beyond lending.

for a deeper understanding of the strategic decisions and privacy considerations behind an applicant's choice to share their banking information, extending the literature on optional data disclosure in financial markets. Second, I assess the direct impacts of open banking on loan application outcomes, shedding light on how these data are used in consumer credit contexts. Third, I test the theoretical model predictions from prior studies by investigating the relationship between data sharing practices and borrower types.

Next, I add to the literature examining the role of alternative data in credit markets. Jagtiani and Lemieux 2019 show, by comparing loans from a FinTech lender and banks, that alternative data-based ratings allowed some borrowers to obtain lower-priced credit. Using a German e-commerce platform's data, Berg et al. 2020 show that online user behaviors can predict default risks. Payment footprints can also have higher predictive performance than credit scores (Rishabh 2022). Using BigTech and bank credit, Gambacorta et al. 2020 emphasize that alternative data could minimize the role of collateral, fostering greater financial inclusion. Similarly, Di Maggio, Ratnadiwakara, and Carmichael 2022 highlight the role of alternative data in spotting invisible primes in the personal loan space; that is, borrowers with low credit scores and short credit histories but also a low propensity to default. This paper provides further evidence of these findings.

My study contributes specifically to the role of payment data in credit risk assessment. Ghosh, Vallée, and Zeng 2021 study the impact of cashless payments by firms on loan application outcomes both at the extensive and intensive margins, using data from a large Indian SME FinTech lender. Exploiting variation in the degree of cashless payments vis-à-vis cash by firms, the authors find that a larger use of cashless payments predicts a higher chance of loan approval, a lower interest rate, and a lower risk-adjusted default rate.[5] This work is the closest to my study in its empirical setting but is different in three ways. First, I use consumer loan data rather than small business loan data. Second, in their loan application, data sharing is mandatory; thus, it does not allow for examining different characteristics among borrowers who do or do not sign up. Last, for the aforementioned reason, their paper does not directly connect to open banking and consumer data rights but rather closely to the value of customer transaction data.

[5] In a similar vein, Ouyang 2021 studies the impact of mobile cashless payment on credit provision to the underprivileged, using a sample of Chinese BigTech *Alipay* users and finds a positive impact of in-person payment flow on credit provision.

Lastly, I contribute to the growing literature discussing the role of technology in reducing market inefficiencies and disparities. Philippon 2016 highlights that the cost of financial intermediation by traditional players remained surprisingly expensive despite technological advances and has thus resulted in the emergence of new players. Big data are often key to their business models, and they can reduce the impact of negative prejudice in the credit market (Philippon 2019), such as racial disparities, by automating the lending processes (Howell et al. 2021). FinTech lenders also serve in areas with less bank presence, lower incomes, more minority households (De Roure, Pelizzon, and Thakor 2022; Erel and Liebersohn 2022), and higher business bankruptcy filings and unemployment rates (Cornelli et al. 2022).

These new players may directly compete with traditional lenders like banks by serving infra-marginal borrowers who value immediacy and have a higher willingness to pay (Buchak et al. 2018; Tang 2019a) or complement bank lending by absorbing unmet demand (Gopal and Schnabl 2022; Sheng 2021; Avramidis, Mylonopoulos, and Pennacchi 2022; De Roure, Pelizzon, and Thakor 2022). Algorithmic lending can also benefit consumers via more efficient loan application processing (Fuster et al. 2019) and mitigating agency conflicts and humans' limited capacity (Jansen, Nguyen, and Shams 2023). Importantly, FinTech loans can greatly alleviate financing constraints faced by SMEs and further improve access to bank financing by providing uncollateralized loans that can be used to acquire pledgeable assets (Beaumont, Tang, and Vansteenberghe 2022; Eça et al. 2022).

While the above studies underscore the broad operational domains of FinTechs and algorithm lending, my study differs in that it delves into the details of data sharing within the open banking framework, analyzing the interaction between how consumers' explicit decisions to share payment data and data-driven screening technology influence the prospects of obtaining a loan.

2 Data

2.1 Institutional setting, descriptive statistics, and evidence of open banking

This section provides the institutional background of open banking and the FinTech lender that supplied the data for this study, descriptive statistics, and descriptive evidence of open banking.

2.1.1 Open banking regulation

Open banking repositions data ownership from banks to customers, thus enabling consumers to access and exert more control over how their financial data are shared. As of October 2021, 80 countries worldwide have at least a nascent government-led open banking effort. Most are still in the early-discussion phase, but 32 countries have fully implemented the policy (Babina, Buchak, and Gornall 2022).[6] The details of open banking regulations differ across jurisdictions. Whereas certain countries impose obligatory data sharing, others merely advise it or offer technical standards and infrastructure to support data sharing.[7] The scope of financial data covered by open banking varies from transaction data to records of savings, lending, and investments. The European Union and the United Kingdom are at the forefront of open banking policies; they are now considering extending the policy. Under the revised Payment Service Directive 2 (PSD2) Access to Account, all European Union institutions offering payment accounts must provide third parties (both banks and non-banks) access to a customer's transactional account information when the customer consents. They are also required to offer dedicated application programming interfaces (APIs)[8] to facilitate secure access. This took effect in January 2016 and was required to be adopted into national laws by January 2018. Given this regulatory environment, Europe serves as a suitable setting to study the impact of customer data sharing on borrowers.[9] For instance, Germany incorporated the directive into its national legal framework on January 13, 2018. As a consequence, this present study considers loan applications from January 13, 2018, to May 22, 2022, ensuring that the legal mandate for open

[6]Babina, Buchak, and Gornall (2022) provide an excellent description of the status of open banking worldwide. In the United States, the Consumer Financial Protection Bureau (CFPB) was tasked under Section 1033 of the Dodd-Frank Act of 2010 to formulate open banking regulations. The CFPB declared in October 2022 its aim to establish definitive regulations by 2024, with the execution phase to follow.

[7]Jurisdictions with mandatory data sharing rules include Australia, Bahrain, Brazil, the European Union, and Israel. By contrast, in Singapore, Malaysia, and Russia, banks are recommended to share and regulators facilitate the process by mediating industry discussion, providing technical standards for APIs, or providing infrastructure for data sharing. For more information, see Babina, Buchak, and Gornall 2022

[8]An API is a software intermediary that allows two applications to communicate with each other. By facilitating customer data sharing among different institutions, APIs play a critical role in securely transferring data and simplifying the customer journey, thus encouraging consumer participation in open banking. Before open banking, sharing bank details was possible, but without an automated process, it was costly and complex for many consumers.

[9]In Europe, open banking is promoted by the European Commission as part of a digital agenda to open up services, provide choice, and foster competition and innovation in the market. For more information, see https://www.openbankingeurope.eu/who-we-are/.

banking-driven data sharing is consistently applied throughout the examined period.

2.1.2 Description of the platform

The original data include approximately 18 million loan applications from the largest German FinTech lending platform, Auxmoney. Founded in 2007, it has originated more than EUR 2.3bn in 319,535 consumer loans between January 2018 and May 2022, and more than EUR 3bn since its inception, making it one of the largest consumer credit marketplace lenders in continental Europe. A prospective borrower can register on the website, enter a desired loan amount anywhere between EUR 1,000 and EUR 50,000, and be guided through an application process during which the applicant is asked to provide a set of personal information and loan details, including loan purpose, employment status, and income and expenses. As a FinTech platform, Auxmoney is not a licensed bank and is thus not subject to banking regulations. To issue loans, it partners with a fully licensed credit institution.

Upon submission of a loan application, the platform evaluates the creditworthiness of the applicant with a platform score: classes AA, A, B, C, D, E, or Z. Those assigned a score of Z are deemed ineligible. In the scoring process, much like traditional banks, the platform initially uses the Schufa score (henceforth the "credit score"), a consumer credit rating generated by Schufa Holding AG, a German credit bureau.[10] Unlike traditional banks, which often exclude specific demographics like students, self-employed individuals, or temporary workers deemed to be risky,[11] the platform does not automatically disqualify specific groups. In the preliminary screening, emphasis is placed on an applicant's historical default records. If applicants meet the criteria in this phase, they advance to a subsequent evaluation, where extensive datasets and digital consumer metrics are employed to compute internal credit ratings, the Auxmoney score (henceforth, the "platform score"). The platform score is primarily derived from five distinct data sources: registration details, credit scores and additional financial information from the credit bureau, behavioral data, web data, and experience data.[12] The entire process is automated, ensuring that in instances of approved applications

[10] In contrast to the United States, Germany assigns credit scores without requiring an extensive credit history; even basic financial activities like maintaining a checking account or paying utility bills will yield a credit score

[11] Under stricter banking regulations such as risk-weighted capital requirements, it is costlier to extend credit to high-risk borrowers since a larger capital buffer has to be set aside to service them. This can result in banks reducing lending to high-risk borrowers (Berger and Udell 1994; Kashyap, Stein, et al. 2004; Popov and Udell 2012; Roulet 2018; Benetton et al. 2021).

[12] For more information, https://www.auxmoney.com/faq/auxmoney-score.

and agreed-upon loan contracts, loan disbursements typically occur within a few days.

Funding for loans comes from both individual and institutional investors. Initially, the platform employed a pure peer-to-peer lending model in which investor and borrower were directly matched. In this disintermediated lending structure, individual lenders selected specific loans to fund, and the platform was not burdened with maturity transformation or information-gathering costs. However, as institutional investors became more involved in the funding process, the platform transitioned to the marketplace model, in which the platform undertakes borrower risk assessment, addressing information asymmetry between retail and institutional lender types and offers diversified loan portfolios (Balyuk and Davydenko 2019; Vallee and Zeng 2019; Braggion et al. 2020). A significant fraction of these loans are now securitized.[13] The study focuses solely on the post-transition period.

2.1.3 Descriptive statistics

As shown in Figure 1, the number of applications on the platform increased steadily over time, except for a noticeable slow down in 2020. Since the beginning of 2021, loan demand on the platform has experienced an uptick, reaching its peak at the end of the sample period. The number of paid-out loans (loan offers accepted by applicants) follows a similar trend.

[Figure 1]

It is important to note it is also possible for an applicant to submit multiple applications. Successful applicants might do this to compare terms across different loan offers. Meanwhile, rejected applicants might return to the platform and apply again. Including multiple applications from the same applicant could lead to over-representation and makes it challenging to discern strategic behaviors regarding data sharing. Thus, only the initial application from each borrower is considered. I also exclude incomplete applications since they lack critical information necessary for the analysis. The final sample consists of 2,484,987 completed loan applications between January 13, 2018, and May 15, 2022.

Table 1 presents the descriptive statistics of the dataset.

[Table 1]

[13] Auxmoney has issued three asset-backed security transactions named Fortuna Consumer Loan ABS, comprising approximately 48,000 loans totaling EUR 350 million in 2023, 25,000 loans totaling EUR 225 million in 2022, and 30,000 loans amassing EUR 250 million in 2021.

The average requested loan amount on the platform was EUR 13,667, with a typical loan term of 55 months. The mean age of applicants was 38 of whom and 65% were male. The platform approved approximately 66% of these loans, with a mean interest rate of 12%. The average credit score stood at 3.12, based on a 4–1 scale (4 being the best credit score group).[14] The median applicant had a monthly income of EUR 1,950, and monthly expenses of EUR 590. A majority (93%) had checking account(s), 63% had one or more credit cards. 24% were homeowners, and 55% had at least one automobile. The variables *Number of current* and *past loan demand* provide a proxy for the number of outstanding and previously held loans, respectively. The average applicant is found to have 1.4 active consumer loans and a historical record of one fully settled loan. The main variable of interest *Signup*, is a dummy variable that takes a value of one if the applicant shared bank transaction data during the loan application process. This indicates that 8% of the applicants during the entire sample period shared their bank account details. Descriptive statistics broken down by data sharing choices can be found in Online Appendix Table A.1.

Panel A in Figure 2 provides a timeline of data sharing rate over time. There is a clear upward trend in open banking participation by borrowers over the period under consideration. This consistent increase is observable across all credit score categories, with those in higher risk brackets (and thus lower credit scores) exhibiting a greater propensity to share information. This observed trend is in line with theoretical expectations that open banking adoption would grow as FinTech lenders refine their business models (He, Huang, and Zhou 2020). The intuition is that over time, FinTech lenders establish their niche markets with improved business models, which enables them to capture more customers.

[Figure 2]

Younger individuals tend to be more comfortable interacting with technology, which may partially explain the rise in the open banking participation rate. However,

[14]Numerical values are assigned to the credit score categories such that high scores correspond to higher implied credit quality: 4 for scores A–D (highest), 3 for scores E–G, 2 for scores H–K, and 1 for scores L–M (lowest). Any applications with a score less than M are excluded from the sample and some applicants had no reported credit scores. There are several issues with this category. For instance, when a person who recently arrived in Germany for the first time applies for a loan on the platform, her credit score will be marked as non-existent by the credit bureau since this person has no credit file registered in Germany. However, if she tries again to apply for a loan, she might be assigned a credit score since her profile had been registered with the credit agency. This may introduce an inconsistency in the data. Therefore, these observations are also dropped.

as shown in the inter-quartile range of age in Panel B from Figure 2, applicant age has stayed fairly constant over time.

