The London School of Economics and Political Science

Essays in Household Finance and Innovation

William Oliver Matcham

A thesis submitted to the Department of Economics of
the London School of Economics and Political Science
for the degree of Doctor of Philosophy

London, October 2023

For my late father, Nicholas Matcham:
my first and foremost academic inspiration

Declaration

I certify that the thesis I have presented for examination for the Ph.D. degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately 55,000 words.

The views expressed in Chapters 1 and 2 are exclusively my own and do not necessarily reflect those of the Financial Conduct Authority. Therefore, this thesis should not be reported as representing the views of the Financial Conduct Authority.

Statement of co-authored work

I confirm that Chapter 3 is co-authored with Mark Schankerman, and I contributed 50% of this work.

Acknowledgments

I thank several individuals who played an important role during my Ph.D. and graciously acknowledge the financial support from the Paul Woolley Centre and Richard Goeltz.

I owe an enormous debt of gratitude to my supervisors, Alessandro Gavazza and Mark Schankerman, for guiding me throughout the writing of this thesis and instructing me in the craft of economic research. They have been extremely generous with their time, and their insights and have contributed enormously to my development as a researcher. I thank Alessandro specifically for his constant encouragement and advice on how to work effectively as a researcher, helping me to focus and simplify my ideas, and defaulting my thinking towards finding the key trade-offs. I thank Mark specifically for teaching me to think like an economist and helping me to realize how to push the envelope of the economic questions in which I am interested. Further, I thank two excellent advisers: Tatiana Komarova, who started me on the research process and Daniel Paravisini, for many stimulating discussions as I produced the first draft of this thesis.

Throughout the years, my colleagues at LSE have provided valuable support. For this reason, I acknowledge those who produced inspiring theses ahead of me: Karun Adusumilli, Matteo Benetton, Svetlana Chekmasova, Patrick Coen, Hao Dong, Dita Eckardt, Alkis Georgiadis-Harris, Krittanai Laohakunakorn, Jay Euijung Lee, Chen Qiu, Luke Taylor, and Mengshan Xu. I acknowledge those who produced outstanding theses alongside me: Heidi Thysen, Arthur Taburet, Bilal Tabti, Aditya Soenarjo, Akash Raja, Lu Liu, Amanda Dahlstrand, Jamie Coen, Thomas Brzustowski, and Daniel Albuquerque. And I acknowledge those who will produce remarkable theses after me: Pete Lambert, Andres Fajardo-Ramirez, Hugo Reichart, Kamila Nowakowicz, Javad Shamsi, and Arnaud Dyèvre. Also, I thank my friends outside of LSE: Hannah Baker, Drummond Clark, Glenn Marshall, Merete Poulsen, David Steynor, Matthew Williams, Greg Wintle, and Charlotte Woodacre. Finally, a special mention

goes to Yimeng Li, who is always willing to discuss my economic research alongside our daily conversations about football.

Chapters 1 and 2 would not have been possible without the outstanding assistance of Cherryl Ng, Jonathan Shaw, Ann Sanders, Karen Croxson, and other staff members at the Financial Conduct Authority. I am especially grateful to Claudia Robles-Garcia for initiating my conversations with the Financial Conduct Authority and to Jeroen Nieboer for his initial assistance with my UK credit card market research. Further, I thank Jakub Drabik, whose research assistance was truly outstanding.

However, my deepest gratitude goes to those who are closest to me. The support and encouragement of my family have been immeasurably important in helping me complete my Ph.D. and follow a career I love. To Emily, for her constant support, immense patience, ever-present optimism, and continuous encouragement. To Shanthi, Theo, Oliver, and to my brother Thomas, for his inspirational academic ability, mathematics and coding advice, and clarity of thought in tackling life's challenges. And to my mother. Mum – you gave up so much to help me become the best version of myself. No Ph.D. acknowledgement could express how thankful I am for that. All of you make every hour of work worthwhile, and I am so grateful to be blessed with such a wonderful family.

Finally, I dedicate this thesis to my father, Nicholas Frederick Matcham. Thank you so much for sowing the seeds of intellectual curiosity throughout my childhood and encouraging me to pursue my academic goals. Your love of academia and mathematics set me on this path. If you did not have the answer to my statistics questions, you owned a book containing the solution. You inspired me to think about academic problems constantly and to challenge everything I was taught. You were my greatest academic supporter, and I am still trying to fill the void left by your absence. I wish you could read this. And, just between us, I hope the spelling, grammar, and punctuation within meet the standards you set in your own Ph.D. thesis.

Abstract

This thesis consists of two essays on household finance and one essay on innovation. Chapter 1 examines descriptively how UK credit card lenders set credit limits and interest rates for customers. First, I summarize the literature on credit card regulation and lenders' choices of credit limits and interest rates. Second, I detail the relevant credit card regulation, focusing on how UK credit card lenders are required to advertize a "representative" interest rate for every credit card product they offer. Third, I describe my credit card dataset together with basic summary statistics. Finally, I offer a set of descriptive findings. My main results describe the limited variation in interest rates within credit card products relative to substantial variation in credit limits.

In Chapter 2, I build on Chapter 1 by presenting and estimating an economic model of the UK credit card market. The modeling novelty is the link between individuals' credit limits and lenders' predictions of customers' risk. With the estimated model, I examine a counterfactual scenario in which credit card lenders are subject to no costs and constraints in individualizing interest rates and credit limits, which the existing environment precludes. In this case, credit card lenders' profits increase, consistent with lenders facing costs and constraints that discourage them from individualizing interest rates.

In Chapter 3, we develop a dynamic structural model of patent screening incorporating incentives, intrinsic motivation, and multi-round negotiation. We use detailed data on examiner decisions and employ natural language processing to create a new measure of patent distance that enables us to study strategic decisions by applicants and examiners. We find that patent screening is moderately effective, *given* the existing standards for patentability. Examiners exhibit substantial intrinsic motivation that significantly improves the effectiveness of screening. A reform that limits negotiation rounds strongly increases screening quality. We quantify the annual net social costs of patent screening at $25.5bn, equivalent to 6.5% of U.S. private sector R&D.

Contents

List of Figures

List of Tables

Chapter 1

Risk-Based Borrowing Limits in the UK Credit Card Market

1.1 Introduction and Summary

Asymmetric information is a pervasive feature of several markets considered essential for the functioning and development of the economy (Kiyotaki and Moore, 1997; Acemoglu, 2001). Two leading examples are insurance and credit markets. The presence of asymmetric information in these markets, specifically in the form of adverse selection, can lead to market inefficiencies and, in extreme cases, market unraveling (Akerlof, 1970; Rothschild and Stiglitz, 1976). The consequences of adverse selection can be severe, with the failure of credit markets described as "one of the major reasons for [economic] under-development" (Akerlof, 2001).

Accordingly, lenders in credit markets attempt to minimize the deleterious effects of adverse selection by tailoring contract characteristics to predictions of customers' default risk. However, governments want such contracts to be simple and transparent so that consumers are not misled and can search effectively across lenders. As a result, regulation has limited the extent to which lenders can tailor certain features of credit contracts according to risk. In the first two chapters of this thesis, I investigate how credit card lenders individualize contracts according to risk in the context of credit card regulation.

The academic literature and policy discourse in this space generally focuses on risk-based pricing, that is, the practice of interest rate discrimination based on a customer's risk. However, in this chapter, I provide evidence that UK credit card lenders

primarily adopt risk-based *credit limits*. In Chapter 2, I estimate a structural model of the credit card market to ascertain whether this empirical feature is a result of lenders' preference for risk-based credit limits over interest rates or the result of costs and constraints that affect lenders' willingness and ability to tailor interest rates according to risk.

Studying credit card lenders' credit limit and interest rate choices is important owing to its standalone economic interest and the credit card market's central role in the economy. It represents the largest unsecured credit market, with most prime and superprime adults owning a credit card. Net lending via credit cards reached £1.5bn in February 2022—the highest monthly amount since records began.[1] For this reason, lenders' credit limit and interest rate choices have material effects on individuals' financial well-being. This is especially true for subprime consumers, who are more likely to revolve a credit card balance and be credit constrained.

To establish my findings in Chapters 1 and 2, I use novel, statement-level administrative data on approximately 80% of all UK credit cards that were active between 2010 and 2015. I observe cardholder demographics and card characteristics for every card, along with monthly card use, borrowing, repayment, and default decisions. Among other advantages, the data contain the lenders' proprietary risk scores for every credit card origination, hence, I can credibly check whether interest rates and credit limits are tailored to predictions of customers' risk.

Using these data, I document how credit limits vary substantially across individuals within lenders and the credit card product, with the highest risk scores corresponding to the lowest credit limits. In contrast, interest rates are almost constant at the card-month level and are generally not risk-based. This fact is best understood in the context of UK credit card regulations, which require (i) lenders to advertise one annual percentage rate (APR) for each credit card and (ii) at least 51% of customers on each card to be granted the advertised APR or lower at origination. However, 80 to 90% of customers are granted the advertised APR at origination. Finally, I report substantial heterogeneity in the shape and scale of credit limit distributions across lenders. This is a primary source of variation that I seek to explain with the economic model in Chapter 2.

This chapter proceeds as follows. Following a review of the germane literature in

[1]Bank of England 2023, Bank of England website, https://www.bankofengland.co.uk/statistics/money-and-credit/2023/february-2022 last accessed 25 May 2023.

Section 1.2 and a summary of the relevant credit card regulations in Section 1.3, I describe my data in Section 1.4 and present my descriptive findings in Section 1.5.

1.2 Related Literature

Chapters 1 and 2 relate to several bodies of literature, and I detail my contributions to the work most closely related to my own in what follows. I describe my relationship to the extensive literature pertaining to credit card markets more generally in Appendix 1.A.

My thesis contributes primarily to the literature concerning the role of credit limits in credit card markets. In this regard, the most relevant article is that of Agarwal, Chomsisengphet, Mahoney, and Stroebel (2017), which shows that average credit limits increase in association with FICO scores in the US. The authors argue that credit limits are the main margin of adjustment for US credit card lenders. Further, the paper reveals that some lenders have FICO thresholds at which average credit limits increase discontinuously. For the authors, risk-based credit limits are a means rather than an end: Their paper focuses on the way in which banks' pass through credit expansions to customers. My contribution to this literature is to explain lender heterogeneity and discontinuities in credit limit schedules by estimating a model of lenders' credit limit choices. In the model, heterogeneous lender screening technologies that provide noisy signals on customers' levels of private risk justify the differences in the shape and scale of lenders' credit limit distributions and can explain discontinuities in the credit limits.[2]

Chapters 1 and 2 also relate to the literature on risk-based pricing. Existing research documents the presence of risk-based pricing in some financial markets (Edelberg, 2006; Magri and Pico, 2011; Magri, 2018; Bachas, 2019) and its absence in others (Benetton, 2021; Robles-Garcia, 2022). Notably, Adams, Einav, and Levin (2009) shows that risk-based pricing mitigates the effects of adverse selection in the US auto market. However, evidence of risk-based pricing in credit card markets is limited.[3] Hence, I contribute to the literature on risk-based pricing by documenting and

[2]On a related theme, Agarwal, Chomsisengphet, Mahoney, and Stroebel (2017) and Gross and Souleles (2002b;a) estimate the causal effect of credit limits on borrowing and default. Aydin (2022) presents an interesting experiment randomizing credit limit shocks across credit card accounts in the US. Fulford (2015) shows that US credit limits vary after origination, with more individuals obtaining credit limit increases than decreases. In the UK, credit limits are less volatile.

[3]Linares-Zegarra and Wilson (2012) argues that cards offered in riskier regions of the United

justifying the lack of risk-based pricing in the UK credit card market.

On the relationship between credit quantity and interest rates in markets with imperfect information, a highly influential and related paper is that of Stiglitz and Weiss (1981). The paper sets forth the notion of *credit rationing*, whereby lenders are not willing to increase interest rates to market clearing rates, because higher interest rates attract riskier borrowers (adverse selection effect) and can lead to more defaults (moral hazard effect). As a result, rather than increasing interest rates to market clearing levels, lenders decide to ration credit. My framework is consistent with that of Stiglitz and Weiss (1981), as I consider a credit market in which lenders set a constant interest rate for each credit card and induce some credit rationing by rejecting some consumers through card-level income thresholds. My contribution is to allow lenders to mitigate the adverse effects of asymmetric information by individualizing the *amount* of credit they offer each individual through the credit limit. Furthermore, whereas lenders in Stiglitz and Weiss (1981) infer default risk based on the willingness of potential borrowers to accept higher interest rates, in my framework, lenders obtain noisy signals on borrowers' risk. Therefore rationing in my framework will only occur as a result of residual imperfect information, given the lender's signal.

Underpinning risk-based credit limits is the use of statistical credit scoring models by lenders to measure risk. Einav, Jenkins, and Levin (2012; 2013) and Paravisini and Schoar (2015) document significant profit increases for lenders following the adoption of risk-scoring methods. A large segment of the literature focuses on the predictive, *statistical* quality of credit scores (Khandani, Kim, and Lo, 2010; Lessmann, Baesens, Seow, and Thomas, 2015; Butaru, Chen, Clark, Das, Lo, and Siddique, 2016; Albanesi and Vamossy, 2019; Fuster, Goldsmith-Pinkham, Ramadorai, and Walther, 2022). However, Einav, Finkelstein, Kluender, and Schrimpf (2016) takes a different approach by focusing on the *economic* content of risk scores. Their paper notes that if risk scores determine contractual terms, then risk scores confound underlying default risk with endogenous responses to those terms. I contribute to this body of literature in Chapter 2 by estimating the underlying screening technologies of lenders, which provide a signal of the underlying unobservable risk on a harmonized scale. By estimating these harmonized scores off credit limits at origination, rather than ex-post default, my measure is not confounded with the potential endogeneity

States have lower APRs on average, though they do not look at the relationship between interest rates and risk within credit cards.

of origination contractual terms and the lender-borrower relationship.

The final primary contribution of my thesis involves the literature on price regulation in credit markets. Two contexts are particularly relevant. The first is Chilean credit markets, studied by, among others, Cuesta and Sepulveda (2021). Their paper shows that tighter interest rate caps decrease surplus, with the welfare costs from loss of credit access outweighing the lower prices in equilibrium. Related to my work, they show that risk-based interest rate caps cause less harm to welfare.

Nelson (2022) and Agarwal, Chomsisengphet, Mahoney, and Stroebel (2014) study the second relevant regulatory context: the 2009 US Credit Card Accountability, Responsibility, and Disclosure (CARD) Act. Agarwal, Chomsisengphet, Mahoney, and Stroebel (2014) documents substantial consumer savings as a result of the Act. Nelson (2022) focuses on how the CARD Act limited lenders' abilities to reprice credit card customers after origination. The estimated economic model implies that consumer surplus rose at the expense of lender profits. In my paper, I focus entirely on *ex-ante* risk-based pricing. While I acknowledge the possible role of ex-post risk-based pricing, it has limited application in the UK, which is a feature I document in the next section. Instead, I show that price regulation limiting ex-ante risk-based pricing coincides with lenders adopting risk-based quantities through credit limits. Further, I consider counterfactual scenarios that allow lenders to freely base prices on risk in the context of endogenous risk-based credit limits, in which risk-based interest rates emerge.

1.3 Interest Rate Regulations

This section provides a brief and non-technical overview of regulations relating to APRs in the UK and the US. For precise details, the interested reader can consult the Consumer Credit Sourcebook (CONC) Section 3.5 for the UK case and the Code of Federal Regulations (CFR) §1022.70 for the US case.[4] The Financial Conduct Authority (FCA, 2015c) offers a general summary of UK credit card regulation.

[4] https://www.handbook.fca.org.uk/handbook/CONC/3/5.html and https://www.consumerfinance.gov/rules-policy/regulations/1022/70/, last accessed 25 May 2023.

1.3.1 Definitions and UK Advertised APR Regulations

A credit card's *purchase balance* is the total amount spent on the card relating to non-cash transactions currently not repaid.[5] A *purchase interest rate* for a credit card is the percentage rate at which interest is added to a credit card purchase balance.

As a prelude to defining the APR, I describe the daily interest compounding method, which many lenders use to add interest to credit cards. At the end of a statement cycle, lenders may allow individuals a grace period of interest-free days to pay their balance. This period is typically between 20 to 40 days. Lenders charge interest for the statement cycle if the total balance is not paid within the grace period. Lenders compound interest on unpaid balances daily by taking each day's average balance and multiplying it by the daily periodic purchase rate. The *daily periodic purchase rate* is the percentage rate at which interest is added to an unpaid balance daily. The consumer is notified of the interest charged on their monthly statement, in which the monthly interest charge is the sum of daily interest across all the days in the month.

The *annual purchase rate* is the daily periodic rate multiplied by the number of days in the year. For example, outside a leap year, if the daily periodic rate is 0.0005, the annual purchase rate is 0.1825 or 18.25%. An *APR* is similar to the annual purchase rate, except it also accounts for all mandatory fees that an individual must pay each year on the card, hence, it represents the total cost of revolving a balance on a credit card each year. If a card has no compulsory fees or charges, its APR equals the annual purchase rate.

Accounting for fees when calculating the total cost of borrowing on a card requires a representative credit limit. The calculation of the APR assumes that the individual pays the fees, spends the entire representative credit limit on the first day of the year, and then pays it back in equal, regular installments over a year without spending anything else. The sum of the charges and interest accruing over a year (as a percentage) when an individual follows this spending pattern and pays the fees defines the APR.

[5]The withdrawal of cash counts towards the cash advance balance and cash advance interest rates are typically higher than purchase interest rates. Transfers of balances from a previous credit card counts towards the balance transfer balance, which may also have a different interest rate to the purchase rate and cash advance rate.

The *representative* or *advertised APR* is defined as "an APR at or below which the firm communicating or approving the financial promotion reasonably expects, at the date on which the promotion is communicated or approved, that credit would be provided under at least 51% of the credit agreements which will be entered into as a result of the promotion." Credit card lenders must include a representative APR on all promotional materials for a credit card, and, by definition, most consumers each month must obtain the representative APR or lower. Before February 2011, the proportion of customers on a given credit card required to obtain the advertised APR or lower was 66%. After February 2011, the UK harmonized its regulations with the EU and the proportion changed from 66% to 51%.

1.3.2 US Credit Card Regulations

US credit card lenders do not have to provide one representative APR for each credit card, but they are still subject to regulation should they use risk-based pricing. Since the introduction of the Truth in Lending Act in 1998, credit card agreements must include a "Schumer" Box, which is a table showing basic information about the card's rates and fees. The box must contain either a list of values or a range of values identifying the APR that the lender will use. The APR values must be in at least an 18-point font size.

Further, lenders must provide a consumer with a "risk-based pricing notice" if they (i) use a consumer credit report in connection with a credit application and (ii) grant or extend credit to that consumer on "material terms that are materially less favorable than the most favorable material terms available to a substantial proportion of consumers from or through the lender." The risk-based pricing notice must inform the consumer that a consumer report includes information about their credit history, that the terms offered have been set based on information from the consumer report, and that the terms offered may be less favorable than the terms offered to consumers with better credit histories, among other information.

Another major addition to recent US credit card regulations is the 2009 CARD Act of 2009. The CARD Act limits lenders' ability to change interest rates after origination and is the subject of the papers by Nelson (2022) and Agarwal, Chomsisengphet, Mahoney, and Stroebel (2014).

1.4 Data and Summary Statistics

In this section, I summarize the novel datasets I employ in my analysis. My primary data source is the FCA Credit Card Market Study (CCMS) Dataset.[6] The FCA used its legal authority as the regulator of UK financial markets to obtain data on all the credit cards active at 14 lenders between 2010 and 2015.[7] The data cover approximately 80% of the universe of cards active in 2010–2015, comprising around 74 million cards. The CCMS databases are only available to restricted staff and associated researchers at the FCA.

1.4.1 Origination Data

The first dataset contains information on cardholders and their cards at *origination*, including the cardholder's demographics (age, income, etc.), their acquisition channel (whether in-branch, online, by post, via telephone, etc.), and the interest rate and credit limit of their cards. The most useful feature of this dataset, however, is the inclusion of each customer's lender-specific risk score at origination.

Documenting that credit limits are based on risk rather than interest rates is the foundation of my analysis. For this reason, I require observations of lenders' measures of customer risk. Furthermore, observations on publicly available risk scores are insufficient because UK lenders generally do not use these scores for credit decisions.[8] As such, it is critical that I have access to observations of lenders' proprietary risk scores.

Lenders' proprietary risk scores are formulated on different numerical scales and, as shown in Figure 1.C.1, vary in how they are distributed over these scales. Further, public risk scores only explain a moderate proportion of the variation in each lender's proprietary risk scores. To provide evidence of this, I regress each lender's proprietary risk scores on percentile dummies of the main publicly available risk score. In these

[6] See (FCA, 2015b) for a detailed summary of the data source.

[7] The FCA chose 11 firms (divided into 14 separate lending entities) as representative of the entire credit card market. For confidentiality reasons, I cannot reveal their identities. In the main analysis, I omit store cards and, where necessary, one other lender for which data submission issues occurred.

[8] For example, suppose a researcher only has access to public credit scores but interest rates are based solely on private risk scores. The researcher would find no relationship between public risk scores and interest rates, and it would be *incorrect* to interpret this as the absence of risk-based pricing.

regressions, public risk scores explain 21% of the variation in private risk scores explained on average.

The use of proprietary risk scores rather than public risk scores in the UK credit card market contrasts with the US, where FICO scores offer a measure of customer creditworthiness that many banks use as part of their lending decisions (Agarwal, Chomsisengphet, Mahoney, and Stroebel, 2017). Recent research has provided some justification for why lenders might create their own risk scores. For example, Albanesi and Vamossy (2019) shows that machine learning (specifically deep learning) methods consistently outperform standard credit scoring models, even when trained on the same data sources. Further, lenders may have more granular customer data than credit reference agencies are able to access.[9]

Table 1.C.4 provides summary statistics on *individuals* at *origination*. The mean age is 43 years. Net monthly individual income is £2,099 at the mean, though the distribution is right-skewed, and the median income is £1604. Four in ten customers have an existing relationship with the credit card lender prior to origination, approximately 52% of cardholders report being female, 57% are homeowners, and 85% are employed. Finally, most customers (53%) originate online, 32% originate in a store, 12% originate via post, and 4% by telephone.

Table 1.C.5 provides summary statistics on individuals' *card features* at origination. The mean credit limit is £3390, and the mean purchase APR at origination is 21.52%. The coefficient of variation in credit limits across all lenders and months is almost 1. The variation in interest rates (purchase and balance transfer) is much smaller at approximately 0.36. Expanding on this finding—reported here across lenders and cards—is the focus of the analysis in Section 1.5. Promotional deal lengths for purchases are short, typically around three or six months where they exist, and around 44% of cards have no purchase promotional deal. Across all cards, 83% of customers obtain the advertised APR, a fact I describe in detail in Section 1.5.1. Finally, 28% of customers transfer a balance from a previous card at origination.

[9]See FCA (2022) for a recent report on the UK credit information market and credit reference agencies.

1.4.2 Statement Data

The second dataset is a monthly panel of statement data for active credit cards. For the 61 months between January 2010 and January 2015, I observe opening and closing balances; repayments; the number and value of transactions, fees, interest; and the evolution of credit limits and interest rates. I also observe the account status, which records the months for which payment is overdue. In the event of repeated failures to repay the minimum repayment, the lender will typically charge off the account, which the dataset also details. Finally, these data contain observations on lenders' costs of servicing the account, including typically unobserved funding costs and provisions for non-repayment of debt at the *statement* level. Observations on lender's funding costs are essential to estimate screening technologies. Without these observations, I cannot separate differences in lenders' costs from differences in the precision of their screening technologies.

Table 1.C.6 provides summary statistics for the statement-level variables. Credit limits are larger, and interest rates are lower relative to origination, as riskier individuals are repriced or eventually close their cards. Over 25% of balances are zero, and the distribution of account balances is heavily right-skewed, with the mean account balance approximately £830 larger than the median. Repayments are much lower than balances, which is unsurprising as many individuals make the minimum monthly repayment. Interest is also highly skewed: over half the statement months carry no interest, but the right tail is long, with a 90^{th} percentile of £26.58. Finally, only 2% of statement months have an overdue payment, and 2% of statement months involve the account being charged off.

Based on these data, I find substantial variation across lenders regarding the proportion of statements in which the entire credit card balance is repaid. The proportion ranges from approximately 20% at one lender to 80% at another.

1.4.3 Card Characteristics Panel

The third CCMS dataset is a monthly panel of card characteristics. For the months between January 2010 and January 2015, the panel collects card rewards (such as cashback and air miles) and income thresholds. Both income thresholds (for choice sets) and rewards (for observable card characteristics) make credible demand estimation feasible. Further, the dataset includes each card's monthly advertised APR. With this variable, I calculate the differences between the advertised and obtained APRs, which provides the intensive and extensive margins of risk-based pricing. As

previously mentioned, the obtained APR must be at most the advertised APR for at least 51% of the originations within a product-month.

Table 1.C.7 provides summary statistics on cards, in which the unit of observation is the card-month. The most important conclusion from this table is that rewards are scant in the UK, with only 9% of card-months offering cashback and 7% offering air miles. This differs from the US, where rewards are generally more readily available. The table also shows the following facts. First, around 88% of cards have no annual fees. Annual fees are also more common in the US. Second, there is significant dispersion across card-months in minimum and maximum credit limits. Third, individuals usually receive around 25 days to repay their bill before interest is added. Fourth, most cards are available to all customers, with only 5% reserved for students and 7% exclusively for those who are employed.

1.4.4 Other Sources

The CCMS data include a credit reference agency (CRA) dataset that matches cards to individuals. The CRA data confirm that, on average, UK adults hold fewer cards relative to the US population, with the majority holding only one card each (see Figures 1.C.2 and and FCA (2015a)). I estimate my model using individuals with one credit card, which circumvents complications arising from (i) balance transfers and (ii) balance-matching heuristics in repayment across multiple cards (Gathergood, Mahoney, Stewart, and Weber, 2019). Finally, I occasionally complement my analysis with an FCA survey of cardholders, detailed in FCA (2015d).

1.5 Central Descriptive Findings

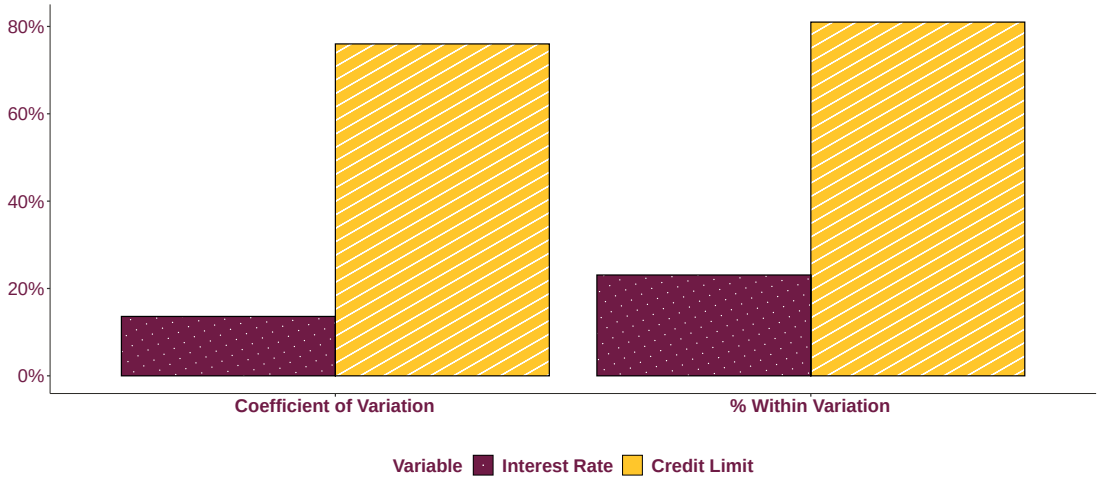
I conclude this chapter by showing that the leading UK credit card lenders individualize credit card contracts through risk-based credit limits rather than interest rates. The previous section revealed that, when pooling across lenders and months, credit limits are substantially more dispersed than interest rates. In what follows, I explore similar statistics within lenders, cards, and months.

1.5.1 Limited Variation in Lenders' Interest Rates

Limited Total Variation

I begin by documenting the limited variation in each lender's interest rates across originations within a month. Table 1.C.1 Column (1) reports the average (over

FIGURE 1.1: Coefficient of variation and proportion of within-card variation in interest rates and credit limits for prime and superprime lenders



Notes: To construct each bar, I calculate the average of the statistic over the months within a lender to create a lender-specific value. Each bar in this plot is a weighted average (weighting by origination share) of the lender-specific averages for the prime and superprime lenders.

months) of the lenders' interest rate coefficient of variation.[10] The values are below 23%, and, as shown in the left-hand dotted maroon bar in Figure 1.1, the average across prime and superprime lenders (weighted by originations) is 14%. This implies that the standard deviation in the interest rate is, on average, one-seventh of the mean at a lender in a given month. Additionally, as detailed in Table 1.C.1 Columns (2) and (3), the across-lender weighted average of the ratio of the 75^{th} to 25^{th} percentile (respectively 90^{th} to 10^{th}) for interest rates is 1.19 (respectively 1.38), further illustrating limited variation in interest rates within lenders.

Limited Within-Card Variation

For the leading UK credit card lenders, a modest proportion of the already minimal total variation in interest rates is within credit cards rather than across them. To expose this feature, I decompose the variation in lenders' interest rates into within-

[10]For lender ℓ in month t, who offers cards $j \in J_{\ell t}$, creating originations $i \in I_{j\ell t}$, I calculate the grand average $\bar{r}_{\ell t}$ and standard deviations $sd_{r,\ell t}$ of interest rates, where $\bar{r}_{\ell t} = \frac{1}{I_{\ell t}} \sum_i \sum_j r_{ij\ell t}$ and $\text{sd}^2_{\ell t} = \frac{1}{I_{\ell t}} \sum_j \sum_i (r_{ij\ell t} - \bar{r}_{\ell t})^2$, and $I_{\ell t}$ is the total number of originations. The value in Column (1) of Table 1.C.1, for lender ℓ is $cv_{r\ell} = \frac{1}{T_\ell} \sum_t \frac{sd_{r,\ell t}}{\bar{r}_{\ell t}}$, where $T_{\ell t}$ is the number of months of observations for each lender. The left-hand dotted maroon bar in Figure 1.1 shows the weighted average (weighted by market share) of $cv_{r\ell}$ over prime and superprime lenders.

card and between-card terms. For each lender ℓ and month t, I divide the total variation $V_{\ell t}^{TOT}$ in interest rates $r_{ij\ell t}$ for cards $j \in J_{\ell t}$ and originations $i \in I_{j\ell t}$ into within-card variation $V_{\ell t}^{W}$ and between-card variation $V_{\ell t}^{B}$

$$\underbrace{\frac{1}{I_{\ell t}} \sum_{j=1}^{J_{\ell t}} \sum_{i=1}^{I_{j\ell t}} (r_{ij\ell t} - \bar{r}_{\ell t})^2}_{V_{\ell t}^{TOT}} = \underbrace{\frac{1}{I_{\ell t}} \sum_{j=1}^{J_{\ell t}} \sum_{i=1}^{I_{j\ell t}} (r_{ij\ell t} - \bar{r}_{j\ell t})^2}_{V_{\ell t}^{W}} + \underbrace{\sum_{j} s_{j\ell t} (\bar{r}_{j\ell t} - \bar{r}_{\ell t})^2}_{V_{\ell t}^{B}}, \tag{1.1}$$

where $I_{\ell t}$ is the total number of originations at lender ℓ in month t, $\bar{r}_{\ell t}$ is the grand mean of interest rates, $\bar{r}_{j\ell t}$ is the card-j-specific interest rate mean, and $s_{j\ell t} = \frac{I_{j\ell t}}{I_{\ell t}}$ is the share of originations on card j at lender ℓ in month t. Intuitively, the decomposition separates the grand variance into an average of within-card variances ($V_{\ell t}^{W}$) and a weighted variance of card averages ($V_{\ell t}^{B}$). As plotted in the right-hand dotted maroon bar in Figure 1.1, within variation for prime and superprime lenders is, on average, 23% of the total variation.[11] Table 1.C.1 Column (4) reports the values of the percentage of within-card variation for all lenders. In the extreme case, two lenders give almost all (99% and 100%) customers on the respective credit card the same interest rate in *all* months, hence, practically all the variation in interest rates at origination is at the card level for these two lenders.

Proportion of Customers Obtaining Advertised APR

To explain the lack of within-card variation in interest rates, I calculate the monthly percentage of customers obtaining the advertised APR and plot its value in Figure 1.C.3. The value across all credit cards in the sample hovers around 80 to 90% and it does not change in February 2011 when regulations on advertised APRs relax. Even though regulation required lenders to give the advertised APR (or lower) to only 51% of their customers after February 2011, most lenders still gave almost all their customers the advertised APR.[12] Further, Figure 1.C.4 plots the proportion of *cards* giving at least 70% (solid) and 90% (dashed) of customers the advertised APR at origination in each month. Each month, around 85% of cards give at least 70% of customers the advertised APR, and in 77% of card-months, over 90% of originations obtain the advertised APR. These statistics confirm that most *cards*, and not just *lenders*, give the majority of their consumers the advertised APR. In Chapter 2 I embed this feature into my economic model by making borrowers'

[11] The weighted average including subprime lenders is 31%. I discuss subprime lenders separately in Appendix 1.B.

[12] I pool over lenders in this case, but the lender-by-lender and card-by-card plots are similar.

credit card preferences dependent on card-level APRs, abstracting from the minimal within-card variation in interest rates.

I summarize the descriptive facts presented thus far in Finding 1.

Finding 1 (**Interest Rate Variation**). *There is limited total variation in interest rates, of which an even smaller part is within-card variation. The fact that 80–90% of customers obtain the advertised APR at origination each month corroborates the limited within-card variation in interest rates.*

1.5.2 Substantial Variation in Credit Limits

Substantial Total Variation

Having confirmed the lack of variation (particularly within-card variation) in interest rates, I turn to credit limits. I provide the average of lenders' credit limit coefficients of variations (weighted by originations) in the left-hand striped gold bar in Figure 1.1. At 78%, it is over five times larger than the interest rate equivalent. As reported in Columns (6) and (7) of Table 1.C.1, the across-lender weighted average of the 75^{th} to 25^{th} (respectively 90^{th} to 10^{th}) credit limit percentile ratios is 3.34 (respectively 9.15), displaying substantial variation in credit limits within each lender.

Substantial Within Variation

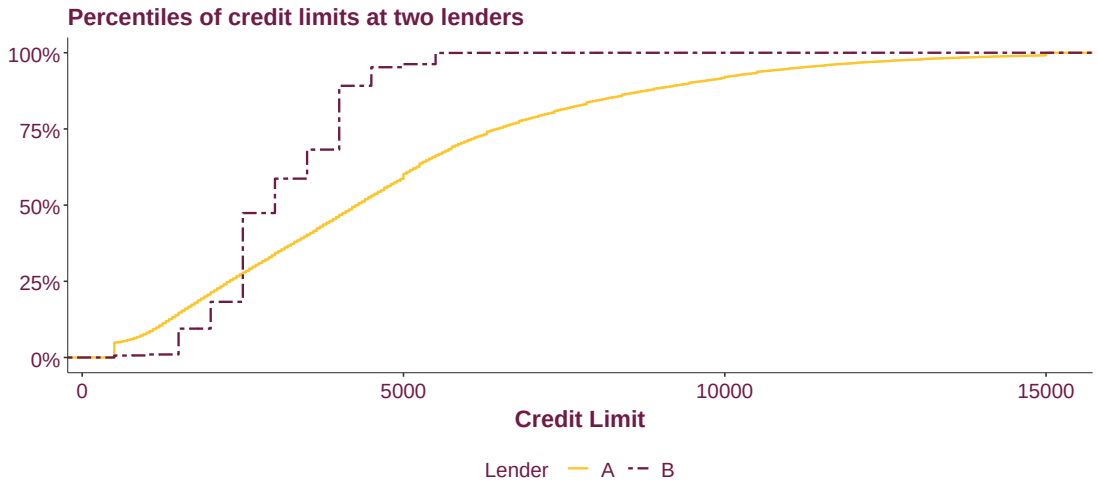
I perform the same within-card and between-card decomposition as in Equation (1.1) for credit limits. Across lenders, as shown in the right-hand gold striped bar in Figure 1.1, the average percentage of total variation found within credit cards is 81%. The dominance of within variation suggests that lenders do not sort individuals onto cards with varying average credit limits. Rather, there is large variation in credit limits across individuals even within a specific credit card product.

Shape and Scale of Credit Limit Distributions

The distribution of credit limits varies substantially across all lenders in both shape and scale.[13] I illustrate this fact in Figure 1.2, where I plot the empirical cumulative distribution function (CDF) of credit limits for two contrasting lenders, labelled

[13]To confirm differences between lenders' credit limit distributions formally, I conduct multiple distribution "Kolmogorov-Smirnov" hypothesis tests. I strongly reject the equality of empirical CDFs across lenders at lower than 0.1% significance levels in all tests. Table 1.C.2 reports the results and further details of the specific tests I conduct.

FIGURE 1.2: Empirical CDFs of two particular lenders' credit limits



Lenders A and B. Two substantial differences are evident. The first relates to the *shape* of the credit limit distributions. Lender B's curve is step-like, implying a coarse process of assigning credit limits to individuals, whereby groups of consumers obtain the same credit limit. Lender A's smooth curve is consistent with a more finely tuned allocation mechanism for origination credit limits. The second difference relates to the *scale* of the credit limit distributions. Lender A has lower values of credit limits than Lender B for the first 25 percentiles; all percentiles after the 25^{th}, however, are larger. The range of Lender A's credit limit distribution is evidently much larger.

Other lenders' credit limit empirical CDFs at origination, plotted in Figure 1.C.5, lie between the two lenders in Figure 1.2. This range in the shape and scale of distributions is consistent with lenders who vary in the coarseness of their credit limit assignment.[14] Some lenders offer large groups of customers the same credit limit, while others with smoother CDFs adjust their credit limits more precisely to each customer. The model I build in Chapter 2 justifies differences in the shape and scale of lenders' credit limit distributions through differences in the coarseness of information they possess on customers' risk levels.

I summarize my descriptive facts on the distributions of credit limits in Finding 2.

Finding 2 (Credit Limit Distributions). *There is substantial within-card variation in credit limits across lenders. The distributions of credit limits differ in shape and scale across lenders.*

[14]These findings are robust to dividing lenders into cards and dividing originations by year and by month.

1.5.3 Risk-Based Credit Limits, Not Risk-Based Prices

Since interest rates at a lender rarely vary within a credit card month, they are unlikely to relate strongly to lenders' predictions of customers' default risk. I confirm this in Figure 1.C.6, in which most lenders' average interest rates are flat across the application risk score support. Exceptions exist for two subprime lenders, who, as described in Appendix 1.B, engage in risk-based pricing.

Lenders could employ risk-based pricing by adjusting interest rates after origination and *repricing* customers according to their evolving risk and behavior.[15] However, in the period I study, limited repricing occurs in the UK credit card markets. As detailed in Table 1.C.3, lenders reprice only 4% of cards within the first year after origination.

Rather, as expected, lenders link each individual's credit limit to an assessment of their risk. In Figure 1.3, I plot the mean of the origination credit limit along application credit scores for two contrasting lenders.[16] Both curves are upward sloping, consistent with risk-based credit limits. Further, the right-hand lender has discontinuities in credit scores at credit score thresholds. If risk is continuously distributed and lenders create finely tuned assessments of customers' risk, discontinuities in credit limits at points of their credit scores are difficult to rationalize. Accordingly, the overarching aim of my model is to rationalize discreteness and discontinuities in lenders' credit limit distributions through coarse (discrete) assessments of customers' risk. Separate and ongoing work exploits these discontinuities to measure the distribution of causal effects of credit limits on borrowing and default, similar to Agarwal, Chomsisengphet, Mahoney, and Stroebel (2017). Several discontinuities in credit limits exist over lenders' credit scores and time. Formally aggregating multiple regression discontinuity design estimates across cards, time, and proprietary risk scores is a detailed procedure and the subject of future work on this topic.

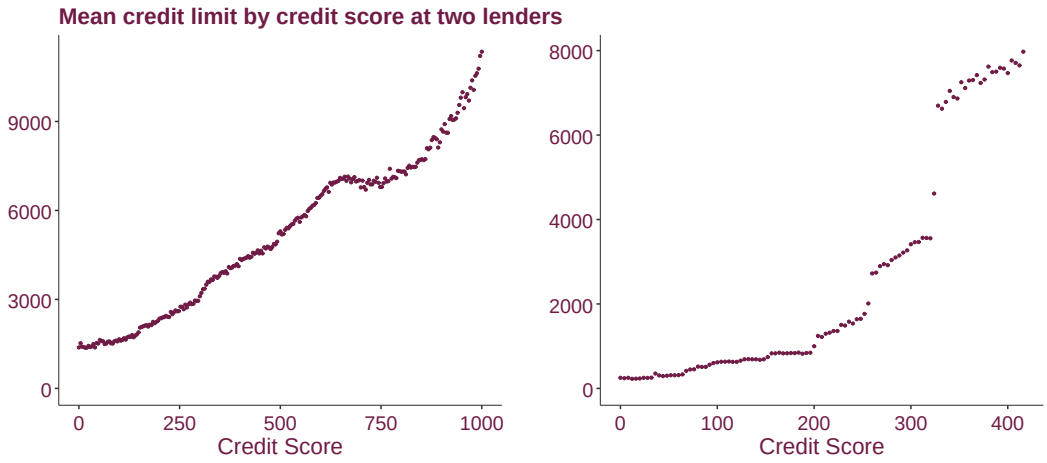
I summarize my descriptive facts on risk-based credit limits in Finding 3.

Finding 3 (**Risk-Based Credit Limits**). *Credit limits vary with lender-specific application credit scores, while interest rates generally do not. Heterogeneity exists in how lenders map their credit scores into credit limits: Some, but not all, lenders*

[15] Nelson (2022) shows that repricing was a relatively common practice in the US credit card market until the (2009) CARD Act essentially outlawed the practice.

[16] In Figure 1.C.7, I plot the mean of the origination credit limit for each lender, along application credit scores. In unreported plots, the same patterns emerge when produced by month.

FIGURE 1.3: Mean credit limits across risk scores for two particular lenders



Notes: Credit score scales differ across lenders so cannot be compared.

exhibit discontinuities in their credit limits at certain credit score thresholds.

1.5.4 Implications of Descriptive Findings

This chapter reveals that the leading UK credit card lenders individualize credit limits according to their assessments of the customer's risk. My empirical facts are best understood alongside UK credit card regulations, which demands a card-level advertised APR that most customers must obtain at origination. The next step is to learn about how lender heterogeneity and the regulatory environment impact lenders' decisions to individualize contract characteristics. For example, the empirical setting is not insightful with regard to how lenders would choose interest rates if they were not required to set and advertise a card-level APR. In the absence of meaningful exogenous variation in the regulatory environment or the makeup of lenders, the best—and perhaps only—way to achieve this aim is to build an economic model of the credit card market. This model and its estimation follows in Chapter 2.

Appendices for Chapter 1

1.A Broader Literature on Credit Card Markets

Through this thesis, I contribute to the vast body of literature in economics and finance examining credit card markets. Several research articles, books, and reports on credit card markets are of note. Agarwal and Zhang (2015) surveys the credit card market literature, while Knight (2010) extensively summarizes the UK credit card market. The FCA produced a UK credit card market study in 2015 (FCA, 2015a), and the Consumer Finance Protection Bureau (CFPB) produces a biennial report on the US credit card market, the most recent appearing in 2021 (CFPB, 2021). Evans and Schmalensee (2005) offers a comprehensive account of the history of credit cards in the US. A separate body of literature studies credit card networks, the most recent of which includes Wang (2023).