2.1.4 Data sharing process and descriptive evidence of open banking

The loan application procedure is divided into three phases: (i) application, (ii) decision, and (iii) loan payout. During the application phase, users submit personal details and specify their desired loan amount and duration. They are also given the option to share their transaction details from a bank account.[15] During this process, applicants are presented with an interface detailing the data sharing option and a message describing the potential benefit of providing bank data (i.e., an average discount on a loan), which is shown to everyone (more details are presented in the Online Appendix Figure A.1). This message may vary over time, albeit at longer intervals. If an applicant opts for data sharing, the platform will gain access to the most recent four months of their banking transactions. Along with presenting the potential advantages of data sharing, the platform also provides applicants with comprehensive and legally mandated information, including a clear outline of how their data will be used. Additionally, the platform discloses that data sharing can have both positive and negative implications. It is stated that while sharing data might offer favorable loan terms for some, it might also lead to negative outcomes for others, such as loan application rejection or a higher proposed interest rate.

When the applicant consents to data sharing, this shared information, combined with other sources like credit bureaus, application details, and digital traces, is used to compute the platform score, a proprietary credit scoring system. In the second phase, the decision phase of the process, this platform score, along with the success or failure of the application, is communicated to the applicant, and successful applicants are also provided with an interest rate. In the final, loan payout phase, the applicant decides to accept or decline the loan offer, leading to either the disbursement of funds or termination of the process.

Panel A in Figure 3 provides a first glimpse of evidence of open banking. It shows simple averages of loan acceptance rates by data sharing decisions across different credit score groups. The difference in approval rate between those who do and do not share appears to be larger for applicants from the lower credit score groups.

[15] To facilitate this, the platform incorporates a secure API interface provided by a third party, enabling applicants to seamlessly log in to their respective banks. Importantly, the platform employs a three-factor authentication process, ensuring that bank login details remain confidential and are never visible to the platform.

This preliminary evidence aligns with expectations that applicants with good credit scores are typically well positioned for loan approvals, rendering additional data less impactful on the decision of whether to grant a loan.

[Figure 3]

Data sharing is also associated with lower interest rates across applicants of all credit scores (panel B in Figure 3). Notably, the largest difference in interest rates is seen among top-tier applicants. In comparison, applicants from the lower credit score brackets see a more modest difference in their interest rates. This preliminary evidence suggests that open banking may favor high-quality borrowers in terms of interest rates, but it offers prospective borrowers across the spectrum some degree of favorable loan prices.

One crucial factor to underscore is that the choice of applicants to share their bank account data is not made at random. Indeed, even within the same credit score group, the attributes of borrowers opting to share might differ *systematically* from those who refrain from sharing. Consequently, directly comparing these two groups without accounting for these underlying disparities might lead to biased conclusions about the impact of open banking. To address this issue, I employ a hybrid matching approach in the subsequent analysis to match borrowers who choose to share data with those who do not, based on several observed characteristics to make the groups comparable.

3 Methodology

This section provides the regression models used for the analysis, matching methods and results, and selection bias corrections.

3.1 Probit Analysis of Data Sharing Choices

To estimate the determinants of open banking participation, I use a probit model and estimate the following:

$$\text{Pr}\ (Signup_i = 1) = \Phi\,(X_i'\beta + G_i'\gamma + Year + \epsilon_i), \tag{1}$$

where *i* indexes an individual and $Signup_i$ is an indicator variable equal to one if the applicant shares data and zero otherwise. X_i are applicant characteristics, including age, credit score, income, dummy variables indicating gender, main earner, homeowner, car owner, the number of outstanding loans, and fully paid loans. G_i are loan characteristics such as loan amount, loan duration, and loan application channel.[16] $Year$ are year dummies, ϵ_i is the error term, and Φ is the standard normal cumulative distribution function. I am mainly interested in the coefficient β, which measures the change in the likelihood of sharing data across different applicant traits. In particular, the main question is how one's observed credit risk as implied by credit scores, is associated with data sharing. In other words, is it *observably* riskier or safer applicants who are more likely to share data? To this end, the coefficients for each credit score group are of central interest. Later, I also explore how borrower type as implied by ex post defaults, is associated with data disclosure. Standard errors are clustered at the zipcode-year level.

3.2 Matching on observables

The next step examines the effect of open banking participation on loan approval and interest rate. It is important to note that applicants who share data may be *systematically* different from those who do not. Therefore, using the full sample to estimate the effect of *Sign up* on the probability of loan approval or the interest rate may introduce bias. To address this issue, I employ a hybrid matching method to address potential selection bias and ensure comparability between the treatment and control groups. This approach combines two matching techniques to achieve optimal balance on observed covariates: exact matching and propensity score matching (PSM). Given that borrower traits may differ substantially across access channels[17] and that the data sharing trend fluctuates over time, exact matching is applied to the variables *Access channel* and *Loan application year*, ensuring that these categorical covariates are precisely matched between the treated and untreated groups. On the other hand, PSM is used for *Age*, *Income decile*, and *Credit score*. Using PSM allows for a degree of flexibility, creating matches based on the similarity of propensity scores, which are computed through logistic regression using the three aforementioned variables as predictors.

3.3 Probit Analysis on Data Sharing and Loan Approval

I use the matched sample to estimate the effect of data sharing on the probability of loan approval using a probit model,

[16] Loan access channel is a categorical variable that indicates the channel through which the user applies for a loan. There are five such channels: 1) directly via the Auxmoney homepage, 2) repeat loan, 3) price comparison websites, 4) brokers, and 5) banks.

[17] See details in Online Appendix Table A.2.

$$\Pr(Approved_i = 1) = \Phi(\rho Signup_i + \sigma_k (Signup_i \times Credit\,score\,group_i) + X_i'\beta + G_i'\gamma + Year + \epsilon_i), \tag{2}$$

where $Approved_i$ is an indicator variable that takes a value of one if the loan application is approved and zero otherwise. $Sign\ up_i$ is an indicator variable that takes a value of one if the person participates in open banking by signing up to share bank account data, and zero otherwise. To examine whether data provision has different effects across credit risk groups, I include the interaction term $Sign\,up_i \times Credit\,bureau\,score_i$. The other variables are the same as in equation (1). The main coefficients of interest are ρ and σ_k which measure the change in the likelihood of loan approval by data sharing decision $Sign\,up_i$, and the differential effect across different credit score categories $k = 4,3,2,1$ (4 (A–D) the best and 1 (L–M) the worst), respectively.[18] It should be noted that matching methods do not account for any unobserved characteristics that may simultaneously determine the selection into treatment and the outcome variable. The omission of such variables may result in endogeneity bias. To address this issue, I focus on a particular group of applicants who submitted multiple applications, initially one without data sharing followed by one or more with data sharing. This method accounts for unobserved individual attributes through fixed effects and distinguishes the data sharing effect. These robustness checks are shown in Section 5. I further apply a Rosenbaum sensitivity analysis to assess the potential influence of unmeasured confounders.

3.4 Data Sharing and Interest Rates

Next, I examine the effect of data sharing on loan interest rates. It is important to note that interest rates are only available for approved loans, leaving a gap in understanding how the decision to share data would have influenced the interest rates of rejected applications. Since the set of approved loans is not a randomly drawn sample, drawing conclusions about interest rates based only on this subset might introduce bias. To rectify this issue, I employ the Heckman correction method to counteract the potential omitted variable bias from this specific sample selection challenge (Heckman 1976; 1979).[19]

[18]Numerical values are assigned to the credit bureau score (*Schufa*) categories such that high scores correspond to higher implied credit quality: 4 for scores A–D (highest), 3 for scores E–G, 2 for scores H–K, and 1 for scores L–M (lowest).

[19]A model explanation appears in Online Appendix B.

I estimate the following equation to assess the effect of data sharing on the interest rate,

$$r_i = \theta \hat{\lambda}_i + \rho Signup_i + \sigma_k (Signup_i \times Creditscoregroup) + X_i' \beta + G_i' \gamma \\ + Year + \epsilon_i, \tag{3}$$

where r_i indexes interest rate, and $\hat{\lambda}_i$ is the inverse Mills ratio. The other variables are the same as in equation (2). The main coefficients of interest are ρ and σ_k, which respectively measure the change in the interest rate by data sharing decision $Signup_i$ and the differential effect across different credit score categories $k = 4, 3, 2, 1$. A negative θ implies a negative correlation between the error terms and proves the presence of downward selection bias. In other words, applicants with a below-average interest rate and are thus safer are selected for the approved pool of applicants. A priori, the sign of θ is unclear. The platform may prefer borrowers with high interest rates so as to maximize its returns or contrarily select relatively safe borrowers. All the other variables are the same as in equation (1).

4 Main Results

In this section, I present the factors influencing data sharing decisions and their impacts on loan approval rates and interest rates. Then, I show the results from the analyses regarding the association between data disclosure and ex post borrower type and introduce economic mechanisms illustrating how data provision influences loan application outcomes.

4.1 Factors influencing data sharing decisions

Table 2 reports the estimation results of equation (1) regarding the factors influencing data sharing decisions. Column (1) only includes credit score variables, column (2) only age, column (3) uses both, and column (4) reports all estimates, including all applicant and loan characteristics, access channel, and year dummies. Column (1)–(3) report marginal effects using probit, and Column (5)–(8) report ordinary least squares estimates.

[Table 2]

The results highlight that applicants with *observably* higher credit risk, as measured by lower credit scores, are more likely to share their bank account data than those with better credit profiles. In economic terms, an applicant in the L–M credit score category, denoting the highest implied risk, is on average 3.9 percentage points more likely to share data than an applicant in the A–D category, representing the lowest implied risk, as illustrated in column (1). The likelihood of data sharing monotonically decreases as credit score improves. This suggests that those with higher scores might be more hesitant to disclose account information. There may be an age-related explanation for this trend. Younger individuals often display a greater willingness to embrace technology and may have shorter credit histories, which results in lower credit scores. To take into account potential confounding factors that could influence the outcome and might be correlated with credit scores, additional controls are included in column (4). Although the magnitudes of the primary coefficients decrease slightly, they remain statistically significant, with a 2.1 percentage point difference between the lowest (L–M) and highest (A–D) credit score categories. This is a sizable magnitude, given that the average rate of data sharing for the main sample is 8%. Initially, these results might seem at odds with the standard theory regarding adverse selection, which holds that those with a better credit standing who often having higher credit scores on average, would be typically more inclined to share this information to stand out from the rest. Yet, these findings indicate that data sharing decisions are much more nuanced.

An important factor to bear in mind is the varying incentives across applicants for sharing their data. Credit scores, while commonly used, may not always accurately capture the actual credit risk of a given applicant. In other words, applicants with lower credit scores might not only have lower credit quality on average, but their scores might also be subject to greater imprecision due to wider variations in the underlying credit factors (Albanesi and Vamossy 2019; Gambacorta et al. 2019; Jansen, Nguyen, and Shams 2023). I use the MSE to test the inference quality of credit scores in predicting defaults. By leveraging ex post default data, I quantify the forecasting error to measure the discrepancy between actual defaults and predictions based on credit scores. A lower mean implies a smaller error in the prediction model.[20]

[Table 3]

As shown in Table 3, credit scores are less precise in predicting default risk for applicants with lower credit scores than for their counterparts with higher scores. This

[20] To mitigate the possibility that the credit score simply captures a mix of other factors that affect defaults, I control for the following borrower and loan characteristics: age, income decile, loan amount, loan duration, female (dummy), main earner (dummy), number of current and past loans, homeowner (dummy), car owner (dummy), and access channel (categorical variable).

suggests that individuals with lower scores might perceive greater potential benefits from data disclosure, particularly if they believe additional information would better represent their self-assessed creditworthiness.

Beyond this empirical observation, there are other potential explanations not directly explored in this study. Sharing data comes with inherent costs, including privacy concerns and the risk of revealing unfavorable information, and these costs could vary across individuals (Lin 2022). Additionally, differences in financial literacy regarding how the shared data is used could influence applicant willingness to disclose information.

The results also highlight heterogeneity in data sharing with respect to gender and age. Female applicants are 0.4 percentage point less likely to sign up than their male counterparts. Holding other factors constant, a 48-year-old applicant is two percentage points less likely to share data than someone who is 38 years old. These observations align with prior studies that suggest women and older individuals tend to have greater privacy concerns (Goldfarb and Tucker 2012). Income, on the other hand, does not appear to be an important factor (Tang 2019b). Individuals with more outstanding consumer loans and fully repaid past loans exhibit a higher propensity to share. These results suggest that these applicants may have reached their maximum debt capacity and are thus more financially constrained. Building on the primary analysis conducted with a probit model, I also estimate ordinary least squares regressions as a supplemental analysis (Columns 5–8). The results are consistent, both qualitatively and quantitatively, with those of the main estimates using a probit model.

4.2 The effect of customer data sharing on loan approval

In this section, I investigate whether data sharing affects loan approval. Before conducting the probit analysis, I perform the matching procedure described in Section 3.2. The goodness of the matching procedure is assessed with *t*-tests for the null hypothesis of equal means for both treatment and control groups. Detailed matching results are reported in Online Appendix Table A.3 and indicate that the matching is successful.[21]

[Table 4]

As Table 4 shows, sharing data improves the probability of loan approval for applicants across all credit score groups, and the magnitudes are both statistically

[21] I test including further matching variables such as loan amount and loan duration; the results are both quantitatively and qualitatively similar.

significant and economically sizable. The results reveal a hump-shaped relationship between data sharing and loan approval. Those who benefit most are from the second-lowest credit score group (H–K) (an 11.7 percentage point increase in the likelihood of loan approval) followed by the second highest (E–G) group (an 8.5 percentage point increase). The magnitudes are sizable, given that the average approval rates in the main sample are 35.3% (H–K) and 61.7% (E–G), respectively. The effects are relatively less pronounced for the highest (a 1.5 p.p increase), and the lowest score group (a 3.8 p.p increase); the average approval rates are 88.5% (A–D) and 13.7% (L–M).[22]

These findings indicate that mid-low tier applicants are more likely to be on the margin of qualifying for a loan. While the effects are quantitatively sizable, the heterogeneity in its impact is not entirely surprising. Applicants with high credit scores already possess a high probability of obtaining a loan, making the impact of any additional shared data less pronounced. Similarly, for applicants with the lowest credit scores, the effect is also relatively smaller, as their ex ante probability of loan approval is already quite low. By contrast, for those on the margin with borderline credit profiles, even a minor positive shift in perceived creditworthiness due to extra information might turn a likely rejection into an approval.

Notably, home ownership positively affects loan approval rates, an indication that the existence of tangible assets can provide an implicit guarantee for fund recovery in default scenarios. However, the coefficient on the interaction term of home ownership and *Signup* is significantly negative, suggesting that data sharing is substantially more valuable for applicants without tangible assets.

It should be noted that matching methods do not account for any unobserved characteristics that may simultaneously determine the selection into treatment and the outcome variable. Therefore, the omission of such variables may result in biased results, giving rise to the endogeneity problem. A standard way to address this empirically is by using individual fixed effects. In Section 5, I conduct robustness checks using a subsample of individuals who filed multiple applications with varying data sharing decisions to account for unobserved individual characteristics. I further implement a Rosenbaum sensitivity analysis. The results are qualitatively and quantitatively consistent.

[22] It is important to note that estimating the marginal effect of interaction terms in a non-linear model is not straightforward (see Ai and Norton 2003). To compute the effect of data sharing by credit score group, I obtain margins with nested designs. In this case, each credit score group is treated as a nesting variable over which margins of data sharing are estimated.

4.3 The effect of customer data sharing on loan interest rate

In this section, I investigate whether data sharing affects loan interest rates. It is important to note that interest rates are revealed conditional on loan approval. Therefore, using only a subset of approved loans to estimate the effect of data sharing on interest rates could introduce bias. To address this issue, I employ the two-stage Heckman selection model (see Online Appendix B).

Table 5 reports the results for the main probit analyses in equation (3). Column (1) reports the baseline results, and column (2) presents estimates after correcting for selection bias using the Heckman two-stage selection model.

[Table 5]

Data sharing leads to lower interest rates, and the effects are heterogeneous by credit score categories. This time, however, the safest borrowers, as measured by the highest credit scores, enjoy the largest reduction in interest rate at 2.26 percentage points. The magnitudes are smaller for the other groups, 2.10, 1.37, and 0.61 percentage point reductions for the E–G, H–K, and L–M groups, respectively. The effects are sizable, given that the average interest rates in the main sample are 8.9%, 12.2% 13.8%, and 14.9% for the A–D, E–G, H–K, and L–M groups, respectively.. The coefficient of the inverse Mills ratio is negative and statistically significant, which suggests that there is a downward selection bias; that is, the platform has selected loans with interest rates lower than the average interest rate of the population, and the unselected loans would have been charged higher interest rates.

Notably, home ownership is associated with a reduction in interest rates by 2.3 percentage points. Even though these loans do not require collateral, such tangible assets may still offer lenders an assurance of potential avenues for debt recovery in case of default. The benefit of home ownership is similar in magnitude to data sharing. However, when interacted with data sharing, home ownership yields a less pronounced reduction of 1.79 percentage points. In essence, the effect of data sharing is less marked for homeowners than for non-homeowners. This observation underscores the economic parallels between data and collateral (Gambacorta et al. 2020), emphasizing the distinct value of data sharing for those without tangible assets.

4.4 Are ex post good-types more likely to share data?

The choice to disclose data can be seen as a form of self-selection, offering insights into an otherwise unobservable borrower type and possibly reflecting their self-assessed creditworthiness. In theory, data disclosure may allow borrowers to distinguish themselves from a group of similar applicants. In this section, I investigate the association between data sharing decisions and underlying borrower type to shed light on any such strategic motivations.

To test this, I use ex post loan payments to infer borrower type[23] and define a loan as in default if payment delay exceeds 90 days. Then, I regress *Default* on data sharing among observably similar borrowers who would otherwise have been pooled in the same risk bracket. I also control for other variables that could directly influence defaults. The results from Panel A in Table 6 confirm that data sharing is associated with lower ex post default rates.

[Table 6]

Borrowers with the highest credit scores (ranging from A–D) who choose data sharing have a 1.1 percentage points lower likelihood of default compared to observably similar borrowers who refrain from disclosing data. For the lower credit score group (E–G), the difference is 1.5 percentage points, for H-K, it is two percentage points. This indicates that individuals with an inherently lower risk of default are more predisposed to data sharing.[24]

Considering the influence of data sharing on loan pricing, I also add the interest rate as a control in a separate regression. This accounts for the potential causal impact of loan prices and defaults, which could arise from moral hazard or inability to pay. Panel B in Table 6 indicates that factoring in the interest rate makes little to no difference in the impact of data sharing on defaults. This observation underscores that the primary application of the shared data is for evaluating credit risk, and the platform correctly calibrates the risk into the interest rates using the shared data.

It is worth noting that adding the interest rate in the regression with a matched sample changes the interpretation of the coefficient of data sharing, *Signup*. If Borrower A shares data and receives the same interest rate as Borrower B, this implies

[23] Data on the payment status of the loans come from the European DataWarehouse, a securitisation repository designated by both the European Securities and Markets Authority and the Financial Conduct Authority. It was established in 2012 as the first securitisation repository in Europe to facilitate the collection, validation, and downloading of standardized loan-level data for asset-backed securities and private whole loan portfolios. For more information, see https://eurodw.eu/. Given the limited sample size due to available loan outcome data, I conduct an additional robustness check using the change in platform scores as a proxy for unobserved borrower type in Online Appendix C.1

[24] The results for borrowers with the lowest scores remain indeterminate due to the constraints of a limited sample size.

that, on average, Borrower A would have received a higher interest rate absent data sharing. Put differently, two borrowers would have been pooled separately without data sharing, with A in a pool with riskier borrowers. By sharing data, Borrower A moves to a better pool with Borrower B and now receives the same interest rate. Thus, the minimal impact of data sharing after controlling for the interest rate suggests that the default probability of borrowers with different observable characteristics, with Borrower A being worse, does not differ significantly. This is consistent with the suggestion based on theoretical models that unobservably good types differentiate themselves by data sharing (He, Huang, and Zhou 2020; Babina, Buchak, and Gornall 2022; Parlour, Rajan, and Zhu 2022).

While data sharing enables lenders to better screen borrowers, its aggregate effect remains ambiguous. As more individuals choose to share their data, those who opt out might be perceived as posing a higher risk, irrespective of their underlying creditworthiness. Notably, applicants with higher credit scores, being less inclined to share data, could face heightened discrimination even as this data transparency might result in reduced credit rationing for those at the lower end of the credit score spectrum. Consequently, while data sharing may seem to optimize the lending process, it introduces new complexities and potential disparities. This paper's results should be interpreted with caution when considering welfare implications, as that is not the primary focus of the study.

4.5 The channels through which data sharing affects loan application outcomes

The findings from the previous sections highlight the heterogeneous effects of data sharing on credit decisions on both the extensive and intensive margins. On the extensive margin (the effect on loan approval), lower credit scores benefit more from data sharing, and this heterogeneity can be intuitively interpreted. Applicants with high credit scores, who inherently have a higher likelihood of loan approval, experience more muted effects from data disclosures. By contrast, for those with marginal credit profiles, the effects are more pronounced, markedly shifting their approval probabilities from potential rejection to acceptance. On the intensive margin (the effect on the interest rate), it is the high-score applicants who experience larger reductions in interest rates, which may appear at first contradictory considering the more prevalent credit score imprecision among low-score applicants, which opens up greater room for interest reduction. Therefore, this outcome necessitates further examination into the underlying dynamics contributing to this heterogeneity of the effect of data sharing on the interest rate.

Thakor and Merton 2018 suggest that FinTech firms might be more susceptible to trust erosion after borrower defaults relative to traditional banks. Meanwhile, Ben-David, Johnson, and Stulz 2021 emphasize the financial constraints that characterize FinTech lenders, which contrasts with the more stable deposit streams of banks. Drawing on this point, I assume the FinTech lender to be risk-averse and that data can help alleviate uncertainties, as highlighted in the data and information literature (Farboodi and Veldkamp 2020). Consequently, I examine two primary channels through which data can influence loan prices: 1) the adjustment in the lender's prior about the borrower type because data can reveal type and 2) the reduction of uncertainties.

To investigate these mechanisms empirically, I estimate the change in the platform score, which is an internal credit score assigned by the FinTech lender once an application is completed. When an applicant decides to disclose her transaction data during the loan application, these data, along with other variables, factor into the platform score calculation. Therefore, should this shared information improve the lender's prior of an applicant's creditworthiness, that would translate to more favorable loan prices. To test this supposition, I use a matched sample consisting of two groups who are comparable in observable traits but diverge in their data sharing choices. In the lender's eyes, these two applicant groups are largely indistinguishable, implying that their financing outcomes should not, in theory, demonstrate marked differences. Yet, if the data-disclosing group consistently achieves higher platform scores than their non-disclosing counterparts, that indicates an improvement in the lender's initial assessment as a result of data sharing and/or that the shared data reduces uncertainties.

[Table 7]

As depicted in panel A in Table 7, the magnitude of improvement in the platform score is not uniform, with a notably larger increase observed for high-score applicants. The increase in the platform score can be attributed to both an improvement in the lender's prior and mitigation of uncertainty. To separately measure the effects of data sharing on risk reduction, I first assess the default forecasting error using MSE and evaluate the degree to which data sharing diminishes risk by measuring the reduction in the standard deviation.

Panel B in Table 7 indicates that the default prediction quality improves with data. Particularly, for borrowers with high scores, data sharing leads to a greater reduction in the variability in the prediction errors (standard deviation); thus, predictions become more consistent and less uncertain. Overall, data sharing results in a greater improvement in the lender's prior regarding borrower type and a further reduction in default prediction uncertainty for high-score borrowers, thus providing a rationale for the greater impact of data sharing on loan pricing.

5 Robustness checks

5.1 Fixed effects to eliminate unobserved characteristics

Throughout this study, the effects of data sharing are estimated using matched samples. That is, the applicants who share their data are matched to a group of individuals who do not share data but are otherwise similar in observable characteristics to minimize the omitted variable bias. However, the principal limitation of matching methods is that they do not account for potential selection on unobservables. In other words, the treatment and control groups may still differ in unobservable characteristics that may simultaneously determine selection into treatment and outcome variables. Such circumstances can lead to biased results. To tackle this potential issue, I exploit individual-day fixed effects to test the robustness of the main findings. On the platform, applicants often file multiple applications on the same day to compare different offers. During this process, a user may first apply without data sharing before changing her mind and deciding to share data. Since there is no change in borrower characteristics in the course of one day, the variation in the user's data sharing decisions within a day allows me to employ stringent individual-day fixed effects. By subsuming away unobserved individual characteristics that may jointly determine selection into treatment and outcome variables, I test the robustness of the effect of data sharing. The sample consists of 34,610 applications from 6,380 users.