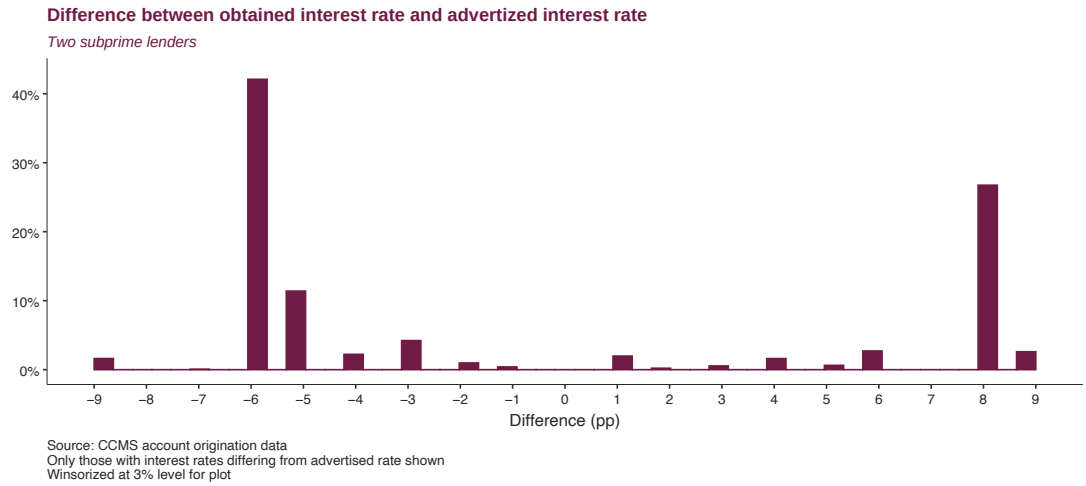
Many papers explore the impact of behavioral biases on the credit card market. The biases include **time inconsistency and present bias** (Ausubel and Shui, 2005; Ausubel, 1991; 1999; Laibson, Repetto, and Tobacman, 2000; Meier and Sprenger, 2010; Kuchler and Pagel, 2021), **self-control and naivete** (Heidhues and Kőszegi, 2010), **anchoring** (Keys and Wang, 2019; Stewart, 2009), **exponential growth bias** (Stango and Zinman, 2009; Adams, Guttman-Kenney, Hayes, Hunt, Laibson, and Stewart, 2022), **over-optimism** (Exler, Livshits, MacGee, and Tertilt, 2021; Yang, Markoczy, and Qi, 2007), **shrouding** (Ru and Schoar, 2016), and **repayment heuristics** (Gathergood, Mahoney, Stewart, and Weber, 2019). Although my model in Chapter 2 does not explicitly account for these features, I base my estimation on a set of linearized equations that are not inconsistent with behavioral biases. Future research could explore the interaction between consumer behavioral biases and lenders' risk-based credit limits and interest rates.

Other papers stress the importance of **search** (Galenianos and Gavazza, 2022; Stango, 2002; Stango and Zinman, 2015; Drozd and Nosal, 2011; Calem and Mester, 1995), **promotional deals** (Drozd and Kowalik, 2019), **learning** (Agarwal, Driscoll, Gabaix, and Laibson, 2008), **minimum repayments** (Druedahl and Jørgensen, 2018), and **information frictions** (Ausubel, 1999) in credit card markets. These topics are relevant features of credit card markets, and, similar to behavioral biases, further work could explore how they interact with risk-based prices and credit limits. In particular, when lenders have to advertise an APR, search becomes less costly for consumers, so the role of consumer search is particularly important.

1.B Pricing by Subprime Lenders

I identify two particular subprime lenders in the sample. These lenders (removed from the solid line to create the higher dashed line in Figure 1.C.3) price differently, giving many customers a rate that differs from the advertised APR. As Table 1.C.1 reveals, in contrast to prime and superprime lenders, most of the interest rate variation for these two lenders is within rather than between cards. I investigate these two lenders' pricing strategies in Figure 1.B.1 by plotting the distribution of percentage point differences (rounded to the nearest integer) between the advertised APRs and the APRs the customers actually received. The differences are minor and often favorable to consumers. In the most commonly occurring case, 42% of customers received an interest rate six percentage points *lower* than that which was advertised. Very few customers (around 2.6%) received interest rates more than eight percentage points above the advertised APR.

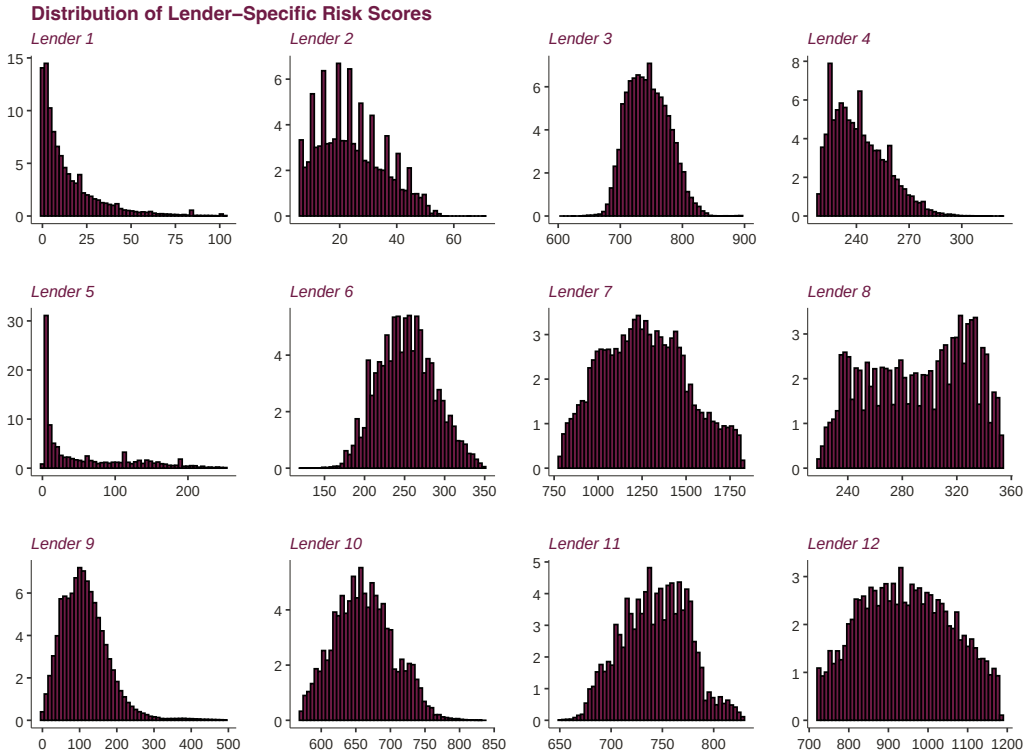
FIGURE 1.B.1: Histogram of differences between obtained and advertised APR at two subprime lenders



1.C Additional Figures and Tables

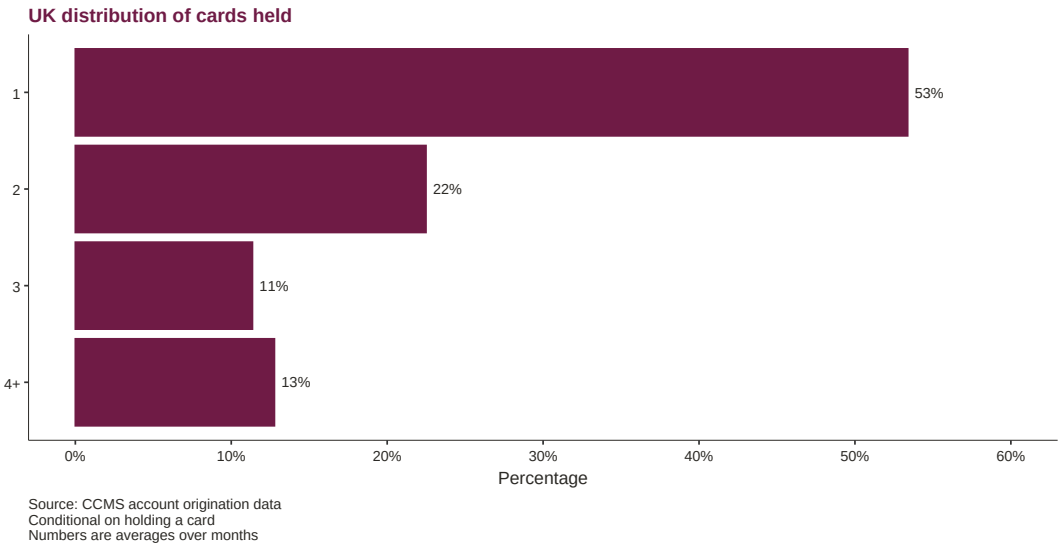
1.C.1 Figures

FIGURE 1.C.1: Distribution of proprietary credit scores across lenders



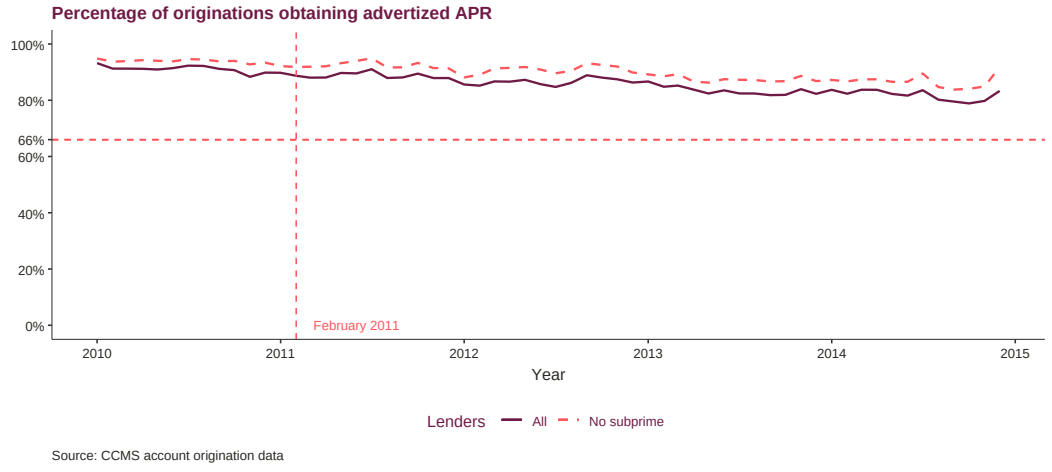
Notes: I scramble lenders' identities to preserve anonymity, so labels do not necessarily match the identities in other tables and figures. Link back to data section

FIGURE 1.C.2: Distribution of the number of cards held by individuals



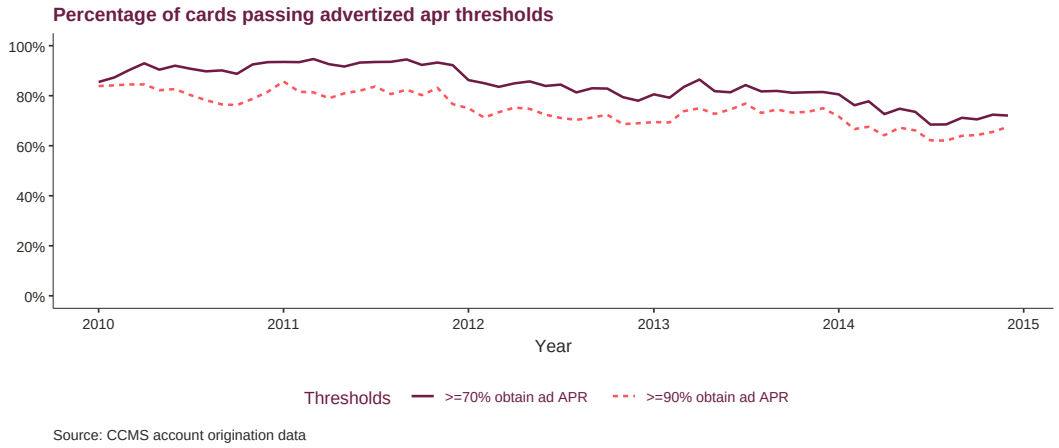
Notes: Distribution of the number of cards held by individuals with at least one credit card in the UK. I calculate the distribution using the CRA dataset described in text. I calculate the distribution of cards held, conditional on holding a card, in each month, and then average over months. The distribution of the number of cards shows total stability over time, justifying the process of averaging the distribution over months. Link back to data section

FIGURE 1.C.3: Proportion of originations each month obtaining the advertised APR



Notes: The solid line includes all lenders; the dashed line removes the two subprime lenders discussed in text. The proportion did not significantly change in February 2011 when regulation on the proportion required to obtain the advertised APR or below fell from 66% to 51%. Link back to descriptive findings

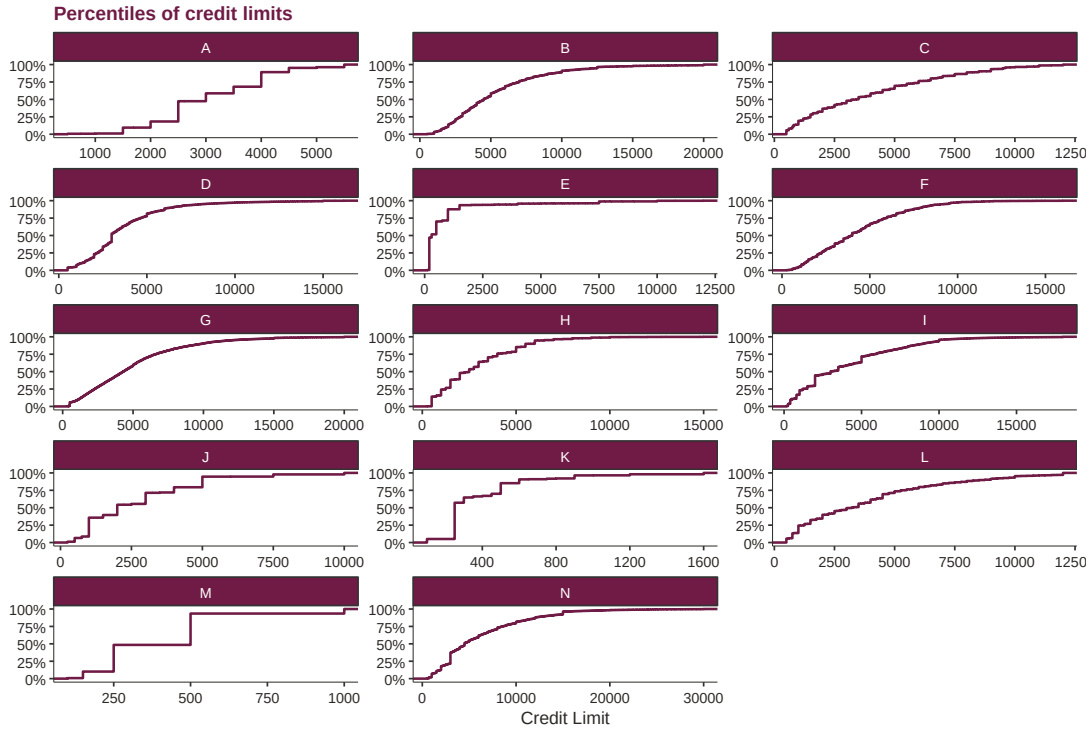
FIGURE 1.C.4: Proportion of cards each month given advertised APR



Source: CCMS account origination data

Notes: Link back to descriptive findings

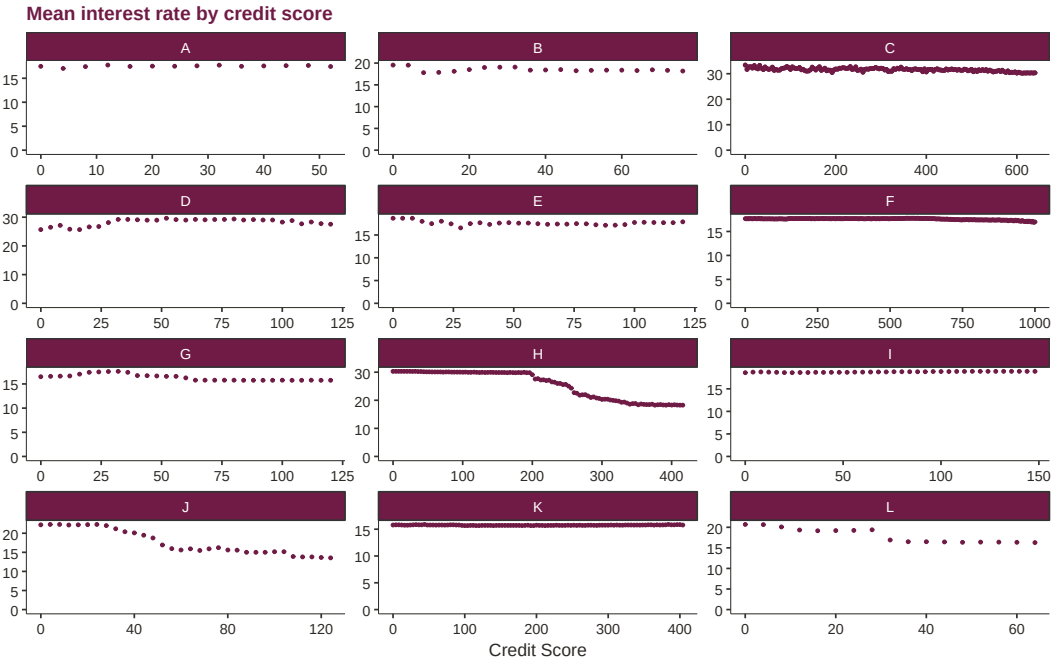
FIGURE 1.C.5: Empirical CDFs of credit limits at all lenders, pooled over time



Source: CCMS account origination data

Notes: I scramble lenders' identities to preserve anonymity, so labels do not necessarily match the identities in other tables and figures. Link back to descriptive findings

FIGURE 1.C.6: Mean interest rates across lenders' risk scores



Notes: I scramble lenders' identities to preserve anonymity, so labels do not necessarily match the identities in other tables and figures. Credit score scales differ across lenders so cannot be compared. Credit scores are not available at two lenders. Link back to descriptive findings

FIGURE 1.C.7: Mean credit limits across lenders' risk scores

Notes: I scramble lenders' identities to preserve anonymity, so labels do not necessarily match the identities in other tables and figures. Credit score scales differ across lenders so cannot be compared. Credit scores are not available at two lenders. Link back to descriptive findings

1.C.2 Tables

TABLE 1.C.1: Interest rate and credit limit variation by lender

	Interest Rate				Credit Limit				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Bank	**C. of V.**	**75/25**	**90/10**	**Within**	**C. of V.**	**75/25**	**90/10**	**Within**	**Share**
A	0.11	1.19	1.32	20.45	0.78	3.28	8.98	88.53	2.2
B	0.15	1.25	1.39	45.62	0.79	4.57	11.74	77.89	8.27
C	0.22	1.29	1.59	18.63	0.84	4.45	16.18	71.11	21.79
D	0.14	1.02	1.66	23.13	0.74	3.87	9.76	73.92	3.16
E	0.11	1.09	1.27	44.72	0.76	3.12	10.36	82.38	8.36
F	0.12	1.11	1.21		0.59	2.65	6.08		5.98
G	0.12	1.06	1.32	0.00	1.64	4.71	9.99	24.97	8.48
H	0.06	1.11	1.15	0.99	0.66	2.07	5.18	98.57	11.35
I	0.23	1.53	1.77	66.07	0.76	4.44	10.83	92.51	5.11
J	0.08	1.03	1.15	19.15	0.66	2.42	5.37	91.31	9.49
K	0.08	1.01	1.17		0.32	1.51	2.39		4.36
Subprime 1	0.19	1.41	1.42	83.68	0.51	2.00	2.68	88.62	8.78
Subprime 2	0.15	1.31	1.49	96.48	0.70	1.77	2.97	97.38	2.66
Mean	0.14	1.19	1.38	38.08	0.75	3.14	7.88	80.65	-
Weight Mean	0.14	1.19	1.38	31.22	0.78	3.34	9.15	78.28	-
NS Mean	0.13	1.15	1.36	26.53	0.78	3.37	8.81	77.91	-
NS Weight Mean	0.14	1.17	1.37	23.09	0.81	3.52	9.98	76.47	-

Notes: "Share" column reports share of originations; "C. of V." columns report coefficients of variation; "75/25" and "90/10" columns report 75^{th} to 25^{th} and 90^{th} to 10^{th} percentile ratios respectively; "within" columns report the ratio of within to total variation, in percentage terms. All values are averages over months. Weighted mean is weighted by number of originations. NS stands for "no subprime", and NS means calculate the mean omitting the subprime lenders. Missing values of within correspond to lenders who only offer one card. Lenders' identities are scrambled for confidentiality reasons and do not necessarily match the identities in other tables and figures. Shares may not add up to 100 because of rounding. Link back to descriptive findings section

TABLE 1.C.2: Tests for equality of lenders' credit limit distributions

Test	**p-value**
Anderson-Darling Version 1	0.00
Anderson-Darling Version 2	0.00
Rank Score Version 1	0.00
Rank Score Version 2	0.00

Notes: p-values from a collection of tests for the equality of lenders' credit limit distributions. p-values are averages over months of the test statistic calculated on the month-by-month credit limit distributions using a random sample of size 1 million. The Anderson-Darling version 1 (respectively 2) test statistic is A^2_{kN} (respectively $A^2_{\alpha kN}$) in Scholz and Stephens (1987). The Rank Score test statistic is QN in Lehmann (2006) and Sidak, Sen, and Hajek (1999), where versions 1 and 2 use integer scores and van der Waerden scores respectively. See Scholz and Zhu (2019) for more details. Link back to descriptive findings section

TABLE 1.C.3: Percentage of cards retaining origination interest rate by month

Month after origination	**Cards not repriced (%)**
6	98.00
9	95.98
12	95.77
15	95.27
18	91.01
21	90.11
24	88.75
27	87.81
30	86.67

Notes: I calculate the proportion of cards that have the same interest rate as they received at origination, for $t = 6, 9, 12, \ldots, 30$ months after origination. Link back to descriptive findings section

TABLE 1.C.4: Summary statistics on credit card originators

Variable	Mean	SD	10%	25%	50%	75%	90%
Age	42.88	14.83	25.00	31.00	41.00	53.00	64.00
Income (£)	2099.26	5185.72	630.00	1058.56	1604.14	2335.00	3393.00
Existing Customer	0.40	0.49					
Female	0.52	0.50					
Homeowner	0.57	0.50					
Employment							
Employed at Company	0.76	0.43					
Self-Employed	0.09	0.29					
Unemployed	0.01	0.10					
Retired	0.12	0.33					
Student	0.01	0.12					
Channel							
Branch	0.32	0.46					
Online	0.53	0.50					
Post	0.12	0.32					
Telephone	0.04	0.20					

Notes: Income is monthly income net of tax. Homeownership is equal to one if the individual owns a house (with a mortgage or without) at origination. Categorical variables' means may not add to 1 because of rounding. Link back to summary statistics description

TABLE 1.C.5: Summary statistics of card features at origination

Variable	Mean	SD	10%	25%	50%	75%	90%
Credit Limit (£)	3390.33	3144.37	300.00	1000.00	2500.00	5000.00	7700.00
Purchase APR (%)	21.52	7.64	15.76	16.90	18.90	23.95	31.11
BT APR (%)	20.24	5.28	15.90	17.50	18.90	20.90	30.33
Purch Promo Length	3.57	4.71	0.00	0.00	3.00	6.00	13.00
BT Promo Length	9.21	8.71	0.00	0.00	9.00	15.00	21.00
Balance Transfer	0.28	0.45					
Get Ad APR	0.83	0.37					

Notes: Unit of observation is the credit card origination (i). "Balance Transfer" is equal to one if the originator transferred a balance from another card onto this newly originated card at origination. Promotional lengths are in months. Purchase (respectively BT) promo are equal to one if the originated card had a purchase (respectively balance transfer) promotional period. "Get Advertised APR" is a dummy equal to one if the individual obtains the APR advertised in the promotional materials. Link back to summary statistics description

TABLE 1.C.6: Summary statistics on credit card statements

Variable	**Mean**	**SD**	**10%**	**25%**	**50%**	**75%**	**90%**
Credit Limit (£)	4213.90	3459.56	500.00	1600.00	3500.00	5900.00	9000.00
Purchase APR (%)	16.46	8.10	0.00	15.70	17.50	18.94	29.90
Account Balance (£)	1224.25	1956.57	0.00	0.00	395.12	1593.46	3669.04
Purchase Balance (£)	611.67	1255.25	0.00	0.00	75.95	660.18	1820.31
Value Transactions (£)	311.19	802.62	0.00	0.00	0.00	259.85	880.38
Repayment (£)	224.69	637.35	0.00	0.00	30.02	150.00	569.40
Total Interest (£)	8.23	20.52	0.00	0.00	0.00	6.01	26.58
Purchase Interest (£)	6.39	17.60	0.00	0.00	0.00	3.30	20.51
Account Status							
Up-To-Date	0.94	0.23					
1 Month Overdue	0.02	0.14					
2 Months Overdue	0.00	0.06					
3 Months Overdue	0.00	0.05					
4 Months Overdue	0.00	0.04					
5+ Months Overdue	0.00	0.06					
Charged Off	0.02	0.15					

Notes: Unit of observation is the statement-month. Account balance includes purchase, cash advance, money transfer, and balance transfer balances. Total interest includes purchase, cash advance, money transfer, and balance transfer interest. The variables 2 Months overdue to 5+ Months Overdue are zero rounded to 2 decimal places. Categorical variables' means may not sum to 1 because of rounding. Link back to summary statistics description

TABLE 1.C.7: Summary statistics for card characteristics

Variable	Mean	SD	10%	25%	50%	75%	90%
Annual fee	10.34	37.37	0.00	0.00	0.00	0.00	24.00
Min income	6463.20	8356.91	0.00	2.08	4000.00	7500.00	20000.00
Min CL (£)	463.09	516.11	100.00	200.00	450.00	500.00	1000.00
Max CL (£)	19881.44	30651.74	1000.00	3000.00	15000.00	20000.00	30000.00
Interest free days	31.29	12.92	20.00	25.00	25.00	46.00	50.00
Eligibility							
Student Only	0.05	0.21					
Employed Only	0.07	0.26					
All	0.88	0.32					
Risk Segment							
Superprime	0.02	0.15					
Prime	0.51	0.50					
Subprime	0.21	0.40					
All	0.26	0.44					
Rewards							
Cashback	0.09	0.29					
Airmiles	0.07	0.26					
Affinity	0.25	0.43					
Credit repair	0.21	0.41					
Purch protection	0.25	0.44					
Contactless	0.48	0.50					
Insurance	0.14	0.35					
Priority	0.12	0.32					

Notes: Unit of observation is the card-month (jt). CL stands for credit limit. Reward variables are all equal to one if the card-month offers the reward. Categorical variables' means may not sum to 1 because of rounding. Link back to summary statistics description

This page is intentionally left blank.

Chapter 2

Regulating Prices in the UK Credit Card Market

2.1 Introduction and Summary

In this chapter, I construct and estimate a structural equilibrium model of the UK credit card market. The primary novelty of my modeling arises through the supply side. Lenders choose credit limits for each customer after they apply for a credit card. I endow each lender with a screening technology that generates the lender a noisy signal on each customer's private type, which represents their risk. Differences in the granularity of these signals across lenders explain the differences in the shape of lenders' credit limit distributions shown in Chapter 1. This chapter offers the first quantitative model of credit card lenders' screening technologies and credit limit choices. I am able to estimate lender-specific screening technologies from lenders' optimizing equations because I have data on typically unobserved marginal costs of lending, which are their funding costs.

On the demand side of the model, I explain borrowers' credit card choices, level of borrowing, and default decisions, allowing for observed and unobserved heterogeneity in all endogenous demand-side variables. For credit card and borrowing choices, preferences over interest rates are heterogeneous, depending on individuals' incomes. I identify demand parameters using a novel source of quasi-experimental price variation. I create an instrument that exploits the cost shock resulting from the April 2011 case in the High Court concerning the mis-selling of payment protection insurance (PPI). Credit card lenders were forced to compensate thousands of consumers

when the court deemed they had mis-sold PPI alongside credit cards.

My estimates imply the following findings. First, I find a positive correlation between unobservables driving the level of borrowing and default, implying adverse selection on the *intensive* borrowing margin. Second, my supply-side estimates indicate that substantial variation exists in lenders' screening technologies, which corresponds with the variation in lenders' credit limit distributions. Third, I find that lenders with more precise screening technologies have a lower proportion of cases in which the customer repays their entire balance. This finding is consistent with a segmentation of credit card lenders in which lenders with the most precise screening technologies serve a riskier, but more profitable, market segment on average. Lenders with more precise screening technologies are more willing to serve customers who will borrow but may default because they can more accurately set lower credit limits for customers they perceive to be riskier.

The lack of interest rate variation, combined with the non-binding regulatory APR constraint, imply either that (i) alternative costs/constraints exist for setting individualized interest rates or (ii) lenders would choose card-level interest rates even in the absence of such frictions. To investigate this further, I analyze a counterfactual scenario in which lenders have the option to use fully individualized interest rates and credit limits, subject to no costs or constraints. The distribution of interest rates moves from a small set of card-level interest rates to a more continuous, individual-level distribution, and interest rate discrimination emerges. The riskiest individuals experience large reductions in consumer surplus, and the consumer surplus of the safest individuals increases. Further, credit limits remain individualized, borrowing increases on average, and lenders' profits increase.

The counterfactual findings suggest that lenders face frictions that limit their willingness to set individualized interest rates. Although I cannot identify the exact source of these frictions, I offer three possibilities. First, lenders may face reputational costs in advertising one APR while giving customers an alternative individualized APR.[17] Individualizing interest rates in a context where interest rates are advertised is also accompanied by the risk of being perceived to discriminate on unfair grounds. Second, overhead and operational costs of tailoring prices optimally

[17]In 2003, the UK House of Commons Treasury Committee described risk-based pricing as an "unacceptable practice," raising "serious transparency issues" (House of Commons Treasury Committee, 2003).

may exist, specifically in the IT infrastructure required to operationalize individualized prices. Lenders may choose to focus their investments on tailoring credit limits if regulations limit their ability to tailor individuals' interest rates. Third, lenders may face behavioral frictions or alternative motives that prevent them from making profit-maximizing decisions. In particular, lenders may opt for parsimonious models for interest rates for practical reasons that are beyond the scope of my economic model.

The chapter proceeds as follows. My structural model follows in Section 2.2. In Section 2.3, I explain how I estimate the model parameters. Section 2.4 discusses my parameter estimates, and Section 2.5 describes the results of the counterfactual analyses. In Section 2.6, I provide potential explanations for the results of the counterfactual analyses. Section 2.7 offers brief remarks to conclude Chapters 1 and 2.

2.2 A Model of the Credit Card Market

This section details my UK credit card market model. To help navigate the model, Tables 2.D.1 and 2.D.2 provide a glossary of notation and Figure 2.1 depicts the timeline within the market.

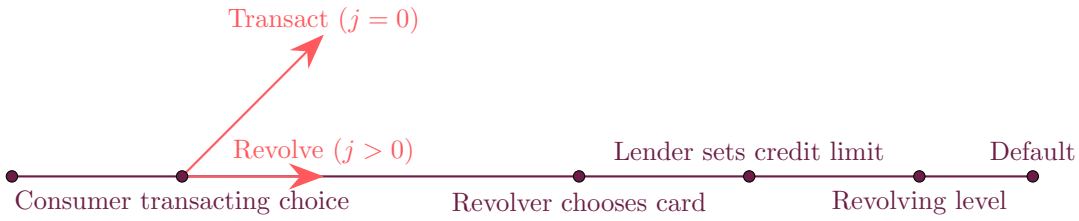
2.2.1 Preliminaries

The market is a pair (m, t). Here t represents an origination month between January 2010 and June 2013, and m represents the distribution channel, divided into originations occurring in the store and out of the store.[18] I describe the model through its three features: the credit card $j \in J_{mt}$, consumers $i \in I_{mt}$ currently without a credit card (who represent demand), and lenders $\ell \in L_{mt}$ (who represent supply). I focus on customers currently without a credit card for two reasons. First, estimating my model on the sample currently without a credit card circumvents complications arising from (i) balance transfers and (ii) balance-matching heuristics in repayment across multiple cards (Gathergood, Mahoney, Stewart, and Weber, 2019). Second, as discussed in Section 1.4.4, most UK adults hold only one credit card.

An alternative option is to microfound my demand model in a typical consumption-

[18] I stop at June 2013 to ensure that I observe 18 months of borrowing and default data on each individual.

FIGURE 2.1: Model timeline within a market



savings setup. However, I prefer to view my demand-side estimating equations as a set of linearized equations, agnostic to most of the behavior that generates them. This is similar to the approach of Einav, Jenkins, and Levin (2012), which focuses on a set of linearized estimating equations from their standard model of consumer choice. The benefit of this approach is that the econometric model becomes a valid approximation of several underlying models of consumer choice, not just the standard model of intertemporal optimization. Though this can limit the extent of welfare analysis, it is a worthwhile concession in modeling credit card borrowing, where standard assumptions about revealed preference, rational expectations, and consumer sophistication are subject to deserved scrutiny. I discuss various departures from rational utility maximizing agents with standard intertemporal preferences in the in credit card market literature in Appendix 1.A.

2.2.2 Credit Card

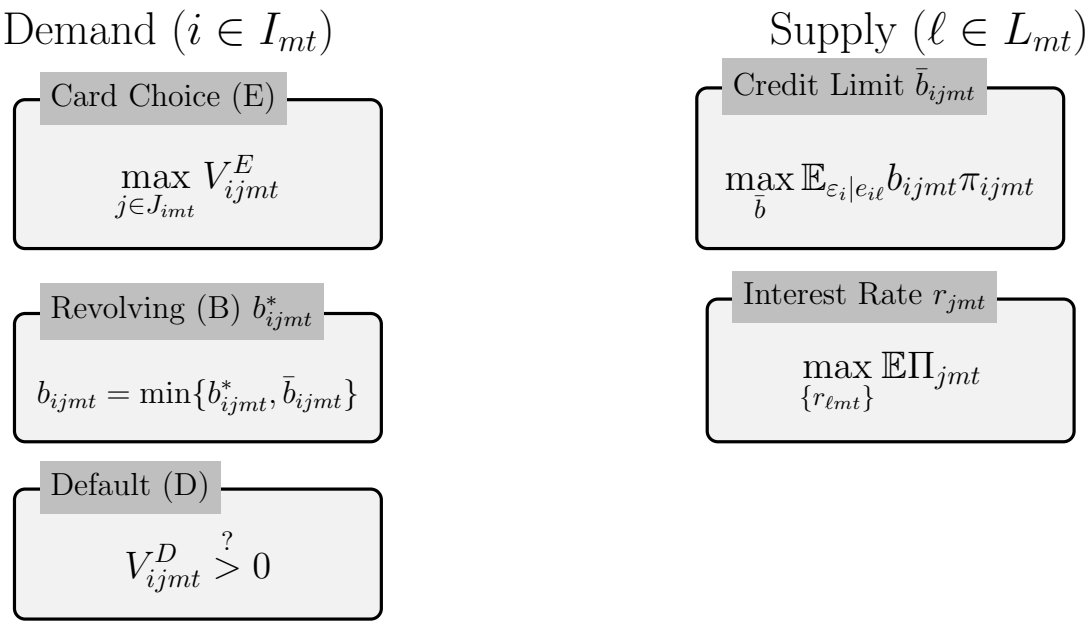
Following Lancaster (1966), I model a credit card as a bundle of features, given by $(r_{jmt}, \underline{Y}_{jmt}, X_{jmt}, \xi_{jmt})$. The first term, r_{jmt}, is the advertised interest rate. The second component, $\underline{Y}_{jmt}$, is the income threshold, explained in Section 2.2.3. The third and fourth are characteristics: those I observe, denoted as X_{jmt} (e.g. cashback and air miles), and those I do not, denoted as ξ_{jmt} (e.g. prestige and brand loyalty).

2.2.3 Consumer

My demand model follows those in the credit market literature, sharing features with Crawford, Pavanini, and Schivardi (2018).[19] The left side of Figure 2.2 depicts the demand-side building blocks. Consumers potentially make three decisions (card choice, borrowing, and default), each of which I detail in turn.

[19] Grodzicki, Alexandrov, Bedre-Defoile, and Koulayev (2022) provides a more general setup of credit card demand.

FIGURE 2.2: Model overview



Card Choice

In the first nest of the model, consumers choose whether they will be transactors or revolvers. Transactors, denoted $j = 0$, either do not use the card, or do use the card but do not use the card's borrowing facility, paying off their balance in full every month. Revolvers leave some of the card balance unpaid, thereby accruing interest.[20] The **revolving** consumer's utility from obtaining card j is

$$V^E_{ijmt} = \bar{V}^E(X^E_{jmt}, \xi^E_{jmt}, r_{jmt}, \eta^E_{mt}, y_i; \theta^E_{mt}) + \nu_{ijmt}.$$

Throughout the model, superscript E represents the ***E***xtensive margin. The term X^E_{jmt} represents the elements of observed card characteristics X_{jmt} that affect card choice; the same convention also applies to ξ. The term ν_{ijmt} represents a random taste shock . I model ν_{ijmt} as *generalized* type-1 extreme value distributed taste

[20]That consumers choose whether they will use the card for revolving or transacting is one of the few substantive behavioral assumptions I require. Though not all consumers commit to transacting or revolving, consumers' use of direct debits (automatic transfers) suggests that many consumers have decided how they intend to use their credit card at origination. In the first three months of originating the card, 28% have set up a direct debit, rising to 34% by six months. Of those who set up a direct debit at origination, around 40% set up a direct debit to automatically pay off their entire balance each month, suggesting they intend to be a transactor. Of the remaining 60% who set up a direct debit for an amount less than the full balance, 77% set up a direct debit to pay the *minimum repayment*, which is usually the maximum of (i) 1-2.5% of the balance, and (ii) £5 (around $6).

shocks. These random taste shocks are independent and identically distributed (iid) across customers and correlated across choices for each consumer. The final components of revolvers' credit card utility are η^E_{mt}, which is a card-utility market fixed effect, y_i, which denotes logged income, and θ^E_{mt}, which denotes market specific parameters that govern the indirect utility function.

To justify my choice concerning the components of $\bar{V}^E$, I draw on the results of a question from the cardholder survey in the FCA Credit Card Market Study. Figure 2.D.1 from the FCA CCMS presents the results to the question "Which of the following applied when you took out your credit card?" The most common response is rewards, provided by 33% of respondents. For this reason, I include X^E_{jmt} in $\bar{V}^E$. Twelve percent of customers mention the card's interest rate, hence I include r_{jmt} in $\bar{V}^E$. Since I focus on individuals currently *without* a credit card, who by definition will not be making a balance transfer, I omit preferences over balance transfer characteristics. Finally, other non-price, non-reward, and non-promotional deal responses comprise some of the remaining survey responses, implying the importance of ξ^E_{jmt}. Such responses include "use abroad" (15%), "low fees" (4%), and "good deal offered" (13%), all of which are examples of unobserved characteristics contained in ξ^E_{jmt}. Finally, there is little to no mention of credit limits, which I omit from $\bar{V}^E$ directly. However, through ξ^E_{jmt}, I do allow for individuals to prefer certain cards because they are aware that these cards have higher average credit limits.

I follow the literature (Berry, Levinsohn, and Pakes (1995) and Nevo (2001) among numerous others) and linearize $\bar{V}^E$ to imply

$$V^E_{ijmt} = \beta^{E'} X^E_{jmt} + \xi^E_{jmt} + \nu_{ijmt} + \alpha^E_{imt} r_{jmt} + \eta^E_{mt}. \tag{2.1}$$

The random coefficient α^E_{imt} represents individual-specific preferences over interest rates. Heterogeneous preferences over interest rates read

$$\alpha^E_{imt} = \alpha^E + \Omega^{E,r}_{mt} \tilde{y}_{imt}. \tag{2.2}$$

The term $\tilde{y}_{imt} = y_i - \bar{y}_{mt}$ denotes log income recentered around the market average, where the market average is given by $\bar{y}_{mt} = I^{-1}_{mt} \sum_{i \in I_{mt}} y_i$. Logged income is centered around the market average so that α^E represents the mean interest rate sensitivity in the card choice equation.

In this version of the model, preferences over rewards, β^E, are constant across individuals. I use random coefficients on interest rates because, on the supply side, I take rewards as exogenous and model lenders' choices of interest rates. Since my counter-

factual scenarios explore how lenders would choose individualized interest rates, it is important that I allow preferences over interest rates to differ across individuals.

I generate choice sets for individuals by comparing an individual's income at origination to the card's income threshold. Individuals qualify for a card if their income Y_i exceeds the income threshold $\underline{Y}_{jmt}$. Consequently, the set of cards available to customer i is

$$J_{imt} = \{j \in J_{mt} | Y_i > \underline{Y}_{jmt}\}.$$

I discuss the rationale for lenders' use of income thresholds in Section 2.2.4.

The utility from **transacting**, also linearized, is $V^E_{i0mt} = \delta_{0mt} + \nu_{i0mt} + \Omega^{E,cons}_{mt}\tilde{y}_{imt}$, where δ_{0mt} is a market-level constant of transacting utility. If the individual chooses to borrow, they choose the card j^* in their choice set corresponding to the maximal value of V^E_{ijmt}. The individual chooses to transact if V^E_{i0mt} exceeds $V^E_{ij^*mt}$.

Revolving

Next, **revolvers** choose their level of borrowing. I denote the *desired* level of borrowing as b^*_{ijmt}, which represents the individual's level of borrowing in the absence of any credit limit. The word "desired" reflects that individuals may wish to revolve a larger balance than their credit limit $\bar{b}_{ijmt}$ allows. The value of b^*_{ijmt} satisfies

$$b^*_{ijmt} = b(X^B_{jmt}, \xi^B_{jmt}, r_{jmt}, \eta^B_{mt}, y_i, \varepsilon^B_{imt}; \theta^B_{mt}).$$

As in card choice utility, the log of borrowing is linear in its parameters:

$$\log(b^*_{ijmt}) = \beta^{B'} X^B_{jmt} + \xi^B_{jmt} + \alpha^B_{imt} r_{jmt} + \eta^B_{mt} + \Omega^{B,cons}_{mt}\tilde{y}_{imt} + \varepsilon^B_{imt}. \tag{2.3}$$

The terms X^B_{jmt}, ξ^B_{jmt}, α^B_{imt}, and η^B_{mt} in (2.3) have analogous definitions to those in (2.1) and (2.2), swapping E for ***B***orrowing. The random variable ε^B_{imt} reflects a revolver's unobserved demand for borrowing. Neither the lender nor I perfectly observe ε^B_{imt}. I define its distribution at the end of this subsection.

In practice, revolvers make different borrowing choices each month, such as those implied by the solution to an intertemporal consumption-savings problem. I do not model the dynamics of borrowing, since the primary aim of the model is to explain lenders' origination credit limit choices. What matters to lenders when choosing origination credit limits are consumers' overall borrowing over the immediate period that they use the card, and less so the dynamics of borrowing within that period. As such, "borrowing" can be interpreted either as the result of a borrowing choice in a two-period consumption-savings model, or as a summary statistic (such as

an average) of multiple choices of borrowing.[21] In either case, my framework does not require a model of multiple values of borrowing across periods as implied by a consumption-savings problem: Modeling a summary statistic of borrowing is a clear profitable abstraction for my context.[22]

Default

Finally, **revolvers** choose whether or not to default on their balance. The net utility from defaulting reads

$$V_{imt}^{D} = V^{D}(\eta_{mt}^{D}, y_i, \varepsilon_{imt}^{D}; \theta_{mt}^{D}),$$

where, again, all terms are analogous to those defined in (2.1) and (2.3), swapping E for ***D***efault. The individual defaults if $V_{imt}^{D} > 0$. I linearize V_{imt}^{D}, implying

$$V_{imt}^{D} = \eta_{mt}^{D} + \Omega_{mt}^{D}\tilde{y}_{imt} + \varepsilon_{imt}^{D}. \tag{2.4}$$

I follow Nelson (2022) by not including the interest rate in default utility. Nelson (2022) and Castellanos, Jiménez Hernández, Mahajan, and Seira (2018) provide empirical evidence that price has an insignificant effect on default in credit markets. Assuming price-invariance of default also follows other structural models of selection markets without moral hazard, for example Cohen and Einav (2007) and Einav, Finkelstein, and Schrimpf (2010b). These findings support research in consumer finance that suggests there are limited channels through which prices can affect default. Much of the research on default implies that short-run liquidity drives default, rather than the long-run value of a loan contract, especially for the relatively small credit lines found on credit cards (Bhutta, Dokko, and Shan, 2017; Guiso, Sapienza, and Zingales, 2013; Ganong and Noel, 2020; Indarte, 2021).

I also follow Nelson (2022) in assuming that default is not a direct function of credit limit. If credit limit does affect default, then, insofar as market fixed effects, income, and the lenders' signal on risk explain individuals' credit limits, my default model in part accounts for the effect of credit limits on default, and my estimates are lower, rather than upper, bounds.[23]

[21] When I take the model to the data, I take the average of individuals' borrowing over 18 months. Since many individuals have only a few interludes of borrowing over 18 months, an alternative choice such as the choice of borrowing at 18 months will not be representative of all 18 monthly borrowing choices made by individuals over the period.

[22] Further evidence supporting an abstraction from the dynamics of borrowing choice is the lack of ex-post repricing, as I discuss in Section 1.5.3.

[23] For example, suppose instead that $V_{imt}^{D} = \eta_{mt}^{D} + \Omega_{mt}^{D}\tilde{y}_{imt} + \Upsilon_1 \bar{b}_{ijmt} + \varepsilon_{imt}^{D}$ and $\bar{b}_{ijmt} = \Upsilon_2 \tilde{y}_{imt} +$

Private Information Structure

I decompose private characteristics $(\varepsilon^B_{imt}, \varepsilon^D_{imt})$ into a common component, $\tilde{\varepsilon}_i$, and an idiosyncratic component, $\tilde{\varepsilon}^h_i$, so that

$$\varepsilon^h_{imt} = \sigma^h_{mt}\tilde{\varepsilon}_i + \tilde{\varepsilon}^h_i$$

for $h \in \{B, D\}$. The common component simplifies the lender signal structure (following in Section 2.2.4) and generates correlation among unobserved private characteristics for each individual. The distribution of unobserved preferences varies over markets through σ^B_{mt} and σ^D_{mt}. Finally, I further simplify by setting $\tilde{\varepsilon}^B_i$ to zero and letting $(\tilde{\varepsilon}_i, \tilde{\varepsilon}^D_i)$ be independently standard normally distributed. Henceforth, I declutter the notation, writing ε_i instead of $\tilde{\varepsilon}_i$.

2.2.4 Lender

My model of supply, specifically lenders' screening technologies and the credit limit optimization problem, comprises the central novelty of my model, though it shares a few similarities with the model of credit limit categories sketched in Agarwal, Chomsisengphet, Mahoney, and Stroebel (2017) and the model in Livshits, Mac Gee, and Tertilt (2016). The right side of Figure 2.2 depicts the supply-side building blocks. Lenders observe individuals' incomes Y_i and take X_{jmt}, ξ_{jmt}, and $\underline{Y}_{jmt}$ as given. I take lenders' choices of card characteristics as given for three reasons. First, in the data, lenders do not individualize rewards and rewards are sticky, rarely changing over the entire five-year period on which I have data. Second, many unobserved characteristics, such as brand prestige and loyalty, cannot be adjusted by a lender in a given month. Third, full-contract pricing introduces issues in equilibrium existence and uniqueness from which it is profitable to abstract, where justified.

The sorting of individuals onto cards based on their income occurs through income thresholds. Lenders use income thresholds because UK lenders must be able to inform consumers of the information used to reject them if they source data from a CRA (Department for Business Innovation and Skills, 2010). Consequently, lenders base decisions on *eligibility* at least in part on income.

To match the institutional environment and my empirical findings in Section 1.5,

$\varepsilon^{\bar{b}}_{imt}$. We expect that $\Upsilon_1, \Upsilon_2 \geq 0$. In this case, my specification will estimate $\Omega^D_{mt} + \Upsilon_1\Upsilon_2$. Therefore, if my estimates are negative, the true value of Ω^D_{mt} will be negative and at most the value of the estimate.

lenders choose credit limits for individuals non-competitively *after they have originated a card*. The regulatory environment requires that, at the beginning of each month, lenders set advertised APRs r_{jmt} at the card-month-market level. This institutional feature usefully circumvents issues of equilibrium existence and uniqueness that are pervasive in the empirical literature on contract pricing in credit markets. I estimate the supply side entirely from lenders' credit limit choices and therefore do not need to take a stance on how lenders set interest rates in the baseline. This avoids the need to model how lenders optimize interest rates around the fiddly regulatory requirements of (i) an advertised APR and (ii) a minimum of 51% of customers obtaining the advertised APR or lower.[24] By not requiring a model of how lenders set interest rates, I also avoid making a specific assumption about the nature of conduct in setting interest rates.

Before presenting the lenders' optimization problem in detail, I describe the main exogenous characteristic of the lender—their screening technology.

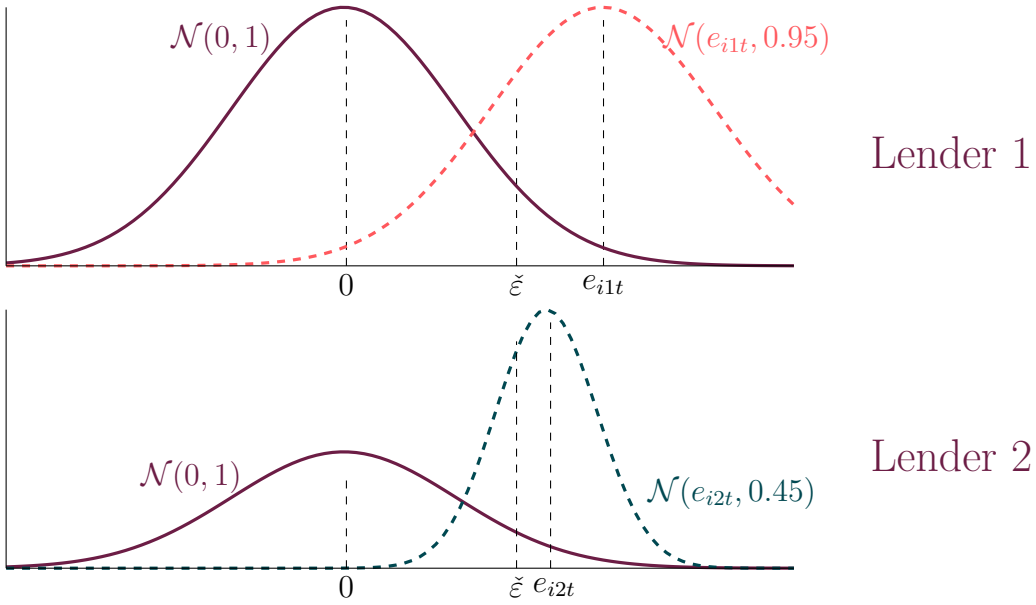
Screening Technology

Each lender has their own screening technology. The screening technology imports data on a customer that is available to the lender and provides the lender with a tailored prediction of possible values for the customer's common risk component ε_i. Without a screening technology, the lender would take expectation over a standard normal for each customer, which is the population distribution of ε_i. The screening technology aims to provide a distribution with a mean closer to each individual's realization of ε_i and a variance of less than one. That is, the screening technology aims to lower the bias and variance in the lenders' estimation of customers' risk.

The lender-specific, tailored distributions that the screening technology delivers are characterized by two features. The first is the set of signals or central points around which the tailored distributions are based. I denote these as $e_{i\ell t}$, which take a finite number of lender-specific values $\{e_{\ell t1}, \ldots, e_{\ell tL_{\ell t}}\}$. The second feature is the precision of the distribution the technology generates. The distribution generated by the screening technology accounts for potential error in the signal. I assume that, for an individual who generated the signal $e_{i\ell t}$, the distribution provided by the screening technology is normal with mean $e_{i\ell t}$ and variance $\sigma^2_{\ell t} \leq 1$, and I call $\sigma_{\ell t}$ the precision parameter. Given the value of $e_{i\ell t}$, the screening technology approximates the

[24] Appendix 2.A.2 presents a yearning reader with one model—the standard Nash-Bertrand pricing model—of how lenders may set advertised APRs competitively.

FIGURE 2.3: Distribution of actual and predicted risk scores at two lenders



Notes: These figures show the distribution of ε (solid) and $\hat{\varepsilon}_i$ (dashed) across two lenders for a customer with unknown value $\varepsilon_i = \check{\varepsilon}$. The bottom lender's screening technology, which delivers the signal e_{i2t}, outperforms the top lender's signal of e_{i1t} for this individual.

possible values of ε_i as $\hat{\varepsilon}_i = e_{i\ell t} + w_{i\ell t}$, where $w_{i\ell t} \sim \mathcal{N}(0, \sigma^2_{\ell t})$. When setting profits, the lender takes expectation using the distribution $\mathcal{N}(e_{i\ell t}, \sigma^2_{\ell t})$, as provided by the screening technology.

Figure 2.3 depicts distributions of ε_i and $\hat{\varepsilon}_i$ for two fictitious lenders. The distribution of risk provided by Lender 1's screening technology for customer i is $\mathcal{N}(e_{i1t}, 0.95)$. The mean of the conditional distribution is relatively far from customer i's true realization of $\varepsilon_i = \check{\varepsilon}$. Lender 2 has a better screening technology. The screening technology distribution given the signal e_{i2t} is much closer to $\check{\varepsilon}$. Furthermore, since σ_2 is smaller than σ_1, the signal errors at Lender 2 are less dispersed around the signal than at Lender 1. When setting credit limits for customer i, Lender 2 will place more weight (relative to Lender 1) on potential values close to $\check{\varepsilon}$ and less weight on incorrect values, such as those close to zero.

Credit Limit

Modeling lenders' credit limit choices requires an expression for their profits, that is, their costs and revenues. Regarding costs, lenders incur some fixed costs such as overheads and operational costs, but the majority of their costs vary with the number of cards they issue and how consumers use the cards they issue. I focus on

charge-off (default) costs and cost of funds, denoted as c. According to statistics from US credit card lenders, these account for over two-thirds of lenders' total costs from issuing credit cards (Evans and Schmalensee, 2005). The remaining third is comprised in most part of fixed costs, which I am free to ignore since they do not affect lenders' margins in choosing credit limits or interest rates. As such, the decision to model cost of funds and charge-off costs is a reasonable counterpart to the lenders' decisions that I observe in the data.

Regarding revenue, I focus entirely on finance charges arising from interest. For US lenders in 2001, this accounted for approximately 70% of their card revenue (Evans and Schmalensee, 2005). The remaining 30% derived from three main factors: interchange, fees, and cash-advances. Each of these three factors are likely to account for a smaller percent of UK lenders' revenue relative to the US, thus, motivating their abstraction. Appendix 2.A.1 describes each of the three factors in more detail and explains why they are less relevant in the UK credit card market than the US.

Each lender's profit from a transacting customer is Π_{i0mt}, which is unrelated to the credit limit and interest rate.[25] Therefore, the credit limit decision is unaffected by whether the individual originating card j is a transactor or a borrower. Let Δ_{imt} denote the probability that borrower i defaults and c_{jmt} denote funding rate. Then the profit per unit of credit borrowed from individual i is the interest rate minus the funding cost if the customer does not default, and $-(1-\psi)-c_{jmt}$ if they do, where ψ is the proportion of the balance that debt collectors are able to recover, which I set to zero in my empirical specification.[26]

Hence, the expected profit per unit credit for individual i on card j is

$$\pi_{ijmt} = (1-\Delta_{imt})(r_{jmt}-c_{jmt}) + \Delta_{imt}(-1-c_{jmt}).$$

[25]The revenue and costs from transactors do not relate to the interest rate, since they do not revolve a balance on which interest accrues. I assume that lenders' variable cost from non-defaulting customers is per-unit credit, and therefore lenders' costs from transactors are unrelated to the credit limit. The credit limit may affect interchange revenue, but I abstract from interchange revenue for revolvers and do so for transactors for the same reason. Resultantly, profits from transactors are not related to credit limit and interest rate choices.

[26]When cardholders default, payment card issuers start debt collection procedures. These cardholders will often have other debts, which may be collected before credit card debt. Debt collection procedures are very costly relative to the size of the loan for credit card lenders. Further, in the US in 2002, 50% of all charge-offs resulted from bankruptcy, in which case debt collection is often futile (Evans and Schmalensee, 2005). These factors considered together, $\psi = 0$ is a reasonable abstraction.

Given the signal $e_{i\ell t}$ and the implied screening technology distribution, the lender chooses the credit limit $\bar{b}_{ijmt}$ to maximize the expected profit from the individual:

$$\begin{aligned}\Pi_{ijmt} &= \max_{\bar{b}_{ijmt}} \mathbb{E}\left[\min\{b^*_{ijmt}, \bar{b}_{ijmt}\}\pi_{ijmt}\right] \\ &= \max_{\bar{b}_{ijmt}} \int \min\{b^*_{ijmt}(e_{i\ell t}, w), \bar{b}_{ijmt}\}\pi_{ijmt}(e_{i\ell t}, w) f_w(w) dw. \end{aligned} \tag{2.5}$$

As derived in Appendix A.2.3, the first order condition for credit limit is

$$\mathbb{E}\left[\pi_{ijmt} | b^*_{ijmt} \geq \bar{b}_{ijmt}\right] = \int_{\omega(\bar{b}_{ijmt})}^{\infty} \pi_{ijmt}(e_{i\ell t}, w_{i\ell t})\phi\left(\frac{w_{i\ell t}}{\sigma_{\ell t}}\right) dw_{i\ell t} = 0, \tag{2.6}$$

where

$$\omega_{i\ell t}(\bar{b}_{ijmt}, e_{i\ell t}) = \frac{\log(\bar{b}_{ijmt}) - \delta^B_{jmt} - u^B_{ijmt}}{\sigma^B_{mt}} - e_{i\ell t} \tag{2.7}$$

is the risk signal uncertainty at which the individual wants to borrow exactly their credit limit, that is, the value of $w_{i\ell t}$ which makes $\log(b^*_{ijmt})$ equal to $\log(\bar{b}_{ijmt})$. The intuition for the first order condition is that at the optimal credit limit, the expected profit per unit credit, over those with unobservables that drive them to use their full credit line, is zero. If the expected profit per unit credit on these types of individuals were positive, the lender should raise the credit limit, because the expected benefit of safer types using the full credit limit exceeds the expected costs of riskier types using the full credit limit. However, if the expected profit per unit credit over those with unobservables that drive them to use the full balance were negative, the types exploiting the full credit line would be too risky, and therefore the lender should lower their credit limit choice in this case, to render the marginal individual using their entire credit line less risky.

My descriptive findings in Section 1.5.2 on the differences in lenders' credit limit distributions motivate the tight relationship between lenders' screening technologies and the shape of the distribution of credit limits. Each unique signal implies a different choice of credit limit for the lender, and, therefore, given income, there is a mapping between the number of unique credit limits at each lender and the number of unique signals provided by their screening technology. Lenders who give observably identical consumers (to the econometrician) a wide range of credit limits must have a wide range of different signals of these consumers' unobserved risk. Conversely, lenders who give consumers who have identical on observables a coarse set of credit limits (or, in the extreme, a single value) do not appear to use a sophisticated screening technology. I use this link between credit limits and signals to estimate the distribution of signals from each of the unique values of credit limits. Consumers

who obtain the maximum credit limit for their income category obtained the lowest signal on their underlying risk ε_i and those obtaining the lowest credit limit for their income category obtained the highest risk signal on the lender's underlying risk scale.

2.3 Estimation

This section outlines how I estimate the model parameters. I start with demand estimation, since the demand estimates serve as inputs into supply estimation. My approach to demand estimation shares features with Benetton (2021), Robles-Garcia (2022) and Benetton, Gavazza, and Surico (2022). Figure 2.4 displays the four steps of the estimation procedure.

2.3.1 Demand

Log-Likelihood Conditional on Borrowing

I start with Step 1 in Figure 2.4, in which I estimate the demand parameters for those who borrow. My demand model for those who borrow consists of equations for consumer card choice (Equation 2.1), borrowing (Equation 2.3), and default (Equation 2.4). The equations map cardholders' characteristics along with lenders' interest rates, credit limits, and card characteristics onto card choice, borrowing level, and default choice. Together with stochastic assumptions on unobservables, the three equations imply a log-likelihood function for observed decisions, enabling maximum likelihood estimation. Appendix 2.B.1 provides detailed expressions for the terms of the log-likelihood. In what follows, I provide its basic structure and intuition for the main components. I focus on how the estimation approach overcomes two primary challenges and discuss the exogenous variation I exploit to identify the parameters.

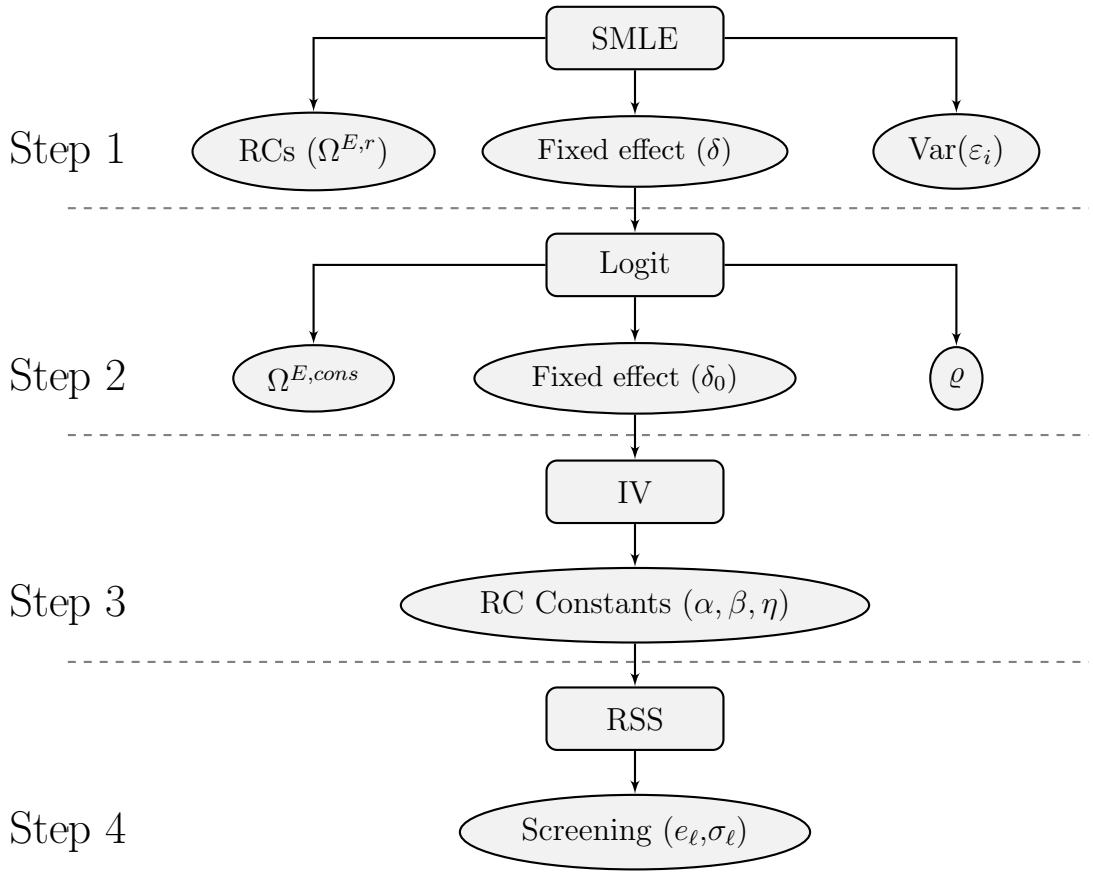
The conditional log-likelihood is the sum of a log-likelihood for card choice, $\log \mathcal{L}_{mt,E}$, and a joint log-likelihood for borrowing and default choices, $\log \mathcal{L}_{mt,BD}$, hence it is equal to

$$\log \mathcal{L}_{mt} = \log \mathcal{L}_{mt,E} + \log \mathcal{L}_{mt,BD}. \tag{2.8}$$

This feature derives from the lack of relationship between the unobservables for card choice and the unobservables driving borrowing and default. I begin by discussing the components relating to borrowing and default and then move to the components relating to card choice.

The first challenge in forming the log-likelihood components relating to borrowing

FIGURE 2.4: Four steps of model estimation



Notes: Step 1 refers to the simulated maximum likelihood estimation of the demand parameters, for those who borrow. Step 2 refers to the choice between transacting and borrowing and the maximum likelihood estimation of the parameters governing the transaction utility. Step 3 refers to instrumental variables estimation of the parameters inside of the fixed effects δ_{jmt}. Step 4 refers to supply estimation.

and default is the truncation in borrowing. Specifically, I observe the *constrained* level of borrowing $b_{ijmt} = \min\{b^*_{ijmt}, \bar{b}_{ijmt}\}$, rather than the *desired* level b^*_{ijmt}. As a result, I do not observe desired borrowing for the revolvers who borrow their entire credit limit. Revolvers either borrow their entire credit line (full utilization) or not (interior utilization), and also default or not. This creates four possible outcomes for revolver i:

1. $i \in I_1$: Interior utilization and default
2. $i \in I_2$: Interior utilization and no default
3. $i \in I_3$: Full utilization and default
4. $i \in I_4$: Full utilization and no default

Let $s_{ijmt}^{(g)}$ denote the likelihood of $i \in I_g$. Then the expression for $\log \mathcal{L}_{mt,BD}$ is

$$\log \mathcal{L}_{mt,BD} = \sum_{i \in I_{mt}} \sum_{j \in J_{imt}} \sum_{g=1}^{4} 1_{ijmt}^{(g)} \log(s_{ijmt}^{(g)}), \tag{2.9}$$

where $1_{ijmt}^{(g)}$ is a dummy equal to one if individual i chooses card j and is in group I_g. I provide expressions for $s_{ijmt}^{(g)}$ in Appendix 2.B.1.

Individuals borrowing their entire credit line (i in I_3 or I_4) create the most complications. Their contribution to the log-likelihood is an integral with no closed form and, as a result, I use simulated maximum likelihood (Pakes and Pollard, 1989; Gouriéroux and Monfort, 1993; 1996; Hajivassiliou and Ruud, 1994; Lee, 1992; 1995) with Halton (1960) draws (Bhat, 2003; Train, 2003).

The second challenge is the endogeneity of interest rates in the card choice and borrowing level equations. Interest rates r_{jmt} are chosen strategically by lenders and are likely to correlate with unobserved card characteristics ξ_{jmt}. For example, interest rates may be high on a given card because its unobserved card characteristics imply high demand for the card. In this example, if I ignore the endogeneity of interest rates, the estimation would deliver a positive value of α^E. Yet, the positive value of α^E would not imply that individuals prefer, or favor, higher interest rates. Instead, it occurs because individuals favor cards with attractive alternative features that *coincide* with high interest rates. In the first of two steps I take to overcome the endogeneity of interest rates, I estimate a full set of product-channel-month fixed effects in the card choice and borrowing equations. Formally, I rewrite Equations (2.1) and (2.3), respectively as

$$\begin{aligned} V_{ijmt}^E &= \delta_{jmt}^E + \nu_{ijmt} + u_{ijmt}^E, && (2.10) \\ \delta_{jmt}^E &= \beta^{E\prime} X_{jmt}^E + \xi_{jmt}^E + \eta_{mt}^E + \alpha^E r_{jmt}, && (2.11) \\ u_{ijmt}^E &= \Omega_{mt}^{E,r} \tilde{y}_{imt} r_{jmt}, \end{aligned}$$

and

$$\begin{aligned} \log(b_{ijmt}^*) &= \delta_{jmt}^B + \varepsilon_{imt}^B + u_{ijmt}^B, \\ \delta_{jmt}^B &= \beta^{B\prime} X_{jmt}^B + \xi_{jmt}^B + \alpha^B r_{jmt} + \eta_{mt}^B, && (2.12) \\ u_{ijmt}^B &= \Omega_{mt}^{B,cons} \tilde{y}_{imt} + \Omega_{mt}^{B,r} \tilde{y}_{imt} r_{jmt}, \end{aligned}$$

where δ_{jmt}^E and δ_{jmt}^B are the card-channel-month fixed effects. As a result of the typical identification issue in discrete choice models, I normalize $\delta_{0mt}^E = 0$ and take interest rates and card characteristics in (2.11) and (2.12) as differences from the

outside option. I provide the second step in overcoming the endogeneity of interest rates at the end of this subsection.

The term in the log-likelihood containing the card choice parameters is

$$\log \mathcal{L}_{mt,E} = \sum_{i \in I_{mt}} \sum_{j \in J_{imt}} 1^E_{ijmt} \log(s^E_{ijmt|j \in J_{imt}}), \tag{2.13}$$

where $1^E_{ijmt} = 1(j^*_{imt} = j)$ is a dummy equal to one if individual i chooses card j in their choice set J_{imt} and $s^E_{ijmt|j \in J_{imt}}$ are logit shares, derived in Appendix 2.B.1. The term $s^E_{ijmt|j \in J_{imt}}$ reflects the probability that individual i chooses card j in channel m and origination month t, *conditional* on individual i choosing to revolve a credit card balance.

To summarize, in the first step of demand estimation, I use market-by-market simulated maximum likelihood estimation on the log-likelihood for card choice, borrowing, and default, *conditional on borrowing*, to estimate scaled versions of the product-market fixed effects (δ^E_{jmt} and δ^B_{jmt}), thereby sidestepping the endogeneity problem for the moment. This step also estimates the variance-covariance matrix of private characteristics ($\varepsilon^B_{imt}, \varepsilon^D_{imt}$), specifically σ^B_{mt} and σ^D_{mt}, and the demographic coefficients ($\Omega^{E,r}_{mt}$, $\Omega^{B,r}_{mt}$, and $\Omega^{B,cons}_{mt}$).

Log-Likelihood for Borrowing and Transacting

In the second step of demand estimation (Step 2 in Figure 2.4) I maximize a log-likelihood for the choice between transacting and borrowing, which estimates δ_{0mt} and outside option utility term $\Omega^{E,cons}_{mt}$, along with the correlation coefficient for the generalized extreme value shocks, ϱ_{mt}. The identification of $\Omega^{E,cons}_{mt}$ derives from differences in incomes between those who transact and those who borrow. I provide more detail and an expression for the log-likelihood of borrowing/transacting in Appendix 2.B.2.

Constant Demand Parameters

In the third and final step of demand estimation (Step 3 in Figure 2.4), I estimate the constant parameters of the card choice and borrowing equations by projecting the estimates of card-channel-month fixed effects ($\delta^E_{jmt}, \delta^B_{jmt}$) onto distribution-channel-month fixed effects, interest rates, and observed characteristics, as in (2.11) and (2.12). The endogeneity problem still exists, hence, I use instrumental variables, the choice of which I now detail.

As an instrument for interest rates, I exploit a cost shock to UK lenders that occurred in mid-2011 relating to the mis-selling of PPI. PPI is a form of insurance designed to cover loan repayments in the event that an individual cannot make credit repayments due to adverse events such as unemployment, illness, or disability. In the late 20^{th} century, UK lenders began bundling PPI with loans and other credit products such as credit cards. In the mid-2000s, claims emerged that PPI was being mis-sold to borrowers. For example, lenders were selling PPI to self-employed individuals who would be unable to use it because of their employment status. In 2006, the Financial Services Authority began imposing fines on financial institutions for the mis-selling of PPI. An important development occurred in January 2011 when the British Bankers' Association (BBA) took the FSA to court over its decision to *retrospectively* impose standards on the correct selling of PPI.[27] The BBA were defeated in the High Court, and in May 2011, banks informed the BBA that they were withdrawing their support for an appeal of the decision. The ruling forced banks to reopen thousands of claims for PPI mis-selling. In total, around 64 million policies were mis-sold between the 1970s and late 2000s, with over £33bn repaid to individuals who complained about the sale of PPI.[28]

The loss of the court case in April 2011 and the reopening of PPI claims led to cost increases, which were spread unevenly amongst banks according to how frequently they had mis-sold PPI. Shortly after, some credit card lenders increased interest rates for all individuals at origination for some of the cards in their portfolios. From this cost shock, I create an instrument for interest rates by interacting lender fixed effects with a "post" treatment dummy.[29] The validity of the instrument requires that the only channel through which the court case ruling affects individual card choice and subsequent borrowing is through the impact of cost increases on card interest rates. I know of no other events in the same period that affected credit card lenders' unobservable card characteristics, and I can find no significant changes in observable characteristics or credit limits in the same period.

[27] See *R (on the application of the British Bankers' Association) vs Financial Services Authority and another [2011] EWHC 999.*

[28] See https://www.fca.org.uk/ppi/ppi-explained, last accessed 6 June 2023.

[29] At the time of writing, I have no data on the proportion of PPI repayments made by each lender over time. If this information were available, I could construct the instrument by constructing a measure of lenders' exposure to the court case decision.

2.3.2 Supply

The final step of estimation (Step 4 in Figure 2.4) concerns the supply parameters. The parameters to estimate in the supply model are the screening technology signals $e_{i\ell t}$ and the standard deviation of the signal noise $\sigma_{\ell t}$. I estimate these by minimizing the residual sum of squares from the first order condition of the credit limit optimization problem in (2.5). As derived in Appendix 2.A.3, for each unique observed credit limit $\bar{b}_{ijmt}$ on card j at lender ℓ in month t, the corresponding signal $e_{i\ell t}$ satisfies

$$\int_{\omega_{i\ell t}(\bar{b}_{ijmt}, e_{i\ell t})}^{\infty} \pi_{ijmt}(e_{i\ell t}, w_{i\ell t}) \phi\left(\frac{w_{i\ell t}}{\sigma_{\ell t}}\right) dw_{i\ell t} = 0. \tag{2.14}$$

Towards an estimation strategy, note that under the distributional assumptions on private characteristics,

$$\Delta_{imt} = \Phi\left(\eta_{mt}^{D} + \Omega_{mt}^{D} \tilde{y}_{imt} + \sigma_{mt}^{D}(e_{i\ell t} + w_{i\ell t})\right).$$

From this expression I can calculate Δ_{imt}—and therefore the integrand—as a function of the (already-estimated) demand parameters and the signal error.

With the demand parameters estimated, Equation (2.14) provides an equation in which, for each observed credit limit and income, the only unknowns are the screening technology $e_{i\ell t}$ and precision $\sigma_{\ell t}$. The basis of the estimation strategy is to estimate the screening technologies as the values that minimize the sum of squared deviations (over individuals) from the integral in (2.14). As in Step 1 of the demand estimation, the integral in (2.14) has no closed form. Therefore, for each lender-month, I simulate the integral using Halton (1960) draws $\omega_{i\ell t}^{h}$, and solve

$$\min_{\{e_{i\ell t}\}, \sigma_{\ell t}} \sum_{i \in I_{\ell t}} \left(\frac{1}{H} \sum_{h=1}^{H} 1\left(\sigma_{\ell t} \omega_{i\ell t}^{h} > \omega_{i\ell t}(\bar{b}_{ijmt}, e_{i\ell t})\right) \pi_{ijmt}(e_{i\ell t}, \sigma_{\ell t} \omega_{i\ell t}^{h}) \right)^2,$$

where $1(A)$ denotes the indicator function, equal to 1 if A is true and 0 otherwise. For estimation, I choose more parsimonious models that pool months within a year (thereby estimating at the lender-year level) or pool over all months (thereby estimating at the lender level).

2.4 Model Estimates

This section discusses the parameter estimates. I begin with demand parameters and then move to my estimates of lenders' screening technologies.

TABLE 2.1: First and second step demand estimates

Variable	Mean	SD
η^D	-1.804	0.125
Ω^D	-0.092	0.088
σ^D	0.532	0.100
$\Omega^{B,cons}$	0.250	0.523
$\Omega^{B,r}$	-0.196	1.515
σ^B	2.909	0.213
$\text{Corr}(\varepsilon^B, \varepsilon^D)$	0.466	0.069
$\Omega^{E,r}$	-0.468	0.717
$\Omega^{E,cons}$	-0.513	2.079
ϱ	0.328	0.182

2.4.1 Demand Estimates

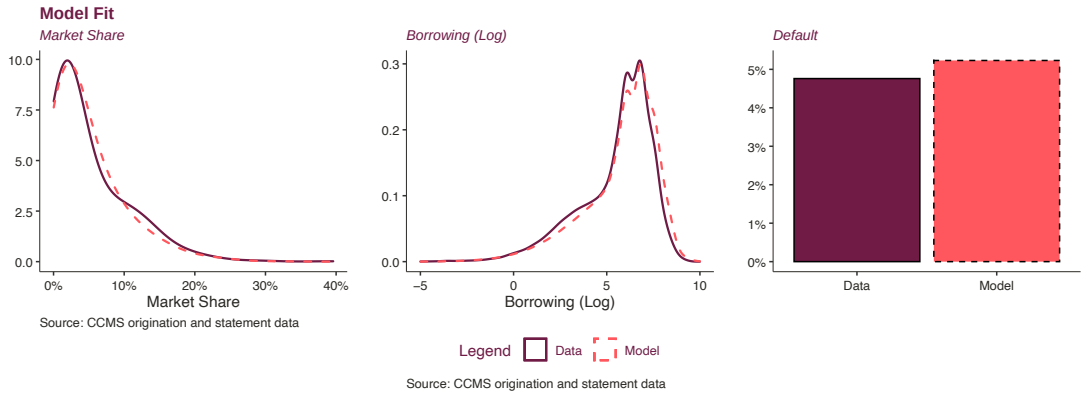
First and Second Stage Estimates

Table 2.1 presents the demand estimates from the first stage (log-likelihood of card choice, borrowing, and default) and second stage (log-likelihood for transacting/revolving) of demand estimation. I report means and standard deviations of estimates over markets.

The signs of the parameters are largely as expected, and some particular parameter estimates warrant discussion. First, I begin with the default equation parameters. The negative value for Ω^D implies that higher income revolvers are less likely to default. The mean value of 0.532 for σ^D indicates unobserved heterogeneity in default, thereby justifying the role of lenders' screening technology.

Second, moving to the borrowing equation, the estimate of 0.250 for $\Omega^{B,cons}$ means that, conditional on revolving, higher income individuals revolve more than lower income individuals. The negative value of $\Omega^{B,r}$ implies that, on average, higher income borrowers' level of revolving is more sensitive to interest rates. The correlation between unobserved preferences for borrowing and default is 0.466, implying that revolvers with a positive unobserved preference for borrowing have a positive unobserved preference for default. I refer to this as evidence of adverse selection along

FIGURE 2.5: Model fit



the intensive borrowing margin.[30] This estimate is larger than the estimate of 0.16 obtained by Crawford, Pavanini, and Schivardi (2018), whose context was the Italian market for small business loans between 1988 and 1998.

Third, I discuss my parameter estimates for the card choice equation and the utility for transacting. I estimate a negative mean value for $\Omega^{E,r}$, implying that higher-income individuals who decide to revolve are more sensitive to interest rates when they choose their card. Finally, the parameter ϱ, estimated at 0.328, indicates a reasonable substitution between transacting and borrowing choices.

Figure 2.5 displays three plots that illustrate how the demand model fits the data on card choice, borrowing, and default. They demonstrate that the fit is good, indicating that the model captures the heterogeneity of the data well.

Third Stage Estimates and Elasticities

Table 2.D.3 reports the estimates and bootstrapped standard errors of the demand parameters recovered in the third stage of demand estimation. The OLS coefficients on interest rates in both the card choice and borrowing equations are positive, whereas instrumental variable estimates are negative, indicating the severity of interest rate endogeneity. Coefficients on dummies for rewards in the card choice equation are generally positive across specifications, though the effect of cashback cannot be estimated precisely. Cashback rewards are rare in the UK and the rate of cashback tends to be low compared to the US, owing to lower interchange fees in the

[30] Lacking data on those without a credit card, I cannot at this point assess the correlation between take-up of a credit card and default, which would be the more traditional form of (extensive margin) adverse selection.

UK. Finally, Figures 2.D.2 and 2.D.3 plot the distribution of random coefficients α_i^E and α_i^B, which are negative almost everywhere and indicate substantial variations in preferences over interest rates.

Next, I turn to interest rate elasticities, where Equations (2.18) and (2.20) provide the formulas for borrowing and card choice price elasticity, respectively. Figures 2.D.4 and 2.D.5 plot the distribution of elasticities over individuals. Three noteworthy features emerge. First, revolvers display much more elasticity to the interest rate in their card choice relative to their borrowing choice. This suggests that individuals are influenced in their card choice by the interest rate, even if the interest rate will not strongly affect their choice of borrowing . Second, there is a very large degree of dispersion in both elasticities: The coefficient of variation of both card choice and borrowing elasticity is approximately one. This implies substantial heterogeneity in responsiveness to changes to interest rates across individuals. Third, both distributions are skewed. The distribution of card choice elasticities has a long left tail and the distribution of borrowing elasticities has a large mass close to zero. Finally, the elasticities are similar, though slightly larger in magnitude to other experimental estimates of interest rate elasticities in credit markets (Alan and Loranth, 2013; Karlan and Zinman, 2018). Estimates of borrowing elasticity are very similar to those in Nelson (2022).

2.4.2 Supply Estimates

My supply estimation delivers two sets of parameter estimates. The first is the variation in signal mismeasurement across lenders, denoted as σ_ℓ. For simplicity, I present estimates from the model pooling over years and consider the nine prime or superprime lenders in the data. Table 2.2 reports summary statistics in the values of σ_ℓ across lenders. The coefficient of variation is 1.699, showing that lenders' screening technologies differ substantially in their precision. While most lenders show a vast improvement in precision relative to the prior distribution (which has a precision of 1), at the 90^{th} percentile, there is less than a 30% improvement over the prior.

The second set of parameter estimates from supply estimation are the lenders' screening technology signals, denoted as e_ℓ. Figure 2.6 shows the estimated screening technologies for two contrasting lenders superimposed onto a standard normal distribution. Each vertical line represents one of the lender's possible signals. I superimpose the values onto a standard normal distribution since the signals partition the standard normally distributed signal, ε_i. The left lender's screening technology

TABLE 2.2: Summary statistics for variation in signal mismeasurement

Variable	Mean	SD	10%	25%	50%	75%	90%
σ_ℓ	0.196	0.333	0.002	0.004	0.004	0.198	0.704

contains many values, and represents a sophisticated screening technology, providing sharp signals on borrowers' types. The right lender's screening technology offers only a few values, implying less precise signals on borrowers' unobservables. Interestingly, the right lender's screening technology also contains a cluster of signals for high values of ε_i, indicating a small degree of specialization towards risky borrowers. Figure 2.D.6 shows the screening partitions for other lenders. Similar to the values of σ_ℓ, there is substantial variation in the values and the coarseness of the screening technology across lenders.

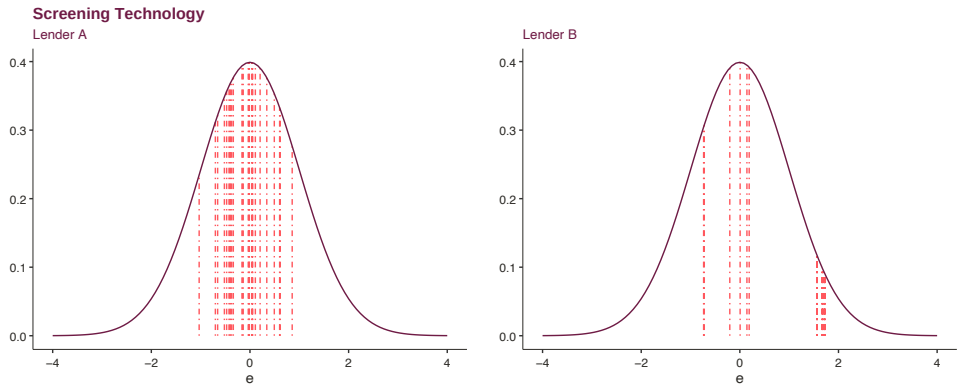
The variation in screening technologies supports the descriptive evidence in Section 1.5.2, showing that different lenders have screening technologies of varying levels of sophistication. Finally, across lenders, the correlation between σ_ℓ and the proportion of periods in which individuals repay the full balance is 0.17. This estimate is consistent with a segmentation of credit card lenders in which lenders with the most precise screening technologies serve a riskier, but more profitable, market segment on average. Lenders with more precise screening technologies are more willing to serve customers who will borrow but may default because they can more accurately set lower credit limits for customers they perceive to be riskier.

2.5 Counterfactual Analysis

2.5.1 Individualizing Interest Rates

The central empirical finding I present and analyze in Chapters 1 and 2 is that lenders individualize credit limits, with minimal within-card variation in interest rates. Related to this empirical fact is the regulatory environment, which requires lenders to set an interest rate for each credit card product they offer. Despite the requirement to *advertise* a card-level interest rate, lenders can still individualize interest rates to some extent. Under the assumption of profit maximization, my empirical findings imply that either (i) it is optimal for lenders to only individualize credit limits, or (ii) there exist costs/constraints exist that restrict lenders' willingness or ability to individualize interest rates. To clarify this, I use my estimated

FIGURE 2.6: Screening technology at two lenders



model to run counterfactual scenarios that change the environment for lenders. In my main counterfactual scenario, I allow lenders to set individualized interest rates subject to no costs or constraints in doing so, then analyze the resulting distribution of interest rates and credit limits. In this case, lenders can charge a higher price to mitigate higher costs from risky individuals, and can manage default risk by lowering borrowing from customers they perceive to be risky *either* through a higher interest rate or a lower credit limit. It is not obvious whether lenders will individualize interest rates, credit limits, or both, in equilibrium. Indeed, elementary economic theory suggests that in a perfect information, monopolistic environment where interest rates and credit limits can be used as screening instruments, credit limits are redundant.

2.5.2 Implementation

I simulate the final market of my previous analysis (June 2013 out of branch) under the new regime, with lenders setting interest rates and credit limits, but keeping income thresholds and card characteristics fixed. Then, cardholders make decisions on card choice, borrowing, and default. In the counterfactual I present, I follow the baseline model by assuming that individuals know their potential interest rate at each lender when choosing their card.[31]

For customer i, lender ℓ now solves simultaneously for all interest rates and credit

[31] I maintain the assumption that consumers do not know their *credit limits* to ensure that I am only changing one part of the environment at a time and also due to the absence of any credible source or way to measure what individuals' preferences concerning credit limits would be, were they known to the consumer.

limits across their cards $J_{i\ell}$ for which consumer i is eligible. This is because the whole vector of interest rate choices affects the probability that the individual chooses each one of the cards that they offer. Formally, given other lenders' optimal interest rate choices $\boldsymbol{r}^*_{-i\ell mt}$, for customer i, lender ℓ solves

$$\max_{\boldsymbol{r}_{i\ell},\bar{\boldsymbol{b}}_{i\ell}} \sum_{j\in J_{i\ell}} s^E_{ij}(\boldsymbol{r}_{i\ell},\boldsymbol{r}^*_{-i\ell})\mathbb{E}\left[\min\{b^*_{ij},\bar{b}_{ij}\}\pi_{ij}\right] \tag{2.15}$$

Similar to the supply estimation, I minimize the residual from the first order conditions to Equation (2.15) to calculate $\boldsymbol{r}_{i\ell}$ and $\bar{\boldsymbol{b}}_{i\ell}$ for all individuals i.[32] Appendix 2.C provides the first order conditions that I use for the calculation of the counterfactual interest rates and credit limits.