[Table 8]

Table 8 shows results from the robustness tests on both the probability of loan approval and the interest rate. The results are both qualitatively and quantitatively similar to the main results from the probit analysis with matched samples. Compared to prime borrowers, low-credit score borrowers enjoy a higher increase in the probability of loan approval, with middle-tier borrowers benefiting the most. The effects are smaller for the highest- and lowest-score borrowers, which is in line with the hump-shaped relationship found in the main results. The magnitude is marginally higher for the highest two credit score groups (A–D and E–G) by approximately one to two percentage points compared to the main results and is slightly attenuated for the two lowest credit score groups (H–K and L–M) by about two percentage points. The effects on the interest rate are also robust quantitatively and qualitatively. Data sharing leads to a larger reduction in the interest rate for high-credit score borrowers, and the effect decreases for low-credit score borrowers. Compared to the main results, the magnitude of the reduction in the interest rate is slightly lower for the highest three groups (A–D, E–G, H–K) and higher for the lowest rating group (L–M). In Online Appendix C.2, I further implement a Rosenbaum sensitivity analysis to quantify how severe unmeasured confounding variables must be between the treated and control units to nullify the treatment effect; those results are robust.

6 Conclusion

This paper provides empirical evidence of open banking, a policy that empowers consumers by giving them greater control of their own data and discretion over sharing their financial data in the consumer credit market. Leveraging highly granular loan application-level data from the largest German online lender, I show that the rate of open banking participation (data sharing) is higher among riskier (lower-credit score) borrowers. I also provide evidence that customer-directed data sharing can benefit borrowers in the form of higher approval rates and lower interest rates. The effect, however, varies across different credit score groups. In terms of economic magnitude, lower-credit score applicants gain the most on the extensive margin (higher increase in the probability of loan approval), while high credit score applicants obtain a larger reduction in the interest rate. Importantly, data sharing is associated with lower ex post defaults. This is suggestive of latent good-type borrowers self-selecting into data sharing based on their self-assessment of creditworthiness. These findings suggest that optional data sharing can lead to more efficient allocation of capital and reduced adverse selection.

Notably, with detailed customer data, standard pricing variables such as credit score, age, and income explain loan application outcomes substantially less fully. Overall, this present study has shown that data sharing can bring substantial benefits to loan applicants and that there is thus far limited evidence of price discrimination from exploiting individuals' preferences and behaviors.

There are a few issues that remain open for future research. Open banking may generate unintended consequences because it limits banks' ability to extract rent from customer data. As open banking is still a relatively new phenomenon, future research may empirically test these predictions; that is, the second-order effects of open banking via its impact on incumbents' profitability and its interactions with borrowers over time. Additionally, this study is related to the effects of open banking and customer-driven data sharing in the lending market. The implications of open banking, however, may be markedly different across a wider range of financial services, which need to be taken into consideration to assess the aggregate impact.

References

Acquisti, Alessandro, Laura Brandimarte, and George Loewenstein (2015). "Privacy and human behavior in the age of information". *Science* 347.6221, pp. 509–514.

Ai, Chunrong and Edward C Norton (2003). "Interaction terms in logit and probit models". *Economics letters* 80.1, pp. 123–129.

Albanesi, Stefania and Domonkos F Vamossy (2019). "Predicting consumer default: A deep learning approach". National Bureau of Economic Research.

Avramidis, Panagiotis, Nikolaos Mylonopoulos, and George G Pennacchi (2022). "The role of marketplace lending in credit markets: Evidence from bank mergers". *Management Science* 68.4, pp. 3090–3111.

Babina, Tania, Greg Buchak, and Will Gornall (2022). "Customer Data Access and Fintech Entry: Early Evidence from Open Banking". *Available at SSRN*.

Balyuk, Tetyana and Sergei Davydenko (2019). "Reintermediation in FinTech: Evidence from online lending". *Journal of Financial and Quantitative Analysis*, pp. 1–75.

Beaumont, Paul, Huan Tang, and Eric Vansteenberghe (2022). "The role of fintech in small business lending". *Available at SSRN 4260842*.

Ben-David, Itzhak, Mark J Johnson, and René M Stulz (2021). "Why did small business Fintech lending dry up during March 2020?" National Bureau of Economic Research.

Benetton, Matteo et al. (2021). "Capital requirements and mortgage pricing: Evidence from Basel II". *Journal of Financial Intermediation* 48, p. 100883.

Berg, Tobias et al. (2020). "On the rise of fintechs: Credit scoring using digital footprints". *The Review of Financial Studies* 33.7, pp. 2845–2897.

Berger, Allen N and Gregory F Udell (1994). "Did risk-based capital allocate bank credit and cause a "credit crunch" in the United States?" *Journal of Money, credit and Banking* 26.3, pp. 585–628.

Braggion, Fabio et al. (2020). "The Value of "New" and "Old" Intermediation in Online Debt Crowdfunding". Centre for Economic Policy Research.

Brunnermeier, Markus and Jonathan Payne (2022). "Platforms, Tokens, and Interoperability". *Unpublished Working Paper*.

Buchak, Greg et al. (2018). "Fintech, regulatory arbitrage, and the rise of shadow banks". *Journal of Financial Economics* 130.3, pp. 453–483.

Chen, Yuxin and Ganesh Iyer (2002). "Research note consumer addressability and customized pricing". *Marketing Science* 21.2, pp. 197–208.

Cornelli, Giulio et al. (2022). "The impact of fintech lending on credit access for us small businesses". FRB of Philadelphia Working Paper.

de Ridder, Maarten (2019). "Market power and innovation in the intangible economy". Faculty of Economics, University of Cambridge.

De Roure, Calebe, Loriana Pelizzon, and Anjan Thakor (2022). "P2P lenders versus banks: Cream skimming or bottom fishing?" *The Review of Corporate Finance Studies* 11.2, pp. 213–262.

Di Maggio, Marco, Dimuthu Ratnadiwakara, and Don Carmichael (2022). "Invisible primes: Fintech lending with alternative data". National Bureau of Economic Research.

Djankov, Simeon, Caralee McLiesh, and Andrei Shleifer (2007). "Private credit in 129 countries". *Journal of Financial Economics* 84.2, pp. 299–329.

Eça, Afonso et al. (2022). "The real effects of fintech lending on SMEs: Evidence from loan

applications". ECB Working Paper.
Eeckhout, Jan and Laura Veldkamp (2022). "Data and market power". National Bureau of Economic Research.
Erel, Isil and Jack Liebersohn (2022). "Can FinTech reduce disparities in access to finance? Evidence from the Paycheck Protection Program". *Journal of Financial Economics* 146.1, pp. 90–118.
Farboodi, Maryam and Laura Veldkamp (2020). "Long-run growth of financial data technology". *American Economic Review* 110.8, pp. 2485–2523.
Fracassi, Cesare and William Magnuson (2021). "Data autonomy". *Vand. L. Rev.* 74, p. 327.
Fuster, Andreas et al. (2019). "The role of technology in mortgage lending". *The Review of Financial Studies* 32.5, pp. 1854–1899.
Gambacorta, Leonardo et al. (2019). "How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm". BIS Working Paper.
Gambacorta, Leonardo et al. (2020). "Data vs collateral". CEPR Discussion Paper No. DP15262.
Ghosh, Pulak, Boris Vallée, and Yao Zeng (2021). "FinTech lending and cashless payments". *Proceedings of Paris December 2021 Finance Meeting EUROFIDAI-ESSEC*.
Goldfarb, Avi and Catherine Tucker (2012). "Shifts in privacy concerns". *American Economic Review* 102.3, pp. 349–53.
Goldstein, Itay, Chong Huang, and Liyan Yang (2022). "Open Banking under Maturity Transformation". *Available at SSRN*.
Goldstein, Itay, Chester S Spatt, and Mao Ye (2021). "Big data in finance". *The Review of Financial Studies* 34.7, pp. 3213–3225.
Gopal, Manasa and Philipp Schnabl (2022). "The rise of finance companies and fintech lenders in small business lending". *The Review of Financial Studies* 35.11, pp. 4859–4901.
He, Zhiguo, Jing Huang, and Jidong Zhou (2020). "Open banking: credit market competition when borrowers own the data". National Bureau of Economic Research.
Heckman, James J (1976). "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models", pp. 475–492.
— (1979). "Sample selection bias as a specification error". *Econometrica: Journal of the Econometric Society*, pp. 153–161.
Hertzberg, Andrew, Jose Maria Liberti, and Daniel Paravisini (2011). "Public information and coordination: evidence from a credit registry expansion". *The Journal of Finance* 66.2, pp. 379–412.
Howell, Sabrina T et al. (2021). "Racial disparities in access to small business credit: Evidence from the paycheck protection program". National Bureau of Economic Research.
Jaffee, Dwight M and Thomas Russell (1976). "Imperfect information, uncertainty, and credit rationing". *The Quarterly Journal of Economics* 90.4, pp. 651–666.
Jagtiani, Julapa and Catharine Lemieux (2019). "The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform". *Financial Management* 48.4, pp. 1009–1029.
Jansen, Mark, Hieu Nguyen, and Amin Shams (2023). "Rise of the machines: The impact of automated underwriting". *Forthcoming at Management Science*.
Kashyap, Anil K, Jeremy C Stein, et al. (2004). "Cyclical implications of the Basel II capital

standards". *Economic Perspectives-Federal Reserve Bank Of Chicago* 28.1, pp. 18–33.
Kirpalani, Rishabh and Thomas Philippon (2020). "Data sharing and market power with two-sided platforms". National Bureau of Economic Research.
Lambrecht, Anja and Catherine E Tucker (2015). "Can big data protect a firm from competition?" *Available at SSRN 2705530*.
Lin, Tesary (2022). "Valuing intrinsic and instrumental preferences for privacy". *Marketing Science* 41.4, pp. 663–681.
Ouyang, Shumiao (2021). "Cashless Payment and Financial Inclusion". *Available at SSRN 3948925*.
Parlour, Christine A, Uday Rajan, and Haoxiang Zhu (2022). "When fintech competes for payment flows". *The Review of Financial Studies* 35.11, pp. 4985–5024.
Philippon, Thomas (2016). "The fintech opportunity". National Bureau of Economic Research.
— (2019). "On fintech and financial inclusion". National Bureau of Economic Research.
Popov, Alexander and Gregory F Udell (2012). "Cross-border banking, credit access, and the financial crisis". *Journal of International Economics* 87.1, pp. 147–161.
Rishabh, Kumar (2022). "Can open banking substitute credit bureaus". Working Paper.
Rosenbaum, Paul R (2002). "Overt bias in observational studies". Springer.
Roulet, Caroline (2018). "Basel III: Effects of capital and liquidity regulations on European bank lending". *Journal of Economics and Business* 95, pp. 26–46.
Sheng, Tianxiang (2021). "The effect of fintech on banks' credit provision to SMEs: Evidence from China". *Finance Research Letters* 39, p. 101558.
Stiglitz, Joseph E and Andrew Weiss (1981). "Credit rationing in markets with imperfect information". *The American economic review* 71.3, pp. 393–410.
Tang, Huan (2019a). "Peer-to-peer lenders versus banks: substitutes or complements?" *The Review of Financial Studies* 32.5, pp. 1900–1938.
— (2019b). "The value of privacy: Evidence from online borrowers". *HEC Paris*.
Taylor, Curtis R (2004). "Consumer privacy and the market for customer information". *RAND Journal of Economics*, pp. 631–650.
Thakor, Richard T and Robert C Merton (2018). "Trust in lending". National Bureau of Economic Research.
Vallee, Boris and Yao Zeng (2019). "Marketplace lending: A new banking paradigm?" *The Review of Financial Studies* 32.5, pp. 1939–1982.

Table 1: **Summary Statistics**

Variable	N	Mean	S.D.	Min	p25	p50	p75	Max
LOAN INFORMATION								
Credit requested	2,484,987	13,669.71	12,979.14	1,000.00	4,000.00	10,000.00	20,000.00	50,000.00
Interest rate*	1,630,862	0.12	0.04	0.00	0.08	0.13	0.15	0.20
Platform score (max 7, min 1)	2,484,987	2.81	1.81	1.00	1.00	2.00	4.00	7.00
Credit score group (max 4, min 1)	2,484,987	3.12	0.85	1.00	3.00	3.00	4.00	4.00
Loan duration	2,484,987	55.10	24.33	0.00	36.00	60.00	84.00	84.00
Application accepted (D)	2,484,987	0.66	0.47	0.00	0.00	1.00	1.00	1.00
Bank account detail shared (D)	2,484,987	0.08	0.26	0.00	0.00	0.00	0.00	1.00
BORROWER CHARACTERISTICS								
Age	2,484,981	37.74	12.62	18.00	27.00	36.00	47.00	69.00
Female (D)	2,484,987	0.35	0.48	0.00	0.00	0.00	1.00	1.00
Married (D)	2,484,987	0.38	0.49	0.00	0.00	1.00	1.00	1.00
Main earner (D)	2,484,987	0.62	0.49	0.00	0.00	1.00	1.00	1.00
No. current loan demand	2,308,526	1.35	1.47	0.00	0.00	1.00	2.00	68.00
No. past loan demand	2,308,526	1.04	1.78	0.00	0.00	0.00	1.00	76.00
INCOME AND EXPENSES								
Total income	2,484,987	3,053.48	2380.33	0.00	1,450.00	1,950.00	2,610.00	50,000.00
Monthly net salary	2,484,979	2,593.52	1874.71	0.00	1,300.00	1,800.00	2,370.00	36,000.00
Child support income	2,484,979	127.21	208.08	0.00	0.00	0.00	204.00	1,638.00
Other income	2,484,979	194.48	553.84	0.00	0.00	0.00	0.00	10,000.00
Total expenses	2,484,987	740.65	617.69	0.00	304.00	590.00	933.00	5980.00
Housing-related expenses	2,484,069	481.37	386.32	0.00	180.00	415.00	645.00	3500.00
Credit installments expenses	2,484,069	166.37	331.21	0.00	0.00	0.00	216.00	3685.00
Other expenses	2,484,069	23.88	123.50	0.00	0.00	0.00	0.00	2000.00
Insurance expenses	2,484,069	49.77	153.69	0.00	0.00	0.00	0.00	1707.00
Child support expenses	2,484,069	19.18	103.03	0.00	0.00	0.00	0.00	1500.00
ASSETS								
Credit card holder (D)	2,484,987	0.63	0.48	0.00	0.00	1.00	1.00	1.00
Checking account owner (D)	2,484,987	0.93	0.25	0.00	1.00	1.00	1.00	1.00
Home owner (D)	2,484,987	0.24	0.43	0.00	0.00	0.00	0.00	1.00
Car owner (D)	2,484,987	0.55	0.50	0.00	0.00	1.00	1.00	1.00

Notes: This table presents summary statistics for the sample. The sample period runs from January 13, 2018, to May 22, 2022. (D) = dummy variable. The monetary unit is EUR. The final sample includes only one application per borrower. In the case of multiple applications, the initial application from each applicant is included. *conditional on loan approval.

Table 2: **What characterizes borrowers who share data?**

	Probit (marginal effects)				Linear Probability Model			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age (10 years)		−0.020***	−0.018***	−0.019***		−0.019***	−0.017***	−0.018***
		(0.0001)	(0.0002)	(0.0002)		(0.0002)	(0.0002)	(0.0002)
Income decile			0.001***	0.000			0.001***	−0.000
			(0.0001)	(0.0001)			(0.0001)	(0.0001)
Credit score (A–D) (base)								
Credit score (E–G)	0.026***		0.013***	0.009***	0.027***		0.014***	0.009***
	(0.0004)		(0.0004)	(0.0004)	(0.0004)		(0.0004)	(0.0004)
Credit score (H–K)	0.039***		0.019***	0.018***	0.038***		0.019***	0.016***
	(0.0006)		(0.0006)	(0.0006)	(0.0006)		(0.0006)	(0.0006)
Credit score (L–M)	0.039***		0.015***	0.021***	0.034***		0.010***	0.013***
	(0.0011)		(0.0010)	(0.0010)	(0.0008)		(0.0009)	(0.0009)
Loan amount requested (ln)				−0.011***				−0.011***
				(0.0002)				(0.0002)
Loan duration (ln)				−0.003***				−0.006***
				(0.0004)				(0.0005)
Female				−0.004***				−0.005***
				(0.0004)				(0.0004)
Married				0.000				−0.000
				(0.0004)				(0.0003)
Main earner				0.009***				0.010***
				(0.0005)				(0.0005)
No. current loan demand				0.006***				0.008***
				(0.0001)				(0.0002)
No. past loan demand				0.005***				0.006***
				(0.0001)				(0.0001)
Homeowner				−0.008***				−0.008***
				(0.0005)				(0.0004)
Car owner				0.012***				0.009***
				(0.0004)				(0.0004)
Access channel = Homepage (base)								
Access channel = Repeat				0.115***				0.070***
				(0.0031)				(0.0025)
Access channel = Price comparison website				−0.077***				−0.067***
				(0.0012)				(0.0011)
Access channel = Broker				−0.108***				−0.098***
				(0.0014)				(0.0013)
Access channel = Bank				−0.127***				−0.131***
				(0.0015)				(0.0017)
Constant					0.020***	0.114***	0.094***	0.261***
					(0.0004)	(0.0007)	(0.0009)	(0.0023)
Dummy	Year	Year	Year	Year	Year	Year	Year	Year
Cluster (Zipcode-Year)	X	X	X	X	X	X	X	X
N	2,237,677	2,237,677	2,237,677	2,237,677	2,237,677	2,237,677	2,237,677	2,237,677
R2	0.0640	0.0724	0.0738	0.1055	0.036	0.040	0.041	0.058

Notes: This table reports the results from equation (1), which estimates the probability that a borrower shares bank data using the full sample. The coefficients (1–3) are marginal effects at means. Clustered standard errors are in parentheses. Columns (1)–(3) reports pseudo R2 and (4)-(6) adjusted R2. The dependent variable is *Signup*, a dummy variable that takes a value of one if the applicant shared data and zero otherwise. In the case of multiple applications, the initial application from each applicant is included.

Table 3: **Credit score predictive accuracy**

(1)	(2)	(3)	(4)	(5)	(6)
Credit score group	N	MSE	S.D.	Min	Max
A–D	26,871	0.0376	0.1705	5.7e-07	1
E–G	22,496	0.0608	0.2029	1.4e-06	0.99
H–K	5,562	0.0818	0.2226	8.9e-06	0.97
L–M	422	0.1065	0.2173	2.4e-07	0.98

Notes: This table demonstrates the imprecision of credit scores in predicting defaults. The imprecision of inference is measured using the mean squared error, denoted as $MSE = E[(Z - E(Z|X))^2]$, where Z represents a binary variable that assumes a value of one if the loan is in default status (i.e., delinquency extending beyond 90 days). A probit model has been used to estimate default probability using credit scores. A lower MSE value suggests a smaller error in the predictive model. Results are presented by credit score groups, with groups A–D representing the highest and groups L–M denoting the lowest credit score categories. The sample comprises loans securitized through the Fortuna Consumer Loan ABS in 2021, 2022, and 2023, totaling EUR 850 million. Loan-level data from European asset-backed securities, sourced from the European DataWarehouse (EDW), are linked with Auxmoney loans using key variables, including income, location, loan amount, loan duration, interest rate, loan disbursement date, occupation type, and loan purpose type. This matching process serves to augment the Auxmoney data with loan payment information from EDW. The final sample includes 55,351 loans. To mitigate the possibility that the credit score simply captures a mix of other factors that affect defaults, I control for the following borrower and loan characteristics: age, income decile, loan amount, loan duration, female (dummy), married (dummy), main earner (dummy), number of current and past loans, homeowner (dummy), car owner (dummy), and access channel (categorical variable).

Table 4: **The effect of the data sharing signup decision on loan approval**

	Probit (marginal effects)			Linear Probability Model		
	(1)	(2)	(3)	(4)	(5)	(6)
Signup	0.063***	0.064***	0.015***	0.033***	0.033***	0.033***
	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
Signup × Credit score (A–D)* (Base)						
Signup × Credit score (E–G)	0.066***	0.078***	0.070***	0.101***	0.102***	0.068***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)
Signup × Credit score (H–K)	0.101***	0.119***	0.102***	0.132***	0.130***	0.106***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Signup × Credit score (L–M)	0.062***	0.077***	0.023***	0.055***	0.053***	0.022***
	(0.009)	(0.010)	(0.009)	(0.006)	(0.006)	(0.006)
Credit score (A–D) (Base)						
Credit score (E–G)	−0.289***	−0.242***	−0.157***	−0.328***	−0.256***	−0.214***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Credit score (H–K)	−0.058***	−0.550***	−0.460***	−0.621***	−0.523***	−0.472***
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)
Credit score (L–M)	−0.073***	−0.760***	−0.734***	−0.811***	−0.707***	−0.617***
	(0.005)	(0.004)	(0.007)	(0.004)	(0.004)	(0.004)
Age		0.009***	0.006***		0.007***	0.005***
		(0.000)	(0.000)		(0.000)	(0.000)
Income decile		0.028***	0.015***		0.022***	0.013***
		(0.000)	(0.000)		(0.000)	(0.000)
Homeowner			0.082***			0.081***
			(0.002)			(0.002)
Signup × Homeowner			−0.082***			−0.074***
			(0.002)			(0.003)
Loan amount requested (ln)			0.013***			0.014***
			(0.001)			(0.001)
Loan duration (ln)			−0.112***			−0.100***
			(0.002)			(0.002)
Female			0.033***			0.033***
			(0.002)			(0.001)
Married			0.040***			0.036***
			(0.002)			(0.001)
Main earner			0.030***			0.020***
			(0.002)			(0.001)
Carowner			0.067***			0.070***
			(0.002)			(0.001)
No. current loan demand			0.019***			0.020***
			(0.001)			(0.001)
No. past loan demand			0.006***			0.007***
			(0.000)			(0.000)
Access channel = Homepage (Base)						
Access channel = Repeat			0.000			−0.073***
			(0.000)			(0.003)
Access channel = Price comp. website			−0.271***			−0.294***
			(0.001)			(0.002)
Access channel = Broker			−0.552***			−0.508***
			(0.005)			(0.004)
Access channel = Bank			−0.465***			−0.450***
			(0.012)			(0.007)
Constant				0.656***	0.642***	1.141***
				(0.002)	(0.004)	(0.006)
Dummy	Year	Year	Year	Year	Year	Year
Cluster (Zipcode-Year)	X	X	X	X	X	X
N	376,852	376,852	376,852	376,852	376,852	376,852
R2	0.1721	0.2545	0.3497	0.2461	0.295	0.352

Notes: This table reports the results from equation (2) which estimates the effect of a prospective borrower's decision to share bank account data (*Signup*, a dummy variable that takes a value of one if the applicant shared data and zero otherwise) on the probability of loan approval using the matched sample. In the case of multiple applications, the initial application from each applicant is included. Each of the 188,453 applicants who shared data is matched one-to-one to create a control group of those who did not share data using hybrid matching. Exact matching is used for the access channel and loan application year, and propensity score matching is used for age, credit score, and income decile. The final sample includes 376,852 loan applications from 375,852 unique applicants.
*The coefficients in columns (1)–(3) show marginal effects at means. Clustered standard errors are in parentheses. Columns (1)–(3) report pseudo R2 and columns (4)–(6) report adjusted R2. It is important to note that estimating the marginal effect of interaction terms in a non-linear model is not straightforward (see Ai and Norton 2003). To compute the effect of data sharing by credit score group, I obtain margins with nested designs. In this case, credit score group is treated as a nesting variable over which margins of data sharing are estimated.

Table 5: **The effect of the data sharing signup decision on interest rates**

	Matched sample	
	(1)	(2)
Signup	−0.0213***	−0.0226***
	(0.0003)	(0.0003)
Signup × Credit score (A-D) (Base)		
Signup × Credit score (E-G)	0.0025***	0.0016***
	(0.0003)	(0.0003)
Signup × Credit score (H-K)	0.0106***	0.0089***
	(0.0004)	(0.0005)
Signup × Credit score (L-M)	0.0187***	0.0165***
	(0.0010)	(0.0011)
Credit score (A-D) (Base)		
Credit score (E-G)	0.0211***	0.0243***
	(0.0002)	(0.0004)
Credit score (H-K)	0.0311***	0.0387***
	(0.0003)	(0.0007)
Credit score (L-M)	0.0388***	0.0517***
	(0.0007)	(0.0014)
Signup × Home owner	0.0041***	0.0048***
	(0.0003)	(0.0003)
Homeowner	−0.0205***	−0.0227***
	(0.0002)	(0.0003)
Age (10 years)	−0.0162***	−0.0181***
	(0.0004)	(0.0005)
Age (10 years) × Age (10 years)	0.0009***	0.0011***
	(0.0001)	(0.0001)
Income decile	0.0019***	−0.0019***
	(0.0000)	(0.0000)
Loan amount requested (ln)	0.0099***	0.0091***
	(0.0001)	(0.0001)
Loan duration (ln)	0.0045***	0.0073***
	(0.0002)	(0.0003)
Married	−0.0054***	−0.0067***
	(0.0001)	(0.0001)
Female	−0.0023***	−0.0024***
	(0.0002)	(0.0002)
Main earner	−0.0032***	−0.0027***
	(0.0002)	(0.0002)
Car owner	−0.0045***	−0.0045***
	(0.0002)	(0.0002)
Credit card holder	−0.0065***	−0.0064***
	(0.0002)	(0.0002)
Checking account owner	−0.0017***	−0.0017***
	(0.0004)	(0.0004)
No. current loan demand	−0.0009***	−0.0009***
	(0.0001)	(0.0001)
No. past loan demand	−0.0001**	−0.0001**
	(0.0000)	(0.0000)
Access channel=Homepage		
Access channelr=Repeat	−0.0252***	−0.0252***
	(0.0005)	(0.0005)
Access channel=Price comp. website	0.0064***	0.0064***
	(0.0003)	(0.0003)
Access channel=Broker	0.0193***	0.0191***
	(0.0005)	(0.0005)
Access channel=Bank	0.0214***	0.0212***
	(0.0012)	(0.0012)
Inverse Mills ratio		−0.0107***
		(0.0009)
Constant	0.0761***	0.0793***
	(0.0011)	(0.0012)
Dummy	Year	Year
Cluster (Zipcode-Year)	X	X
N	249,240	249,240
Adjusted R2	0.4473	0.4476

Notes: This table reports the results of equation (3), which estimates the effect of a prospective borrower's decision to share bank account data (*Signup*, a dummy variable that takes a value of one if the applicant shared data and zero otherwise) on the interest rate conditional on loan approval, using the matched sample. In the case of multiple applications, the initial application from each applicant is included. Column (2) shows the results after correcting for selection bias using the Heckman selection model discussed in Section 3.4 (see Online Appendix B for more detail). Each of the 125,889 approved loan applicants that shared data is matched one-to-one to create a control group of those who did not share data using hybrid matching. Exact matching is used for the access channel and loan application year, and propensity score matching is used for age, credit score, and income decile. The final sample includes 249,240 loan applications from 249,240 unique applicants.