In the counterfactual, I measure changes to the distributions of several endogenous variables of interest. The first set I describe is interest rates and credit limits. Then I consider changes to consumers' levels of borrowing and consumer surplus. I calculate individuals' card choices and borrowing using indirect card utility (2.1) and borrowing Equation (2.3), respectively, replacing r_{jmt} with r_{ijmt}. I define consumer surplus as

$$CS_i = \frac{1}{\alpha_i}\log\left(\sum_{j\in J_i}\exp\left(\bar{U}^E_{ij}\right)\right),$$

where $\bar{U}^E_{ij}$ is equal to $\bar{V}^E_{ij}/\varrho$, a scaled version of indirect utility. Ex-post profit from borrower i is given by

$$\pi^{\text{post}}_{ij} = b_{ij}\Big[\mathcal{D}_i(r_j - c_j) + (1-\mathcal{D}_i)(-1-c_j)\Big],$$

where $\mathcal{D}_i$ is equal to 1 if borrower i defaults. Finally, I measure concentration using the combined market share of the largest three, four, and five lenders.

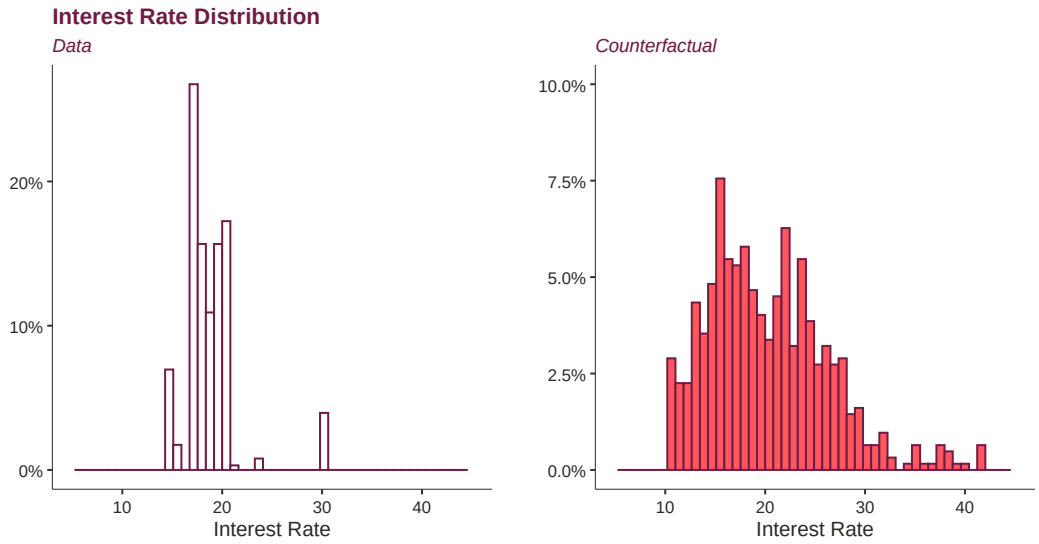
2.5.3 Counterfactual Results

Interest Rates and Credit Limits

The two variables driving all changes in the counterfactual are lenders' new choices of interest rates and credit limits. Figure 2.7 displays the distribution of interest rates in the data and separately in the counterfactual. The distribution of interest rates becomes individualized in the counterfactual, where there are over 500 unique

[32] This is a computationally intensive procedure because I have to solve the optimization problem for each consumer separately. Consequently, I use a random sample of 1000 consumers.

FIGURE 2.7: Distributions of interest rates in baseline and counterfactual

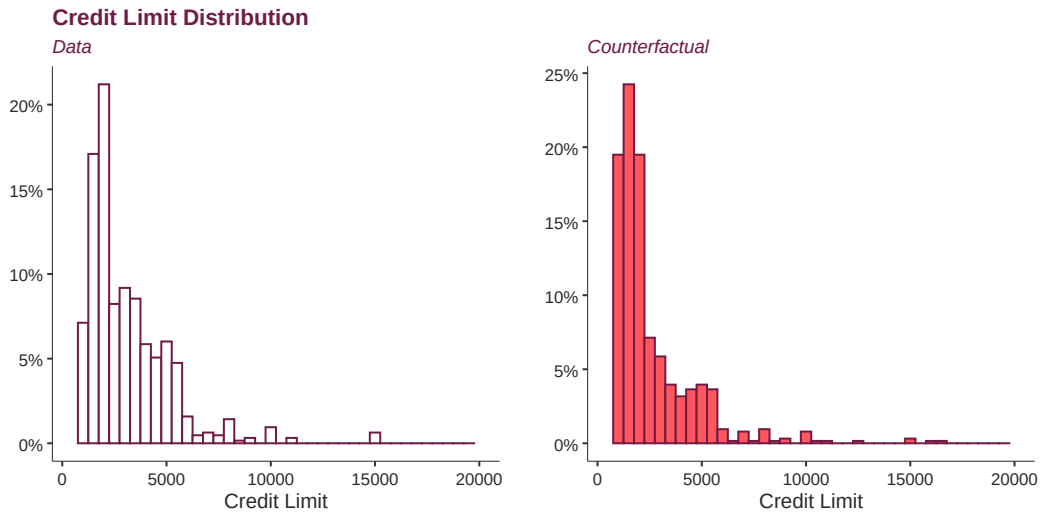


values of interest rates. Lenders with finer screening technologies use a larger set of interest rate values, as expected. This contrasts with the observed data shown in the left panel of Figure 2.7, which features 21 distinct interest rate values across 22 cards. The coefficient of variation in interest rates increases from 15.0% in the data to 32.9% in the counterfactual, and the standard deviation increases from 0.028 to 0.068. These together imply a large increase in the dispersion of interest rates.

However, the net directional effect of the counterfactual on the values of interest rates is ambiguous. Interest rates may increase because lenders can now price discriminate, but interest rates may decrease because lenders need not pool interest rates across risk types. The former dominates in the counterfactual, with interest rates increasing by 1.9 percentage points, equivalent to a 10.0% increase. This result is consistent with some success in using credit limits to manage default risk in the baseline: if lenders were unable to manage default risk using credit limits alone, we would expect them to set large pooled interest rates in the baseline, which would fall, on average, once lenders had the option to individualize them.

The net increase in interest rates in the counterfactual masks vast heterogeneity in interest rate changes across borrowers. In the counterfactual, lenders practice traditional third-degree price discrimination. Individuals with the most inelastic demand receive an average interest rate increase of 7.5 percentage points, equivalent to a 39.4% increase. In contrast, interest rates fall by 2.5 percentage points for the most elastic individuals. Further, interest rates become risk-based. I create two groups of consumers representing high-risk (income below the 25^{th} percentile and

FIGURE 2.8: Distributions of credit limit in baseline and counterfactual

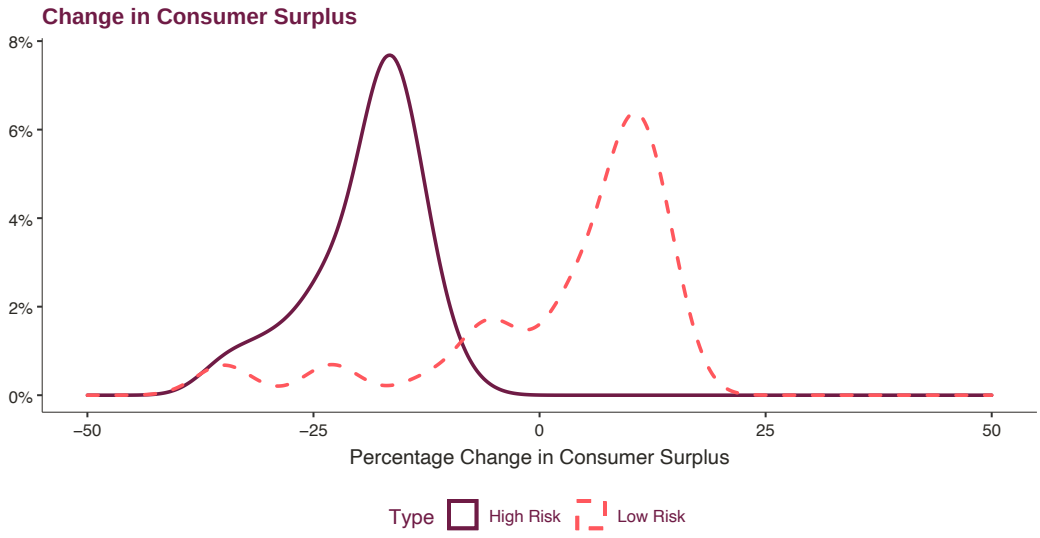


risk above the 75^{th} percentile) and low-risk (income above the 75^{th} percentile and risk below the 25^{th} percentile) borrowers. The proportion of borrowers defaulting in the high-risk group is 5.3%, compared to 2.6% in the low-risk group. Interest rates rise by 12.3 percentage points for the high-risk group and fall by 4.7 percentage points for the low-risk group.

The second screening instrument available to the lender is the credit limit. Figure 2.8 displays the distribution of credit limits in the data and the counterfactual scenario. Credit limits remain individualized in the counterfactual and become more dispersed, with the coefficient of variation in credit limits increasing by 11.8% and standard deviation increasing by 7.8%. Credit limits fall by 15.9% on average in the counterfactual. The concurrence of rising interest rates and falling credit limits follows the intuition of downward sloping cost curves in Einav, Finkelstein, and Cullen (2010a) and Einav and Finkelstein (2011). The set of individuals receiving an increase in interest rates reduce their borrowing, therefore, the set of individuals using their entire credit limit becomes riskier. To rebalance this and make the marginal profit over those using the entire credit limit zero, credit limits should fall.

The intuition for why lenders combine individualized interest rates and credit limits is that interest rates also affect an individual's choice of card through the term s_{ij}^E in the profit function for individual i, whereas credit limits do not. Individualized prices are also a device for standard third-degree price discrimination, along with their role as a tool for competition among lenders. Credit limits do not affect individuals' card choices and therefore serve as a tool for managing downside risk from default only.

FIGURE 2.9: Distributions of consumer surplus in baseline and counterfactual



This intuition explains why lenders use a combination of individualized interest rates and credit limits in the counterfactual scenario.

Demand-Side Variables

Next, I explore changes to borrowers' outcomes. In the counterfactual, borrowing increases for revolvers by 13.8% on average. The increase occurs because the most elastic borrowers obtain reductions in interest rates, which they respond to with significant increases in borrowing. This contrasts with the least elastic borrowers, who react to interest rate increases with smaller borrowing reductions. The net effect is, therefore, an increase in borrowing on average.

Relative to the baseline, consumer surplus falls by 6.6% on average in the counterfactual. It is intuitive that once lenders obtain the freedom to individualize interest rates without cost, consumers are disadvantaged on net. However, as with interest rates, this decrease in the average masks vast heterogeneity across borrowers. In Figure 2.9, I plot the distribution of percentage changes in consumer surplus for high- and low-risk individuals. Consumer surplus generally *increases* in the counterfactual for the low-risk group—a 2.6% increase on average—because they benefit from lower interest rates. Consumer surplus falls by 19% on average for the high-risk group. In sum, the counterfactual induces discrimination in interest rates and credit limits, which benefits individuals with the lowest probability of default.

Supply-Side Variables

Finally, I explore changes to lenders' outcomes. The market shares of the largest three, four, and five firms all increase by approximately eight percentage points, implying an increase in concentration in the counterfactual relative to the baseline. Further, lenders' ex-post profits increase by 25% on average. These significant changes to profit imply that the gains from tailoring interest rates alongside credit limits are substantial, in the absence of any frictions that dissuade lenders from individualizing interest rates. If lenders do not face material costs or constraints in individualizing interest rates, my findings suggest they are leaving money on the table by failing to do so. I discuss potential reasons for this finding in the section that follows.

2.6 Implications of Counterfactual Findings

The results of the previous section suggest that, in the absence of any costs or constraints involved in individualizing interest rates, profit-maximizing lenders would tailor interest rates and credit limits. However, in the data, interest rates are set at the card level and not individualized. These findings, together with the sizable increases in profits available from individualizing interest rates, imply that frictions restrict lenders' willingness to adopt individualized prices. Identifying the exact sources of these frictions is beyond the current scope of this thesis. Nevertheless, in what follows, I discuss four potential contributing factors.

First, as described in Section 1.4, UK regulations requires that at least 51% of customers originating a card must obtain the advertised APR or lower. This constraint directly impedes from fully individualizing prices. If there is a sufficiently large fixed cost in individualizing *any* interest rate, which can only be recovered if over 51% of interest rates are set above the advertised APR, it may be optimal not to individualize *any* interest rates, even if the regulatory constraint allows 49% to be tailored individually. These fixed costs may include administrative costs related to setting up the infrastructure and software to optimally set individualized prices optimally.[33] Given that restrictions on the ability to individualize interest rates already exist, lenders may have focused their investments on optimal individualized credit limits instead.

[33]Conversations with industry and policy experts suggest that lenders point to sizeable infrastructure investments as the reason why they currently do not individualize interest rates.

Second, lenders may encounter significant reputational costs if they advertise a particular APR but then provide customers with a differing, individualized APR, especially if the individualized rate is set after the individual signs the contract. In fact, members of the UK Government have already expressed their disapproval for such practices (House of Commons Treasury Committee, 2003). In April 2022, the UK Chancellor of the Exchequer stated that it was "important that advertised APRs reflect the rate the consumer is likely to receive."[34] His statement was made in response to a report on advertised APRs by the largest consumer website in the UK, MoneySavingExpert.com.[35] As part of their report, the website conducted two nationally representative surveys of over 2,000 British adults. The findings revealed that 35% of customers who were offered a higher rate than advertised stated that it had a negative effect on their financial well-being, and the same percentage claimed the higher rate had a detrimental impact on their emotional well-being.

This issue is a focal point for lenders, as they recognize that negative attention arising from unpopular business practices generates reputational risk. There is a substantial body of literature that discusses the importance of reputational risk in the banking sector (Fiordelisi, Soana, and Schwizer, 2013; Scandizzo, 2011; Xifra and Ordeix, 2009). My dataset spans the years immediately following the global financial crisis—an event that greatly impaired the public's attitude towards the banking industry (Bennett and Rita, 2012). Therefore, in the short term, avoiding further reputational damage was likely to have been a primary objective of credit card lenders at this time. Hence, though it may be challenging to quantify, the long-term reputational cost resulting from routinely deviating from the advertised interest rate may outweigh the immediate increases in profit from the lenders' perspective.

Third, the concept of lenders augmenting their profits by personalizing interest rates and credit limits is predicated on the assumption that these lenders possess the ability and tools to efficiently implement profit-maximizing individualized rates in real world situations. This is not a trivial task, as it requires a robust understanding of each customer's financial behavior and risk profile, along with the ability to accurately translate this understanding into corresponding interest rates. The theoretical framework for individualizing interest rates is intricate, necessitating a delicate balance between risk assessment and profit optimization. Consequently, any missteps

[34] https://on.ft.com/3uKGZ92 last accessed 6 June 2023.

[35] https://www.moneysavingexpert.com/news/2022/03/chancellor-ask-regulator-credit-card-loan-aprs-martin-lewis/ last accessed 6 June 2023.

in this process could result in undesirable repercussions. Such mistakes could be the result of several factors, such as inaccurate risk assessment, the dynamic nature of customer behavior, market fluctuations, or unforeseen changes in the economic climate. These errors could potentially leave lenders in a worse position than if they had adopted simpler, card-level interest rates.

Finally, a potential concern for lenders could stem from the legal implications tied to the individualization of interest rates, particularly in a landscape that requires the advertisement of APRs. Deploying machine learning and deep learning technologies in the creation of refined risk scores could inadvertently lead to reliance on protected characteristics, such as race or gender. This potential issue may place lenders at the precipice of significant legal repercussions if interest rates are individualized based on these scores. An illustrative example of this possible pitfall can be seen in the recent case of Amazon.com Inc., who had to abandon an AI recruiting tool that had inadvertently "learned" to view the male gender as a more desirable characteristic for job candidates.[36] This incident serves as a stark reminder that the implementation of "advanced technologies" with the aim of enhancing firm outcomes, can have unintended consequences.

2.7 Concluding Remarks

Chapters 1 and 2 investigate how credit card lenders in the UK manage customers' unobserved default risk by individualizing contracts through risk-based credit limits. I use novel microdata to estimate a structural model of the UK credit market. The critical innovation in the model is the lender screening technology that provides noisy signals on borrowers' unobserved types. Lenders make credit limits contingent on these signals, and the coarseness of the set of potential signals offered by the screening technology corresponds to the coarseness of their equilibrium credit limit distribution. I use the estimated model to evaluate a counterfactual scenario in which lenders can freely individualize interest rates and credit limits, which the existing regulatory environment precludes. As a result, individualized, risk-based interest rates and credit limits emerge. The induced interest rate discrimination results in consumer surplus gains for low-risk individuals and losses for high-risk individuals. Lenders' profits increase on average. My findings imply either that lenders are not

[36] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, last accessed 6 June 2023.

maximizing profits, or that the current environment imposes meaningful restrictions on lenders' willingness to adopt risk-based pricing, hence, motivating lenders to use risk-based credit limits instead.

There are also several important extensions to the content in these two chapters. For example, my model considers screening technologies as exogenous. Endogenizing screening technologies is a natural and interesting extension that may provide additional insights into lenders' interest rates and credit limit choices and their investments into financial technologies. Future work could analyze counterfactuals that change lenders' screening technologies. One example would be a scenario in which lenders share their screening technologies. This would create a setting closer to the US environment, where many lenders use FICO scores to make decisions about consumers. Furthermore, building on the empirical work of Panetta, Schivardi, and Shum (2009) I could analyze the welfare effects of mergers in which the merging lenders combine their screening technologies. Along with the typical trade-off between cost synergies and increased concentration, mergers would gain an advantage from shared and improved screening technologies. The profit increases resulting from improved screening technologies would gauge the private benefits of screening technologies. The model could also *measure* an element of the cost synergies from the merger, which is typically challenging.

There are two other avenues for extensions concerning this part of the thesis. The first is the role of consumer search and inattention. Throughout this study, I assume that consumers are fully aware of the interest rates at all lenders and are aware of all the cards for which they qualify, implying that their consideration set (Abaluck and Adams-Prassl, 2021) is equal to their choice set. The role of consumer search in this context is nuanced by lenders who currently impose heterogeneous costs on consumers to learn their interest rates and credit limits. Some lenders allow consumers to learn their contractual terms before origination. In contrast, other lenders will not divulge them until after the credit card origination. A second extension relates to behavioral biases. Consumers may have incorrect expectations or be overly optimistic about their interest rate at each lender. These biases may affect lenders' optimal use of risk-based credit limits and interest rates. These extensions warrant particular attention in work that quantifies consumer welfare in this context.

Regarding the external validity of my findings, financial products in developed economies use a variety of risk-based prices and quantities. For example, mortgages and credit cards across UK and US markets all feature different combinations

of risk-based contractual characteristics. No general theory exists to explain how product features and regulatory environments interact to influence lenders' choices among multiple screening instruments. Understanding the product characteristics and regulatory conditions that result in risk-based prices or quantities (or both) is a natural sequel to this work.

Appendices for Chapter 2

2.A Additional Modeling Details

2.A.1 Focus on Interest Revenue

I focus on interest revenue in lenders' profit functions because it comprises the vast majority of revenue for lenders. Evans and Schmalensee (2005) reports that 70% of US credit card lenders' revenue is interest revenue. The remaining 30% consists of revenue sources that are likely to be proportionally smaller in the UK relative to the US. I describe the three largest alternative revenue sources below and explain why they are likely to be smaller in the UK compared to the US.

The first is interchange revenue, which accounts for 15% of US lenders' revenues on average (Evans and Schmalensee, 2005). Interchange revenues are the funds lenders receive from merchants when individuals use their cards for purchases. When a customer makes a purchase using a credit card, the merchant pays a percentage of the transaction amount, known as the interchange fee, to the credit card company. Owing to EU regulation, interchange fees were much lower in the UK than in the US between 2010-2013, making it likely that interchange accounted for a lower proportion of UK lenders' revenue than was the case in the US.[37]

The second major portion of the remaining 30% of non-interest revenue derives from cash-advance fees. Cash-advance fees are the charges that consumers pay for using a credit card to withdraw cash or conduct other non-standard card uses such as gambling. Cash-advance revenues became a negligible part of UK lenders' revenue in April 2011, when new credit card regulations forced lenders to use customers' repayments towards high-interest cash-advance balances first rather than last, as was the practice of most lenders before the regulation.

The final main source of revenue is fee revenue. Over 75% of cards have no annual fee in the UK, hence, I focus on fees other than annual fees. In 2003, the Office of Fair Trading (OFT) began an inquiry into the "default charges" levied by credit card companies when, for example, a cardholder exceeded their credit limit or was late

[37] In 2015, the European Parliament and the Council of the European Union adopted the Interchange Fee Regulation (IFR), which set the default interchange fee cap at 0.3% of the transaction for credit cards. The UK adopted these changes in late 2015.

in making the minimum monthly payment.[38] In 2006, the OFT stated that many of the charges were "unlawful," and disclosed that it would act upon receiving notice of any fee over £12 (Office of Fair Trading, 2006). From 2010 to 2015, all fees apart from annual fees (including late, dormancy, over-limit, and foreign transaction) were at most £12, approximately 50% lower than in 2003 (House of Commons Treasury Committee, 2003). Fees are generally more common and are usually larger than £12 in the US, once more suggesting that fees accounted for a smaller proportion of UK lenders' revenues. These arguments imply that interest revenue accounts for the vast majority of UK credit card lenders' revenue, thus, justifying its use as the sole source of lenders' revenue in my model.

2.A.2 Interest Rate Model

The following subsection offers one possible model for how lenders set advertised APRs. I provide it merely to give one such example of how these rates may be set, rather than specifying that it accurately represents the method used by lenders.

In this model, lenders choose rates strategically so that interest rates form a Bertrand Nash equilibrium. Let $\boldsymbol{r}^*_{-\ell mt}$ denote the equilibrium interest rates on cards at lenders other than ℓ. Then, for every lender ℓ, the vector of interest rate $\boldsymbol{r}^*_{\ell mt}$ solves

$$\max_{\boldsymbol{r}_{\ell mt}} \sum_{i \in I_{mt}} \sum_{j \in J_{i\ell mt}} s^E_{ijmt}(\boldsymbol{r}_{\ell mt}, \boldsymbol{r}^*_{-\ell mt}) \Pi_{ijmt}(r_{jmt}). \tag{2.16}$$

The term s^E_{ijmt} denotes the probability of individual i originating card j as a borrower. The term $J_{i\ell mt} = J_{imt} \cap J_{\ell mt}$ is the set of cards offered by lender ℓ with income thresholds lower than Y_i. I define the term Π_{ijmt} in Equation (2.5).

2.A.3 First Order Condition Derivation

Now I derive Equation (2.6) from the first order condition of the lender's profit maximization problem. The first step is to replace ε_i with $e_{i\ell t} + w_{i\ell t}$. The second—and main—step is to note that for every $\bar{b}$, there exists a threshold signal error $\omega_{i\ell t}(\bar{b})$ such that if the signal error is larger (respectively smaller) than $\omega_{i\ell t}$, the individual's desired borrowing will be larger (respectively smaller) than $\bar{b}$.[39] The value of $\omega_{i\ell t}$

[38] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/284445/oft842.pdf, last accessed 5 June 2023.

[39] This version assumes that σ^B_{mt} is positive, a condition I impose in estimation without loss of generality. The sign of σ^B_{mt} is not identified so I normalize it as positive. The sign of σ^D_{mt} then

sets $\log(b^*_{ijmt})$ equal to $\log(\bar{b}_{ijmt})$ and is therefore

$$\omega_{i\ell t}(\bar{b}_{ijmt}, e_{i\ell t}) = \frac{\log(\bar{b}_{ijmt}) - \delta^B_{jmt} - u^B_{ijmt}}{\sigma^B_{mt}} - e_{i\ell t}.$$

From this, I split the objective function into

$$\int_{-\infty}^{\omega_{i\ell t}} b^*_{ijmt}\pi_{ijmt}(e_{i\ell t}, w_{i\ell t})\phi\left(\frac{w_{i\ell t}}{\sigma_{\ell t}}\right) dw_{i\ell t} + \bar{b}_{ijmt}\int_{\omega_{i\ell t}}^{\infty} \pi_{ijmt}(e_{i\ell t}, w_{i\ell t})\phi\left(\frac{w_{i\ell t}}{\sigma_{\ell t}}\right) dw_{i\ell t}.$$

By L'Hopital's rule, the first derivative with respect to $\bar{b}_{ijmt}$ is equal to

$$\int_{\omega_{i\ell t}}^{\infty} \pi_{ijmt}(e_{i\ell t}, w_{i\ell t})\phi\left(\frac{w_{i\ell t}}{\sigma_{\ell t}}\right) dw_{i\ell t} \tag{2.17}$$

and the second derivative

$$-\frac{d\omega_{i\ell t}}{d\bar{b}_{ijmt}}\pi(e_{i\ell t}, \omega_{i\ell t})\phi\left(\frac{\omega_{i\ell t}}{\sigma_{\ell t}}\right),$$

which is negative provided that $\pi(e_{i\ell t}, \omega_{i\ell t}) > 0$. In this region, the objective is concave and the first order condition is necessary and sufficient for a maximum.

2.A.4 Derivation of Elasticities

I derive formulas of the demand elasticities, both for the intensive borrowing quantity b_{ijmt} and extensive product choice s^E_{ijmt}. I start with the intensive borrowing quantity. The elasticity for individual i is

$$\frac{\partial \log(b_{ijmt})}{\partial \log(r_{jmt})} = r_{jmt}\frac{\partial \log(b_{ijmt})}{\partial r_{jmt}}.$$

The right-hand side derivative is the marginal effect from a Tobit model with top censoring at $\log(\bar{b}_{ijmt})$. The marginal effect in this model is (Greene, 2017)

$$\frac{\partial \log(b_{ijmt})}{\partial r_{jmt}} = \alpha^B_{ijmt}\Phi\left(\frac{\bar{\mathcal{Q}}^B_{ijmt}}{\sigma^B_{mt}}\right),$$

where

$$\bar{\mathcal{Q}}^B_{ijmt} = \log(\bar{b}_{ijmt}) - \delta^B_{jmt} - u^B_{ijmt}.$$

Hence the elasticity of intensive borrowing is

$$\frac{\partial \log(b_{ijmt})}{\partial \log(r_{jmt})} = r_{jmt}\alpha^B_{ijmt}\Phi\left(\frac{\bar{\mathcal{Q}}^B_{ijmt}}{\sigma^B_{mt}}\right). \tag{2.18}$$

determines the sign of the correlation between ε^B_{imt} and ε^D_{imt}. If I normalize σ^B_{mt} as negative, the first order condition bounds would swap to $(-\infty, \omega_{i\ell t}]$ but the equation is otherwise unchanged.

The elasticity for the extensive product choice is more involved. By definition, the probability that an individual chooses card j as a borrower is

$$s^E_{ijmt} = (1 - s^E_{i0mt}) s^E_{ijmt|j\in J_{imt}},$$

where $s^E_{ijmt|j\in J_{imt}}$ is the probability of individual i choosing card j, conditional on revolving, and s^E_{i0mt} is the probability that individual i chooses to transact. From this,

$$\frac{\partial s^E_{ijmt}}{\partial r_{jmt}} = (1 - s^E_{i0mt}) \frac{\partial s^E_{ijmt|j\in J_{imt}}}{\partial r_{jmt}} - s^E_{ijmt|j\in J_{imt}} \frac{\partial s^E_{i0mt}}{\partial r_{jmt}}.$$

The standard logit derivative for the inside options is

$$\frac{\partial s^E_{ijmt|j\in J_{imt}}}{\partial r_{jmt}} = s^E_{ijmt|j\in J_{imt}} (1 - s^E_{ijmt|j\in J_{imt}}) \frac{\alpha^E_{ijmt}}{\varrho_{mt}}$$

and derivative of the outside option probability is

$$\frac{\partial s^E_{i0mt}}{\partial r_{jmt}} = -\alpha^E_{imt} s^E_{ijmt|j\in J_{imt}} s^E_{i0mt} (1 - s^E_{i0mt}) = -\alpha^E_{imt} s^E_{i0mt} s^E_{ijmt}.$$

Putting these together yields

$$\frac{\partial s^E_{ijmt}}{\partial r_{jmt}} = \alpha^E_{ijmt} s^E_{ijmt} \left[\frac{1 - s^E_{ijmt|j\in J_{imt}}}{\varrho_{mt}} + s^E_{ijmt|j\in J_{imt}} s^E_{i0mt} \right]. \tag{2.19}$$

Multiplying (2.19) by $\frac{r_{jmt}}{s^E_{ijmt}}$ provides the product choice price elasticity of demand for individual i, given by

$$\frac{\partial \log(s^E_{ijmt})}{\partial \log(r_{jmt})} = r_{jmt} \alpha^E_{ijmt} \left[\frac{1 - s^E_{ijmt|j\in J_{imt}}}{\varrho_{mt}} + s^E_{ijmt|j\in J_{imt}} s^E_{i0mt} \right]. \tag{2.20}$$

2.B Additional Estimation Details

2.B.1 Conditional Log-Likelihood

Recall that the demand model (conditional on revolving) is a system of three equations: (i) a logit equation for card choice, (ii) a Tobit equation for borrowing choice (with censoring at the credit limit), and (iii) a Probit equation for default. The estimating equations for individual i, card j, in channel m, and origination month t are

$$\begin{aligned} V^E_{ijmt} &= \delta^E_{jmt} + \nu_{ijmt} + u^E_{ijmt}, \\ \log(b^*_{ijmt}) &= \delta^B_{jmt} + \varepsilon^B_{imt} + u^B_{ijmt}, \\ V^D_{imt} &= \eta^D_{mt} + \Omega^D_{mt} \tilde{y}_{imt} + \varepsilon^D_{imt}, \end{aligned}$$

where

$$\begin{aligned}
\delta^E_{jmt} &= \beta^{E\prime} X^E_{jmt} + \xi^E_{jmt} + \eta^E_{mt} + \alpha^E r_{jmt}, \\
u^E_{ijmt} &= \Omega^{E,r}_{mt} \tilde{y}_{imt} r_{jmt}, \\
\delta^B_{jmt} &= \beta^{B\prime} X^B_{jmt} + \xi^B_{jmt} + \eta^B_{mt} + \alpha^B r_{jmt}, \\
u^B_{ijmt} &= \Omega^{B,cons}_{mt} \tilde{y}_{imt} + \Omega^{B,r}_{mt} \tilde{y}_{imt} r_{jmt},
\end{aligned}$$

with all terms defined as in the main text and in the notation tables 2.D.1 and 2.D.2.[40] The system's endogenous dependent variables are borrowing utility V^E_{ijmt}, desired borrowing b^*_{ijmt}, and default net utility V^D_{imt}. Interest rates r_{jmt} correlate with unobserved card characteristics ξ_{jmt}, creating additional endogeneity along with the simultaneity. The exogenous variables are card characteristics X_{jmt} and individual logged income y_i. I never observe utilities V^E_{ijmt} and V^D_{ijmt}. I observe card choice j^*_{imt}, constrained borrowing b_{ijmt}, and default choice for revolvers. Constrained borrowing b_{ijmt} is equal to $\min\{b^*_{ijmt}, \bar{b}_{ijmt}\}$, implying that I only observe desired borrowing b^*_{ijmt} for those who borrow less than their credit limit $\bar{b}_{ijmt}$. Unobservables ε^B_{imt} and ε^D_{imt} satisfy

$$\begin{aligned}
\varepsilon^B_{imt} &= \sigma^B_{mt} \varepsilon_i, \\
\varepsilon^D_{imt} &= \sigma^D_{mt} \varepsilon_i + \tilde{\varepsilon}^D_i,
\end{aligned}$$

where $(\varepsilon_i, \tilde{\varepsilon}^D_i) \sim \mathcal{N}(0, I_2)$. I require no distributional assumption on ξ^E_{jmt} and ξ^B_{jmt}.

Expressions for $s^{(g)}_{ijmt}$

I derive the expressions $s^{(g)}_{ijmt}$ in Equation (2.9) for $g = 1, \ldots, 4$. The first term $s^{(1)}_{ijmt}$, which is for an individual who borrows $b < \bar{b}_{ijmt}$ and then defaults, is

$$\begin{aligned}
s^{(1)}_{ijmt} &= \mathbb{P}(\text{Default} | \log(b^*_{ijmt}) = \log(b)) \cdot f_{\log(b^*_{ijmt})}(\log(b)) \\
&= \frac{1}{\sigma^B_{mt}} \mathbb{P}(\varepsilon^D_{imt} > -\mathcal{Q}^D_{imt} | \varepsilon^B_{imt} = \mathcal{Q}^B_{ijmt}(b)) \phi\left(\frac{\mathcal{Q}^B_{ijmt}(b)}{\sigma^B_{mt}}\right) \\
&= \frac{1}{\sigma^B_{mt}} \Phi^{BD,1}_{ijmt} \phi\left(\frac{\mathcal{Q}^B_{ijmt}(b)}{\sigma^B_{mt}}\right),
\end{aligned}$$

[40] As described in text, because of the typical identification issue in discrete choice models, I normalize $\delta^E_{0mt} = 0$ and take interest rates and card characteristics in the card choice equation as differences from the outside option.

where

$$\begin{aligned}\Phi_{ijmt}^{BD,1} &= \Phi\left(\mathcal{Q}_{imt}^{D} + \frac{\sigma_{mt}^{D}}{\sigma_{mt}^{B}}\mathcal{Q}_{ijmt}^{B}(b)\right),\\ \mathcal{Q}_{ijmt}^{B}(b) &= \log(b) - \delta_{jmt}^{B} - u_{ijmt}^{B},\\ \mathcal{Q}_{imt}^{D} &= \eta_{mt}^{D} + \Omega_{mt}^{D}\tilde{y}_{imt}.\end{aligned}$$

By a similar derivation,

$$s_{ijmt}^{(2)} = \frac{1}{\sigma_{mt}^{B}}\left[1 - \Phi_{ijmt}^{BD,1}\right]\phi\left(\frac{\mathcal{Q}_{ijmt}^{B}(b)}{\sigma_{mt}^{B}}\right).$$

The third and fourth terms are slightly more complicated, because of the full utilization of credit limit. The third term $s_{ijmt}^{(3)}$ is

$$\begin{aligned}s_{ijmt}^{(3)} &= \mathbb{P}\left(\log(b_{ijmt}^{*}) > \log(\bar{b}_{ijmt})\right)\mathbb{P}\left(V_{imt}^{D} > 0 | \log(b_{ijmt}^{*}) > \log(\bar{b}_{ijmt})\right)\\ &= \mathbb{P}\left(\varepsilon_{imt}^{B} > \bar{\mathcal{Q}}_{ijmt}^{B}\right)\mathbb{P}(\varepsilon_{imt}^{D} > -\mathcal{Q}_{imt}^{D} | \varepsilon_{imt}^{B} > \bar{\mathcal{Q}}_{ijmt}^{B})\\ &= \mathbb{P}\left(\varepsilon_{imt}^{B} > \bar{\mathcal{Q}}_{ijmt}^{B}\right)\int_{\bar{\mathcal{Q}}_{ijmt}^{B}}^{\infty}\mathbb{P}(\varepsilon_{imt}^{D} > -\mathcal{Q}_{imt}^{D} | \varepsilon_{imt}^{B} = a) f_{\varepsilon_{imt}^{B} | \varepsilon_{imt}^{B} > \bar{\mathcal{Q}}_{ijmt}^{B}}(a | \varepsilon_{imt}^{B} > \mathcal{Q}_{ijmt}^{B}) da\\ &= \frac{1}{\sigma_{mt}^{B}}\int_{\bar{\mathcal{Q}}_{ijmt}^{B}}^{\infty}\Phi\left(\mathcal{Q}_{imt}^{D} + \frac{\sigma_{mt}^{D}}{\sigma_{mt}^{B}}a\right)\phi\left(\frac{a}{\sigma_{mt}^{B}}\right)da\\ &= \int_{\bar{\mathcal{Q}}_{ijmt}^{B}/\sigma_{mt}^{B}}^{\infty}\Phi\left(\mathcal{Q}_{imt}^{D} + \sigma_{mt}^{D}\tilde{a}\right)\phi\left(\tilde{a}\right)d\tilde{a},\end{aligned}$$

where

$$\bar{\mathcal{Q}}_{ijmt}^{B} = \mathcal{Q}_{ijmt}^{B}(\bar{b}_{ijmt}).$$

Similarly,

$$s_{ijmt}^{(4)} = \int_{\bar{\mathcal{Q}}_{ijmt}^{B}/\sigma_{mt}^{B}}^{\infty}\left[1 - \Phi\left(\mathcal{Q}_{imt}^{D} + \sigma_{mt}^{D}\tilde{a}\right)\right]\phi\left(\tilde{a}\right)d\tilde{a}.$$

Expressions for $s_{ijmt|j\in J_{imt}}^{E}$

Next, I write out the expression for $s_{ijmt|j\in J_{imt}}^{E}$ in Equation (2.13). It is

$$s_{ijmt|j\in J_{imt}}^{E} = \frac{\exp\left(\bar{U}_{ijmt}^{E}\right)}{\sum_{k\in J_{imt}}\exp\left(\bar{U}_{ikmt}^{E}\right)},$$

where

$$\bar{U}_{ijmt}^{E} = \frac{\bar{V}_{ijmt}^{E}}{\varrho_{mt}},$$

ϱ_{mt} is the parameter of the generalized type-1 distributed terms ν_{ijmt}, and the indirect utility term $\bar{V}_{ijmt}^{E}$ is

$$\bar{V}_{ijmt}^{E} = \delta_{jmt}^{E} + u_{ijmt}^{E}.$$

The first step yields estimates of the following parameters

$$\frac{\delta^E_{jmt}}{\varrho_{mt}},\ \frac{\Omega^{E,r}_{mt}}{\varrho_{mt}},\ \delta^B_{jmt},\ \Omega^{B,r}_{mt},\ \Omega^{B,cons}_{mt},\ \Omega^D_{mt}, \eta^D_{mt},\ \sigma^B_{mt},\ \sigma^D_{mt}.$$

The next subsection derives the log-likelihood of borrowing/transacting, which delivers estimates of δ_{0mt}, ϱ_{mt} and $\Omega^{E,cons}_{mt}$.

2.B.2 Log-Likelihood For Transacting

An individual transacts if the utility from transacting V^E_{i0mt} exceeds the maximal utility from borrowing. The probability that this occurs for individual i is

$$s^E_{i0mt} = \frac{1}{1 + \exp\left(\varrho_{mt} F_{imt} - \bar{V}_{i0mt}\right)},$$

where

$$F_{imt} = \log \sum_{k \in J_{imt}} \exp\left(\bar{U}^E_{ikmt}\right)$$

is the inclusive value and $\bar{V}_{i0mt} = \delta_{0mt} + \Omega^{E,cons}_{mt} \tilde{y}_{imt}$. Let ζ_{imt} be a dummy equal to one if the individual chooses to transact. Then the log-likelihood for transacting is

$$\log \mathcal{L}^{tr}_{mt} = \sum_{i \in I_{mt}} \zeta_{imt} \log(s^E_{i0mt}) + (1 - \zeta_{imt}) \log(1 - s^E_{i0mt}).$$

Maximizing $\log \mathcal{L}^{tr}_{mt}$ market-by-market provides estimates of δ_{0mt}, ϱ_{mt} and $\Omega^{E,cons}_{mt}$, from which I recover $\Omega^{E,r}_{mt}$ and δ^E_{jmt}.

2.C Additional Counterfactual Details

I derive the first order conditions to the optimization problem in Equation (2.15). First, I define

$$\mathcal{E}_{ij} = \mathbb{E}_{\varepsilon_i | e_{i\ell}}\left[\min\{b^*_{ij}, \bar{b}_{ij}\}\pi_{ij}\right]$$

and rewrite the objective function by separating out card j as

$$s^E_{ij}(\boldsymbol{r}_{i\ell}, \boldsymbol{r}^*_{-i\ell})\mathcal{E}_{ij} + \sum_{k \neq j} s^E_{ik}(\boldsymbol{r}_{i\ell}, \boldsymbol{r}^*_{-i\ell})\mathcal{E}_{ik}. \tag{2.21}$$

Since $\bar{b}_{ij}$ only affects the lender's profit for card j, the first order condition with respect to $\bar{b}_{ij}$, after cancelling $s^E_{ij}(\boldsymbol{r}_{i\ell}, \boldsymbol{r}^*_{-i\ell}) > 0$, is

$$\frac{\partial}{\partial \bar{b}_{ij}} \mathbb{E}_{\varepsilon_i | e_{i\ell}}\left[\min\{b^*_{ij}, \bar{b}_{ij}\}\pi_{ij}\right] = \frac{\partial \mathcal{E}_{ij}}{\partial \bar{b}_{ij}} = 0.$$

The equation is exactly the same first order condition for credit limits as in the baseline model. However, because interest rates change in equilibrium, even if the individual stays on the same card, their credit limit may change.