Table 6: **Data sharing decisions and borrower type using ex post defaults**

A.

	Default = 1 if payment is more than 90 days late			
Credit score group	(A–D)	(E–G)	(H–K)	(L–M)
Signup	−0.011***	−0.015***	−0.020**	−0.037
	(0.003)	(0.004)	(0.010)	(0.037)
Controls	Y	Y	Y	Y
Cluster (Zipcode-Year)	Y	Y	Y	Y
N	15,466	15,126	3,190	58
Pseudo R2	0.0541	0.0564	0.0494	0.0501

B. Controlling for interest rate

	Default = 1 if payment is more than 90 days late			
Credit score group	(A–D)	(E–G)	(H–K)	(L–M)
Signup	−0.005*	−0.002	−0.006	0.000
	(0.003)	(0.004)	(0.010)	(–)
Interest rate (%)	0.004***	0.007***	0.010***	0.000
	(0.000)	(0.001)	(0.002)	(–)
Controls	Y	Y	Y	Y
Cluster (Zipcode-Year)	Y	Y	Y	Y
N	15,466	15,126	3,190	58
Pseudo R2	0.0817	0.0801	0.0658	–

Notes: Panel A shows the association between data sharing decisions (*Signup*, a dummy variable that takes a value of one if the applicant shared data and zero otherwise) and *Default* (a dummy variable that takes a value of one if the loan has been late over 90 days), and panel B adds interest rate as a control to account for the causal impact of interest rate on defaults via moral hazard or inability to pay. A probit model is used for the analysis. Each column represents a credit score group, with (A–D) being the highest and (L–M) being the lowest. The sample comprises loans securitized through the Fortuna Consumer Loan ABS in 2021, 2022, and 2023, totaling EUR 850 million. Loan-level data from European asset-backed securities, sourced from the European DataWarehouse (EDW), are linked with Auxmoney loans using key variables, including income, location, loan amount, loan duration, interest rate, loan disbursement date, occupation type, and loan purpose type. This matching process serves to augment the Auxmoney data with loan payment information from EDW. The final sample includes 33,840 loans, of which 16,920 borrowers who shared data are matched one-to-one to create a control group of those who did not share data using hybrid matching. Exact matching is used for the access channel and loan application year, and propensity score matching is used for age, credit score, and income decile. Control variables include age, income decile, loan amount, loan duration, female (dummy), married (dummy), main earner (dummy), number of current and past loans, homeowner (dummy), car owner (dummy), and access channel (categorical variable). Results for the group L–M for Panel B are indeterminate due to the constraints of a limited sample size.

Table 7: **Channels through which data sharing affects loan application outcomes**

A. Data reveals type (change in platform scores)

	Dependent variable = *Platform score*			
	(A–D)	(E–G)	(H–K)	(L–M)
Signup	0.7732***	0.6619***	0.3798***	0.0999***
	(0.0088)	(0.0072)	(0.0077)	(0.0094)
Dummy	Year	Year	Year	Year
Controls	X	X	X	X
Cluster (Zipcode-Year)	X	X	X	X
N	122,906	155,868	64,180	14,362
Adjusted R2	0.3449	0.3223	0.3333	0.4658

Notes: This table reports the results of data sharing (*Signup*, a dummy variable that takes a value of one if the applicant shared data and zero otherwise) on the change in the platform score by credit score group, using the matched sample. Control variables include age, income decile, loan amount, loan duration, female (dummy), married (dummy), main earner (dummy), number of current and past loans, homeowner (dummy), car owner (dummy), and access channel (categorical variable). The dependent variable, *Platform score*, ranges from 7 (highest) to 1 (lowest and rejected). Each of the 178,658 loan applicants who shared data is matched one-to-one to create a control group of those who did not share data using hybrid matching. Exact matching is used for the access channel and loan application year, and propensity score matching is used for age, credit score, and income decile. The final sample includes 357,316 loan applications (357,316 unique applicants). Each column represents a credit score group with (A–D) being the highest and (L–M) being the lowest credit score group.

B. Data mitigates uncertainty

	Shared	Not shared	Shared	Not shared	Shared	Not shared
Credit score group	A–D		E–G		H–M	
N	7,733	7,733	7,563	7,563	1,628	1,628
MSE	0.033	0.0413	0.0529	0.066	0.0757	0.094
Std.Dev	0.1553	0.1761	0.1914	0.2087	0.2131	0.2257
Reduction in Std.Dev	13.39%		9.03%		5.91%	

Notes: This table presents the imprecision of platform scores in predicting defaults and the reduction in variance by credit score group. The imprecision of inference is measured using the mean squared error, denoted as $MSE = E[(Z - E(Z|X))^2]$, where Z represents a dummy variable that takes a value of one if the loan is in default status (delinquency extending beyond 90 days) and zero otherwise. A probit model has been used to estimate the default probability using platform scores. A lower MSE value suggests a smaller error in the predictive model. Results are presented by credit score groups, with groups A–D representing the highest and group H–M denoting the lowest group (due to insufficient observations of the previous denoted L-M group, H–K and L–M are combined for this analysis). The sample comprises loans securitized through the Fortuna Consumer Loan ABS in 2021, 2022, and 2023, totaling EUR 850 million. Loan-level data from European asset-backed securities, sourced from the European DataWarehouse (EDW), are linked with Auxmoney loans using key variables, including income, location, loan amount, loan duration, interest rate, loan disbursement date, occupation type, and loan purpose type. This matching process serves to augment the Auxmoney data with loan payment information from EDW. A matched sample is used for the analysis to ensure that the two groups are observably similar and comparable but differ by data sharing choices. The final sample includes 33,838 loans. To mitigate the possibility that the credit score simply captures a mix of other factors that affect defaults, I control for the following borrower and loan characteristics: age, income decile, loan amount, loan duration, female (dummy), married (dummy), main earner (dummy), number of current and past loans, homeowner (dummy), car owner (dummy), and access channel (categorical variable).

Table 8: **Robustness checks using fixed effects**

A. The effect of data sharing decision on loan approval

	Credit score			
	(A–D)	(E–G)	(H–K)	(L–M)
Signup	0.035***	0.094***	0.092***	0.042***
	(0.012)	(0.008)	(0.008)	(0.011)
Controls	Y	Y	Y	Y
Individual-day FE	Y	Y	Y	Y
N	4766	15922	11313	2609
Adjusted R2	0.068	0.077	0.089	0.080

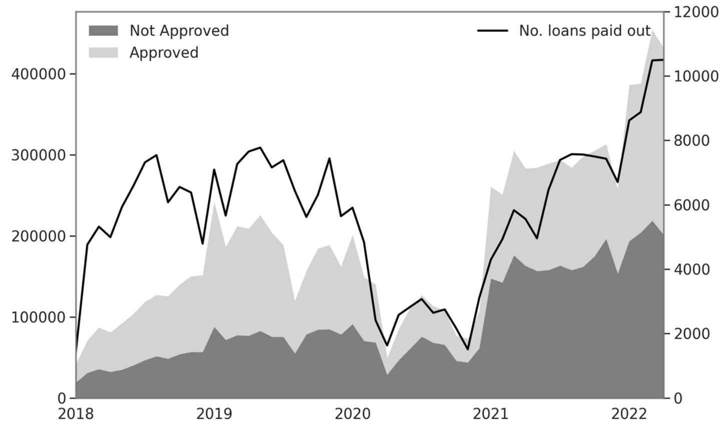
Notes: This table shows the effect of data sharing on the probability of loan approval. The sample includes multiple applications filed by the same individuals on the same day. These individuals do not share their data in initial applications but do share it in others. Given that there is within-individual variation in data sharing decisions while borrower characteristics do not change in the course of one day, I employ individual-day fixed effects to test the effect of data sharing on loan approval as a robustness check. Control variables include loan amount and loan duration. The dependent variable is a dummy variable that takes a value of one if the loan application is approved and 0 otherwise. *Signup* is a dummy variable that takes a value of one if the applicant shared data and zero otherwise. Each column represents a credit score group, with (A–D) the highest and (L–M) the lowest.

B. The effect of data sharing decision on interest rate

	Credit score			
	(A–D)	(E–G)	(H–K)	(L–M)
Signup	−0.017***	−0.014***	−0.007***	−0.007**
	(0.001)	(0.001)	(0.001)	(0.003)
Controls	Y	Y	Y	Y
Individual-day FE	Y	Y	Y	Y
N	3523	5625	1580	135
Adjusted R2	0.217	0.181	0.098	0.068

Notes: This table shows the effect of data sharing on the interest rate. The sample includes multiple applications filed by the same individuals on the same day. These individuals do not share their data in initial applications but do share it in others. Given that there is within-individual variation in data sharing decisions while borrower characteristics do not change in the course of one day, I employ individual-day fixed effects to test the effect of data sharing on loan approval as a robustness check. Control variables include loan amount and loan duration. The dependent variable is the loan interest rate conditional on the application being approved. *Signup* is a dummy variable that takes a value of one if the applicant shared data and zero otherwise. Each column represents a credit score group, with (A–D) the highest and (L–M) the lowest.

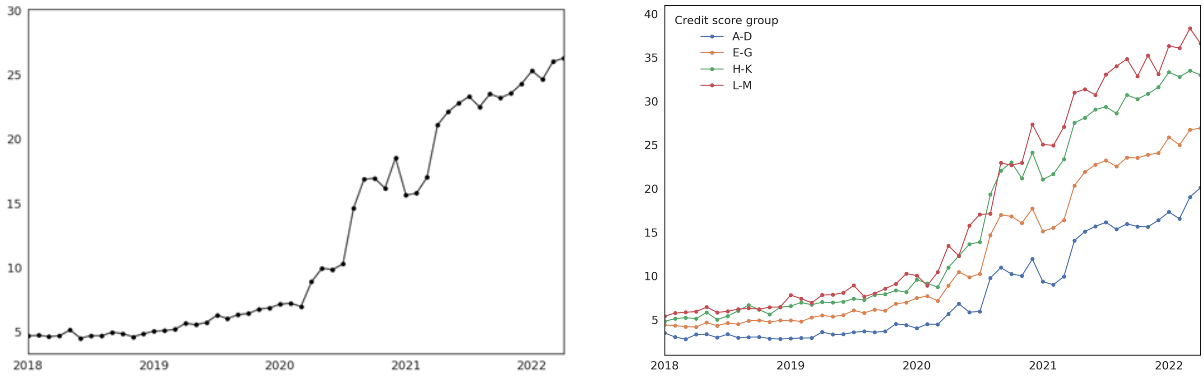
Figure 1: Number of applications and disbursed loans, measured monthly



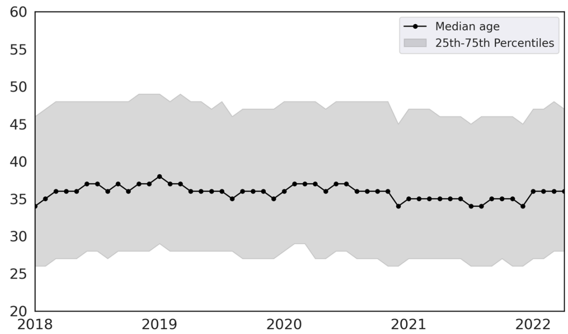
Notes: The figure depicts the monthly count of loan applications, differentiated by approval status. The dark gray bars represent the number of non-approved applications, while the light gray bars indicate approved applications (both plotted on the first *y*-axis). The second *y*-axis displays the count of disbursed loans among the approved applications. The sample period is from January 13, 2018 to May 22, 2022.

Figure 2

A. Data sharing over time (overall vs. by credit score, measured monthly)



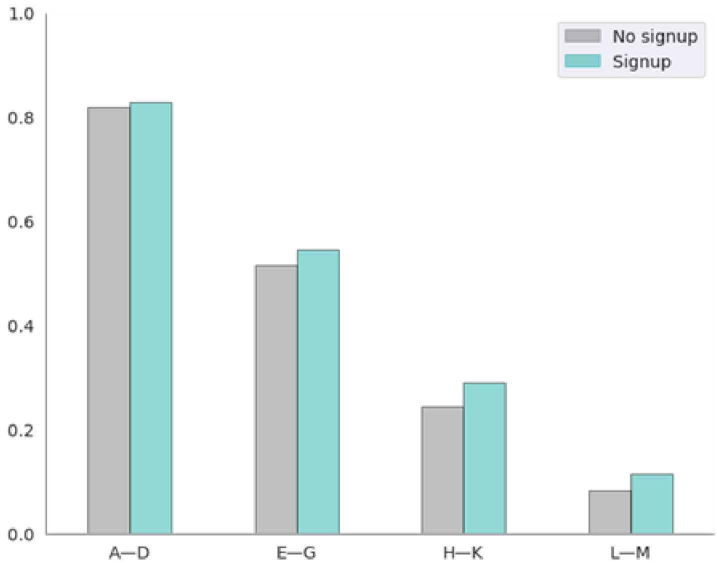
B. Age of applicants, measured monthly



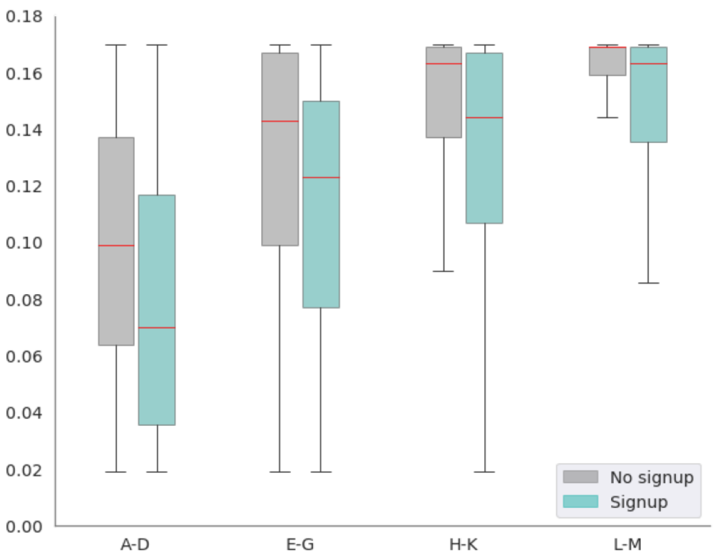
Note: The left side of panel A shows the percentage of loan applications in which applicants shared their data, calculated as a fraction of the total number of loan applications (including multiple applications per borrower). The right side of panel A illustrates these percentages by applicants' credit score group (A–D: highest, L–M: lowest). Panel B shows the inter-quartile range of applicant age. The sample period is from January 13, 2018, to May 22, 2022.

Figure 3

A. Loan approval rate by data sharing decision



B. Loan interest rate by data sharing decision



Note: Panel A displays the average loan approval rate, and panel B shows the inter-quartile range of interest rates. Green bars represent data sharing applicants, and gray bars represent non-sharing applicants across credit score groups from A–D (highest) to L–M (lowest).

Open Banking and Customer Data Sharing: Implications for FinTech Borrowers

Online Appendix

Rachel J. Nam

A Additional Tables and Figures

Table A.1: **Descriptive statistics by data sharing decision**

Variable	Signup	No signup	*p*-value
Credit requested	12,608.19	13,756.30	0.00
Interest rate*	0.10	0.12	0.00
Platform score (max 7, min 1)	2.84	2.44	0.00
Credit score group (max 4, min 1)	3.05	3.12	0.00
Loan duration	52.47	55.32	0.00
Application accepted (D)	0.69	0.65	0.00
Flagged for quality check (D)	0.25	0.28	0.00
Bank account detail shared (D)	1.00	0.00	0.00
Age	33.78	38.07	0.00
Female (D)	0.34	0.35	0.00
Married (D)	0.33	0.39	0.00
Main earner (D)	0.62	0.62	0.00
No. current loan demand	1.56	1.33	0.00
No. past loan demand	1.28	1.02	0.00
Total income	2,658.31	3,086.17	0.17
Total expenses	728.44	741.79	0.86
Credit card holder (D)	0.78	0.62	0.00
Checking account owner (D)	0.96	0.93	0.00
Homeowner (D)	0.19	0.25	0.00
Car owner (D)	0.60	0.55	0.00

Notes: This table presents summary statistics separately by data sharing choice, *Signup*, and for those who opt out, *No signup*. (D) = dummy variable. The monetary unit in EUR. The sample period runs from January 13, 2018, to May 22, 2022. The final sample includes only one application per applicant. In the case of multiple applications, the initial application from each applicant is included. *conditional on loan approval.

Table A.2: **Descriptive statistics by access channels**

	Access channel				
Variable	Directly via homepage	Repeat Borrower	Price comp. website	Broker	Bank
Credit requested	8,280.71	11,331.79	14,887.52	11,437.16	4,772.35
Interest rate*	0.12	0.09	0.11	0.14	0.13
Platform score (max 7, min 1)	2.39	4.62	3.04	1.79	1.55
Credit score group (max 4, min 1)	2.87	3.15	3.21	2.85	2.49
Loan duration	26.14	53.61	56.68	62.89	68.68
Application accepted (D)	0.57	0.97	0.72	0.37	0.31
Bank account detail shared (D)	0.11	0.17	0.08	0.03	0.03
Age	34.23	43.17	38.35	37.06	29.54
Female (D)	0.38	0.40	0.34	0.38	0.22
Married (D)	0.28	0.41	0.41	0.35	0.19
Main earner (D)	0.12	0.32	0.66	0.69	0.87
No. current loan demand	1.20	1.84	1.39	1.23	0.73
No. past loan demand	1.02	1.81	1.03	1.12	0.53
Total income	1,700.00	2,001.50	2,000.00	1,750.00	1,832.00
Total expenses	660.00	903.50	600.00	450.00	786.04
Credit card holder (D)	0.40	0.65	0.69	0.38	0.93
Checking account owner (D)	0.83	0.97	0.96	0.82	0.99
Homeowner (D)	0.16	0.30	0.27	0.15	0.16
Car owner (D)	0.49	0.67	0.61	0.26	0.37

Notes: This table presents descriptive statistics on borrower and loan characteristics by access channel. (D) = Dummy variable. The monetary unit is EUR. The sample period runs from January 13, 2018, to May 22, 2022. The final sample includes only one application per borrower. In the case of multiple applications, the initial application from each borrower is included. *conditional on loan approval.

Table A.3: **Matched variables and matching results**

A. Sample to estimate the effect of data sharing on loan approval rates (equation 2)

	Mean treated	Mean control	Mean p-value difference
Age	33.778	33.762	0.654
Credit score	3.0552	3.0579	0.306
Income decile	5.3476	5.3476	0.997
Access channel		——— exact matching ———	
Application year		——— exact matching ———	

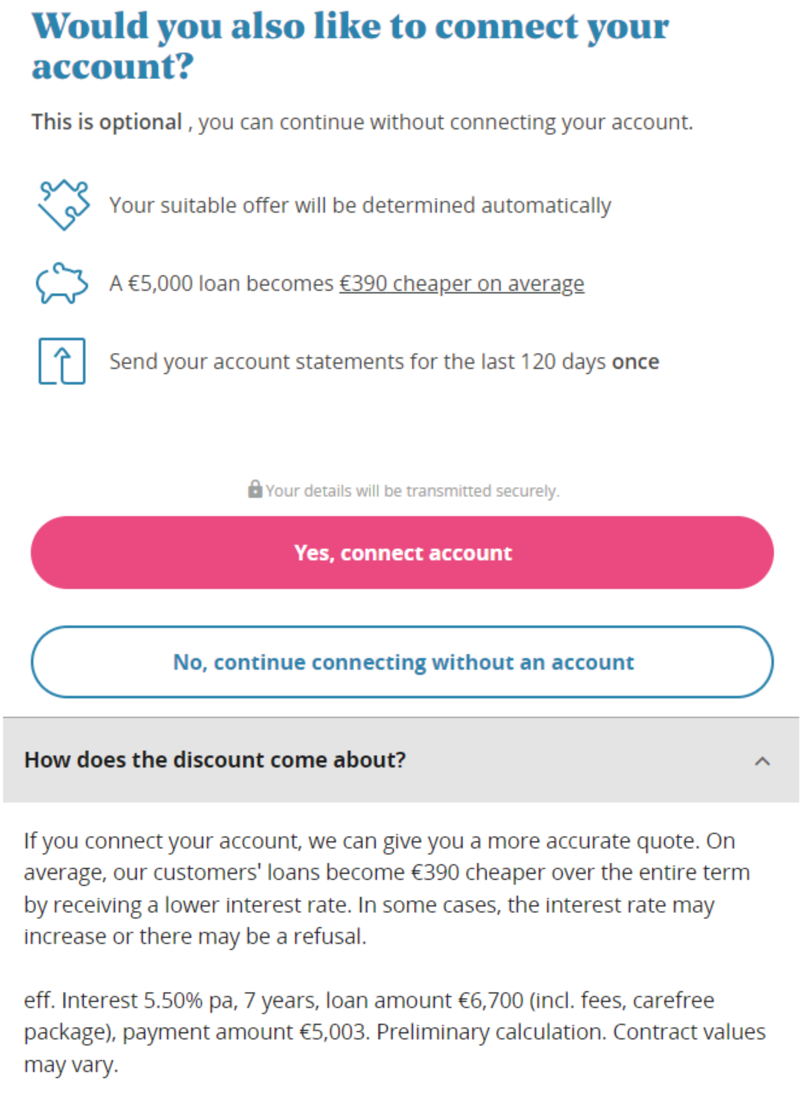
Notes: This table shows t-tests for the null hypothesis of equal means for both treatment and control groups. This sample is used to compute the effect of data sharing on the probability of loan approval (equation 2). Each of the 188,453 applicants who shared data is matched one-to-one using age, credit score, income decile, access channel, and loan application year to create a control group of those who did not share data but is observably similar. Propensity score matching is used for age, credit score, and income decile, and exact matching is used for access channel and loan application year. The final sample includes 376,852 loan applications from 375,852 unique applicants.

B. Sample to estimate the effect of data sharing on interest rates (equation 3)

	Mean treated	Mean control	Mean p-value difference
Age	36.496	36.477	0.662
Credit score	3.3081	3.3081	1.000
Income decile	5.8796	5.8908	0.331
Access channel		——— exact matching ———	
Application year		——— exact matching ———	

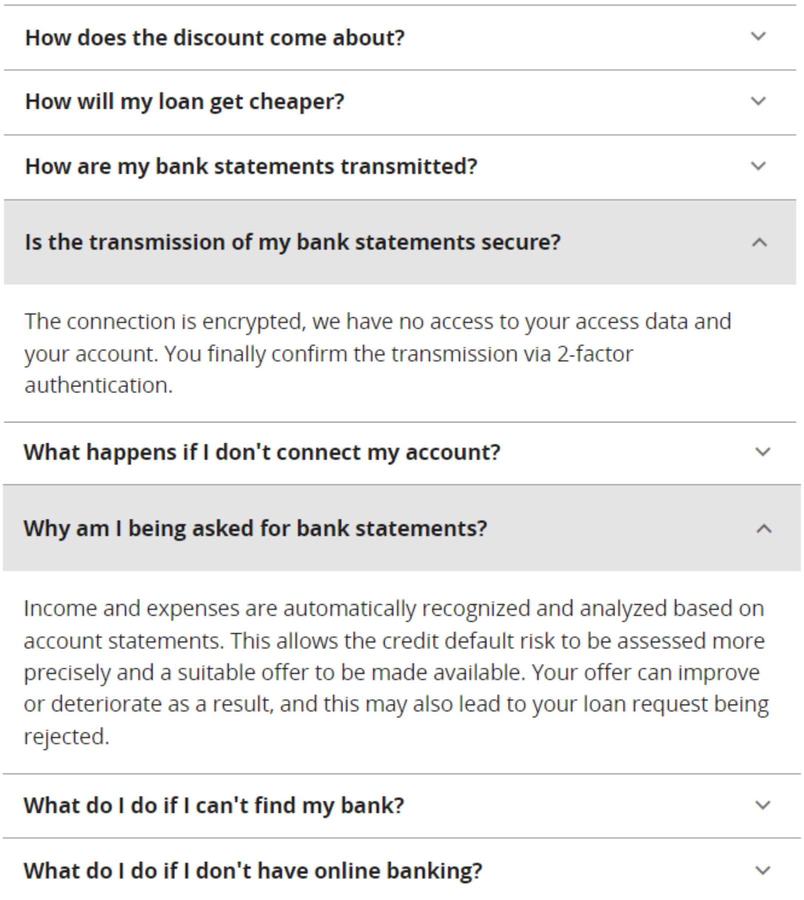
Notes: This table shows t-tests for the null hypothesis of equal means for both treatment and control groups. This sample is used to compute the effect of data sharing on the interest rate (equation 3). Interest rates are revealed only for successful loan applications. Each of the 125,889 *approved* applicants is matched one-to-one with *approved* applicants (to ensure interest rate information is available for all units), using age, credit score, income decile, access channel, and loan application year to create a control group of those who did not share data but is observably similar. Propensity score matching is for age, credit score, and income decile and exact matching is used for access channel and loan application year. The final sample includes 249,240 loan applications from 249,240 unique loan applicants.