The first order condition with respect to r_{ij} is

$$\frac{\partial s_{ij}^E}{\partial r_{ij}}\mathcal{E}_{ij} + s_{ij}^E\frac{\partial \mathcal{E}_{ij}}{\partial r_{ij}} + \sum_{k\neq j}\frac{\partial s_{ik}^E}{\partial r_{ij}}\mathcal{E}_{ik} = 0.$$

Equation (2.19) provides an expression for $\dfrac{\partial s_{ij}^E}{\partial r_{ij}}$. To finish this section, I provide expressions for $\dfrac{\partial \mathcal{E}_{ij}}{\partial r_{ij}}$ and $\dfrac{\partial s_{ik}^E}{\partial r_{ij}}$ when $k \neq j$. The former of these two terms is

$$\frac{\partial \mathcal{E}_{ij}}{\partial r_{ij}} = \int_{-\infty}^{\omega_{i\ell}} \left[b_{ij}^*(1-\Delta_i) + \alpha_i^B b_{ij}^* \pi_{ij}\right] \phi\left(\frac{w_{i\ell}}{\sigma_\ell}\right) dw_{i\ell} + \bar{b}_{ij}\int_{\omega_{i\ell}}^{\infty}(1-\Delta_i)\phi\left(\frac{w_{i\ell}}{\sigma_{\ell t}}\right) dw_{i\ell}.$$

The expression for $\dfrac{\partial s_{ik}^E}{\partial r_{ij}}$ is more involved. To start,

$$\frac{\partial s_{ik}^E}{\partial r_{ij}} = (1-s_{i0}^E)\frac{\partial s_{ik|k\in J_i}^E}{\partial r_{ij}} - \frac{\partial s_{i0}^E}{\partial r_{ij}} s_{ik|k\in J_i}^E.$$

The standard logit cross-derivative yields

$$\frac{\partial s_{ik|k\in J_i}^E}{\partial r_{ij}} = -s_{ij|j\in J_i}^E s_{ik|k\in J_i}^E \frac{\alpha_i^E}{\varrho}$$

and

$$\frac{\partial s_{i0}^E}{\partial r_{ij}} = -\alpha_i^E s_{i0}^E s_{ij}^E.$$

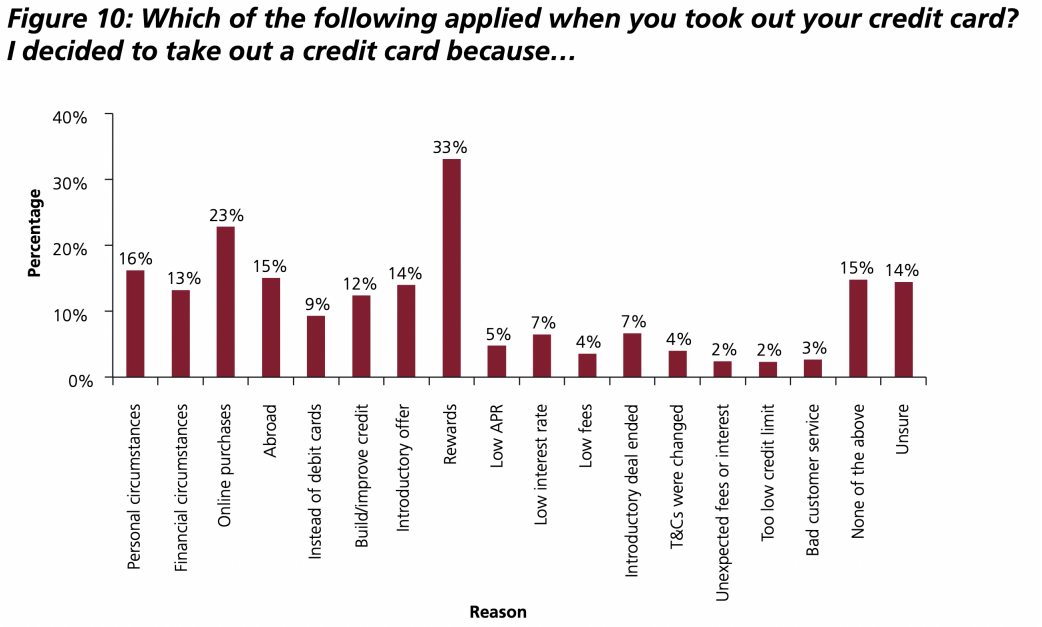
Putting these together yields

$$\frac{\partial s_{ik}^E}{\partial r_{ij}} = s_{ij}^E s_{ik|k\in J_i}^E \alpha_i^E \left[s_{i0}^E - \frac{1}{\varrho}\right].$$

2.D Additional Figures and Tables

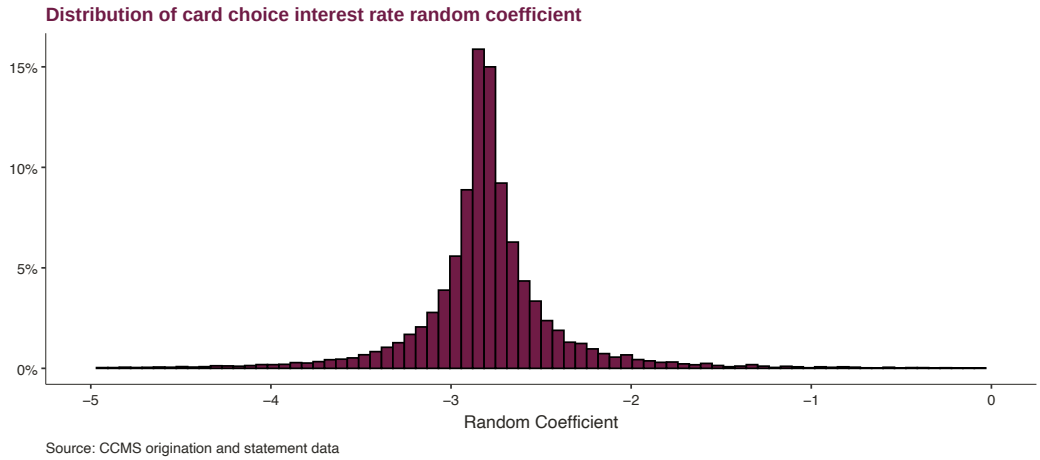
2.D.1 Figures

FIGURE 2.D.1: Reasons for taking out a credit card



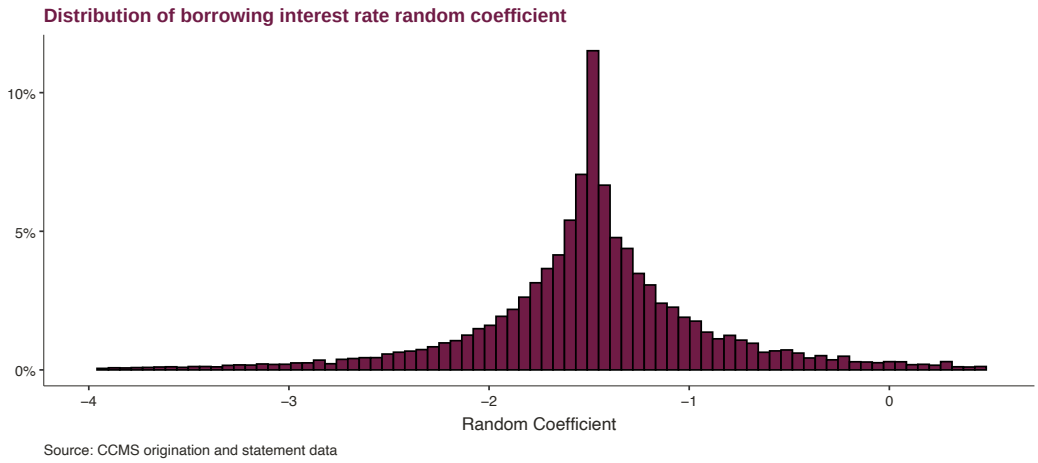
Notes: Link back to card utility discussion

FIGURE 2.D.2: Histogram of card choice interest rate random coefficient



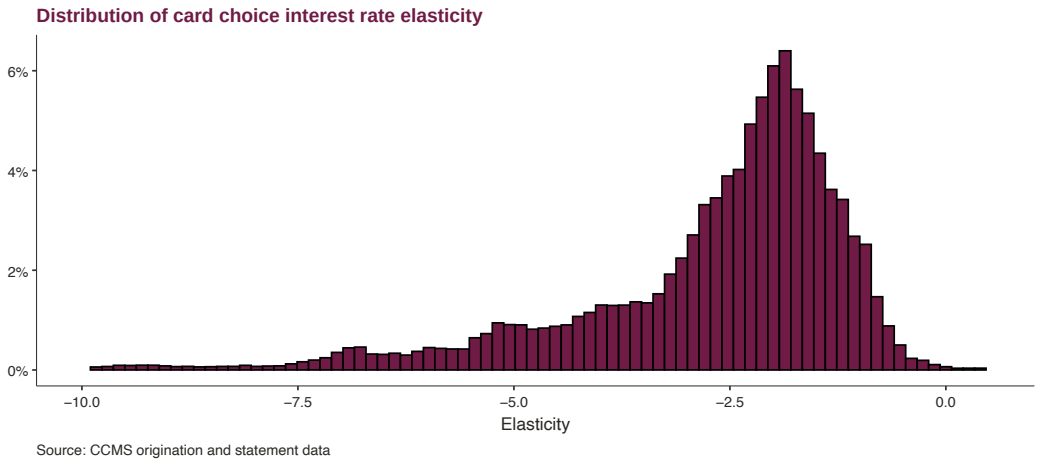
Notes: I plot the estimated distribution of α_{imt}^{E}, defined in Equation (2.2). Link back to demand estimates discussion

FIGURE 2.D.3: Histogram of borrowing interest rate random coefficient



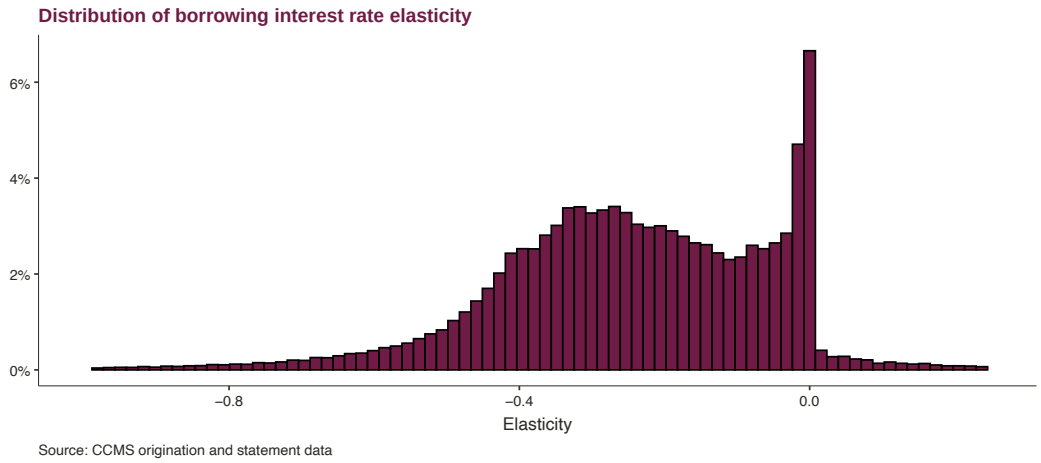
Notes: I plot the estimated distribution of α_{imt}^{B}. Link back to demand estimates discussion

FIGURE 2.D.4: Histogram of revolvers' interest rate elasticity for card choice



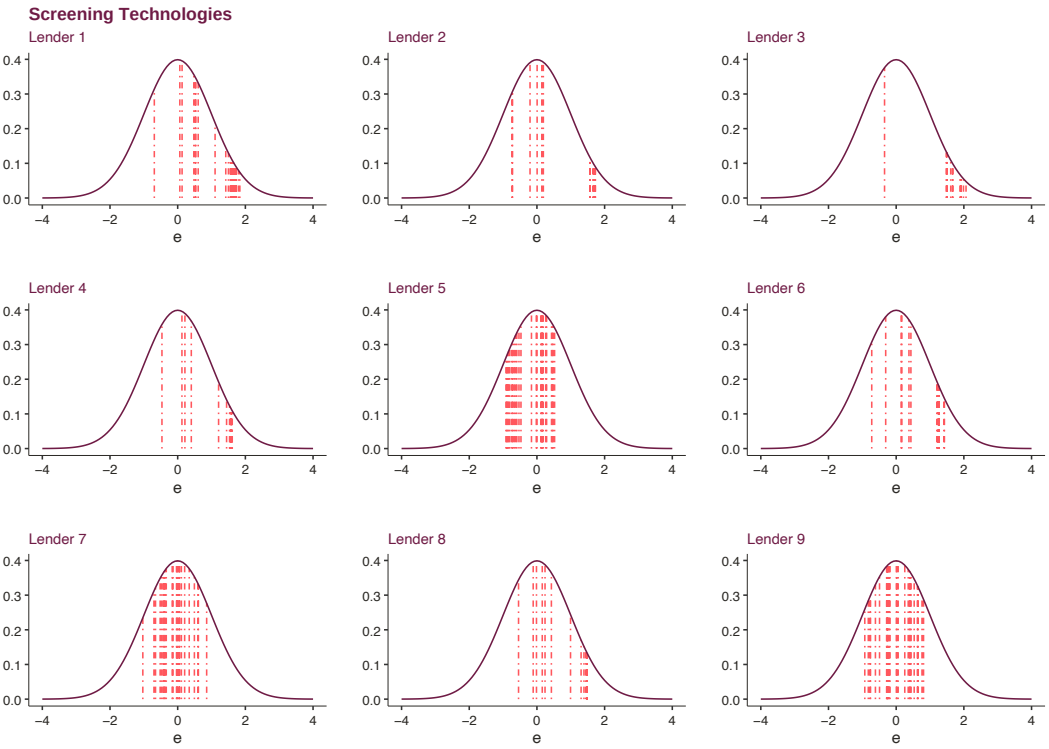
Notes: Equation (2.20) defines card choice elasticity. Link back to demand estimates discussion

FIGURE 2.D.5: Histogram of revolvers' interest rate elasticity for borrowing levels



Notes: Equation (2.18) defines borrowing elasticity. Link back to demand estimates discussion

FIGURE 2.D.6: Screening technologies at prime and superprime lenders



Notes: I scramble lenders' identities to preserve anonymity, so labels do not necessarily match the identities in other tables and figures. Link back to supply estimates

2.D.2 Tables

TABLE 2.D.1: Variable glossary: Latin

Letter	**Meaning**
b	Observed borrowing
b^*	Desired borrowing
$\bar{b}$	Credit limit
B	Borrowing symbol
c	Funding rate (marginal cost)
D	Default symbol
e	lender signal
E	Extensive margin symbol
F	Inclusive value
h	Halton draw dummy
H	Number of Halton draws
i	Credit card origination
I	Number of originations
j	Card
J	Number of cards
ℓ	Lender
L	Number of lenders
m	Distribution channel
M	Number of channels
r	Interest rate
s	Market share
t	Origination month
T	Number of origination months
u	Individual-specific terms in indirect utility
$\bar{U}$	Scaled indirect utility
$\bar{V}$	Indirect utility
V	Utility
w	Signaling error
X	Card characteristics
y	Logged income
$\tilde{y}$	Centered logged income
$\underline{Y}$	Minimum income threshold

Notes: Link back to model section

TABLE 2.D.2: Variable glossary: Greek

Letter	**Meaning**
α	Interest rate sensitivity
β	Rewards sensitivity
δ	Card-market fixed effect
Δ	Default probability
ε	Individual unobserved characteristics
ζ	Transactor dummy
η	Market fixed effect
ν	Generalized Type-1 EV shocks
ξ	Unobserved card characteristics
π	Profit per unit credit
Π	Total profit
ρ	Correlations
ϱ	ν substitution parameter
σ	Standard deviations
ϕ	Standard normal PDF
Φ	Standard normal CDF
ψ	Proportion of default debt recovered
Ω	Demographic random coefficient

Notes: Link back to model section

TABLE 2.D.3: Third step demand estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent Variable	δ^B	δ^B	δ^E	δ^E	δ^E	δ^E	δ^E	δ^E	δ^E
Price Sensitivity (α)	2.626	-1.489	1.083	-1.277	-0.934	-1.238	-3.264	-0.901	-2.825
	(0.369)	(1.71)	(0.269)	(0.804)	(0.831)	(0.793)	(0.904)	(0.815)	(0.834)
Airmiles (β_{airmiles})					0.121			0.124	0.266
					(0.048)			(0.049)	(0.042)
Cashback (β_{cashback})						0.059		0.072	-0.026
						(0.069)		(0.070)	(0.056)
Contactless ($\beta_{\text{contactless}}$)							0.178		0.270
							(0.035)		(0.075)
Estimation	OLS	IV	OLS	IV	IV	IV	IV	IV	IV
First-stage F	-	22.870	-	21.912	20.562	22.416	19.540	21.508	20.007
Wu-Hausman	-	30.120	-	13.410	4.653	9.196	32.177	4.699	22.316

Notes: This table provides the estimates and bootstrapped standard errors of the demand parameters recovered in the third stage of demand estimation. In IV specifications I use a cost shifter as excluded instrument for interest rate. I include distribution-month, and network fixed effects in all regressions. Link back to parameter estimates section

This page is intentionally left blank.

Chapter 3

Screening Property Rights for Innovation

3.1 Introduction

Public institutions play a central role in promoting innovation. The two most important channels are government support for public and private research, both in the form of direct funding and indirect fiscal subsidies, and the allocation of property rights, in the form of patents, to enhance innovation incentives for private sector R&D. To give a sense of the scale of investment, in 2015 the U.S. federal government financed 54.3% of overall R&D expenditures, or $151.5 billion (2023 U.S.D.), and 34.1% of university research. At the same time, the U.S. Patent and Trademark Office (hereafter, *Patent Office*) issued nearly 400,000 new patents. These property rights promote innovation by increasing the private returns to R&D, facilitating access to capital markets, and underpinning the market for technology, especially for small, high-technology firms (Hall and Lerner, 2010; Galasso and Schankerman, 2018). Moreover, the aggregate economic impact of these investments and property rights for innovation is magnified by the extensive knowledge spillovers they generate (Bloom, Schankerman, and Van Reenen, 2013).

Despite their evident importance, little is known about whether these innovation-supporting public institutions allocate resources efficiently and how organizational changes affect agency performance. The aim of this chapter, as part of a broader research program, is to show how structural models can be used to study and improve the efficiency of resource allocation by innovation-related public agencies. We

study this topic in the context of the U.S. patent system, focusing on the quality of screening—that is, the allocation of property rights for innovation—by the Patent Office.

We develop a dynamic structural model of the patent screening process, which incorporates incentives, intrinsic motivation, and the actual structure of multi-round negotiation in the current system. We estimate the model using novel negotiation-round-level data on examiner decisions and text data from 20 million patent claims. From the claim text data, we use modern natural language processing (NLP) methods to develop a new measure of distance between patents, a key ingredient for characterizing strategic decisions by patent applicants and examiners. We conduct counterfactual analyses of how reforms to incentives, fees, and the structure of negotiations affect the quality and speed of patent screening, and we develop an approach to quantify these impacts and thus construct a "pseudo-welfare" measure of the quality of patent screening.

The effectiveness of patent screening and its implications for the quality of patents is a hotly debated policy issue. Academic scholars and policymakers have argued that patent rights have increasingly become an impediment to innovation rather than an incentive. These concerns have been prominently voiced in public debates (The Economist, 2015; Federal Trade Commission, 2011), recent U.S. Supreme Court decisions (eBay Inc. v. MercExchange L.L.C., 547 U.S. 338, 2006), and the major statutory reform of the patent system, the Leahy-Smith America Invents Act of 2011.

Critics of the patent system claim that the problems arise in large part from ineffective patent office screening, where patents are granted to inventions that do not represent a substantial inventive step—especially in emerging technology areas such as business methods and software (Jaffe and Lerner, 2004). The issue is important because granting "excessive" patent rights imposes static and dynamic social costs: higher prices and deadweight loss on patented goods, greater enforcement (litigation) costs, and higher transaction costs of R&D and the potential for retarding cumulative innovation (Galasso and Schankerman, 2015).

The patent prosecution process is an advantageous context to study the effects of incentives and motivation on screening for two primary reasons. First, the patent application process has a clear and well-documented structure that can be modeled. The multi-round negotiation between the applicant and examiner fits naturally into a dynamic game, which forms the basis of our model. The model involves an ap-

plicant who "pads" their patent application, attempting to extract more property rights than their invention truly entails. The examiner's role is to grant or reject the application based on the existing judicial interpretation of statutory criteria as applied to each claim in the patent application.

The fundamental trade-off for the applicant when choosing the level of padding is between the benefits of increased patent scope and the costs of engaging in a lengthy and costly negotiation with the examiner. The trade-off for the patent examiner for each specific application is between the incentives to grant patents quickly and the intrinsic utility cost of awarding an inappropriate degree of "patent scope"—i.e., granting only patent claims (after narrowing) that satisfy the patentability criteria.[41] The patent examiner searches prior art to estimate the appropriate scope of patent protection for the invention, but this estimate contains error. Allowing for examiner error is important because it implies that negotiation between the applicant and examiner, while costly, may not always be socially wasteful.

The second advantage of the patent context is the quality of data. The Patent Office collects detailed and extensive data on all *applications*, not just granted patents. We constructed a dataset covering around 55 million patent application decisions across 20 million patent claims between 2010–2015 and we observe the examiner's decisions on each patent claim over all rounds of the negotiation. These data allow us to formulate and estimate a structural model that reflects the actual patent application process.

Our estimates imply several key empirical findings. First, intrinsic motivation plays a significant role in contributing to the accuracy of patent screening. Junior examiners are more motivated than seniors on average, but both groups display substantial heterogeneity. Further, using the estimated parameters, counterfactual analysis shows that turning off intrinsic motivation increases the frequency of examiners granting invalid patents four-fold. This finding highlights the importance of designing human resource policies that effectively select examiners with high intrinsic motivation and ensure examiners sustain this motivation over their entire careers.

Second, we find that innovators substantially pad their patent applications, claiming (typically) greater property rights than are warranted by the true "inventive step" of their innovation. Moreover, there is a large degree of heterogeneity in the extent

[41] For a discussion of the economics and legal doctrines of patent scope, see Merges and Nelson (1990).

of padding across patent applications. This result highlights the importance of effective screening. An essential feature of our model is that the extent of padding is endogenous and thus is affected by various counterfactual policy reforms, which we detail later. We estimate the average level of padding at about 8%, rising to 10% when we weight by the value of the patent. This exaggerated scope of the patent applications is reflected in the fact that more than 80% of claims start below the distance threshold for patentability—as measured by the minimum required distance to claims in prior patents—and thus should be rejected.

However, the multi-round screening process is relatively effective at narrowing the scope of patent rights sought and, in so doing, reducing the number of invalid claims to about 7% among granted claims, but still, nearly one in five granted patents contains at least one claim that does not meet the threshold. One implication of this finding is that the proportion of patent applications that are granted—a commonly used indicator of the effectiveness of screening—is a misleading measure because it does not capture the extent to which granted property rights are narrowed during the screening process.

We evaluate counterfactual reforms involving changes to fees for the patent applicant, the structure of the negotiation process (e.g., limiting the number of rounds allowed), and the degree of intrinsic motivation of patent examiners. We quantify the effects of counterfactual reforms along three distinct dimensions. The first two relate to the accuracy of screening, meaning the degree of alignment between the scope of property rights granted and the scope justified by the invention. We assess accuracy in terms of granting claims that are not justified (false grants, or "type 1" error) and not granting claims that should be (false rejections, or "type 2" error).

Both errors carry their own social costs and benefits. Incorrect grants impose ex post welfare costs (deadweight loss) from higher prices and litigation costs associated with enforcing these patents, but at the same time may raise innovation incentives. False rejections dilute ex ante innovation incentives and discourage the development of new inventions that would contribute positive social value, but at the same time they reduce ex post deadweight loss. The last dimension is the speed of patent examination, measured by the number of negotiation rounds in equilibrium. We develop a method to quantify these impacts in terms of the associated net social costs and thus construct a "pseudo-welfare" measure of the quality of patent screening. We estimate the total net social cost of patent screening at $25.5bn per annual cohort of applications. This figure represents 6.5% of total R&D performed by business

enterprises in the United States.

The counterfactual analysis highlights two key conclusions. First, restrictions on the number of allowable rounds of negotiation (currently absent in the U.S. patent system) significantly reduce the net social costs of screening, with a reduction of 45% in the case of allowing only one round. We show that these outcomes can be replicated through an equivalent fee per round for the applicant, but the required fees are too high to be politically feasible. Second, given the high levels of intrinsic motivation we estimate, extrinsic incentives are largely ineffective, leading to almost no change in net social costs. Extrinsic incentives do affect outcomes in a scenario with low intrinsic motivation, but they are counterproductive in that they raise the net social costs of screening.

We organize the chapter as follows. Section 3.2 briefly summarizes the related literature. Section 3.3 describes the datasets and summarizes key descriptive features. The structural model is presented in Section 3.4. Section 3.5 describes our estimation methods. Section 3.6 presents the empirical estimates. Section 3.7 analyzes the impact of counterfactual reforms on the accuracy and speed of patent screening, and Section 3.8 describes our quantification of the net social costs and benefits associated with these counterfactual reforms. Section 3.9 concludes.

3.2 Related Literature

Intrinsic Motivation in Public Agencies

We contribute to the literature that studies how intrinsic motivation affects the optimal design of incentives in mission-oriented agencies. On the theoretical side, Benabou and Tirole (2003; 2006) show conditions under which extrinsic rewards may crowd out intrinsic motivation. Particularly relevant to ourwork, Besley and Ghatak (2005) emphasize how intrinsic motivation—which they define as the alignment between worker and agency objectives—induces welfare-improving sorting of workers across entities with different goals and also affects the optimal design of incentives and authority.

Empirical studies use field experiments to analyze intrinsic motivation and public agency performance. These rely on various proxies for motivation. Leading examples include Ashraf, Bandiera, and Jack (2014), which evaluates the impact of extrinsic rewards on agents' performance in a public health organization in Zambia, and Ashraf, Bandiera, Davenport, and Lee (2020), which studies whether career benefits

induce sorting at the expense of "pro-social" motivation. Both papers find that extrinsic rewards and intrinsic motivation are complementary.

Despite their interesting findings, these empirical studies cannot be used for counterfactual policy analysis, for which structural models are more appropriate. Our project is the first structural model of a public agency that incorporates intrinsic motivation.[42] In doing this, we follow Besley and Ghatak's definition of intrinsic motivation—alignment of workers' objectives and the public agency mission. In our context, the Patent Office's mission is to award inventors property rights over their invention, consistent with statutory and judicial prescriptions. We model intrinsic motivation as an inherent disutility that examiners incur if they grant more intellectual property rights than they believe the inventor deserves, based on the information the examiners have. We show that patent examiners sometimes award patents to applications they believe are invalid due to strategic considerations and the extrinsic pay scheme they face.

Finally, recent papers study how screening mechanisms affect the performance of public agencies. Adda and Ottaviani (2023) develop a model of nonmarket allocation of resources, including but not limited to the award of grants to research projects. The authors study how the design of allocation rules and informational noise in the evaluation process affect the optimal design. In two empirical papers, Li and Agha (2015) and Li (2017) analyze the allocation of research grants at the National Institutes of Health (NIH) and show that peer review increases the effectiveness of grants in terms of post-grant citations. Azoulay, Graff Zivin, Li, and Sampat (2018) study the economic impact of these NIH grants, linking screening outcomes to publication citations and other innovation outcomes. Our contribution is to quantify some of the forces these papers identify and evaluate the equilibrium effects of various counterfactual reforms in the patent context.

Patents and Innovation

We also contribute to the limited empirical literature on patent screening. In a first paper on the topic, Cockburn, Kortum, and Stern (2003) show that patent examiner characteristics affect the "quality" of issued patents, measured by subsequent citations and the frequency of litigation. Frakes and Wasserman (2017) use data on

[42] Egan, Matvos, and Seru (2023) develop a structural model of consumer arbitration in which arbitrators differ in their idiosyncratic degrees of "slant" (or bias), which can be interpreted as a form of intrinsic motivation.

promotions of patent examiners (which are accompanied by lower extrinsic incentives) and show that promotions are associated with sharp increases in grant rates. They interpret this result as less rigorous screening and lower quality patents. While this is a striking finding, their analysis does not pin down whether it is driven by differences in extrinsic incentives, intrinsic motivation, or examiner opportunity costs, which our structural model will do.

Perhaps the most closely related paper is Schankerman and Schuett (2022), who develop an integrated framework to study patent screening, encompassing the patent application decision, examination, post-grant licensing, and litigation in the courts. They calibrate the model on data for the U.S. and use it to evaluate various counterfactual patent and court reforms. Their model estimates the effectiveness of patent examination, but they treat this as an exogenous parameter, but they do not model the prosecution process. In contrast, we develop and estimate the first equilibrium model of the patent examination process itself, which in turn allows us to investigate how various reforms to the incentives and design of patent screening affect the performance of this public agency.

Before turning to the data, we summarize a few key features of patents that guide our modeling choices. The critical feature of the patent document is the list of independent claims, which delineate the "metes and bounds," or scope, of the property right. The examination process involves assessing the patentability of *each* claim, not the patent as a single entity. In a departure from most existing literature, we treat a patent as a collection of claims that differ in both their private value and their similarity to previous patented claims. These two dimensions are a critical feature in the model. This heterogeneity is a first-order feature necessary to match the actual process of patent examination and to develop accurate statements about the potential effects of regime changes on the patent examination process.

3.3 Data and Descriptive Results

In this section, we describe our primary data sources, focusing on datasets not previously used in empirical studies of patents. We also present summary statistics and describe reduced-form evidence. Appendix 3.B provides hyperlinks to all publicly available datasets and data sources we use in our empirical work.

Distance Metric

We construct a new measure of independent claim distance. To create this, we exploit

the *U.S.PTO Patent Application Claims Full Text Dataset* and the *Granted Patent Claims Full Text Dataset*. The first dataset contains the full text for all U.S. patent application claims between 2001 and 2014 and an indicator for whether the claim is independent. The Granted Patent Claims Full Text Dataset records the full text for all U.S. patent claims granted between 1976 and 2014.

We summarize our approach to creating a distance measure here (see Appendix 3.C for more detail). The approach calculates distances by representing a patent claim's text as a numerical vector and calculating a metric on that vector space.[43] We adopt the *Paragraph Vector* approach of Le and Mikolov (2014), which uses an unsupervised algorithm to "learn" the meaning of words by studying the context in which they appear and forming a vector representation for each word, picking up the meaning of paragraphs as a by-product.[44] As is common in the NLP literature, we measure distances between numerical vectors using the angular distance metric. To reflect distance to prior art, we compute the distance from each independent claim to every previously granted independent claim.[45]

Rounds Data

Since we estimate a model of the patent prosecution process over multiple rounds, comprehensive and reliable *round-level* data on the patent process are essential. We use the *Transactions History* data in the *Patent Examination (PatEx) Research Dataset* to create a dataset on the round-by-round evolution of utility patent applications between 2007 and 2014. In total, the transactions dataset includes 275.6 million observations covering 9.2 million unique applications. For every patent application, these data record examiner and applicant decisions at each round of the examination process.

[43] Kelly, Papanikolaou, Seru, and Taddy (2021) use similar methods to calculate patent similarity.

[44] The standard method (*bag-of-words*) for representing the patent claim text as a numerical vector has two significant weaknesses: it ignores the *ordering* and *semantics* of words.

[45] We conduct two falsification tests on our distance measure. First, we put independent claims into twenty, five-percentile bins of the distance measure and then calculate the proportion of claims rejected on novelty/obviousness grounds in each bin. We would expect that examiners are more likely to reject claims with a small distance to existing claims based on novelty/obviousness criteria. Thus the proportion of first-round rejections should be a declining function of the distance metric and the results confirm this prediction. Second, we conduct a similar test on the average number of examination rounds for each granted *patent*, by five-percentile bins of average distance of independent claims. Patents with higher average distance should be granted faster, and this is what we find.

Sources Matched to Round Data

We match the round-level data to three other datasets on patent applications. The first is the *Application Data* in the *PatEx* Dataset, which contains features of the patent application, such as the identities of the applicant and examiner, the patent art unit (narrow technology classifications), and a binary indicator of the size of the applying firm (below or above 500 employees). Second, we match our data to renewal decisions by patent holders using the *U.S. PTO Maintenance Fee Events Dataset.* Third, since we focus on novelty/obviousness rejections, we require data on the *types* of rejections of each claim at each stage of the process. We obtain this from the *U.S.PTO Office Action Research Dataset for Patents.*

Legal Fees

For attorney fees, we use data from the *2017 American Intellectual Property Law Association (AIPLA) Report of the Economic Survey.* The survey reports means and percentiles of the distribution of hourly fees for different tasks, such as preparing and filing an application, issuing, paying renewal fees, and amending applications, split into three broad technology areas (biotechnology/chemical, electrical/computer, and mechanical). We use these moments to estimate the distributions of application and fighting costs for each patent application, adjusted for inflation.

Seniority and Technology Complexity Credit Adjustments

We obtained data on examiner seniority from Frakes and Wasserman (2017), who provide a panel of General Schedule (GS) grades for examiners, including each examiner's promotion dates. Using this, we work out the seniority of the examiner for each application. Finally, we received information on examiner extrinsic rewards from the Patent Office at the disaggregated U.S. Patent Classification level and then aggregated them to the technology center level.

Descriptive Statistics

Several features of the data are worth noting (Table 3.A.1 in the Appendix provides details). First, 70% of applications resulted in the issuance of a patent. However, this is a misleading measure of the fraction of *content* granted because, as we will see, most applications are heavily narrowed during the examination process. Second, the prosecution time varies across applications—the mean duration is 2.96 years, and the mean number of rounds is 2.40. Third, the mean, median, and modal number of independent claims is three. Fourth, 24% of applications were by firms with fewer than 500 employees (a so-called "small entity"). Lastly, 46% of granted patents were renewed to the statutory limit, and only 13% were not renewed at the first renewal

date.

Existing studies show that patent grant rates vary widely across technology centers and examiner seniority, with more senior examiners granting more frequently (Frakes and Wasserman, 2017; Sampat and Williams, 2019). In Appendix 3.D, we confirm these findings about grant rates using our data, and we also show that the likelihood of multi-round negotiation (lasting beyond one round) is much lower for senior examiners and varies substantially across technology centers. In addition, small entities are less likely to negotiate. We also analyze the variation in these outcomes for *each* examiner, decomposing variation in examiner-specific outcomes (such as their grant rate) within and between technology center-seniority pairs.[46] This decomposition shows that 80% of the variation in examiner grant rates and 81% of the variation in each examiner's average number of rounds is within-group variation.

Our model allows for several factors that can explain the substantial variation in examiner statistics even within seniority-technology-center dyads: we allow for a different distribution of intrinsic motivation for junior and senior examiners, we incorporate differences in the examiner credit structure across seniorities and technology centers, and we allow for heterogeneous legal (fighting) costs for applicants across technology centers. Our parameter estimates will enable us to disentangle the effects of these factors in explaining the variation in outcomes.

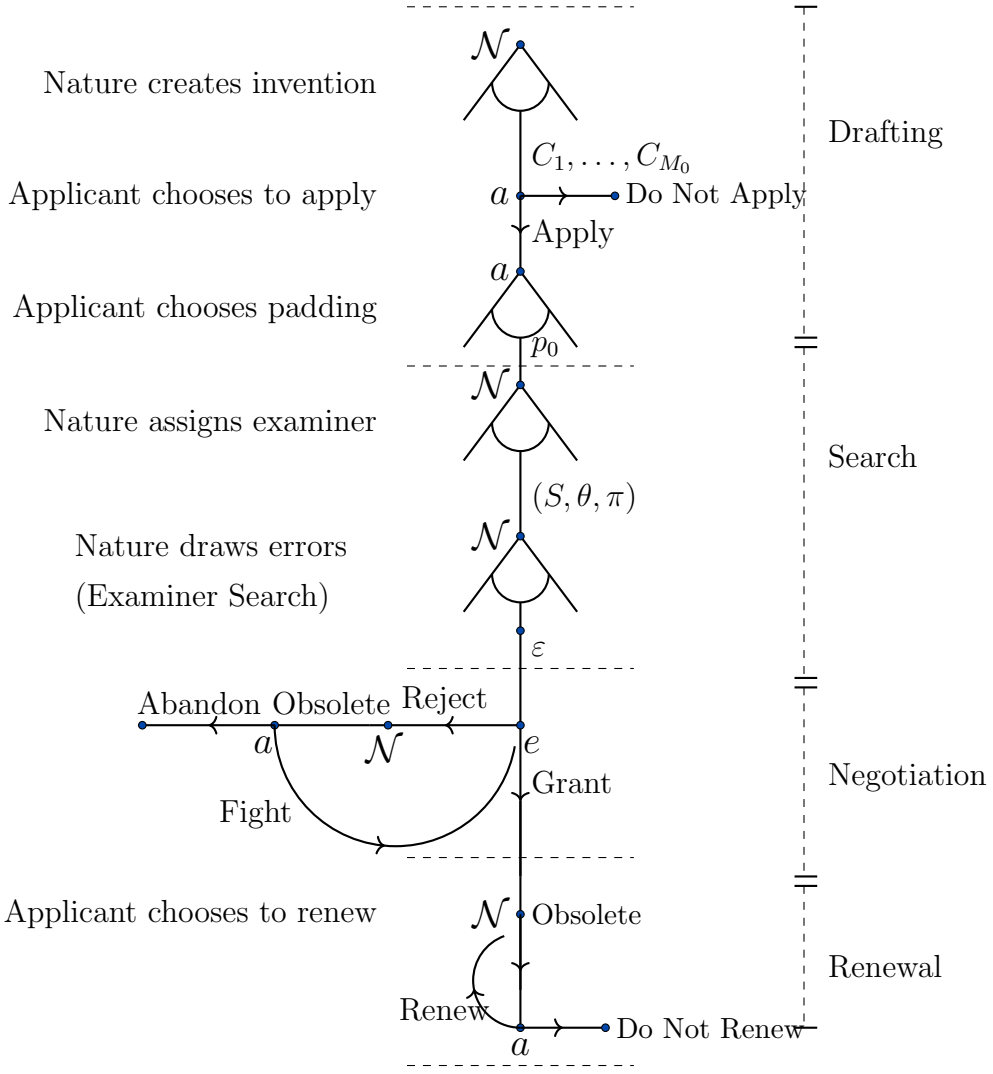
3.4 Model of the Patent Screening Process

We model the patent screening process as a dynamic game in technology center T, between an inventor, a, and a realization of the examiner, e. There are four potential stages: (1) Application Decision and Patent Drafting, (2) Examiner Search, (3) Negotiation, and (4) Renewal. Figure 3.1 depicts the extensive form of the model.

In the baseline model, we analyze patent screening *conditional* on the invention being developed. For the validity of the structural model (and the counterfactual analysis), we do not need to model the potential inventor's decision whether to invest to develop their idea into an invention. However, to quantify the net social costs associated with these errors, we need to model the decision to develop (as well

[46] Table 3.D.2 provides more detail, along with the proportion of within-group variation for other dependent variables, such as mean examination length, mean number of rounds, etc.

FIGURE 3.1: Extensive Form of the Model



as how the patentee licenses their invention), which we do in Section 3.8.

Regarding the examiner, we present a model of how they act on *one specific application*. Therefore, we focus on intra-application incentives and costs for the examiner, rather than inter-application incentives induced by factors such as meeting their quarterly credit targets. A model in which examiners make decisions over time with consideration of the complete set of examinations in their docket would introduce significant complications and is not necessary to meet the aims of our model.

3.4.1 Application Decision and Patent Drafting

Inventor Type

An inventor is endowed with a developed invention they are considering patenting. The patent application for the invention consists of M_0 initial independent claims

$(C_1, ..., C_{M_0})$.[47] We characterize an independent claim C_j by the pair (D_j^*, v_j^*) where $D_j^* \sim G_D(\cdot)$ is the distance of the true version of claim j to the nearest claim in any existing invention and $v_j^* \sim G_v(\cdot)$ denotes the initial flow net returns generated by the true version of claim j once it is commercialized.[48] We define the returns v_j^* as relative to the inventor's outside option, e.g., protecting the invention by trade secrecy.[49]

Application Decision

First, the inventor decides whether to apply. If they do not, the game ends, and their payoff is zero. If they do, they become an *applicant*, and the game continues. The inventor, a risk-neutral expected utility maximizer, chooses to apply if the expected utility of the game that follows applying is positive (because flow returns are defined relative to the next best alternative).

Padding

After deciding to apply, the applicant chooses the amount to exaggerate the claims on their patent application. We refer to this as the initial choice of *padding*, denoted p. Padding obfuscates the true "metes and bounds" of the invention, thereby concealing the inventive step and expanding the property rights claimed in the application. Padding allows the patent owner to extract potentially more revenue, by working it themselves or licensing it. However, greater padding also entails some obfuscation in defining the relationship between the actual invention and the boundaries of the patent rights claimed and necessarily moves the application closer to the prior art. Figure 3.2 illustrates the concepts of independent claims and padding.

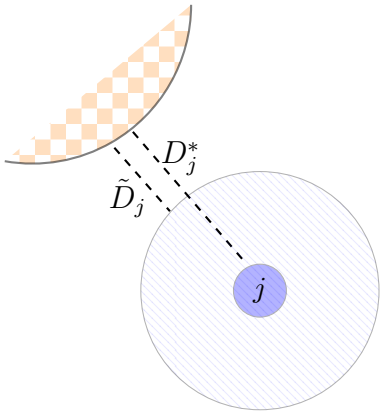
There is a tradeoff for the applicant in the choice of padding. The advantage is that it increases the initial returns of claim j for the applicant from v_j^* to $\tilde{v}_j^0 =$

[47] As we want to focus on the economic incentives for the applicant, we do not consider any agency issues between the inventor and the patent attorney who actually drafts the application.

[48] We assume that distances and values are uncorrelated. Based on the theoretical literature on differentiated products, the relationship is ambiguous. Other things equal, being further from rivals (in product space, which we assume is correlated with claim distances) softens price competition and thus increases private value – implying a positive correlation between distance and value. However, the distribution of demand will typically vary with location, with firms endogenously locating (patenting) in areas of high demand. This implies a negative correlation between distance to rivals and values.

[49] Table 3.A.2 provides our choices for parameterized distributions of distances $G_D(\cdot)$ and values $G_v(\cdot)$ (along with all other distributions in the model).

FIGURE 3.2: Distances and Padding



Notes: The orange semicircle in the top left corner represents the closest existing invention to the independent claim j, which is the small full blue circle in the bottom right corner. The applicant pads the true independent claim to create the larger cross-hatched circle. The distance between the true independent claim and the nearest existing invention is D_j^*, whereas the distance between the padded claim and nearest point is $\tilde{D}_j$.

$\mathcal{V}(v_j^*, p)$, where the padded value function $\mathcal{V}(\cdot,\cdot)$ is increasing in both arguments. On the other hand, padding increases the likelihood of examiner rejections during the examination process on the grounds of non-obviousness (closeness to existing patents) and indefiniteness. Padding shrinks independent claim distances from D_j^* to $\tilde{D}_j^0 = \mathcal{D}(D_j^*, p)$, where the padded distance function $\mathcal{D}(\cdot,\cdot)$ is increasing in D_j^* and decreasing in p. For simplicity, we assume that value (distance) is proportional (inversely proportional) to the degree of padding: $\tilde{v}_j^0 = p \cdot v_j^*$ and $\tilde{D}_j^0 = D_j^*/p$.

Finally, there is a direct cost of padding in the form of legal costs, which we assume is proportional to padding because heavily padded applications require more time to craft.[50] In particular, we specify legal costs as $F_{\text{app}} = f_{\text{app}} \cdot (1 + |p - 1|)$, where f_{app} is the attorney fees associated with patent drafting (which is log-normally distributed across applicants). The motivation for this specification is that it takes additional time for the attorney either to under-pad ($p < 1$) or over-pad ($p > 1$); writing down the truth ($p = 1$) is quickest. We assume symmetry for simplicity.

[50]The applicant may choose to *understate* the true scope of the invention ($p < 1$) and thus earn lower returns, as it reduces the likelihood of rejection by the examiner (especially if there is a restriction of the number of rounds allowed). We find some evidence of such under-padding in the empirical results.

Applicant Expected Utility

The applicant decides the initial level of padding without knowing the identity of the examiner the Patent Office will assign. This feature is relevant because examiners differ in types (seniority, time cost, and intrinsic motivation) and, thus, in their strategies. As a result, applicants make initial padding decisions in light of the distribution of examiner types. The applicant chooses initial padding to maximize their expected utility less application legal costs, where the expectation is taken first over the roster of potential examiners $e = 1, \dots \underline{E}$ (where the random assignment of applications implies an equally likely chance of each examiner in the relevant technology center), over the error of the examiner $\varepsilon \sim G_{e,\varepsilon}(\cdot)$, and potential obsolescence of their invention $\boldsymbol{\omega}$ (all described later).

Formally, the applicant's optimal padding choice p_0 maximizes the ex ante value of patent rights $\Gamma(p)$, defined[51]

$$\Gamma(p) = \mathbb{E}_{e,\varepsilon,\boldsymbol{\omega}}\left[U_a^0(e,\varepsilon,\boldsymbol{\omega},p)\right] - F_{\text{app}}(p),$$

where

$$\mathbb{E}_{e,\varepsilon,\boldsymbol{\omega}}\left[U_a^0(e,\varepsilon,\boldsymbol{\omega},p)\right] = \frac{1}{\underline{E}}\sum_{e=1}^{\underline{E}}\int \mathbb{E}_{\boldsymbol{\omega}}U_a^0(e,\varepsilon,\boldsymbol{\omega},p)\, dG_{e,\varepsilon}(\varepsilon),$$

and we define $\mathbb{E}_{\boldsymbol{\omega}}U_a^0(e,\varepsilon,\boldsymbol{\omega},p)$, the applicant expected utility (over the full vector of obsolescence) for a given examiner e and error ε, later in Equation (3.4).[52] The applicant applies if

$$\Gamma^* \equiv \Gamma(p_0) \geq 0. \tag{3.1}$$

3.4.2 Examiner Search

Examiner Assignment

The patent office assigns the application randomly to an examiner within the relevant art unit of the technology center. We characterize an examiner by the tuple (S, θ, π).

[51] Throughout, we use the notation $\mathbb{E}_{\boldsymbol{\omega}}$ to denote expectations taken over the vector of obsolescence shocks that are not yet realized. Before applying, this is the full vector of 20 possible shocks that could occur, one each year after application. As the process continues, obsolescence shocks occur, and fewer shocks are left to be realized. With a slight abuse of notation, whenever we use $\mathbb{E}_{\boldsymbol{\omega}}$ with an emboldened $\boldsymbol{\omega}$, it refers to the sub-vector of $\boldsymbol{\omega}$ that have not yet occurred. The notation $\mathbb{E}_{\omega_r}$ refers to an expectation over ω only in round r.

[52] We simplify notation by using e to denote both the random variable reflecting the (unknown) examiner prior to application its realization after applying. The same holds for examiner errors.

The first term S represents examiner seniority. The type $\theta \sim G_{S,\theta}(\cdot)$ corresponds to the level of intrinsic motivation. Intrinsically motivated workers incur a disutility from awarding patent rights that do not meet the patentability standard, based on the information available to them (see Section 3.4.4 for formalization of how this enters the examiner's payoff). We let the distribution of θ depend on seniority S. Finally, $\pi \sim G_{\pi}(\cdot)$ corresponds to the examiner's cost of delay (i.e., the extra effort cost for going another round plus any pressure costs associated with timely docket management). The effort cost component will reflect the examiner's productivity.

Examiner Grounds for Rejection

Once assigned, the examiner learns the applicant's identity and thus their fighting costs. The examiner also knows the padded value of the application to the applicant. The examiner reads the application and independently searches the existing prior art to assess the grounds for rejection throughout the negotiation process. There are three main grounds for rejection: *novelty*, *non-obviousness*, and *indefiniteness*. Novelty requires that the claim has not been in use for one year before filing. Non-obviousness requires that the claim makes an inventive step beyond the closest existing invention that would not be immediate to anyone skilled in the relevant area. Indefiniteness requires that the claim is precise and clear on the exact boundaries of claimed property rights. In this chapter, we focus on novelty/non-obviousness.[53]

After searching the prior art, the examiner assesses the obviousness/novelty of *each claim* j, with their assessment denoted by $\hat{D}_j$ and equal to

$$\hat{D}_j = \mathcal{D}(D_j^*, p) \cdot \varepsilon,$$

where ε denotes the drawn examiner error in assessing obviousness/novelty, which is assumed to be independent of the true distances D_j^*. The distribution of search errors depends on the seniority of the examiner and may also depend on the technology center since the number and complexity of patents and other prior art vary across technology fields.

The distribution of search errors also depends on the intrinsic motivation of the

[53] Using the Office Action Research Dataset described in Section 3.3, which identifies the reasons the examiner rejects claims in a patent at each round, we analyzed the overlap between novelty/non-obviousness (102/103) and indefiniteness (112) rejections. We find that 73% of office actions containing a 112 rejection also contain a 102/103 rejection. Thus, novelty/non-obviousness rejections cover most of the observed indefiniteness rejections, so omitting indefiniteness from the baseline model is a profitable abstraction.

examiner. We specify that the *mean* of the search error distribution satisfies two criteria. The first is that the mean of the error tends to one (the unbiased case) as $\theta \to \infty$. The second is that for all $\theta < \infty$, the mean of the search error distribution is greater than one. We specify the second feature because examiners who are not perfectly intrinsically motivated do not scour the literature so thoroughly, thereby missing relevant prior art. When they miss relevant prior art, they perceive distances to be larger than they are and hence have errors greater than one. However, these requirements do not force one-sided examiner error since some draws may still be below one, even if the mean is above one. Our functional form choice satisfying this assumption is $\mu_\varepsilon = 1 + \frac{1}{\theta}$.

We say the examiner has *grounds* for an obviousness rejection if $\hat{D}_j$ is less than an obviousness threshold τ. However, having grounds for rejection will not necessarily mean the examiner will reject the claim. The examiner's decision will be the one that maximizes their utility, taking into account their explicit incentives (credits) and intrinsic motivation. This is a crucial point as it implies that examiners' decisions in the data may not align with decisions made solely on legal grounds.

Finally, examiner errors are specified to be constant throughout the negotiation stage. In this sense, there is no updating of examiner error. However, the grounds for rejection will be recalculated at every negotiation round as the applicant narrows the extent of padding in response to a rejection by the examiner.

3.4.3 Information Structure

The information structure for the applicant and examiner is as follows. The inventor knows the set of claims covered by the invention (given by nature), their true distance to all prior art, and the private value of each claim. Before deciding whether to apply for a patent, the inventor does not know which examiner will be assigned to the application. After assignment, the applicant knows the characteristics of the examiner, including the level of intrinsic motivation, productivity, seniority, and structure of patent office incentives the examiner faces. The applicant also knows the structure of the process and the fees imposed by the patent office at each stage.

The assigned examiner does not observe the true claim distances or the applicant's extent of padding, only the padded distances, contaminated by examiner error, for each claim in the application. The examiner does not know the error she makes in determining the claim distances during the search of prior art. The examiner

observes the fighting costs and padded private value of the applicant's claims.[54] Since the examiner reports their assessment of the padded distance to the applicant, the applicant knows the examiner's error.

3.4.4 Negotiation

The Negotiation Stage is a finitely repeated version of the stage game shown in the "Negotiation" section of Figure 3.1. At round r, if required to act, first, the examiner chooses whether to grant or abandon and, if rejected, the applicant chooses whether to abandon or fight. In between the examiner's and applicant's decision, the applicant's invention can become obsolete, in which case the applicant abandons it immediately. The applicant and examiner discount each stage at rate β.

Let $\mathbf{x}_a$ and $\mathbf{x}_e$ be the vector of strategies of the applicant and examiner, respectively, if the invention is not obsolete.[55] We detail the actions and payoffs obtained at the two decision nodes, starting at the point at which the examiner has just rejected in round r so that $x_e^r = \text{REJ}$.

Obsolescence and Credits

First, pre-grant obsolescence, denoted by ω_r, is realized. If $\omega_r = 1$, the applicant's invention becomes obsolete. In this case, all returns shrink to zero permanently, and trivially, the applicant abandons and obtains a period payoff of zero.[56] In this case, the examiner obtains a period payoff of credits $g_{ABN}^r(S,T)$. If the invention does not become obsolete, then $\omega_r = 0$, and the applicant makes a non-trivial decision. Formally, obsolescence is a Markov process, where, for all r, if $\omega_r = 1$, then $\omega_{r+1} = 1$ (an absorbing state). Otherwise, if $\omega_r = 0$, ω_{r+1} is a Bernoulli random variable with parameter $P_{\omega,\text{pre}}$ if we are still in the application process, and parameter $P_{\omega,\text{post}}$ if a patent has been granted and we are in the renewals process.

[54] We could assume that the examiner does not perfectly observe the private value, but instead obtains an unbiased signal of the value. This feature would not deliver any additional insights and would increase computational burden.

[55] Of course, the vectors include a rejection/acceptance decision and abandonment/fight decision for *every* round. To check whether a strategy is optimal, we must specify what each player would do in every round, even if the prior parts of the strategy dictate that this round will not be reached on the equilibrium path.

[56] The applicant obtains a period payoff of zero because the Patent Office reveals all applications (after 18 months), so their potential for appropriation of innovation returns (e.g., by trade secrecy as an alternative) has essentially vanished.

We provide the full schedule of examiner credits in Appendix 3.E. The most important feature to note is that credits weakly decline as the applicant enters subsequent requests for continued examination, which make early granting more attractive to the examiner.

Applicant Decision

Upon receiving a rejection, if the invention has not become obsolete, the applicant has two choices. They can abandon (x_a^r = ABN), in which case the applicant's and examiner's payoffs are as described in the event of obsolescence. Instead of abandoning, the applicant can continue the application (x_a^r = FIGHT). Continuing involves narrowing rejected claims, which we model as a reduction in padding p by proportion η.[57] Hence for all *rejected* claims j, the padding becomes $p_{j,r+1} = \eta p_{j,r}$. The padding level remains the same for all accepted claims.

Continuing involves a fighting cost to the applicant. The applicant must pay the attorney the fee for amending the application, F_{amend}. In the case of a Request for Continued Examination, the applicant must pay the associated patent office fee, F_{round}^r. Continuation involves delay costs for the examiner, denoted by π. After narrowing occurs, the applicant pays fighting costs, and we move to round $r+1$.

Formally, let the value function for the applicant *upon being rejected in round* r be $U_a^r(\omega_r, \mathbf{x}_e)$. Clearly, the value function for the applicant is a function of the future strategies of the examiner. Further, because ω is a Markov process, the value function for the applicant only depends on the realization of ω in period r. The term $U_a^r(\omega_r, \mathbf{x}_e)$ is defined as follows. If the invention becomes obsolete, so that $\omega_r = 1$, we have (for all $\mathbf{x}_e$)

$$U_a^r(1, \mathbf{x}_e) = 0. \tag{3.2}$$

[57] We could extend the model to allow the applicant to choose whether to narrow by proportion η with some probability or respond by arguing that the examiner is in error and not narrow at all. However, our data on patent word counts imply that this extension is empirically unimportant. To see this, we look at word counts on patents granted with one rejection after publication and calculate the proportion of cases whether the applicant resubmits an application with the same word count. This happens only 7% of the time, so we view the choice to ignore the possibility of no narrowing as a profitable abstraction in the baseline.

Otherwise,

$$U_a^r(0,\mathbf{x}_e) = \max\Bigg\{0, -F_{\text{amend}} \quad - \quad F_{\text{round}}^{r+1} + \beta\Big(1(x_e^{r+1}=\text{GR})[V^{r+1}-\phi] \quad (3.3)$$
$$+ \quad 1(x_e^{r+1}=\text{REJ})\mathbb{E}_{\omega_{r+1}}U_a^{r+1}(\omega_{r+1},\mathbf{x}_e)\Big)\Bigg\},$$

where $1(A)$ is the indicator function, equal to one if statement A is true and zero otherwise, V^{r+1} defines the ex post net expected benefits from patent rights if granted in round $r+1$, as given in Equation (3.9) in Section 3.4.5, and ϕ is the finalizing fee. Equation (3.3) says that the value for the applicant in round r, provided they are not obsolete, is either zero if it is optimal for them to abandon or the sum of fighting costs, plus either the payoff of being granted in the next round (if the examiner will grant them) or the expected value from round $r+1$ if the examiner will reject them in round $r+1$ (both discounted by β).

If $x_e^{r+1} = \text{GR}$, and $\omega_r = 0$, the applicant abandons in round r if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta[V^{r+1}-\phi]$$

and if $x_e^{r+1} = \text{REJ}$, the applicant abandons in round r if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta\mathbb{E}_{\omega_{r+1}}U_a^{r+1}(\omega_{r+1},\mathbf{x}_e).$$

At this point, we can define the expected utility for the applicant before applying, for a given choice of padding, as

$$\mathbb{E}_{\boldsymbol{\omega}}U_a^0(e,\varepsilon,\boldsymbol{\omega},p) = 1(x_e^1=\text{GR})[V^1-\phi] + 1(x_e^1=\text{REJ})\mathbb{E}_{\omega_1}U_a^1(\omega_1,\mathbf{x}_e), \quad (3.4)$$

where all four terms on the right-hand side are (implicitly) functions of the level of padding.

Examiner Grant/Rejection

If the applicant fights ($x_a^r = \text{FIGHT}$), we move to a new round $r+1$, and the examiner obtains updated assessments $\hat{D}_j^{r+1} = \mathcal{D}(D_j^*, p_0\eta^r)$ on previously rejected claims. Based on their updated assessment, the examiner recalculates the grounds for rejection and decides whether to grant the patent.

Granting

Granting a patent in round $r+1$ ($x_e^{r+1} = \text{GR}$) ends the negotiation game and moves the applicant into the renewal stage. Let $\mathcal{R}^{r+1} \in [0,1]$ denote the proportion

of claims the examiner thinks they should reject on obviousness/novelty grounds. Then the immediate payoff to the examiner from granting is

$$\mathcal{G}^{r+1} = g_{GR}^{r+1}(S,T) - \theta\mathcal{R}^{r+1}.$$

Here $g_{GR}^{r+1}(S,T)$ is the credit received by the examiner for granting at stage $r+1$. The term $\theta\mathcal{R}^{r+1}$ captures the intrinsic utility cost for the examiner. For intuition on this term, consider the extreme cases. When $\mathcal{R}^{r+1} = 0$, the examiner believes there are no independent claims on which they have grounds to reject and therefore feels no intrinsic disutility in granting the application. On the other hand, when $\mathcal{R}^{r+1} = 1$, the examiner believes that they should reject every independent claim, so the examiner is going against the organization's mission statement in granting a patent. The examiner's intrinsic penalty from premature granting is the product of the proportion of strategically incorrect claim acceptances and their intrinsic motivation parameter.

One might be concerned that our specification of intrinsic motivation also captures examiner career concerns within the Patent Office. Even if the examiner were not intrinsically motivated, their internal career prospects may depend on the frequency with which they grant invalid claims. However, while the Office does have a "random review" of examiners' decisions by a senior panel, these reviews are very rare, they do not come with explicit punishments, and Patent Office data confirm that decisions are frequently successfully appealed by the head examiner in the art unit.

Rejecting

If the examiner chooses not to grant in round $r+1$ ($x_e^{r+1} = \text{REJ}$) they get credits $g_{REJ}^{r+1}(S,T)$, and the stage game continues. The examiner follows this choice by rejecting any claim on which they believe there are grounds to reject. Hence, the examiner rejects any independent claim j if $\hat{D}_j^{r+1} < \tau$. After this, the application moves back into the hands of the applicant, at which point another obsolescence realization occurs, and then the applicant decides again whether to abandon or continue.

Formally, we define the value function for *the examiner after rejecting in round* r as $W_e^r(\omega_r, \mathbf{x}_a)$. The value function for the examiner satisfies

$$W_e^r(\omega_r, \mathbf{x}_a) = \begin{cases} g_{ABN}^r & \text{if} \quad x_a^r = \text{ABN or } \omega_r = 1 \\ -\pi + \beta \max\left\{\mathcal{G}^{r+1}, g_{REJ}^{r+1} + \mathbb{E}_{\omega_{r+1}} W_e^{r+1}(\omega_{r+1}, \mathbf{x}_a)\right\} & \text{if} \quad x_a^r = \text{FIGHT} \end{cases} \tag{3.5}$$

In the bottom branch of Equation (3.5), where the applicant fights, the value to the

examiner of rejecting in round r is the cost π plus either the (discounted) benefits of granting in round $r+1$ or the net benefits of rejecting in round $r+1$, whichever is larger.

Given the applicant's strategy $\mathbf{x}_a$ the examiner grants in round r if

$$\mathcal{G}^r > g^r_{REJ} + \mathbb{E}_{\omega_r} W^r_e(\omega_r, \mathbf{x}_a).$$

This says that the examiner grants if the period payoff from granting exceeds the credits from rejecting plus the expected continuation value from the point of having rejected in round r, with expectation taken over obsolescence outcomes.

3.4.5 Renewal

We enter the renewals stage if the examiner grants the patent and the applicant pays the finalizing fee. Our renewal model adapts Schankerman and Pakes (1986) to the United States context, adding a probability of post-grant obsolescence in addition to deterministic depreciation. Suppose the patent is granted in round r. The returns for each granted claim j start at $\tilde{v}_{j,r} = v^*_j \cdot p_r$ and depreciate at rate δ each period after grant. With probability $P_{\omega,\text{post}}$, the invention becomes obsolete, at which point the returns shrink to zero permanently. To keep the patent rights, the applicant must pay renewal fees F_4, F_8, and F_{12} at years four, eight, and twelve after grant. The patent life ends at $L = 20$ years after submission of the patent application, at which point the invention enters the public domain.

The renewal decisions by the applicant are those that maximize their expected utility from retaining patent rights. Formally, define the expected returns from years t_1 to t_2 as

$$\mathbb{E}_{\boldsymbol{\omega}} V_{t_1,t_2} = \sum_{t=t_1}^{t_2} \left[\beta(1-\delta)(1-P_{\omega,\text{post}})\right]^{t-t_1} \sum_j \tilde{v}_{j,r}$$

and let I_t be equal to one if the applicant will renew at year t (provided the patent is not obsolete) and zero otherwise. Then, the applicant will renew at year four if the net expected benefit after year four is positive:

$$V^{N,r}_4 \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{4,7} - F_4 + I_8 \beta^4 V^{N,r}_8 > 0, \tag{3.6}$$

where $V^{N,r}_8$ is the net returns from patent rights after year eight, which is defined analogously. The renewal decision at year eight is analogous, and the decision at year

12 is similar, except there is no future renewal decision post year 12.[58] Finally, we define the ex post net expected benefits from patent rights, when granted in round r, denoted as V^r (as in Equation (3.3), as

$$V^r = \mathbb{E}_{\boldsymbol{\omega}} V_{1,3} + I_4 \beta^4 V_4^{N,r}. \tag{3.9}$$

Characterizing the Equilibrium

For every given parameter vector and choice of padding, the negotiation game is a finite game of perfect information, and hence has a subgame-perfect equilibrium that can be found through backward induction. The equilibrium strategies ($\mathbf{x}_a^*$ and $\mathbf{x}_e^*$) are characterized by (for all r):[59]

1. $x_e^{r,*} = \text{GR}$ if and only if

$$\mathcal{G}^r > g_{REJ}^r + \mathbb{E}_{\boldsymbol{\omega}} W_e^r(\omega_r, \mathbf{x}_a^*).$$

2. If $x_e^{r+1,*} = \text{GR}$, $x_a^{r,*} = \text{ABN}$ if and only if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta[V^{r+1} - \phi].$$

3. If $x_e^{r+1,*} = \text{REJ}$, $x_a^{r,*} = \text{ABN}$ if and only if

$$F_{\text{amend}} + F_{\text{round}}^{r+1} > \beta \mathbb{E}_{\omega_{r+1}} U_a^{r+1}(\omega_{r+1}, \mathbf{x}_e^*).$$

4. The terms $U_a^r(1, \mathbf{x}_e^*)$, $U_a^r(0, \mathbf{x}_e^*)$, and $W_e^r(\omega_r, \mathbf{x}_a^*)$, and $W_e^r(0, \mathbf{x}_a^*)$ satisfy Equations (3.2), (3.3), and (3.5), respectively.
5. I_4, I_8, and I_{12} are equal to one if and only if inequalities (3.6), (3.7), and (3.8), respectively, are satisfied.

We want to highlight the important advantages of modelling patents as comprised of multiple claims rather than as a single object. First, a model with multiple claims

[58] To be precise, conditional on not becoming obsolete, the applicant renews at year eight if

$$V_8^{N,r} \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{8,11} - F_8 + I_{12} \beta^4 V_{12}^{N,r} > 0, \tag{3.7}$$

and, conditional on not becoming obsolete, the applicant renews at year 12 if

$$V_{12}^{N,r} \equiv \mathbb{E}_{\boldsymbol{\omega}} V_{12,20-r} - F_{12} > 0. \tag{3.8}$$

[59] In practice, we limit the process to six rounds (around 95% of applications last at most three rounds of negotiation and the modal number is two) so the characterization holds for $r < 6$. In the sixth round, if rejected, we force the applicant to abandon. The examiner's continuation value is therefore only $g_{ABN}^6(S, T)$.

allows a more realistic description of the patent prosecution system and thus a tighter link to the data on which the model parameters will be estimated. Second, endowing applicants with multiple claims allows for specific claims to be narrowed only up to the round at which they are granted. Third, a multiple claim model enables us to specify the examiners' intrinsic motivation disutility as a function of the proportion of granted claims they judge invalid.

3.5 Estimation

Our primary estimation method is simulated method of moments (SMM), though we estimate some parameters outside the model. For reference, Appendix Table 3.A.2 summarizes all parameters and their associated distributional assumptions.

3.5.1 External Estimation

Discount Rate (β)

The data lack some detailed information for identifying all parameters. Specifically, discount rates are traditionally difficult to identify. Hence, we set $\beta = 0.95$—as in most of the literature (Pakes, 1986).

Distance Threshold (τ)

We estimate the distance threshold externally using observations on claim distances and examiners' grant decisions. For every examiner e, we calculate the minimum value of the distances among claims they grant. This number corresponds to their "personal distance threshold," denoted as $\tau_e = \min_{j \in J_e} \tilde{D}_j$, where J_e is the set of claims granted across all applications by examiner e. Since examiners are not perfectly intrinsically motivated, some examiners' personal thresholds are below the true threshold τ in cases where they knowingly grant patents with relatively small distances. However, we assume that the most intrinsically motivated examiner will have a personal threshold τ_e equal to the true threshold τ. This assumption allows us to estimate the distance threshold as the maximum of the distribution of examiners' individual thresholds. The validity of this approach relies on $\max_{e=1,\ldots,\underline{E}} \tau_e \to \tau$ as the number of examiners in the technology center $\underline{E} \to \infty$.[60] We calculate these

[60] In practice, we experiment with the first and fifth percentiles as robustness checks. We also remove examiners who have conducted fewer than a threshold number of examinations. We experiment with values of 50 and 100 for this threshold and find only minor differences in all of these cases.

thresholds separately for each technology center to create technology-center-specific thresholds. Notably, our estimates of the threshold in different technology centers are very similar, ranging from 0.48 to 0.52.

Applicant Fighting Costs (f.)
We have data on the quantiles of the distributions of *amendment*, *maintenance*, and *issuance* hourly fees charged by lawyers. We assume these three costs are log-normally distributed. Since these moments directly correspond to the elements of applicant fighting costs and do not identify any other parameters in the model, we estimate the mean and variances of the log of fighting costs using the optimal two-step generalized method of moments estimation procedure for each of the three negotiation-based fighting costs. The data allow us to estimate different negotiation fighting cost distributions for simple applications (less than ten claims) and complex applications in chemical, electrical, and mechanical fields.[61]

Depreciation of patent returns (δ)
Bessen (2008) estimates the combined effect of depreciation and the probability of obsolescence at 0.14, using U.S. renewal data. In our context, this corresponds to $(1 - P_{\omega,\text{post}}) \cdot \delta + P_{\omega,\text{post}} \cdot 1$. Hence, for each parameter guess of $P_{\omega,\text{post}}$, we extract the implied pure depreciation rate from this relationship.

3.5.2 Simulated Method of Moments

We estimate the remaining set of model parameters using SMM. The model does not admit an analytic solution for endogenous variables as a function of all the model primitives. Hence, the goal is to choose the parameters that best match the moments of the data with the corresponding moments computed from the model's numerical solution. We estimate the model using moments from the data described in Section 3.3, assuming the model's equilibrium generates the data.

We denote the full vector of parameters to estimate as $\boldsymbol{\psi} = (\boldsymbol{\psi}_e, \boldsymbol{\psi}_a)$. The vector of applicant parameters is $\boldsymbol{\psi}_a = \left(\eta, P_{\omega,\text{pre}}, P_{\omega,\text{post}}, \alpha_D, \beta_D, \mu_v, \sigma_v, \boldsymbol{\mu}_{\boldsymbol{f}_{\text{app}}}, \boldsymbol{\sigma}_{\boldsymbol{f}_{\text{app}}}\right)$. We

[61] On application fighting costs, though we have similar moments on lawyers' application drafting fees, because application fighting cost is proportional to padding in the model, its distribution is contaminated by the endogenous choice of padding (which is a function of all model parameters). This feature means that we cannot estimate the distribution of application fighting costs outside the model: we must estimate these parameters as part of the simulated method of moments procedure described in the next subsection.

described the narrowing and obsolescence parameters in Section 3.4. For distances, we assume D_j^* is Beta distributed with parameters (α_D, β_D). The Beta distribution is a natural choice as it provides a flexible distribution on the interval $[0, 1]$, which coincides with the interval of our distance metric. Further, we use a multivariate normal distribution copula to correlate claim distances within an application.[62] Motivated by Schankerman and Pakes (1986), the log of initial *claim* flow returns is normally distributed with mean μ_v and variance σ_v^2. Finally, we assume that the log of application drafting legal fees per unit padding, f_{app}, are normally distributed with mean $\mu_{f_{app}}$ and variance $\sigma_{f_{app}}^2$, with different parameters for simple and complex applications in chemical, electrical, and mechanical fields.

The vector of examiner parameters is $\boldsymbol{\psi}_e = (\mu_{\theta,\text{junior}}, \mu_{\theta,\text{senior}}, \sigma_\theta, \mu_\pi, \sigma_\pi, \sigma_\varepsilon)$. The first three parameters $(\mu_{\theta,\text{junior}}, \mu_{\theta,\text{senior}}, \sigma_\theta)$ correspond to log-normal parameters for the distribution of examiner intrinsic motivation. We estimate different μ parameters for "junior" (pre-GS-14 grade) and "senior" examiners. Though we constrain the σ parameter to be the same for juniors and seniors, given the log-normal specification, this does not force the variances (or even the variance relative to the mean) to be the same for juniors and seniors. The log of examiner delay costs, π, are normally distributed with mean μ_π and variance σ_π^2. Finally, examiner errors are normally distributed, with mean $1 + \dfrac{1}{\theta}$ and variance σ_ε^2.

We estimate $\boldsymbol{\psi}$ using a minimum-distance estimator that matches moments of the data with the corresponding moments implied by the model. More specifically, for any value of $\boldsymbol{\psi}$, we solve the model for several simulated draws from the distributions of exogenous variables. Then, we calculate moments of the endogenous variables across the simulated observations. The minimum-distance estimator minimizes the SMM objective function:

$$\hat{\boldsymbol{\psi}} = \arg\min_{\psi} \left(\boldsymbol{m}(\psi) - \boldsymbol{m}_{\mathcal{S}}\right)' \Omega \left(\boldsymbol{m}(\psi) - \boldsymbol{m}_{\mathcal{S}}\right),$$

where $\boldsymbol{m}(\psi)$ is the vector of simulated moments computed from the model when the parameter vector is ψ, $\boldsymbol{m}_{\mathcal{S}}$ is the vector of corresponding sample moments, and

[62] Specifically, in the simulation, for each application, we draw a vector of size M_0 from a standard multivariate normal with correlation coefficient ρ. We apply the quantile function of the normal to the draws to create correlated uniform random variables. Then for the estimation guess $(\tilde{\alpha}_D, \tilde{\beta}_D)$, we apply the inverse CDF of a Beta distribution with these parameters to the uniform draws to generate correlated beta distributed initial distances. For ρ, we use the empirical correlation of granted distances. Simulations confirm that the correlation of the multivariate copula is very close to the correlation of the distances. See Nelsen (2007) for details.

Ω is a symmetric, positive-definite weighting matrix.[63]

3.5.3 Choice of Moments

We now briefly describe our choice of moments for the SMM estimation. In Appendix 3.F, we provide some intuition about how these moments aid in identifying the parameters we estimate.

The number of moments we can calculate on endogenous variables in the model far exceeds the number of model parameters. To select a subset of moments for our estimation procedure, we followed a rigorous, data-driven methodology, based on the sensitivity matrix of parameter estimates to the inclusion of particular moments (Andrews, Gentzkow, and Shapiro, 2017), along with plots of how estimated model moments (and separately, the value of the SMM objective) vary with parameter values. We provide details on the complete set of moments we considered and our pruning procedure in Appendix 3.F. Through this procedure, we pruned the set of moments down to 40 that clearly assist in estimating the parameters.

The selected moments corresponding to outcomes for examiners are the proportion of applications granted by round and seniority, the standard deviation of examiner rejection rates by seniority, and the proportion of patents granted containing an invalid claim (again, by seniority and round). The selected moments corresponding to outcomes for applicants are the proportion of abandonments by round and examiner seniority, patent renewal rates, means and standard deviations of granted claim distances by round granted, and means and medians of legal application fees by technology class.

3.6 Empirical Results

In this section, we present and interpret our parameter estimates and briefly discuss model fit and robustness. For model estimates, we bootstrap standard errors. Standard errors are negligible for all parameters, which is unsurprising since we calculate data moments using millions of observations.

[63]For the weighting matrix we use a diagonal matrix that scales moments to a uniform scale. We cannot use the optimal two-step weight matrix because we do not have application-specific data on fighting costs that can allow us to compute the correlation between these moments and others. Details on computation and numerical optimization are available on request.

TABLE 3.1: Applicant Parameter Estimates

Parameter	**Symbol**	**Estimate**	**S.E.**
Per-round narrowing	$1-\eta$	0.25	0.000
Pre-grant obsolescence	$P_{\omega,\text{pre}}$	0.14	0.001
Post-grant obsolescence	$P_{\omega,\text{post}}$	0.04	0.000
Initial returns log-mean	μ_v	10.55	0.077
Initial returns log-sigma	σ_v	1.32	0.022
Initial distance alpha	α_D	4.57	0.003
Initial distance beta	β_D	7.74	0.004
Simple application fighting cost log-mean	$\mu_{f,\text{simple}}$	8.53	0.011
Simple application fighting cost log-sigma	$\sigma_{f,\text{simple}}$	0.87	0.054

Notes: This table provides the applicant's model parameters. Standard errors are bootstrapped. Table 3.A.3 provides fighting cost parameters by technology area.

3.6.1 Applicant Parameters

Table 3.1 presents the estimates for parameters relating to the applicant. First, we estimate the proportion of narrowing per round as $1-\eta = 0.25$. This estimate indicates that screening substantially narrows over-claiming by the applicant. Second, we estimate two probabilities of obsolescence: a pre-grant probability during the application process and post-grant obsolescence during the patent's life. The estimated pre-grant obsolescence probability is 14% for each negotiation round. The post-grant rate is 4% per year, which is broadly similar to other estimates in the literature.[64] The probability of obsolescence is higher during the application process for two reasons. First, applicants are more likely to discover their invention to be obsolete earlier in its life cycle (e.g., discovering that commercialization costs make the project unviable). Second, the prosecution stage contains applications that are eventually granted and those who abandon, and many of those who abandon do so precisely because they become obsolete.

Third, the distribution of initial returns from an unpadded independent claim is highly skewed. Though the mean is $91,046, the median is $38,069, and the modal

[64]Using data for three European countries, pooled across technology fields, Pakes (1986) calculates values of 6%, 4% and 1% for the likelihood of obsolescence in the first, second and third year after grant, respectively. Using German data, disaggregated by four technology areas, Lanjouw (1998) estimates a range of 7-12%.

value of initial returns for an unpadded independent claim is \$6,656. To understand the distribution of unpadded initial returns on the *application*, we take the distribution of the number of independent claims and use it to construct sums of draws from the distribution of claim returns. For example, the first patent application in our dataset has two independent claims. Hence, we draw two values from the distribution of claim initial unpadded returns and add them to get the total initial unpadded returns on that application. The median initial unpadded returns from a patent application are \$129,659.

It is difficult to compare our estimates of initial returns to existing estimates in the literature on total patent returns since we estimate the distribution of initial returns for (a) all *applications* (not just granted ones) and (b) *unpadded* claims. Nonetheless, it is worth noting that Bessen (2008) estimates the mean net present value of patents (adjusted to 2018 U.S.D) for all U.S. patentees as \$78,168 and \$113,067 for just U.S. public firms in manufacturing.

Next, we discuss the implied distribution of initial unpadded distances and fighting costs. The mean distance is 0.37, and the distribution is approximately symmetric. Given that our estimated thresholds are between 0.48 and 0.52, these estimates imply that about 83% of application claims have distances below the threshold. Despite this, many applications are eventually granted because of extensive narrowing and examiners granting invalid claims. Fighting costs for simple applications are lower than all other categories. Recall that legal costs per application are specified as $F_{\text{app}} = f_{\text{app}} \cdot (1 + |p - 1|)$, where f_{app} is the attorney fees associated with patent drafting. Evaluated at the mean levels of p and f_{app}, we estimate these transaction costs at \$7,920 for simple applications and \$12,333 for electrical applications.

Padding (overclaiming property rights)
We compute statistics on the model's endogenous variables by simulating the model at our estimates. Relating to the applicant, we calculate the distribution of optimal initial padding for those who apply. The mean padding level is 8%, with 70^{th} and 90^{th} percentiles equalling 18% and 31%, respectively. These results suggest that many applicants substantially exaggerate the true extent of their invention when they apply for patent rights.

We also compute two weighted averages of padding, where our weights are either the mean (over claims) of initial unpadded distances ($\bar{D}_s^*$) or initial unpadded values ($\bar{v}_s^*$). The weighted average of padding rises to 10% when weighted by values and 9% when weighted by distances, indicating that inventors increase padding for applications

TABLE 3.2: Examiner Parameters Estimates

Parameter	Symbol	Estimate	Standard Error
Junior intrinsic motivation log-mean	$\mu_{\theta,\text{junior}}$	3.92	0.004
Senior intrinsic motivation log-mean	$\mu_{\theta,\text{senior}}$	3.38	0.005
Intrinsic motivation log-sigma	σ_θ	0.77	0.055
Delay cost log-mean	μ_π	0.19	0.006
Delay cost log-sigma	σ_π	0.27	0.015
Error standard deviation	σ_ε	0.02	0.000

Notes: This table provides the model parameters relating to the examiner. Standard errors are bootstrapped.

with claims that are more valuable and distant from the prior art (where such padding is less likely to induce the examiner to reject).

3.6.2 Examiner Parameters

Table 3.2 presents the estimates of the examiner parameters. To understand examiner costs and intrinsic motivation, we provide a slight digression on the units of examiner payoffs in the model, which we call "normalized credits."[65] The Office adjusts each examiner's credits based on their seniority and the technological complexity of applications. We use the same adjustments when we model payoffs for examiners.[66] These normalized credits are the unit of examiner payoffs. This ensures that payoffs are in the same units for all examiners, regardless of their seniority and technology center.

We start by interpreting the parameters of intrinsic motivation. To our knowledge, these are the first structural estimates of intrinsic motivation in a public agency. We estimate σ_θ as 0.77, which implies, by the properties of the log-normal distribution,

[65] Appendix Section 3.E provides a detailed derivation of the examiners' credit structure.

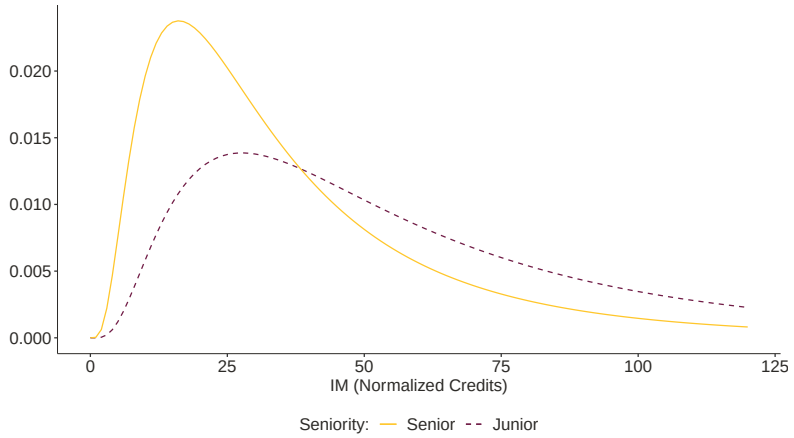
[66] For example, an examiner receives two credits for granting a patent in the first negotiation round. We adjust these credits by dividing by a seniority factor (for example, by 1.25 for a senior GS-14 examiner) and multiplying by a technology correction (say, 29 for the relatively complex category of computer networks). Therefore, a GS-14 examiner in technology center "computer networks" receives 46.4 normalized credits for granting a patent in the first round. Tables 3.E.1 and 3.E.2 report the values of seniority and technology corrections across all seniorities and technology centers, respectively.

a coefficient of variation of 0.82 (82%). This estimate implies substantial variation in intrinsic motivation across examiners, even within seniority category. We estimate $\mu_{\theta,\text{junior}} = 3.92$ and $\mu_{\theta,\text{senior}} = 3.38$, which implies that, on average, junior examiners are more intrinsically motivated than senior examiners. Figure 3.3 plots the distribution of intrinsic motivation for junior and senior examiners as implied by the log-normal assumptions. It is clear that the distribution of senior examiners' intrinsic motivation (yellow solid) is generally lower than that of junior examiners (maroon dashed). At least two countervailing forces influence the relationship between seniority and intrinsic motivation. Intrinsic motivation will fall with seniority if examiners become "jaded" with experience. However, selection cuts the other way since the least intrinsically motivated examiners are likelier to move to the private sector with higher remuneration. The evidence thus indicates that the jading effect dominates the selection effect.

To interpret the magnitude of intrinsic motivation, we calculate the associated cost for a median intrinsically motivated GS-12 (junior) examiner in a selected technology center 36 ("Miscellaneous" category). For this examiner, the seniority correction is one and the technology correction is 22.4. Recall that intrinsic motivation cost (in terms of normalized credits) is $C_{IM} = \theta\mathcal{R}$, where θ is the intrinsic motivation parameter, and $\mathcal{R}$ is the proportion of claims the examiner believes invalid. We divide C_{IM} by 22.4 to change the units back to pure credits. Hence, in terms of raw credits, this examiner's intrinsic motivation cost is $2.25R$, which means that the examiner faces a cost of 2.25 credits for knowingly granting a patent with *100%* of its claims as invalid. This cost is equivalent to the credits the examiner obtains for making *three* final rejections. This example is only an illustration, but our estimates generally imply that intrinsic motivation costs are sizeable relative to extrinsic rewards.

Next, we consider examiner delay costs. The coefficient of variation of examiner costs is 0.08, ten times smaller than examiner intrinsic motivation. Moreover, delay costs are estimated to be small, with the median cost for a GS-12 (junior) examiner in technology center 36 paying an equivalent of 0.05 credits to go an extra round on this particular application. The fact that these costs are so small suggests that examiners are not pressured explicitly to finish applications fast and that the opportunity cost of devoting more time to this application relative to the next on their desk is small. This finding is intuitive since the most time-consuming activity for the examiner is their initial literature search. Hence, continuing to make decisions on an application they have already reviewed is less time-intensive than starting a new application (though also less compensated).

FIGURE 3.3: Density of examiner intrinsic motivation



Notes: Orange solid curve represents the distribution for senior examiners; maroon dashed curve represents the distribution for junior examiners. To interpret the x-axis, consider an examiner in technology center 36, where the technology correction is 22.4. Dividing the values on the x-axis by 22.4 yields the number of credits the examiner pays as an intrinsic motivation cost to a GS-12 examiner for granting a patent for an application on which every claim is invalid.

Finally, we discuss examiner error parameters. Recall that examiner errors are normally distributed, with an estimated standard deviation and mean equal to $\mu_\varepsilon = 1 + \frac{1}{\theta}$, where θ is the examiner's intrinsic motivation. The error that an examiner draws multiplies padded distances to create the examiner's distance assessment. Since junior examiners are more intrinsically motivated on average, the mean of the junior examiners' error distribution is closer to one. We estimate the standard deviation of examiner errors to be 0.02, indicating that errors are modest, typically within 4% of the examiner-specific mean.

Calculating Examiner Errors

We compute statistics on two kinds of examiner errors by simulating the model with our baseline estimates. The first error occurs when an examiner grants a patent with invalid claims. We refer to this as a "type 1" error. We calculate that this happens for 19% of grants, suggesting that while examiners are screening out some invalid patents, nearly one in five applications contain some claims that should not have been granted. The last statistic represents the "extensive" margin of this type of examiner error; we can also calculate an "intensive margin" error. Among all granted claims, 7% are invalid (compared to 83% of claims whose unpadded distance is below the threshold), implying that most invalid patents contain only a few invalid claims.

We also calculate the weighted errors (focusing on the intensive margin), where

weights reflect the distance of the claim from the patentability threshold. Among simulations, indexed by s, let S_G be the set granted and j represent a claim. We calculate the measure

$$\sum_{s \in S_G} \sum_{j} \frac{w_{sj}}{\sum_{s' \in S_G} \sum_{j'} w_{s'j'}} E_{1sj,\text{int}}, \tag{3.10}$$

where $E_{1sj,\text{int}}$ is equal to one if claim j on simulation s is invalid (has a distance below the threshold) and zero otherwise, and the weight $w_{sj} = |\tilde{D}_{sj} - \tau_s|$, where τ_s is the threshold relevant to simulation s. The idea is to put more weight on errors where claims are further away from the threshold (making the error more "egregious"). If the weighted average is lower than the unweighted average, it implies that errors occur in marginal cases in which it is not obvious whether the patent is valid. Indeed, the weighted error is 2%, much lower than the unweighted value of 7%, suggesting that most errors occur in cases of marginal validity.[67]

The other kind of "error" (or "undesirable" outcome) occurs when an applicant abandons an application that contains valid claims. We refer to these as "type 2" errors. Approximately 36% of abandonments have at least one valid claim. Strictly speaking, these are not a mistake by the examiner since they should only grant patents to applications on which *all* claims are valid. At the intensive margin, among all claims the applicant abandons, 18% are valid.

Similar to Equation (3.10), we calculate a weighted average of type 2 intensive margin errors, where the weights are the same as before—the distance of a claim from the threshold. In this case, the weighted error falls to 6%, again implying that abandonments occur on marginally valid claims rather than clearly valid ones. When we compute the extensive margin weighted error, the proportion of abandoned applications with at least one valid claim is 11%.

Model Fit and Robustness

We compare the values of simulated moments, calculated at the estimated parameters, with moments in the data.[68] As expected, we match most of the internal moments well (two exceptions are described in Appendix 3.F). The real test of model fit, however, is how well we match moments that are not used in the estimation procedure. Appendix Figure 3.A.2 displays these comparisons for the excluded

[67] We also calculate weighted averages for the extensive margin type 1 errors (details available on request). The weighted error in this case is 5.1%, similarly suggesting that most errors occur in marginal cases.

[68] Appendix Figure 3.A.1 displays the full set of comparisons.

moments described in Appendix 3.F, which include percentiles on granted distances in each round, the mean of distances for latter rounds, and means and percentiles of round one rejection rates across seniority categories. We match all of these moments well.

We run a series of robustness checks on our baseline model (Appendix Table 3.A.5 summarizes the results). First, we examine changes to how we define the distance threshold for patentability. In the baseline, we define each examiner's "revealed" threshold as the minimum distance they grant and then take the threshold as the maximum of those values over examiners. We experiment with using the first and fifth percentile of distances granted for each examiner, which allows for measurement error in their personal threshold. We also check robustness to a discount factor of 0.99 and a broader definition of examiner seniority. In all cases, the parameter estimates are generally robust.

3.7 Counterfactual Analysis

We use the estimated model to conduct a series of counterfactual analyses to examine the impacts of various reforms on the speed and quality of the screening process and the degree of padding in patent applications. The counterfactual scenarios we examine include removing intrinsic motivation, and changing the level of patent office fees, the number of allowable rounds in the process, and examiner extrinsic incentives (credits).

Table 3.3 presents the results. We focus on four endogenous outcomes. The first is the proportion of applicants who choose not to apply for a patent on their developed invention. The second is the applicant's choice of how much to pad the application. The third set of outcomes is the proportion of grants in round one and the average number of rounds (speed of resolution). The fourth set, relating to screening quality, is the proportion of granted patents with at least one invalid claim (type 1 error at the extensive margin) and abandoned applications with at least one valid claim (type 2 error at the extensive margin). We note but do not report that the changes in these errors at the intensive margin errors are similar.

Fees

In the baseline, there are relatively low fees for applicants throughout the prosecution process. In the first counterfactual, we introduce a substantial per-round fee that

TABLE 3.3: Counterfactual Experiments

Counterfactual	**Not Apply** (%)	**Pad** (%)	**Rounds**	**R1 Gr** (%)	**T1** (%)	**T2** (%)
Baseline	6.3	8.0	2.5	11.5	18.8	36.5
25K Round Fee	8.4	6.6	2.4	12.8	18.0	37.9
50K Round Fee	12.2	5.9	2.4	14.2	17.7	39.8
Three Rounds	27.2	3.6	2.1	15.1	15.9	49.2
Two Rounds	51.1	0.8	1.6	25.8	11.9	56.1
One Round	79.6	-2.3	1.0	98.4	0.5	91.7
15% IM	3.9	8.0	2.1	30.2	89.3	22.4
Credit$\searrow$	6.3	7.9	2.5	11.5	18.7	36.3
Credit$\searrow$ + 15% IM	3.4	18.1	2.1	32.8	88.9	17.7

Notes: "Not Apply' is the percent of inventors who do not apply for a patent; Pad is the mean level of padding. Rounds is the mean number of rounds. "R1 Gr" is the percent of applications granted in Round 1. T1 represents the proportion of granted patents with some invalid claims. T2 represents the proportion of abandoned applications with some valid claims.

the applicant must pay for *each* negotiation (not just for an RCE).[69] This fee acts as a marginal cost per round of negotiation. Since each round is now more expensive, applicants have increased incentives to exit the patent process as soon as possible, and less incentive to apply in the first place. A substantial $50,000 fee for every extra round reduces padding by a quarter (from 8.0% to 5.9%) and slightly reduces the mean number of rounds, from 2.5 to 2.4. The proportion of grants in round one increases from 11.5% to 14.2%, reflecting the reduced padding, and the fraction of granted patents with some invalid claims (type 1 error) falls slightly. However, the rounds fee increases type 2 error—rising from 36.5% to 39.8%. The trade-off between these two types of errors is a feature of many of the counterfactuals we analyze.