Figure A.1: Data sharing during the application



Notes: This figure shows the exact manner in which data is shared during loan applications. Loan applicants are also supplied with information regarding data usage, and data security.

Figure A.1: Data sharing during the application (continued)



How does the discount come about?

How will my loan get cheaper?

How are my bank statements transmitted?

Is the transmission of my bank statements secure?

The connection is encrypted, we have no access to your access data and your account. You finally confirm the transmission via 2-factor authentication.

What happens if I don't connect my account?

Why am I being asked for bank statements?

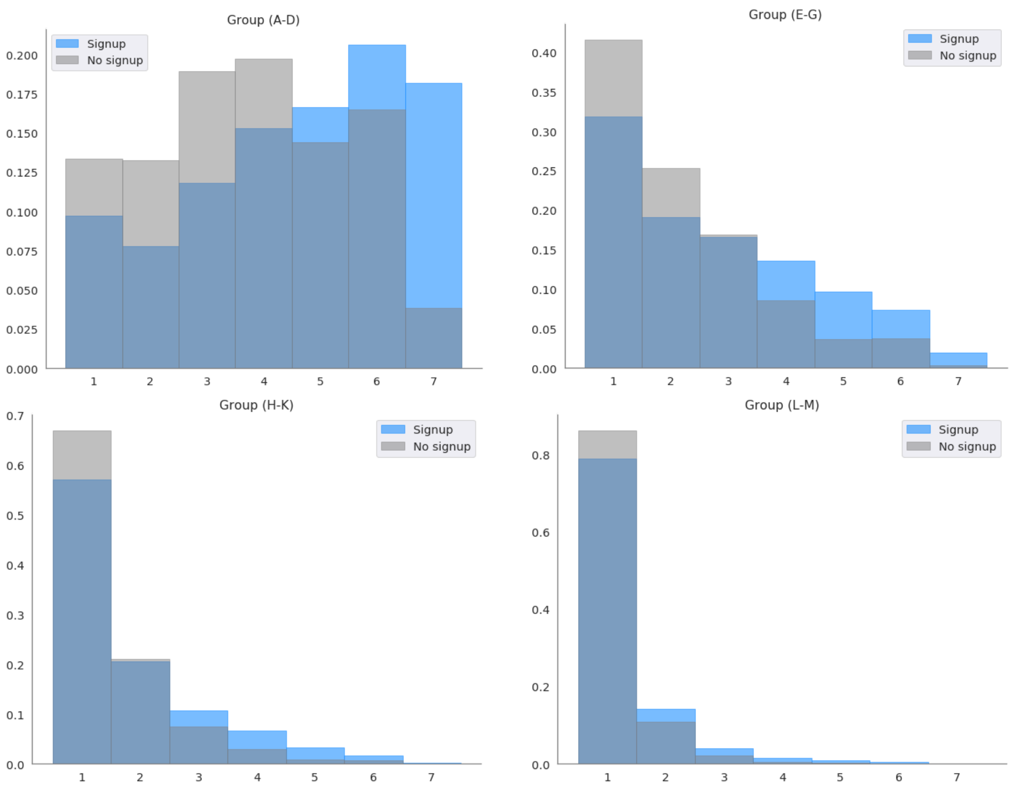
Income and expenses are automatically recognized and analyzed based on account statements. This allows the credit default risk to be assessed more precisely and a suitable offer to be made available. Your offer can improve or deteriorate as a result, and this may also lead to your loan request being rejected.

What do I do if I can't find my bank?

What do I do if I don't have online banking?

Notes: This figure shows the exact manner in which data is shared during loan applications. Loan applicants are also supplied with information regarding data usage, and data security.

Figure A.2: Distribution of platform-provided credit score by signup decision (using the matched sample)



Note: These figures depict the range of scores assigned by the platform (x-axis) after the application is completed, with 7 the highest and 1 the lowest and indicating rejection. Applicants choose to share their data prior to obtaining the loan approval decision, platform score, and interest rate. The y-axis measures the share of applicants.

B Heckman's two-stage correction to address selection bias

Let the loan approval and interest rate functions be given by

$$L_i^* = Z_i'\gamma + \epsilon_i,$$

$$r_i = X_i'\beta + u_i.$$

First, I begin by introducing the basic Heckman model in a first stage and estimate the probability of being accepted for all applicants,

$$\begin{aligned} Prob(L_i^* > 0|Z) &= Prob(\epsilon_i > -Z_i'\gamma) \\ &= \Phi(Z_i\gamma), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function with the variable of ϵ normalized to one. Interest rates are observed for those whose $L_i^* > 0$ so that the expected interest rate of a borrower is given by

$$\begin{aligned} E(r_i|L_i^* > 0, Z) &= X_i'\beta + E(u_i|\epsilon_i > -Z_i'\gamma) \\ &= X_i'\beta + \theta\lambda_i, \end{aligned}$$

where $\theta = \rho\sigma_u, \lambda_i = \frac{\phi(Z_i'\gamma)}{\Phi(Z_i'\gamma)}$, and $\phi(\cdot)$ is the standard normal density function. In the second stage, the interest rate equation for those who are accepted can then be expressed as

$$r_i|L_i^* > 0 = X_i'\beta + \theta\hat{\lambda}_i + e_i,$$

where $\theta\hat{\lambda}_i = \rho\sigma_u\hat{\lambda}_i$ represents the correction term. Here, ρ is the correlation between the unobserved determinants of the probability of being accepted ϵ and unobserved determinants of interest rate u, σ_u is the standard deviation of u, and $\hat{\lambda}$ is the inverse Mills ratio evaluated at $Z_i\gamma$.

C Additional robustness checks

C.1 Data sharing and borrower type: Using platform scores

Given the restricted sample size in Section 4.4 due to the limited availability of the loan outcome variable (*Default*), I conduct an additional analysis regarding the association between data sharing and borrower type using the platform score, which is the credit risk computed by the platform and incorporates the information derived from the shared transaction data. Therefore, the platform score captures traits that are unobservable ex ante but still relevant for credit risk.

[Figure A.2]

First, the distribution of platform-provided scores for each credit score group is shown in Figure A.2. If applicants who disclose data are of a good type conditional on credit score, I expect to see a rightward shift of the distribution for those who share because the platform score is provided after data sharing. The critical assumption is that the signup and no-signup population have *ex-ante* an identical distribution. Thus, I use the matched sample. A quick look at the graphs indicates that the distribution of the platform score shifts toward the right.

Empirically, I regress the data sharing decision *Signup*, on a dummy variable *Good type*, which takes a value of one for platform scores three through seven and zero for platform score two. Platform score one (rejected) is excluded. Assuming both the signup and no-signup population ex ante have an identical distribution, and if the decision to signup was random in terms of unobserved borrower type, it is expected that there will be no significant shift in the distribution. I estimate this for each credit score group. Table C.1 shows that for the highest credit score group, on average, it is 16 percentage points more likely that the good type signs up, and the effect is even larger for the lower credit score groups: 20.6 and 19.9 percentage points for (E–G) and (H–K), respectively. The magnitude, however, is attenuated for the lowest credit score group, with only an 8.7 percentage point increase. The results are qualitatively consistent with the main analysis using ex post defaults.

[Table C.1]

There are, however, limitations to using platform scores as proxies for borrower types. For instance, the data sharing decision itself may lead to a better score regardless of borrower type and the information contained in the data. Thus, there is a possibility that the rightward shift of the distribution is partially driven by the signup decision itself rather than positive information content that suggests a good borrower type. Importantly, even with access to such granular payment data, the borrower type will only be known ex post.

Table C.1: **Data sharing decisions and the borrower type using platform scores**

	DV = 1 if data is shared Matched sample			
Credit score group	(A–D)	(E–G)	(H–K)	(L–M)
Goodtype (=1 if platform score 7-3)	0.160***	0.206***	0.199***	0.087**
	(0.006)	(0.004)	(0.007)	(0.021)
Controls	Y	Y	Y	Y
Cluster (Zipcode-Year)	Y	Y	Y	Y
N	110,968	107,012	27,950	3,148
Pseudo R2	0.0306	0.0362	0.0322	0.0193

Notes: This table shows the probability of data sharing among *Goodtype* borrowers (a dummy variable equal to one for platform scores from three through seven and zero for platform score two). A probit model with the matched sample is used for the analysis. Each column represents a risk group with (A–D) the highest and (L–M) the lowest. Each of the 124,539 loan applicants who shared data is matched one-to-one using age, credit score, income decile, access channel, and loan application year to create a control group of those who did not share data but is observably similar. Propensity score matching is used for age, credit score, and income decile, and exact matching is used for access channel and loan application year. The final sample includes 249,078 loan applications from 249,078 unique applicants. Control variables include age, income decile, loan amount, loan duration, female (dummy), married (dummy), main earner (dummy), number of current and past loans, homeowner (dummy), car owner (dummy), and access channel (categorical variable).

C.2 Selection on unobserved variables: Rosenbaum sensitivity analysis

To further address hidden bias from unobserved variables that may affect both the assignment to treatment and the outcome variable, I follow the method proposed by Rosenbaum 2002 and test the size of the quantitative deviation from a random assignment that would result in a statistically insignificant treatment effect. It is a useful tool to test the sensitivity of causal inferences by allowing researchers to quantify how severe unmeasured confounding variables must be between the treated and control units in order to nullify the treatment effect. Rosenbaum bound explicitly allow the odds of treatment to vary between the treated and control individuals by a parameter, $\Gamma \geq 1$, when the two groups have similar observable characteristics $X_t = X_c$; that is,

$$\frac{1}{\Gamma} \leq \frac{\dfrac{\pi_t}{(1-\pi_t)}}{\dfrac{\pi_c}{(1-\pi_c)}} \leq \Gamma \qquad \text{when} \quad X_t = X_c, \tag{4}$$

where $\pi_i = Pr(D_i = 1|X_i) = F(\beta x_i + \gamma u_i)$ is the probability of data sharing that can be expressed is a logistic function *F*, and x_i and u_i are the observable and unobservable variables, respectively. The *i*'s odds of data sharing are $\frac{\pi_i}{1-\pi_i} = e^{\beta x_i + \gamma u_i}$. If $\Gamma = 1$, $\pi_t = \pi_c$, which means the odds of treatment (sharing data) are the same for the treatment and control groups, who share similar observable characteristics. By setting the value of Γ to be greater than one, the degree of hidden bias can be varied. If $\Gamma = 2$, the treatment group is twice as likely as the control group to share data due to unobservable differences.

I first match individuals who share data on all observable characteristics to create a control group and examine the bounds at which the treatment effect is no longer significant. Table C.2 reports the bounds parameter Γ. The statistically significant effect of data sharing on the extensive margin will be challenged only if the unobserved biased selection into sharing were so high to cause the odds ratio of data sharing to differ between the two groups by around five times for the highest credit rating bracket (A–D). While the results for the second-highest credit score group (E–G) are less pronounced, the selection on unobservables would still have to be more than 50% as high. For the rest of the groups, the effect is statistically significant at all levels of the sensitivity parameter gamma. This means that even if there is a large amount of hidden bias due to unobserved covariates (i.e., a 20 times larger odds ratio), data sharing still has a statistically significant effect. Overall, this evidence suggests that selection on unobservables would have to be very large to eliminate the effects of data sharing.

Table C.2: **Rosenbaum bounds sensitivity analysis**

	Credit score groups			
	(A–D)	(E–G)	(H–K)	(L–M)
$\Gamma_{p>.01}$	4.98	1.53	20^{+}	20^{+}
$\Gamma_{p>.05}$	5.04	1.54	20^{+}	20^{+}
$\Gamma_{p>.10}$	5.08	1.55	20^{+}	20^{+}

Notes: This table shows how much higher the odds of data sharing based on unobservables would need to be for the treatment group compared to the control group such that the treatment effect is no longer significant at the 1%, 5%, and 10% level. Each of the loan applicants who shared data is matched one-to-one to create a control group of those who did not share data using all observable variables.