[69]We also consider substantially increasing the *application* fees to as much as $50,000. However, because this is a fixed fee paid upon application, provided it is still profitable to apply, applicants will not change their padding decision. Even at this level, the fee does not materially alter average padding and, since there is practically no change to the proportion of inventors who choose to apply, introducing an application fee acts mainly as a transfer from applicants to the Patent Office, with minimal changes to quality or speed of prosecution. If the additional resources from the higher application fee were reinvested in patent office examination, there would be improvements. This finding—that application fees only really help if they are reinvested—is similar to the findings in Schankerman and Schuett (2022), who use a completely different theoretical model and data.

It is at first surprising that per-round fees as high as $50,000 do not substantially change the speed or quality of patent prosecution. The explanation is that the private value of patent rights is large enough to make applying for a patent on many of these inventions worthwhile, even with high per-round fees. Further, the applicant has the option to apply and then abandon after the first rejection, without paying any negotiation fee. Fees would have to be much higher to substantially impact outcomes.[70]

Restricting the Number of Rounds

Instead of using fees, we consider limiting the maximum number of rounds of negotiation between the applicant and examiner. We consider a maximum of three rounds, then a maximum of two rounds (equivalent to removing all RCEs, allowing only one round of interaction between applicant and examiner), and finally, we allow for only one round (that is, no negotiation between applicant and examiner so that the examiner's first decision is final). These counterfactuals are motivated by a 2007 U.S.PTO proposal to restrict the number of RCEs. The proposed rulemaking was challenged in federal court, which judged the restrictions as an overreach of Patent Office authority.[71] The court decision did not consider the quantitative impact of such changes on patent office screening quality or its welfare effects, which our work makes possible.

Round restrictions have material consequences on screening outcomes. Removing all RCEs (allowing only two rounds) would lead to half of all inventors not applying for a patent and would virtually eliminate padding. In this case, 25.8% of applicants are granted in the first round, and because applicants respond to the restriction by reducing padding, the proportion of patents granted with invalid claims falls. In particular, with only one opportunity for negotiation, type 1 error falls sharply, from 18.8% to 11.9%.

[70]Of course, these fees would be significant for small firms or single inventors who may be cash constrained. However, round fees for small and micro entities could be reduced, as the Patent Office already does for other types of fees.

[71]The proposed changes are in *U.S.PTO Changes to Practice for Continued Examination Filings, Patent Applications Containing Patentably Indistinct Claims, and Examination of Claims in Patent Applications—the "New Rules"* (SmithKline Beecham Corp. v. Dudas, 541 F. Supp. 2d 805, 2008). The court decided that the "New Rules" were substantive and that the Patent Office did not have the rulemaking authority to make substantive changes, though the Court noted that the Patent Office could make procedural changes, such as fees. As we will show, one can achieve the same equilibrium number of rounds with an "equivalent" fee, so from an economic point of view, this distinction is problematic.

The disadvantage of limiting the scope for negotiation is that it increases the proportion of abandoned applications with valid claims. With no RCEs, this proportion rises from 36.5% to 56.1%. As with fees, making the process tougher for applicants through fewer allowable rounds generally reduces the granting of invalid claims and speeds up the process but leads to the abandonment of valid claims. As we discuss in the next section, granting invalid claims and not granting valid claims each imposes social costs, and we need to measure these to evaluate the overall impact of the reforms.

Finally, we compare the effectiveness of fees and round restrictions ("price versus quantity" instruments) by computing the equivalent per-round fee—in the sense of equalizing the mean number of rounds in equilibrium—to restrictions on the number of RCEs. The simulations show that the fee equivalent to removing all RCEs is a massive $600,000 per round. Using fees generally produces lower type 1 and type 2 errors than their rounds equivalents, but such fee levels are politically unpalatable.

Removing Intrinsic Motivation

Next, we evaluate the impact of removing intrinsic motivation by reducing it for every examiner to 15% of its original value.[72] Knowing that examiners will be more unwilling to grant invalid patents, only 3.9% of inventors do not apply, the number of rounds falls from 2.5 to 2.1, and the proportion of applications granted in round one almost triples, increasing from 11.5 to 30.2. Not surprisingly, type 1 error jumps sharply to 89.3%, while type 2 error declines. This counterfactual highlights the quantitative importance of intrinsic motivation on the quality of patent screening and confirms its potential salience for economic analyses of other public agencies.[73]

Removing Credits

Finally, we consider changes to the structure of credits for examiners. We remove all credits for the examiner after the first round. If it is the case that examiner costs of delay represent the marginal cost of an extra examination round, then such a policy change could be justified on efficiency grounds of "marginal cost pricing"

[72]We cannot fully remove intrinsic motivation because our specification of mean error being inversely related to IM implies that IM cannot be exactly zero. We provide the reason behind our choice of 15% in Section 3.8.

[73]Interestingly, increasing intrinsic motivation (not reported) does not have much impact in reducing padding or type-1 error. The explanation for this outcome is that examiners are already sufficiently intrinsically motivated to get most of the benefits, so further increases do not have much bite.

since we estimated examiner delay costs to be small.[74] When we remove all credits after the first round in the baseline model, there are minimal, if any, impacts on any of the outcome variables. This result suggests that our baseline estimates of intrinsic motivation are sufficiently large for examiners to want to avoid granting invalid patents even in a context where they will receive no further extrinsic reward if they do so. This striking finding reflects the extent to which patent office examiners are intrinsically motivated. The results are not consistent with extrinsic incentives crowding out intrinsic incentives.

To complement this exercise, we also analyze the effect of removing all credits after the first round alongside reducing intrinsic motivation to 15% of its value (at any higher value of intrinsic motivation, removing credits has no material effect). In this case, we find non-trivial impacts of credits consistent with economic intuition. First, padding doubles, up to 18% (relative to 8.0% when only intrinsic motivation is changed) and first-round grants increase from 30.2% to 32.8%. Type 2 error declines because the increased padding means that abandonments are less likely to include valid claims. These results indicate that extrinsic incentives and intrinsic motivation are *substitutes*, not complements, as sometimes found in the experimental literature (see Section 3.2 for citations): credits only work as an effective device to incentivize examiners when examiners are not intrinsically motivated (and even then, as we show in the next section, credits do *not* reduce social costs of screening).

In summary, these counterfactual experiments show that no reform we consider unambiguously improves both prosecution speed and quality. There is typically a trade-off: policies that make prosecution stricter lead to fewer grants of invalid patents but increased abandonments of valid applications. Evaluating reforms requires converting these outcomes into social costs, which we do in the following section.

[74]This counterfactual has limitations that the others do not because our model is focused on optimal decisions on a given patent application. It does not incorporate any interactions between different applications the examiner faces, such as optimizing docket management across applications (including meeting quarterly or annual targets). This counterfactual is best thought of informing an examiner that for one of the new applications in their docket, they will only receive credits for the first round.

3.8 Quantifying the Social Costs of Screening

We classify net social costs into three categories: *type 1*, *type 2*, and *prosecution* costs. Type 1 costs refer to the costs induced by granting invalid claims. Type 2 costs refer to the social value of inventions that are not developed ex ante because of the potential threat of not being granted valid claims. Type 1 benefits refer to the social value of inventions that would not be developed ex ante without type 1 error. Type 2 benefits refer to the ex post deadweight loss *not* incurred when inventors abandon valid claims. Prosecution costs are the Patent Office's costs of examining applications plus the legal fees incurred by the applicants during the negotiation process. In what follows, we summarize our quantification approach; full details are in Appendix 3.G. We start with the costs of each type of error and then discuss the benefits.

3.8.1 Type 1 Costs

There are two sources of costs from type 1 error: the deadweight loss associated with the royalties extracted by the patentee and the litigation costs associated with legal challenges against invalid patents that are granted (and that are valuable enough to warrant a challenge).

Deadweight loss from royalties

We assume that the patentee charges the Arrow royalty equal to the unit cost savings due to the invention, Δc. The deadweight loss from royalties depends on the market structure of licensees. Our baseline specification is perfect competition among licensees, with a linear demand and constant unit cost.[75] In this case, the deadweight loss is

$$DWL = \frac{1}{2}\Delta_{\wp}\Delta q = \frac{\lambda}{2}\frac{\Delta_{\wp}}{\wp}\tilde{V},$$

where $\wp$ is the initial price (without the royalty associated with the claim), $\Delta_{\wp} = \Delta c$ with perfect competition, $\tilde{V} = q\Delta_{\wp}$ denotes total *royalty payments*, and λ is the elasticity of product demand (in absolute value).[76] To calibrate this expression, we follow Schankerman and Schuett (2022), who estimate the ratio of corporate licensing

[75] In Appendix 3.G, we extend the approach to Cournot competition. Our calibration indicates that this extension yields quantitatively very similar results.

[76] For invalid patents, we cannot use the model estimates of values of patent rights V to represent royalty payments for invalid patents $\tilde{V}$, since our estimates of V are contaminated with potential legal costs (explained in the next subsection). In Appendix 3.G, we explain how we overcome this challenge to calculate type 1 costs.

revenue from intangible industrial property to R&D at 39.3%. Multiplying this ratio by the ratio of R&D to sales in manufacturing in 2002 (4.1%), we take $\frac{\Delta\wp}{\wp} = 1.61\%$. We do the computation for values of the demand elasticity $\lambda \in (1, 3)$ and report $\lambda = 2$ in the main analysis (qualitative conclusions hold for the other values).

Cost of litigation on invalid patents

The social cost of type 1 error also involves litigation costs on invalid patents. Not all invalid patents are "exposed" to litigation because their private value is not large enough to justify the litigation expense. Letting $G_{\tilde{V}}(\cdot)$ denote the distribution of the value at stake $\tilde{V}$, we take the proportion of patents not exposed to litigation from Schankerman and Schuett (2022) ($\check{v} = 89.6\%$) and calculate the $\check{v}^{st}$ percentile of the value at stake distribution, $\check{V} = G_{\tilde{V}}^{-1}(\check{v})$. Then, all patents with $\tilde{V}$ exceeding the threshold $\check{V}$ are exposed to litigation.

The social cost for invalid patents not exposed to litigation is only the deadweight loss from royalties. From Schankerman and Schuett (2022), exposed invalid patents have a 16.3% probability of being litigated, in which case, we assume that courts are perfect and thus always invalidate wrongly granted claims. In this case, the social cost is the sum of litigation costs for the patentee and challenger, each denoted $\mathcal{C}(\tilde{V})$.[77] The remaining 83.7% of exposed invalid patents are not litigated and only impose the deadweight loss.[78]

In summary, the expected social cost of granting an invalid patent of value $\tilde{V}_s$ is

$$S_{1s} = I_s DWL_s + (1 - I_s)\left[0.837 \cdot DWL_s + 0.163 \cdot 2\mathcal{C}(\tilde{V}_s)\right], \tag{3.11}$$

where $I_s = 1(\tilde{V}_s \leq \check{V})$ is an indicator equal to one if the patent is not exposed to litigation. Then, the total type 1 cost is

$$T_1 = \sum_{s \in S_G} E_{1s} S_{1s} \tag{3.12}$$

where E_{1s} is equal to one if a granted application $s \in S_G$ is invalid and zero otherwise.

[77] We take $\mathcal{C}(\tilde{V})$ as linear in $\tilde{V}$ and calibrate the coefficients using AIPLA data.

[78] Patentees with invalid patents can pre-empt a challenge by charging a royalty payment (typically a lump sum) equal to the cost of litigation for the challenger (this is commonly referred to as "trolling" behavior). For these cases, the social cost is only the deadweight loss associated with the patent, since the payment is a pure transfer from the licensee to the patentee (we ignore possible R&D incentive effects of the transfer). See Schankerman and Schuett (2022) for more discussion.

3.8.2 Type 2 Costs

From the ex post perspective, there is *no social cost* from type 2 errors because the innovation has already been produced and the R&D cost is sunk (this is essentially ex post hold-up). Therefore, it only makes sense to analyze the social cost of type 2 errors from the ex ante (incentive) perspective. Type 2 error reduces the expected value of patent protection for the inventor and, thus, the ex ante decision of inventors to develop their (exogenous) ideas. We want to calculate the social value of the set of socially valuable inventions that are *not* developed when there is the possibility of type 2 error but which *would be* developed in the absence of type 2 error. This task requires us to construct a simple model of development. We emphasize that we do not require this extension to estimate the screening model, nor to calculate padding, the number of equilibrium rounds, type 1 and type 2 errors, type 1 social costs, and prosecution costs.

The decision to develop an idea into an invention depends on three things: the ex ante value of patent rights (Γ^*), the value of the invention without patent rights (π), and the development cost (κ). To compute Γ^*, we use our model to calculate the ex ante value of patent rights (net of all costs), as in Equation (3.1). To calculate the private value of the invention without patent rights, we define the patent premium (ξ) as the percentage increase in private value due to patent protection. Hence, for positive Γ^*, by definition $\Gamma^* = \xi\pi$, implying a set of values of π. We assume that the patent premium is constant across inventions and calibrate it based on existing estimates from the literature on patent renewal models (Schankerman, 1998).[79] For the cost of developing an idea into an invention, κ, we draw values from the distribution estimated by Schankerman and Schuett (2022).[80]

An inventor *does not* invest to develop an idea i if

$$ND_i \equiv \mathcal{B}_i - \kappa_i \leq 0,$$

where $\mathcal{B}_i \equiv \pi_i + \max\{\Gamma_i^*, 0\}$ is the private benefit of development. An idea is socially

[79] This is a strong assumption, but it is not feasible to identify π_s if we allow the patent premium to vary. The reason is that we do not have any information on who develops their ideas, which might allow us to back out π from the decision to develop and our estimated value of Γ^*. Furthermore, we must specify π for inventions with negative ex ante value of patent rights. To do this, we draw from the distribution of π created from positive values of patent rights.

[80] An alternative approach is to assume that inventors do not know their development cost, and thus use the mean cost $\bar{\kappa}$. We experimented with this approach and qualitative conclusions are robust.

valuable to develop if the net social benefit of development,

$$S_{2i} \equiv \frac{\rho_{\text{soc}}}{\rho_{\text{priv}}} \mathcal{B}_i - \kappa_i,$$

is positive (where ρ_{priv} and ρ_{soc} denote the private and social rates of return). We use a conservative estimate of $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}} = 2$ from Bloom, Schankerman, and Van Reenen (2013).

Let Υ_0 denote the set of ideas that are socially beneficial to develop ($S_{2i} > 0$) but which are not developed ($ND_i \leq 0$). To calculate type 2 social cost, we compute the subset of Υ_0, which we denote Υ_1, that *would* develop in the absence of type 2 error. To do this, we simulate the outcome from a "counterfactual" patent prosecution where, at the point of patent abandonment, the inventor obtains the value of all valid claims in that patent. By definition, in this scenario, all abandoned claims are invalid, so there is no type 2 error. Let Γ' denote the expected value of patent rights in this new scenario. The idea i would be developed in this scenario if

$$ND_i' \equiv \pi_i + \max\{\Gamma_i', 0\} - \kappa_i > 0.$$

We then compute type 2 costs as

$$T_2 = \sum_{i \in \Upsilon_1} S_{2i}, \tag{3.13}$$

where Υ_1 is the subset of Υ_0 with $ND_i' > 0$. This is the set of ideas that are socially beneficial to develop, that are not developed in the scenario with type 2 error, but that would be developed in the absence of any type 2 error.

3.8.3 Patent Prosecution Costs

The social cost of patent prosecution for each application s consists of two components: applicant legal costs of amending the application each round and Patent Office administrative costs. The amendment cost is the per-negotiation cost $F_{\text{amend},s}$ drawn from the estimated distribution, multiplied by the equilibrium number of negotiations for application s (equal to the number of rounds r_s minus 1). For the administrative cost, we calculate the patent operations budget per application as \$4,117 (in 2018 dollars). This value excludes patent office fees, as these are transfers from the applicant to the patent office, as well as loss in patent value associated with pre-grant obsolescence since that, too, is a transfer from the applicant to the owner of the invention that superseded it. We divide the operations budget per application

by the average number of rounds across all simulations and by the average number of independent claims in an application, to create the average patent office cost per round and claim, denoted by RCC. Then, the total social cost of patent prosecution is

$$T_3 = \underbrace{\sum_s (r_s - 1) F_{\text{amend},s}}_{\text{Applicant Fighting Costs}} + \underbrace{\sum_s M_{0,s} r_s RCC,}_{\text{Office Costs}} \tag{3.14}$$

where $M_{0,s}$ is the initial number of claims in application s.

3.8.4 Benefits of Type 1 and Type 2 Errors

There are also benefits from errors. In the type 1 case, when invalid patents are incorrectly granted, the ex ante incentives for inventors to develop and patent their ideas are increased. This is analogous to the *costs* of type *2* error. We compute these benefits as the sum of social development benefits from welfare-enhancing projects that would not be developed without type 1 error but that are developed with type 1 error.[81] The method is similar to the approach described in Section 3.8.2.

Further, there are benefits from type 2 errors. Not granting valid patents saves the deadweight loss on those patents. We compute these benefits as described in Section 3.8.1. Note that there is no benefit associated with litigation cost savings since, under our assumption of costly but perfect courts (always upholding valid patents and overturning invalid ones), valid patents that are granted would not be challenged.

One important point to note is that the quantification of net social costs in this section is based on the presumption that the patentability threshold used by the Patent Office corresponds to the social optimum, that is, the threshold that only grants patents to inventions that are welfare-enhancing but would not be developed without patent rights. To see this, suppose the threshold is too low (the conventional wisdom) so that some patents are considered "valid" and granted despite not being welfare-enhancing. We would incorrectly not count these as a type 1 error, so they would not contribute to our measure of type 1 social costs. Thus, we would understate type 1 costs (and type 2 benefits). By an analogous argument, we would overstate type 2 costs (and type 1 benefits). Therefore, if the threshold is too low, the consequence is that we would understate net type 1 social costs and overstate

[81] The "counterfactual" patent prosecution in this case is one where, at the point of patent grant, the inventor only obtains the value of the *valid* claims in the patent.

net type 2 social costs. It remains an open and important research question to determine the "optimal" distance threshold, that is, the one that grants patents only to inventions that are welfare-enhancing and not otherwise developed (Schankerman and Schuett, 2022).

3.8.5 Social Costs in Counterfactual Reforms

Table 3.4 summarizes the three components of net social costs for the baseline model and the set of counterfactual reforms.[82] The baseline row approximates the net social costs associated with a yearly cohort of ideas, averaged over 2011–2013 (Appendix 3.G explains how we calibrate the annual number of ideas). Subsequent rows provide the net social costs in that counterfactual scenario. All values are adjusted for inflation, presented in 2023 U.S. dollars.

In the baseline, total type 1 net costs equal \$6.4bn, total type 2 net costs are \$1.5bn, and prosecution costs equate to \$17.6bn. In the final column, we sum these three net costs and estimate the total net social cost of patent screening at \$25.5bn. This total constitutes 6.5% of total R&D performed by business enterprises in the U.S. in 2011.[83]

Introducing a per-round fee lowers type 1 and prosecution costs because it discourages applications and lowers padding for those that do apply. This, in turn, implies that fewer grants are invalid and that grants occur in fewer rounds. However, a round fee increases type 2 costs as applicants are more likely to abandon with some valid claims in a scenario with high negotiation fees. With a \$25,000 round fee, the latter effect dominates, so the total net social cost increases by a very modest 1.9%. As mentioned earlier, for sufficiently large rounds fees (likely to be politically infeasible), the reductions in type 1 and prosecution costs eventually dominate. Further, in these counterfactuals, the extra revenues generated by the fees are *not* reinvested in more intensive or faster examinations. If they were reinvested, social costs from

[82]The table presents the values of net social costs for $\lambda = 2$, $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}} = 2$, $\xi = 0.1$, and development costs drawn. The qualitative conclusions are similar for a range of other parameter values. In Appendix 3.9, we provide results for the cases of a 5% patent premium with $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}$ equal to 1.5 and 2, and a 10% patent premium with $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}$ equal to 1.5. We do not present results for different values of λ because quantitative values in this case are very similar to the baseline.

[83]It is worth noting that this is at the lower end of estimates of the private value of patent rights (Pakes, 1986; Schankerman, 1998). This suggests that the patent system, as it is currently configured, generates net positive social value. For similar findings in a different framework, see Schankerman and Schuett (2022).

TABLE 3.4: Net Social Costs of Patent Prosecution

Counterfactual	T_1	T_2	T_3	**Total**
Baseline ($Bn)	6.4	1.5	17.6	25.5
25K Round Fee	5.9	3.7	16.4	26.0
50K Round Fee	6.1	6.3	15.1	27.5
Three Rounds	4.9	10.1	10.2	25.1
Two Rounds	2.9	15.6	4.7	23.1
One Round	0.0	13.4	0.7	14.1
15% IM	25.8	2.3	15.0	43.0
Credit$\searrow$	6.4	1.5	17.6	25.5
Credit$\searrow$ + 15% IM	15.9	4.0	15.8	35.7

Notes: Equation (3.12) defines T_1; Equation (3.13) defines T_2, respectively; Equation (3.14) defines T_3; Total sums the three kinds of costs. The "baseline" row provides the total social costs in billions of 2023 U.S. dollars.

introducing fees would be mitigated or even converted to social gains.

Restrictions on the allowable number of negotiation rounds have qualitatively similar effects on social costs as rounds fees, but the impacts are much larger. Removing all RCEs (two rounds) yields a 10.4% fall in total social costs relative to the baseline. Restricting the process to one round reduces net social costs by 45%.

Removing intrinsic motivation (down to 15% of its original level) increases the total social cost by 68.6%. When examiners have almost no intrinsic motivation, they are willing to grant applications fast, even if they are padded. As a result, administrative costs fall when intrinsic motivation is removed.[84] However, the grants of patents with invalid claims cause type 1 net costs to triple and consequently lead to an overall rise in net social costs. This finding confirms the critical role that intrinsic motivation plays in this public agency.

Finally, with the baseline level of intrinsic motivation, removing all examiner credits after the first round for one examination has almost no effect on social costs – precisely as we would expect, given the negligible changes to any endogenous variables.

[84]The decrease in prosecution costs is countervailed by the fact that when intrinsic motivation is low, there is an extensive margin increase in the number of inventors applying for patent rights.

In fact, examiners' intrinsic motivation must be as low as 15% of original values for credits to have any effect on net social costs. With 15% intrinsic motivation, type 2 gross (and net) costs, prosecution costs, and type 1 *gross* costs all increase when credits are removed. As a result, when intrinsic motivation is lowered by 85%, removing credits increases total *gross* social costs. Yet, total *net* social costs *decrease*, suggesting that credits are counter-productive even when intrinsic motivation is low. This finding is driven by a large increase in type 1 *benefits* (and hence a decrease in type 1 *net* social costs) from removing credits. This result highlights the importance of accounting for the increased development from relaxed patent granting, as opposed to just the ex post social costs that arise through deadweight losses and litigation.

3.9 Conclusion

In this chapter, we develop and estimate a structural model of the patent screening process. The model incorporates incentives, intrinsic motivation, and multi-round negotiation between the examiner and applicant. We show how structural modeling of the incentives and organization of innovation-supporting public agencies can be used to design reforms to improve agency performance. Our work highlights the fact that, to analyze the impact of reforms on the effectiveness of screening, it is critical to incorporate *both* the agency's decision-making and the endogenous responses of applicants being screened.

Our findings show that patent screening is moderately effective *given* the statutory and judicial standards for patentability within which the Patent Office is required to operate. This effectiveness is driven by substantial intrinsic motivation of examiners. We find that restrictions on the number of allowable rounds of negotiation reduce the social costs of screening. This outcome can be replicated through an equivalent round fee for the applicant, but the required fees are too high to be politically feasible. Finally, we estimate the total net social cost of patent screening at $25.5bn per annual cohort of applications. This figure represents 6.5% of R&D in the United States performed by business enterprises.

This chapter studies patent screening and instruments to improve its effectiveness at the *pooled* technology level. A fruitful extension would be to estimate the model for individual technology fields, such as biotechnology and software, which would allow for the evaluation of the differential effectiveness of various instruments in different areas. More generally, we hope this project illustrates the value of using

structural models to inform decisions on how to reform public agencies, particularly those that affect the allocation of R&D resources, including leading institutions like the National Institutes of Health and National Science Foundation, and similar institutions in other countries.

Appendices for Chapter 3

3.A Additional Tables and Figures

TABLE 3.A.1: Summary Statistics

Variable	Observations	Mean	Median	Std. Dev.
Issued	4,846,053	0.70	1.00	0.46
Duration of Prosecution (years)	4,846,053	2.96	2.67	1.57
Number of Rounds	4,608,833	2.40	2.00	1.45
Independent Claims	3,838,553	2.99	3.00	2.94
Small Entity	4,781,012	0.24	0.00	0.43
Not Renewed at 4	410,667	0.13	0.00	0.33
Renewed at 4, not at 8	410,667	0.19	0.00	0.39
Renewed at 8, not at 12	410,667	0.23	0.00	0.42
Renewed at 12	410,667	0.46	0.00	0.50

Notes: Sample sizes are lower for rounds, claims, and examiner variables since the datasets containing these variables cover a subset of the years 2001-2017. On renewal variables, we restrict attention to patents granted before 2006 to ensure that we have full renewal data on all granted patents. Categorical variables may not sum to one due to rounding.

TABLE 3.A.2: Estimated and Assigned Parameters

Estimated Parameters			
Variable	**Notation**	**Distribution**	**Parameters**
Examiner			
Intrinsic motivation	$\theta \sim G_{S,\theta}(\cdot)$	Log-normal	σ_θ, $\mu_{\theta,\text{junior}}$ or $\mu_{\theta,\text{senior}}$
Examiner Delay Cost	$\pi \sim G_\pi(\cdot)$	Log-normal	μ_π, σ_π
Error	$\varepsilon \sim G_{e,\varepsilon}(\cdot)$	Normal	σ_ε
Applicant			
Initial claim returns	$v_j^* \sim G_v(\cdot)$	Log-normal	μ_v, σ_v
Initial claim distances	$D_j^* \sim G_D(\cdot)$	Beta	α_D, β_D
Obsolescence	$\boldsymbol{\omega}$	Bernoulli	$P_{\omega,\text{pre}}$ or $P_{\omega,\text{post}}$
Application legal costs	f_{app}	Log-normal	$\mu_{f,\text{app}}, \sigma_{f,\text{app}}$
Issuance legal costs	f_{iss}	Log-normal	$\mu_{f,\text{iss}}, \sigma_{f,\text{iss}}$
Maintenance legal costs	f_{main}	Log-normal	$\mu_{f,\text{main}}, \sigma_{f,\text{main}}$
Amendment legal costs	f_{amend}	Log-normal	$\mu_{f,\text{amend}}, \sigma_{f,\text{amend}}$
Narrowing	η	-	-

Assigned Parameters		
Variable	**Notation**	**Values**
Discount rate	β	0.95
Depreciation	δ	$\frac{0.14-P_{\omega,\text{post}}}{1-P_{\omega,\text{post}}}$
Threshold by technology center	τ	Range from 0.48 to 0.52
Credits	$g^r(S,T)$	-
Finalizing fee	ϕ	\$2,268
RCE fees	$F^3_{\text{round}} = F^5_{\text{round}}$	\$1,034
	F_4	\$1,685
Renewal fees	F_8	\$3,791
	F_{12}	\$7,792

TABLE 3.A.3: Application Fighting Costs by Technology Area

Parameter	**Symbol**	**Estimate**	**S.E.**
Chemical application fighting cost log-mean	$\mu_{f,\text{chem}}$	9.15	0.008
Chemical application fighting cost log-sigma	$\sigma_{f,\text{chem}}$	0.38	0.010
Electrical application fighting cost log-mean	$\mu_{f,\text{elec}}$	9.18	0.010
Electrical application fighting cost log-sigma	$\sigma_{f,\text{elec}}$	0.57	0.014
Mechanical application fighting cost log-mean	$\mu_{f,\text{mech}}$	9.02	0.008
Mechanical application fighting cost log-sigma	$\sigma_{f,\text{mech}}$	0.47	0.011

Notes: Standard errors are bootstrapped.

TABLE 3.A.4: Applicant Fighting Costs by Technology Area

Parameter	**Symbol**	**Estimate**
Simple amendment fighting cost log-mean	$\mu_{f,\text{amend,simp}}$	7.60
Simple amendment fighting cost log-sigma	$\sigma_{f,\text{amend,simp}}$	0.37
Chemical amendment fighting cost log-mean	$\mu_{f,\text{amend,chem}}$	8.13
Chemical amendment fighting cost log-sigma	$\sigma_{f,\text{amend,chem}}$	0.45
Electrical amendment fighting cost log-mean	$\mu_{f,\text{amend,elec}}$	8.07
Electrical amendment fighting cost log-sigma	$\sigma_{f,\text{amend,elec}}$	0.38
Mechanical amendment fighting cost log-mean	$\mu_{f,\text{amend,mech}}$	7.95
Mechanical amendment fighting cost log-sigma	$\sigma_{f,\text{amend,mech}}$	0.43
Issuance cost log-mean	$\mu_{f,\text{iss}}$	6.54
Issuance cost log-sigma	$\sigma_{f,\text{iss}}$	0.62
Maintenance cost log-mean	$\mu_{f,\text{main}}$	5.67
Maintenance cost log-sigma	$\sigma_{f,\text{main}}$	0.46

TABLE 3.A.5: Robustness of Estimates

Parameter	**Symbol**	**Baseline**	1% τ	5% τ	$\beta = 0.99$	**Definition of Seniority (GS13 + GS14)**
Junior intrinsic motivation log-mean	$\mu_{\theta,\text{j}}$	3.92	3.96	3.96	3.90	4.16
Senior intrinsic motivation log-mean	$\mu_{\theta,\text{s}}$	3.38	2.90	2.73	3.18	2.93
Intrinsic motivation log-sigma	σ_θ	0.77	0.82	0.79	0.90	0.99
Examiner delay cost log-mean	μ_π	0.19	0.16	0.18	0.49	0.12
Examiner delay cost log-sigma	σ_π	0.27	0.37	0.42	0.10	0.60
Error standard deviation	σ_ε	0.02	0.02	0.02	0.03	0.02
Initial returns log-mean	μ_v	10.55	10.59	10.88	10.07	10.28
Initial returns log-sigma	σ_v	1.32	1.13	1.61	2.94	0.57
Initial distance alpha	α_D	4.57	3.92	3.90	4.56	3.75
Initial distance beta	β_D	7.74	6.72	6.22	7.79	7.15
Narrowing probability	η	0.75	0.73	0.74	0.75	0.72
Application obsolescence probability	$P_{\omega,\text{pre}}$	0.14	0.13	0.13	0.12	0.14
Renewal obsolescence probability	$P_{\omega,\text{post}}$	0.04	0.04	0.04	0.04	0.04
Simple application fighting cost log-mean	$\mu_{f,\text{simple}}$	8.53	8.43	8.56	8.60	8.53
Simple application fighting cost log-sigma	$\sigma_{f,\text{simple}}$	0.87	0.97	0.79	0.74	0.95
SMM Objective		1.23	1.47	1.29	1.25	1.33

Notes: This table provides estimates of the model parameters across various model alternatives. The baseline model defines senior examiners as those at the GS14 level. The last column expands this to include GS13 and GS14.

TABLE 3.A.6: Net Social Costs of Patent Prosecution: Robustness

	Patent Premium (ξ) = 0.10				Patent Premium (ξ) = 0.05					
Counterfactual	$\boldsymbol{T_1}$	$\boldsymbol{T_2}$ **(1.5)**	$\boldsymbol{T_3}$	**Total**	$\boldsymbol{T_1}$	$\boldsymbol{T_2}$ **(1.5)**	$\boldsymbol{T_2}$ **(2.0)**	$\boldsymbol{T_3}$	**Total (1.5)**	**Total (2.0)**
Baseline ($Bn)	6.4	0.7	17.6	24.7	6.6	0.0	0.2	20.6	27.2	27.4
25K Round Fee	5.9	1.8	16.4	24.1	6.3	0.7	1.4	19.1	26.1	26.8
50K Round Fee	6.1	3.1	15.1	24.2	5.5	1.7	3.5	17.1	24.7	26.1
Three Rounds	4.9	4.8	10.2	19.8	5.4	1.9	3.9	11.5	18.8	20.8
Two Rounds	2.9	7.4	4.7	14.9	2.9	3.2	6.6	5.2	11.4	14.8
One Round	0.0	6.3	0.7	7.0	0.0	1.6	3.3	0.8	2.4	4.1
15% IM	29.0	1.1	15.0	45.1	31.6	0.4	0.8	17.3	50.1	49.8
Credit$\searrow$	6.4	0.7	17.6	24.7	6.5	0.0	0.2	20.6	27.2	27.3
Credit$\searrow$ + 15% IM	24.3	1.9	15.8	42.0	23.7	0.7	1.5	18.2	47.8	43.3

Notes: This table provides the values of net social costs for alternative values of the patent premium and social multiplier. Columns denoted $\boldsymbol{T_2}$ **(1.5)** and $\boldsymbol{T_2}$ **(2.0)** provide values of type 2 net social costs when $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}$ is equal to 1.5 and 2.0, respectively. Columns **Total (1.5)** and **Total (2.0)** provide the total net social costs when $\frac{\rho_{\text{soc}}}{\rho_{\text{priv}}}$ is equal to 1.5 and 2.0, respectively.

FIGURE 3.A.1: Match of internal data and model moments

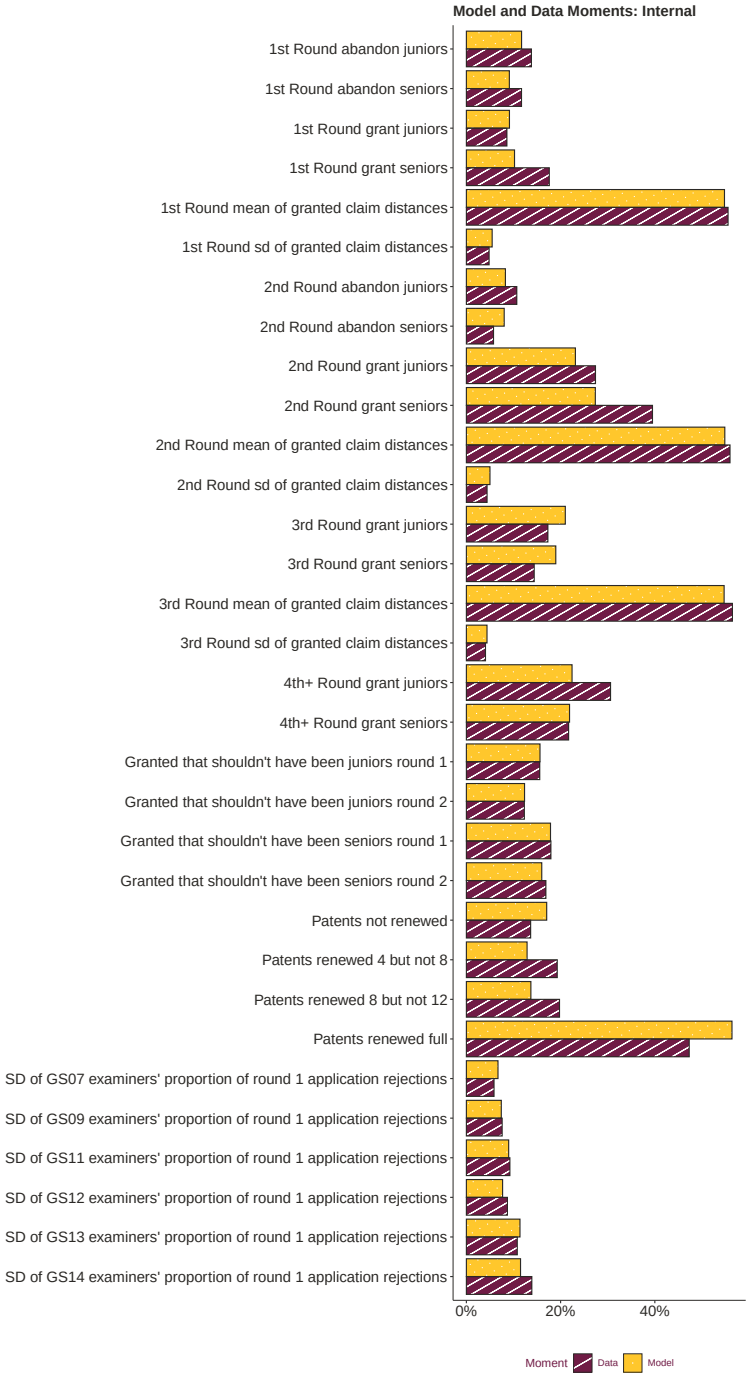
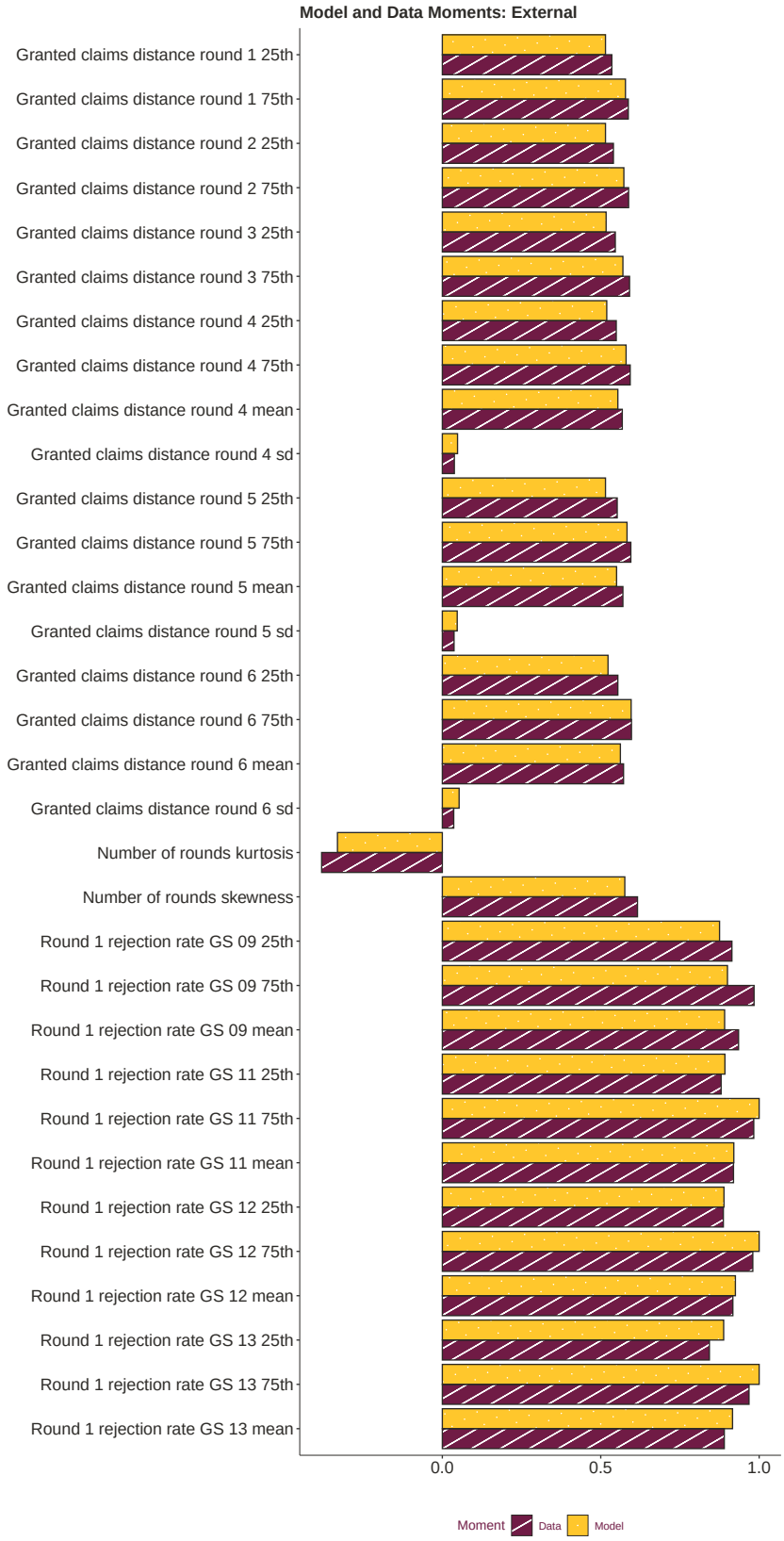


FIGURE 3.A.2: Match of external data and model moments



3.B Data Sources

If the links are broken, the documents are available upon request.

3.B.1 Publicly Available Datasets

1. *U.S.PTO Patent Application Claims Full Text Dataset* and *U.S. PTO Patent Claims Full Text Dataset*: https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-claims-research-dataset
2. *Patent Examination Research Dataset*: https://www.uspto.gov/ip-policy/economic-research/research-datasets/patent-examination-research-dataset-public-pair
3. *U.S.PTO Maintenance Fee Events Dataset*: https://developer.uspto.gov/product/patent-maintenance-fee-events-and-description-files
4. *U.S.PTO Office Action Research Dataset*: https://www.uspto.gov/ip-policy/economic-research/research-datasets/office-action-research-dataset-patents
5. *Frakes and Wasserman (2019)*: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ABE7VS

3.B.2 Data from Public Documents

6. GDP Deflator: https://fred.stlouisfed.org/series/GDPDEF.
7. *AIPLA Report of the Economic Survey*: See https://www.aipla.org/detail/journal-issue/economic-survey-2017 for 2017.
8. Industry concentration: https://www.census.gov/content/dam/Census/programs-surveys/economic-census/data/archived_tables/2007/sector31/2007_31-33_Con_Ratios_US.zip.
9. Patent Office fees: https://www.govinfo.gov/content/pkg/CFR-2011-title37-vol1/pdf/CFR-2011-title37-vol1.pdf or from https://www.uspto.gov/sites/default/files/aia_implementation/AC54_Final_Table_of_Patent_Fee_Changes.pdf.
10. Patent operations costs:

    2005: https://www.uspto.gov/sites/default/files/about/stratplan/ar/USPTOFY2005PAR.pdf

2010: https://www.uspto.gov/sites/default/files/about/stratplan/ar/USPTOFY2010PAR.pdf
2015: https://www.uspto.gov/sites/default/files/documents/USPTOFY15PAR.pdf

11. Patent applications: https://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm.

12. R&D expenditures: https://www.nsf.gov/statistics/infbrief/nsf14307/.

3.C Distance Measure

This section provides details on how we construct our patent distance metric. We describe our preferred choice, the paragraph vector approach.[85] The method consists of four steps: (1) standardizing the independent claim text, (2) turning the text into a numerical vector, (3) calculating the distances between a focal patent claim on an application to all existing granted patent claims and (4) calculating the distance to the closest existing independent claim.

The first step before converting text into a numerical vector is text standardization. We perform basic changes to the content of the text and remove words that carry no informational content. Once we standardize the text, we drop any claims with fewer than two words or illegible text.

We use the paragraph vector approach to represent the text of a patent claim as a numerical vector. The paragraph vector approach is an improvement of the word vector approach. We implement the Paragraph Vector approach using Gensim's `Doc2Vec` Python model (Řehůřek and Sojka, 2010).

The step above converts all patent claims, including those on applications and those granted, into a numerical vector. The next step involves taking every focal application patent claim vector and calculating its distance to every *existing* granted claim at the point of application. After representing a patent claim's text as a numerical vector, we use cosine similarity and angular distance, both of which are standard in the text matching and the NLP literature. We compute the cosine similarity (CS)

[85] At the time of writing this chapter, we used the state-of-the-art approach, but there is a fast-moving frontier. The most recent approaches use GPT-4 or BERT word embeddings integrated directly into Neural Networks. See Elliot and Hansen (2023) for details on text algorithms.

between claim text vectors x and y as

$$cs(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_j y_j^2}}.$$

Then, we calculate the angular distance (AD) metric, $AD(x,y) = \arccos(cs(x,y))/\pi$ and then double AD to obtain a normalized distance in the interval $[0,1]$.

With all distances computed, it is a simple step to find the closest 50 claims to each application. We experiment with different choices on which percentile of the closest 50 distances to use. We also experimented with taking an average of the five closest distances for example, and the resulting distances were similar.

3.D Descriptive Results

We show how patent application outcomes vary with technology center and examiner seniority. First, we regress a binary variable equal to one if the application process lasts more than one round against fixed effects for examiner seniority grade, technology center, year of application, and a small entity indicator (applying firm having fewer than 500 employees). The results in Column (1) of Table 3.D.1 reveal substantial variation across technology centers; e.g., Computer Networks (TC-24) has a 12 percentage point higher likelihood of multi-round negotiation than the reference category, Biotechnology (TC-16). Further, the likelihood of any negotiation decreases with the seniority of the examiner, with senior (GS-14) examiners nine percentage points less likely to require negotiation relative to the most junior, holding technology center and application year fixed. Further, small entities are 12 percentage points less likely to negotiate (all else fixed).

In Column (2), we do the same analysis for the dependent variable equal to one if the examiner grants a patent. We match the findings of Frakes and Wasserman (2017) – senior examiners are more likely to grant and grant rates vary substantially across technology centers. In our model, we explain this variation by letting the distribution of intrinsic motivation vary with seniority level, by incorporating differences in the credit structure for examiners that vary across seniority and technology centers, and by allowing fighting costs to differ for applicants, with technology category-specific distributions. Our parameter estimates enable us to disentangle the effects of these factors in explaining the variation in outcomes, as we discuss in the text.

Table 3.D.1: Regression Results

Variable	(1) Negotiation	(2) Grant
Intercept	0.7433 (0.006)	0.542 (0.005)
GS-7	-0.002 (0.004)	0.003 (0.005)
GS-9	-0.016 (0.004)	0.035 (0.004)
GS-11	-0.020 (0.004)	0.066 (0.004)
GS-12	-0.034 (0.004)	0.092 (0.004)
GS-13	-0.045 (0.004)	0.126 (0.004)
GS-14	-0.091 (0.004)	0.178 (0.004)
Chemicals (17)	0.064 (0.002)	0.067 (0.002)
Comp. Software (21)	0.105 (0.002)	0.196 (0.002)
Comp. Networks (24)	0.123 (0.002)	0.192 (0.002)
Communications (26)	0.047 (0.002)	0.198 (0.002)
Electronics (28)	-0.010 (0.001)	0.244 (0.001)
Other (36)	0.065 (0.002)	0.136 (0.002)
Mech Engineering (37)	0.042 (0.002)	0.139 (0.001)
Small Entity	-0.120 (0.001)	-0.170 (0.001)
Year FE	Yes	Yes
N	1,641,333	1,759,313

Notes: Omitted grade is GS-5 and omitted technology center is Biotechnology and Organic Fields (16). Technology center "Other" refers to Center 3600, which is "Transportation, Electronic Commerce, Construction, Agriculture, Licensing and Review." Following Frakes and Wasserman (2017), we omit GS-15 grade examiners. We report heteroskedasticity robust (HC1) standard errors in parentheses.

These results show stark differences in average grant rates and likelihood of negotiation across technology centers and examiner seniority grades. Next, we investigate the variation in *examiner-specific* decisions within and between seniority grades and technology center pairs. To do this, we calculate examiner-specific outcomes (average grant rates, number of rounds, length of examination period, probability of negotiation, etc.) within each seniority grade examiners are in at the time. We decompose the variation in these examiner averages into within and between seniority grade-technology center pairs by introducing dummies for each seniority-grade-technology-center dyad in Table 3.D.2. The proportion of within-group variation in examiner

TABLE 3.D.2: ANOVA Results

Variable	**Grade × TC Fixed Effects**
Grant rate	79.84
Duration of examination (years)	75.79
No negotiation (one round)	89.53
Independent claims granted	74.93

Notes: For each variable y, and an examiner e when they are in seniority grade S and technology center T, we calculate $\bar{y}_{eST}$. Then we regress $\bar{y}_{eST}$ on a set of interactive dummies for seniority grade and technology center. We report $1 - R^2$ (as a percentage) for these regressions, thereby providing the proportion of within group variation.

grant rates is 80%, implying substantial variation in examiner grant rates not explained by seniority and technology centers. Our model explains this variation in examiner-specific grant rates within the technology center and seniority groups by incorporating group-specific *distributions* of examiner intrinsic motivation and costs of delay.

Here we provide expressions for $g^r_{GR}(S,T)$, $g^r_{ABN}(S,T)$, $g^r_{RCE}(S,T)$ and $g^r_{REJ}(S,T)$. For $y \in \{GR, ABN, REJ, RCE\}$, we write $g^r_y(S,T) = \nu^r_y \cdot c(S,T)$, and give expressions for ν^r_y and $c_(S,T)$ separately.

3.E.1 Credits

Granting in the first round gives the examiner a payoff of $\nu^1_{GR} = 2$ credits. Rejecting in the first round gives $\nu^1_{REJ} = 1.25$. If the applicant abandons in round one, the examiner obtains $\nu^1_{ABN} = 0.75$. Granting in the second round gives $\nu^2_{GR} = 0.75$ credits. Rejecting in the second round gives $\nu^2_{REJ} = 0.25$ credits, with an extra $\nu^2_{ABN} = \nu^2_{RCE} = 0.5$ credits whether the applicant abandons or continues to an RCE. Ultimately, the examiner obtains two credits irrespective of what happens in the first two rounds. The only difference is whether they obtain the credits immediately (say, from an immediate grant) or spread out over two rounds.

The structure of the payoffs in the first RCE are the same, except $\nu^3_{REJ} = 1$ and $\nu^3_{GR} = 1.75$. In this case, irrespective of what happens in the RCE, the examiner will obtain 1.75 credits. The difference comes from whether they receive all 1.75 credits at once by granting, or 1 credit from their non-final rejection and $\nu^4_{REJ} = 0.25$ plus $\nu^4_{ABN} = \nu^4_{RCE} = 0.5$ credits from the applicant's response.

TABLE 3.E.1: Seniority Corrections

Seniority Grade	**Signatory Authority**	$c_{SEN}(S)$
GS-5	None	0.55
GS-7	None	0.7
GS-9	None	0.8
GS-11	None	0.9
GS-12	None	1.0
GS-13	None	1.15
GS-13	Partial	1.25
GS-14	Partial	1.25
GS-14	Full (primary examiner)	1.35

Notes: This table provides the seniority factors for credit adjustment. In the empirical work, we use 1.15 for GS-13 and 1.25 for GS-14.

In the second and any subsequent RCEs, the structure of the payoffs is still the same, except $\nu_{REJ}^{2r+1} = 0.75$ and $\nu_{GR}^{2r+1} = 1.5$ ($r > 1$). As before, the examiner will receive 1.5 credits from second and subsequent RCEs. The difference comes from whether they receive all 1.5 credits at once from granting, or 0.75 credits from their non-final rejection and $\nu_{REJ}^{2r+2} = 0.25$ plus $\nu_{ABN}^{2r+2} = \nu_{RCE}^{2r+2} = 0.5$ credits from the applicant's response.

3.E.2 Seniority and Technology Complexity Adjustments

The seniority and technology complexity adjustment term is

$$c(S,T) = \frac{c_{TECH}(T)}{c_{SEN}(S)}.$$

Table 3.E.1 gives the values of $c_{SEN}(S)$ across the GS categories. Higher seniority factors imply larger values of c_{SEN}, and therefore lower values of credits. Table 3.E.2 gives the values of $c_{TECH}(T)$ we created for the different technology centers and use in the estimation of the model. The Patent Office does not have adjustments at the technology center level, but rather at the more detailed U.S. Patent Class (USPC) level. We obtained the adjustments at the USPC level from the Patent Office and constructed a patent-application weighted average for each technology center.

TABLE 3.E.2: Technology Center Adjustments

Technology Center T	**U.S.PTO Number**	**Correction** ($c_{TECH}(T)$)
Chemical and Materials Engineering	17	22.2
Computer Architecture Software and Information Security	21	31
Computer Networks, Multiplex, Cable and Cryptography/Security	24	29
Communications	26	26.5
Semiconductors, Electrical and Optical Systems and Components	28	21.4
Transportation, Electronic Commerce, Construction, Agriculture...	36	22.4
Mechanical Engineering, Manufacturing and Products	37	19.9

3.F Moment Selection and Identification Intuition

First, we provide further details on the possible moments we could use to estimate our model. Then, we provide some information on our methods to prune moments from the full set. Finally, we provide some intuition on how the moments identify the model parameters.

3.F.1 Available Moments

We have seven sets of moments available, which we describe in turn.

Our first group of moments corresponds to examiners' issuance and applicants' abandonment decisions. For each round in the model and each seniority level, we calculate the proportion of applications examiners grant and the proportion that applicants abandon. Since there are nine seniority grade-signatory authority pairs, and we observe at least six rounds, this implies at least 108 moments on grants and abandonments.

Second, we observe the distribution of the proportion of claims rejected, both by round (six) and by seniority grade-signatory authority pair (nine). These observations generate another 54 moments. Third, we observe the proportion of granted patents that renew at four, eight, and twelve years after issuance. These observations generate four moments on patent renewals (don't renew at four, renew at four but not eight, renew at eight but not twelve and renew at twelve).

Fourth, we calculate the distribution of claim distances by round. We calculate the mean and standard deviation of the distance distribution by round for at least six rounds, implying at least 12 moments on distance. Another moment comes from the within-application distance correlation. Fifth, at each of the nine seniority grades, we calculate each examiner's *leniency*, which is their average rejection rate across all

the applications they examine. Hence for each seniority grade-signatory authority pair, we obtain a distribution of examiner rejection rates, for which we can calculate the mean and standard deviation of the distribution of examiner fixed effects. From this we obtain another 18 moments.

Next, given that we can identify the distance threshold externally, we calculate the proportion of granted patents containing at least one invalid claim (that is, a claim whose distance is below the distance threshold). Hence, for each round and each seniority level, we calculate the proportion of patents granted containing an invalid claim, implying another 54 moments.

Finally, we observe the distribution of application fighting costs. We have six moments on the distribution of legal application fees for four technology categories (simple, chemical, electrical and mechanical), which we match to the technology centers on which we estimate the model. This implies another 24 moments.

3.F.2 Choosing Moments

We have more than two hundred data moments that we can calculate from endogenous variables in the model. Since we have 21 model parameters to estimate with simulated method of moments, in principle, we are over-identified. However, not all moments will aid the estimation procedure in identifying the parameters, so we begin by pruning the set of moments for estimation.

We follow a rigorous, data-driven methodology to create a subset of the moments that best estimate the parameters. To do this, we calculate the sensitivity matrix described in Andrews, Gentzkow, and Shapiro (2017). As the authors explain, "sensitivity gives a formal, quantitative language in which to describe the relative importance of different moments for determining the value of specific parameters." If a moment had a small value in the sensitivity matrix for all parameters, we considered it as not useful in estimating our model. Further, as described in Jalali, Rahmandad, and Ghoddusi (2015), for each parameter and moment, we plot the value of the moment for different values of the parameter, fixing the other parameters at their estimates. If this curve is flat, this parameter does not influence on the value of the moment. For a given moment, if the curve is flat across *all* parameters, it suggests that the moment offers no useful variation to identify the parameters.

For each parameter, we also plot the value of the SMM objective across all values of the parameter, fixing other parameters at their estimates. Ideally, the SMM will

be U-shaped in each parameter to ensure a well-defined global minimum exists. By doing this, we learn how well we pin down parameters based on the set of moments we have available.

By combining the sensitivity matrix with moment and SMM plots, we pruned the set of moments down to those that offer some assistance in estimating the parameters. Since we split many parameters into two seniority groups (junior and senior), we split some of our moments into the same seniority categories.

3.F.3 Full Set of Moments

The full set of moments we use for estimation is as follows. The selected moments corresponding to outcomes for examiners are:

(i) The proportion of applications granted in each round for juniors and seniors, for rounds one, two, three, and all rounds after four combined [eight moments]

(ii) The standard deviation of the distribution of examiner rejection rates for the six seniority categories used by the Patent Office (GS levels 7, 9, 11, 12, 13, and 14) [six moments]

(iii) The proportion of patents granted containing an invalid claim (for juniors and seniors) for rounds one and two [four moments]

The moments corresponding to outcomes for applicants are:

(i) The proportion of abandonments in each round, when the assigned examiner is junior and senior, for rounds one and two [four moments]

(ii) The proportion of granted patents not renewed, renewed at year four but not eight, renewed at year eight but not twelve, and renewed at year twelve [four moments]

(iii) The mean and standard deviation of the distribution of granted claim distances for rounds one, two, and three [six moments]

(iv) Mean and median of legal application fees for simple applications and complex applications in electrical, mechanical, and chemical technologies [eight moments]

3.F.4 Identification

A model is either point identified or not, and technical conditions on the required variation in exogenous variables determine whether a model is identified (Andrews,

Gentzkow, and Shapiro, 2017). Due to our model's complicated and nonlinear nature, we cannot calculate these conditions. Identification with simulated method of moments is based on how different moments are affected by specific parameters. While we cannot identify this link exactly, we provide some intuition of how moments aid in pinning down specific parameters of the model.

We start with the parameters relating to the applicant. The renewal rates, together with first-round abandonment decisions, aid in identifying the parameters of the distribution of flow returns, i.e., μ_v and σ_v. This is because, all else equal, an applicant with higher returns is less likely to abandon after learning their examiner and more likely to renew their patent, conditional on being granted. The renewal moments also aid in identifying the post-grant obsolescence probability $P_{\omega,\text{post}}$. Similarly, the ex post claim distribution of padded distances, as calculated using the distance between text vectors, aids in identifying the parameters of the distribution of ex ante unpadded distance, i.e., α_D and β_D. Moments on application fighting costs directly pin down the distribution of application fighting costs, $\mu_{f_{app}}$, and $\sigma_{f_{app}}$.

Regarding pre-grant obsolescence $P_{\omega,\text{pre}}$, the only case in which an applicant abandons in interim rounds two to four is when they become obsolete. If an applicant, upon learning their examiner calculates that they will want to abandon in any round after the first, they will abandon immediately in round one. Therefore, interim round abandonments offer substantial assistance in identifying the obsolescence probability in the application process.

Intuition for examiner parameters is more complicated. Observing that examiners grant several invalid patents could result from low intrinsic motivation, high examiner error, or high examiner delay costs. Three factors make this challenge less formidable. First, since we assume that only intrinsic motivation varies by seniority, differences in grant rates and examiner errors by seniority pick up the value of intrinsic motivation, μ_{im} by seniority, and differences in the variation in examiner-specific grant rates by seniority capture the variation in intrinsic motivation, σ_θ by seniority.

Second, we assume that each examiner has the same delay cost across all applications and rounds but faces varying intrinsic motivation costs at each round of every application (because $\mathcal{R}^r$, the proportion of invalid independent claims varies across rounds and applications). This implies that the proportion of invalid patents granted in rounds one and two offer the best assistance in identifying the mean examiner intrinsic motivation and mean examiner delay costs. Third, examiner error is two-sided and symmetric. This feature creates cases where examiners do not grant valid

patents, whereas intrinsic motivation and delay costs only incentivize examiners to grant when they should not. Otherwise, we know that an examiner, making no mistake, and facing a fully valid patent, will always issue it. Together, this implies that we can use the residual variation in grant rates (valid and invalid) by round and seniority to learn about the distribution of examiner error.

3.F.5 Details on Model Fit

As shown in Figure 3.A.1, we match most of the internal moments well, though there are two exceptions. The first is the proportion of fully renewed patents, which we overestimate. The other exception is the second-round grant rate. This moment is difficult to match with our model because examiners have incentives to wait until the third round and obtain RCE credits if they do not choose to grant in the first round. Since examiners have incentives and targets across applications on their desks (docket management), they are more likely to grant in the second round than our baseline model predicts.

3.G Quantification of Social Costs

3.G.1 Implementing Type 1 Social Cost Calculation

As indicated in the text, a key challenge in implementing our calculation of type 1 social costs comes from the fact that the estimates of the value of patent rights for invalid patents include potential litigation costs. To impute the "value at stake" in litigation for these patents, we need to adjust our methodology to exclude these costs.

To do this, we make two assumptions:

A1: Valid patents are not litigated. This assumption holds in a model with perfect courts, where a competitor knows (or can pay a fee to discover) whether a patent is valid or not, and then choose whether to litigate based on the result.[86] This assumption allows us to calculate the value of patent rights for valid

[86]This assumption is *not* at odds with Schankerman and Schuett (2022), where *high types* are litigated with some probability even though they will not be invalidated. The important point is that high types in their model (patents that would not be developed without patent rights) are not the same as valid patents in our model, which are defined as those with distance larger than the threshold.

patents, $\tilde{V}$, as equal to the observed value since there are no litigation costs to net out.

A2: The *distribution* of the value at stake, $G_{\tilde{V}}(\cdot)$, is the same for valid patents as invalid patents. The basis for this assumption is that initial distances and values are uncorrelated in the model. This assumption allows us to draw values from the observed distribution of $\tilde{V} = V$ for valid patents and use them as draws from the distribution of $\tilde{V}$ for invalid patents.

Given A1 and A2, the procedure for calculating type 1 social costs is as follows:

1. Estimate the parameters of a log-normal distribution for the value at stake for *valid* patents.[87] Let the estimated distribution be denoted as $\hat{G}_{\tilde{V}}(\cdot)$.
2. Let $\bar{P}$ be the total number of *invalid* patent grants for the given period we simulate. Then, for each $p = 1, \ldots, \bar{P}$:
   (a) Take a draw from the estimated distribution of *valid* patents' value at stake (ex post value), $\hat{G}_{\tilde{V}}(\cdot)$, to represent the value at stake for the invalid patent p
   (b) Using the draw, calculate S_{1p} from Equation (3.11).
3. Calculate the total social cost of type 1 error as $\sum_{p=1}^{\bar{P}} S_{1p}$.

Finally, note that we calculate the threshold for exposure to litigation from the *empirical* distribution of the value at stake for valid patents, $\hat{G}_{\tilde{V}}(\cdot)$.

3.G.2 Implementing Type 2 Social Cost Calculation

The primary challenge in implementing our calculation of type 2 social costs comes from calibrating the value of the invention without patent rights (π), particularly for inventions with $\Gamma^* \leq 0$, where we cannot use the patent premium. In a similar vein to our approach to type 1 social costs, we assume that the distributions of π for those with positive and negative Γ^* are the same and then draw values of π from this distribution for those inventions.

To be precise, our specific implementation is as follows:

[87] The sum of log-normal distributions is approximately log-normal (Dufresne, 2004), which our simulation here exhibits.

1. Draw a pilot set of potential inventions, used to calculate a distribution of π. Run these set of potential inventions through the model and calculate Γ^*. For those with positive Γ^*, create a distribution of π using the relationship $\Gamma = \xi\pi$.

2. Now start the simulation for type 2 social costs by drawing a new set of potential inventions (returns, distances, number of claims, fighting costs, examiner etc.). For each potential invention i, calculate Γ_i^*. If $\Gamma_i^* > 0$, calculate $\pi_i = \dfrac{\Gamma_i^*}{\xi}$. If $\Gamma_i^* \leq 0$, draw a value of π_i from the distribution calculated in 1. Also, draw a development cost κ_i.

3. For each of the potential inventions i, work out the set $i = 1, \ldots, \mathcal{I}_{\text{no dev}}$ that do not develop as those with $\max\{\Gamma_i^*, 0\} + \pi_i < \kappa_i$

4. For $i = 1 \ldots, \mathcal{I}_{\text{no dev}}$, run the potential invention through a model where, at the point of abandonment, the inventor obtains all valid claims they have, and so obtains the patent value of their valid claims, instead of a payoff of 0. By definition, this scenario has the property that all abandoned claims are invalid, so that there is no type 2 error. Let Γ_i' denote the expected value of patent rights in this new scenario.

5. For $i = 1 \ldots, \mathcal{I}_{\text{no dev}}$, calculate the set $i = 1, \ldots, \mathcal{I}_{\text{now dev}}$ who have $\max\{0, \Gamma_i'\} + \pi_i \geq \kappa_i$. This is the set who do not develop because of type 2 error but do develop in the absence of type 2 error.

6. For $i = 1, \ldots, \mathcal{I}_{\text{now dev}}$, calculate $S_{2i} = \dfrac{\rho_{\text{soc}}}{\rho_{\text{priv}}}\Big(\max\{0, \Gamma_i\} + \pi_i\Big) - \kappa_i$ and calculate the total type 2 social cost as

$$T_2 = \sum_{i=1}^{\mathcal{I}_{\text{now dev}}} S_{2i}.$$

3.G.3 Calibrating Deadweight Loss

In the derivation of deadweight loss, note that

$$DWL = \frac{1}{2}\Delta_{\wp}\Delta q = \frac{1}{2}\frac{\Delta q}{q} q \Delta_{\wp} = \frac{\lambda}{2}\frac{\Delta_{\wp}}{\wp}\tilde{V},$$

by the definitions of $\tilde{V}$ and λ. Further, note that

$$\frac{\Delta_{\wp}}{\wp} = \frac{q\Delta_{\wp}}{q\wp} = \frac{\text{lic. rev}}{\text{sales}} = \frac{\text{lic. rev}}{\text{R\&D}} \cdot \frac{\text{R\&D}}{\text{sales}}$$

As described in the text, we use Schankerman and Schuett (2022) for the ratio of licensing revenue to R&D, and data from the Bureau of Economic Analysis for the ratio of R&D to sales.

3.G.4 Deadweight Loss Under Cournot Competition

In the main text, we compute deadweight loss from a patented invention assuming symmetric licensees operate in a perfectly competitive industry. Suppose instead that the licensees compete in a Cournot setting. By standard calculations, the equilibrium price-cost margin is $\frac{\wp - c}{\wp} = \frac{m^*}{\lambda}$ where $m^* = \frac{1}{N}$ is the average market share and λ is the demand elasticity. We write this as $\frac{\wp - c}{\wp} = \frac{H^e}{\eta}$ where H^e is the symmetric-equivalent Herfindahl index of concentration. Thus for $H^e < 1$

$$\wp = \frac{c}{1 - \frac{H^e}{\lambda}}.$$

With imperfect competition, the change in equilibrium price is larger than the Arrow royalty due to double marginalization: $\Delta\wp = \frac{\Delta c}{1 - \frac{H^e}{\lambda}} > \Delta c$. The associated deadweight loss with Cournot competition is

$$DWL_{\text{cournot}} = \frac{1}{2}\Delta\wp\Delta q = \frac{1}{2}\frac{\Delta c}{1 - \frac{H^e}{\lambda}}\Delta q = DWL_{\text{pc}} \cdot \frac{1}{1 - \frac{H^e}{\lambda}},$$

where it should be noted that in this case $\tilde{V} = q\Delta c$ denotes total *royalty payments.* Since $H^e \in (0, 1)$ and we require that $|\lambda| > 1$, deadweight loss in this imperfect competition setting is larger than in perfect competition case.

Using U.S. Census data for 2007, the value added weighted-average Herfindahl index for manufacturing industries (based on the 50 largest firms), H, for manufacturing sectors is 0.05. As is well-known,the Herfindahl index can be decomposed as $H = \frac{1}{N} + N \cdot \text{Var}(m) = H^e + N \cdot \text{Var}(m)$, where m is the market share of each firm. Thus, the observed H overstates the unobserved H^e, so the computed deadweight loss will be an upper bound to the true value of DWL. Despite this, the upper bound for the Cournot setting is not materially different from the competitive case in the text.

The value of H varies widely across industries. We do not compute deadweight loss using industry-specific values because it is difficult to assign patents in different patent classes to industries, and the existing Patent Office concordance is problematic (e.g., the mapping is not unique).

3.G.5 Calibrating Litigation Costs

To calibrate litigation costs, $\mathcal{C}(\tilde{V})$, we use data from the American Intellectual Property Law Association (AIPLA) surveys on litigation costs as a function of (intervals) of the value at stake, which we assume is the same for the patentee and challenger. We use the linear specification

$$\mathcal{C}(\tilde{V}) = \ell_0 + \ell_1\tilde{V}$$

Using this same specification, Schankerman and Schuett (2022) estimate the value of ℓ_0 as $\$624,000$ and $\ell_1 = 0.162$ (2018 USD). Note that this calibration of legal costs is at the patent, not claim, level.

3.G.6 Calibrating Development Costs

We apply the estimates from Schankerman and Schuett (2022) to our context. They assume that development costs κ are exponential, with mean equal to $k_0 + k_1 s$, where s is the size reduction of the invention and k_0 and k_1 are estimated as 254.6×10^3 and 2.33×10^{10}, respectively. Regarding the size reduction, they assume that s is log-logistic distributed with parameters $\beta_0 = 1.02$ and $\beta_1 = 1.14 \times 10^{-6}$. We use the mean value of s in our calibration.

In the baseline quantification, we draw values of κ from the distribution described above, which assumes that development costs are independent of Γ^* and π. In this model, inventors know their development costs prior to their decision to develop their idea. We also experiment with another model, which makes the opposite assumption that inventors do not know their development costs and thus use the mean value, $\bar{\kappa} = k_0 + k_1 \bar{s}$, to make their development decision. Both models produce similar conclusions; results are available upon request.

3.G.7 Calibrating the Number of Ideas

To compute the number of ideas, we start with the average annual number of utility patent applications in the period 2011–2013. We convert this number into the number of ideas in two steps. First, we use the estimates from Schankerman and Schuett (2022) that about two-thirds of applications are "low type" inventions (defined by them as those that would have been developed even without patent protection), and second, that one-third of ideas become a low type patent application. Together, this implies about one million ideas for potential inventions for each cohort of applications.

This page is intentionally left blank.

Bibliography

The numbers at the end of every reference link to the pages citing the reference.

Abaluck, J. and A. Adams-Prassl (2021): "What do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses," *The Quarterly Journal of Economics*, 136, 1611–1663. 73

Acemoglu, D. (2001): "Credit market imperfections and persistent unemployment," *European Economic Review*, 45, 665–679. 12

Adams, P., B. Guttman-Kenney, L. Hayes, S. Hunt, D. Laibson, and N. Stewart (2022): "Do Nudges Reduce Borrowing and Consumer Confusion in the Credit Card Market?" *Economica*, 89, S178–S199. 29

Adams, W., L. Einav, and J. Levin (2009): "Liquidity Constraints and Imperfect Information in Subprime Lending," *American Economic Review*, 99, 49–84. 14

Adda, J. and M. Ottaviani (2023): "Grantmaking, Grading on a Curve, and the Paradox of Relative Evaluation in Nonmarkets," *forthcoming in The Quarterly Journal of Economics*. 95

Agarwal, S., S. Chomsisengphet, N. Mahoney, and J. Stroebel (2014): " Regulating Consumer Financial Products: Evidence from Credit Cards," *The Quarterly Journal of Economics*, 130, 111–164. 16, 18

——— (2017): "Do Banks Pass through Credit Expansions to Consumers Who want to Borrow?" *The Quarterly Journal of Economics*, 133, 129–190. 14, 20, 27, 50

Agarwal, S., J. C. Driscoll, X. Gabaix, and D. Laibson (2008): "Learning in the Credit Card Market," Working Paper 13822, National Bureau of Economic Research. 29

Agarwal, S. and J. Zhang (2015): "A review of credit card literature: perspectives from consumers," *Unpublished Working Paper*. 29

Akerlof, G. A. (1970): "The Market for "Lemons": Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*, 84, 488–500. 12

——— (2001): "Behavioral Macroeconomics and Macroeconomic Behavior," *Nobel Prize Committee, Nobel Prize lecture*. 12

Alan, S. and G. Loranth (2013): "Subprime Consumer Credit Demand: Evidence from a Lender's Pricing Experiment," *The Review of Financial Studies*, 26, 2353–2374. 63

Albanesi, S. and D. F. Vamossy (2019): "Predicting Consumer Default: A Deep Learning Approach," *NBER Working Paper Series*. 15, 20

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 132, 1553–1592. 115, 150, 151

Ashraf, N., O. Bandiera, E. Davenport, and S. S. Lee (2020): "Losing Prosociality in the Quest for Talent? Sorting, Selection, and Productivity in the Delivery of Public Services," *American Economic Review*, 110, 1355–94. 94

Ashraf, N., O. Bandiera, and K. Jack (2014): "No margin, no mission? A field experiment on incentives for public service delivery," *Journal of Public Economics*, 120, 1 – 17. 94

Ausubel, L. M. (1991): "The Failure of Competition in the Credit Card Market," *The American Economic Review*, 81, 50–81. 29

——— (1999): "Adverse selection in the credit card market," *Unpublished Working Paper*. 29

Ausubel, L. M. and H. Shui (2005): "Time Inconsistency in the Credit Card Market," *Unpublished Working Paper*. 29

Aydin, D. (2022): "Consumption Response to Credit Expansions: Evidence from Experimental Assignment of 45,307 Credit Lines," *American Economic Review*, 112, 1–40. 14

Azoulay, P., J. S. Graff Zivin, D. Li, and B. N. Sampat (2018): "Public R&D Investments and Private-sector Patenting: Evidence from NIH Funding Rules," *The Review of Economic Studies*, 86, 117–152. 95

Bachas, N. (2019): "The Impact of Risk-Based Pricing in the Student Loan Market: Evidence from Borrower Repayment Decisions," *Unpublished Working Paper*. 14

Benabou, R. and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *The Review of Economic Studies*, 70, 489–520. 94

——— (2006): "Incentives and Prosocial Behavior," *American Economic Review*, 96, 1652–1678. 94

Benetton, M. (2021): "Leverage Regulation and Market Structure: A Structural Model of the U.K. Mortgage Market," *The Journal of Finance*, 76, 2997–3053. 14, 55

Benetton, M., A. Gavazza, and P. Surico (2022): "Mortgage Pricing and Monetary Policy," *Unpublished Working Paper*. 55

Bennett, R. and K. Rita (2012): "Public attitudes towards the UK banking industry following the global financial crisis," *The International Journal of Bank Marketing*, 30, 128–147. 71

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63, 841–890. 47

Besley, T. and M. Ghatak (2005): "Competition and Incentives with Motivated Agents," *American Economic Review*, 95, 616–636. 94

Bessen, J. (2008): "The value of U.S. patents by owner and patent characteristics," *Research Policy*, 37, 932–945. 113, 117

Bhat, C. R. (2003): "Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences," *Transportation Research Part B: Methodological*, 37, 837–855. 57

Bhutta, N., J. Dokko, and H. Shan (2017): "Consumer Ruthlessness and Mortgage Default during the 2007 to 2009 Housing Bust," *The Journal of Finance*, 72, 2433–2466. 49

Bloom, N., M. Schankerman, and J. Van Reenen (2013): "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, 81, 1347–1393. 90, 130

Butaru, F., Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique (2016): "Risk and risk management in the credit card industry," *Journal of Banking & Finance*, 72, 218–239. 15

Calem, P. S. and L. J. Mester (1995): "Consumer Behavior and the Stickiness of Credit-Card Interest Rates," *The American Economic Review*, 85, 1327–1336. 29

Castellanos, S. G., D. Jiménez Hernández, A. Mahajan, and E. Seira (2018): "Expanding Financial Access Via Credit Cards: Evidence from Mexico," *National Bureau of Economic Research*. 49

CFPB (2021): "The Consumer Credit Card Market," *Consumer Finance Protection Bureau Research Publications*. 29

Cockburn, I., S. Kortum, and S. Stern (2003): *Are All Patent Examiners Equal? Examiners, Patent Characteristics, and Litigation Outcomes*, Washington, DC: The National Academies Press. 95

Cohen, A. and L. Einav (2007): "Estimating Risk Preferences from Deductible Choice," *American Economic Review*, 97, 745–788. 49

Crawford, G. S., N. Pavanini, and F. Schivardi (2018): "Asymmetric Information and Imperfect Competition in Lending Markets," *American Economic Review*, 108, 1659–1701. 45, 62

Cuesta, J. I. and A. Sepulveda (2021): "Price Regulation in Credit Markets: A Trade-Off between Consumer Protection and Credit Access," *Unpublished Working Paper*. 16

Department for Business Innovation and Skills (2010): "Guidance on the regulations implementing the Consumer Credit Directive updated for EU Commission Directive," *Guidance Note*. 50

Drozd, L. and M. Kowalik (2019): "Credit Cards and the Great Recession: The Collapse of Teasers," *Society for Economic Dynamics 2019 Meeting Papers*. 29

Drozd, L. and J. B. Nosal (2011): "Competing for Customers: A Search Model of the Market for Unsecured Credit," *Unpublished Working Paper*. 29

Druedahl, J. and C. N. Jørgensen (2018): "Precautionary borrowing and the credit card debt puzzle," *Quantitative Economics*, 9, 785–823. 29

Dufresne, D. (2004): "The Log-Normal Approximation in Financial and Other Computations," *Advances in Applied Probability*, 36, 747–773. 154

Edelberg, W. (2006): "Risk-based pricing of interest rates for consumer loans," *Journal of Monetary Economics*, 53, 2283–2298. 14

Egan, M. L., G. Matvos, and A. Seru (2023): "Arbitration with Uninformed Consumers," *conditionally accepted in The Review of Economic Studies*. 95

Einav, L. and A. Finkelstein (2011): "Selection in Insurance Markets: Theory and Empirics in Pictures," *The Journal of Economic Perspectives*, 25, 115–138. 68

Einav, L., A. Finkelstein, and M. R. Cullen (2010a): "Estimating Welfare in Insurance Markets Using Variation in Prices," *The Quarterly Journal of Economics*, 125, 877–921. 68

Einav, L., A. Finkelstein, R. Kluender, and P. Schrimpf (2016): "Beyond Statistics: The Economic Content of Risk Scores," *American Economic Journal: Applied Economics*, 8, 195–224. 15

Einav, L., A. Finkelstein, and P. Schrimpf (2010b): "Optimal Mandates and the Welfare Cost of Asymmetric Information: Evidence From the U.K. Annuity Market," *Econometrica*, 78, 1031–1092. 49

Einav, L., M. Jenkins, and J. Levin (2012): "Contract Pricing in Consumer Credit Markets," *Econometrica*, 80, 1387–1432. 15, 45

——— (2013): "The impact of credit scoring on consumer lending," *The RAND Journal of Economics*, 44, 249–274. 15

Elliot, A. and S. Hansen (2023): "Text Algorithms in Economics," *Unpublished Working Paper*. 144

Evans, D. and R. Schmalensee (2005): *Paying with Plastic: The Digital Revolution in Buying and Borrowing*, MIT Press Books, The MIT Press, 2nd ed. 29, 53, 75

Exler, F., I. Livshits, J. MacGee, and M. Tertilt (2021): "Consumer Credit with Over-Optimistic Borrowers," *Bank of Canada Staff Working Paper*. 29

FCA (2015a): "Credit Card Market Study: Final Findings Report," *Financial Conduct Authority Reserach Publications*. 22, 29

——— (2015b): "Credit Card Market Study Interim Report: Annex 10 - Account Level Data," *Financial Conduct Authority Reserach Publications*. 19

——— (2015c): "Credit Card Market Study Interim Report: Annex 2 - How credit card products are regulated," *Financial Conduct Authority Reserach Publications*. 16

——— (2015d): "Credit Card Market Study Interim Report: Annex 3 - Results from the consumer survey," *Financial Conduct Authority Reserach Publications*. 22

——— (2022): "Credit Information Market Study: Interim Report," *Financial Conduct Authority Reserach Publications*. 20

Federal Trade Commission (2011): *The Evolving IP Marketplace: Aligning Patent Notice and Remedies with Competition*, Washington D.C.: Government Printing Office. 91

Fiordelisi, F., M.-G. Soana, and P. Schwizer (2013): "The determinants of reputational risk in the banking sector," *Journal of Banking & Finance*, 37, 1359–1371. 71

Frakes, M. D. and M. F. Wasserman (2017): "Is the Time Allocated to Review Patent Applications Inducing Examiners to Grant Invalid Patents? Evidence from Microlevel Application Data," *The Review of Economics and Statistics*, 99, 550–563. 95, 98, 99, 145, 146

Fulford, S. L. (2015): "How important is variability in consumer credit limits?" *Journal of Monetary Economics*, 72, 42–63. 14

Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022): "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *The Journal of Finance*, 77, 5–47. 15

Galasso, A. and M. Schankerman (2015): " Patents and Cumulative Innovation: Causal Evidence from the Courts," *The Quarterly Journal of Economics*, 130, 317–369. 91

——— (2018): "Patent rights, innovation, and firm exit," *The RAND Journal of Economics*, 49, 64–86. 90

Galenianos, M. and A. Gavazza (2022): "Regulatory Interventions in Consumer Financial Markets: The Case of Credit Cards," *Journal of the European Economic Association*. 29

Ganong, P. and P. Noel (2020): "Liquidity versus Wealth in Household Debt Obligations: Evidence from Housing Policy in the Great Recession," *American Economic Review*, 110, 3100–3138. 49

Gathergood, J., N. Mahoney, N. Stewart, and J. Weber (2019): "How Do Individuals Repay Their Debt? The Balance-Matching Heuristic," *American Economic Review*, 109, 844–75. 22, 29, 44

Gouriéroux, C. and A. Monfort (1993): "Simulation-based inference: A survey with special reference to panel data models," *Journal of Econometrics*, 59, 5–33. 57

——— (1996): *Simulation-based Econometric Methods*, Oxford University Press. 57

Greene, W. (2017): *Econometric Analysis: Eighth Edition*, Econometric Analysis, Pearson. 77

Grodzicki, D., A. Alexandrov, O. Bedre-Defoile, and S. Koulayev (2022): "Consumer Demand for Credit Card Services," *Journal of Financial Services Research*. 45

Gross, D. B. and N. S. Souleles (2002a): "Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data," *The Quarterly Journal of Economics*, 117, 149–185. 14

——— (2002b): "An Empirical Analysis of Personal Bankruptcy and Delinquency," *The Review of Financial Studies*, 15, 319–347. 14

Guiso, L., P. Sapienza, and L. Zingales (2013): "The Determinants of Attitudes toward Strategic Default on Mortgages," *The Journal of Finance*, 68, 1473–1515. 49

Hajivassiliou, V. A. and P. A. Ruud (1994): "Chapter 40: Classical estimation methods for LDV models using simulation," *Handbook of Econometrics*, 4, 2383–2441. 57

Hall, B. and J. Lerner (2010): *The Financing of R&D and Innovation*, vol. 1, Elsevier. 90

Halton, J. H. (1960): "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals," *Numerische Mathematik*, 84–90. 57, 60

Heidhues, P. and B. Kőszegi (2010): "Exploiting Naïvete about Self-Control in the Credit Market," *American Economic Review*, 100, 2279–2303. 29

House of Commons Treasury Committee (2003): "Transparency of Credit Card Charges," *Stationery Office by Order of the House*. 43, 71, 76

Indarte, S. (2021): "Moral Hazard versus Liquidity in Household Bankruptcy," *Unpublished Working Paper*. 49

Jaffe, A. and J. Lerner (2004): *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It*, Princeton University Press. 91

Jalali, M., H. Rahmandad, and H. Ghoddusi (2015): *Using the method of simulated moments for system identification*, MIT Press. 150

Karlan, D. and J. Zinman (2018): "Long-Run Price Elasticities of Demand for Credit: Evidence from a Countrywide Field Experiment in Mexico," *The Review of Economic Studies*, 86, 1704–1746. 63

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3, 303–20. 97

Keys, B. J. and J. Wang (2019): "Minimum payments and debt paydown in consumer credit cards," *Journal of Financial Economics*, 131, 528–548. 29

Khandani, A. E., A. J. Kim, and A. W. Lo (2010): "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, 34, 2767–2787. 15

Kiyotaki, N. and J. Moore (1997): "Credit Cycles," *Journal of Political Economy*, 105, 211–248. 12

Knight, H. J. (2010): "An empirical investigation of pricing and competition in the UK credit card market," *PhD Thesis, University of Nottingham*. 29

Kuchler, T. and M. Pagel (2021): "Sticking to your plan: The role of present bias for credit card paydown," *Journal of Financial Economics*, 139, 359–388. 29

Laibson, D., A. Repetto, and J. Tobacman (2000): "A Debt Puzzle," Working Paper 7879, National Bureau of Economic Research. 29

Lancaster, K. J. (1966): "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, 132–157. 45

Lanjouw, J. O. (1998): "Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value," *The Review of Economic Studies*, 65, 671–710. 116

Le, Q. and T. Mikolov (2014): "Distributed representations of sentences and documents," in *International conference on machine learning*, PMLR, 1188–1196. 97

Lee, L.-F. (1992): "On Efficiency of Methods of Simulated Moments and Maximum Simulated Likelihood Estimation of Discrete Response Models," *Econometric Theory*, 8, 518–552. 57

Lee, L.-f. (1995): "Asymptotic Bias in Simulated Maximum Likelihood Estimation of Discrete Choice Models," *Econometric Theory*, 11, 437–483. 57

Lehmann, E. (2006): *Nonparametrics: Statistical Methods Based on Rank*, Springer. 37

Lessmann, S., B. Baesens, H.-V. Seow, and L. C. Thomas (2015): "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, 247, 124–136. 15

Li, D. (2017): "Expertise versus Bias in Evaluation: Evidence from the NIH," *American Economic Journal: Applied Economics*, 9, 60–92. 95

Li, D. and L. Agha (2015): "Big names or big ideas: Do peer-review panels select the best science proposals?" *Science*, 348, 434–438. 95

Linares-Zegarra, J. and J. Wilson (2012): "Risk Based Pricing in the Credit Card Industry: Evidence from US Survey Data," *Unpublished Working Paper*. 14

Livshits, I., J. C. Mac Gee, and M. Tertilt (2016): "The Democratization of Credit and the Rise in Consumer Bankruptcies," *The Review of Economic Studies*, 83, 1673–1710. 50

Magri, S. (2018): "Are lenders using risk-based pricing in the consumer loan market? The effects of the 2008 crisis," *Unpublished Working Paper*. 14

Magri, S. and R. Pico (2011): "The rise of risk-based pricing of mortgage interest rates in Italy," *Journal of Banking and Finance*, 35, 1277–1290. 14

Meier, S. and C. Sprenger (2010): "Present-Biased Preferences and Credit Card Borrowing," *American Economic Journal: Applied Economics*, 2, 193–210. 29

Merges, R. P. and R. R. Nelson (1990): "On the Complex Economics of Patent Scope," *Columbia Law Review*, 90, 839–916. 92

Nelsen, R. B. (2007): *An introduction to copulas*, Springer Science & Business Media. 114

Nelson, S. (2022): "Private information and price regulation in the us credit card market," *Unpublished Working Paper*. 16, 18, 27, 49, 63

Nevo, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69, 307–342. 47

Office of Fair Trading (2006): "Calculating fair default charges in credit card contracts," *OFT Statement*, 1–36. 76

Pakes, A. (1986): "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks," *Econometrica*, 54, 755–784. 112, 116, 132

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057. 57

Panetta, F., F. Schivardi, and M. Shum (2009): "Do Mergers Improve Information? Evidence from the Loan Market," *Journal of Money, Credit and Banking*, 41, 673–709. 73

Paravisini, D. and A. Schoar (2015): "The Incentive Effect of Scores: Randomized Evidence from Credit Committees," *NBER Working Paper Series*. 15

Řehůřek, R. and P. Sojka (2010): "Software Framework for Topic Modelling with Large Corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. 144

Robles-Garcia, C. (2022): "Competition and Incentives in Mortgage Markets: The Role of Brokers," *Unpublished Working Paper*. 14, 55

Rothschild, M. and J. Stiglitz (1976): "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *The Quarterly Journal of Economics*, 90, 629–649. 12

Ru, H. and A. Schoar (2016): "Do Credit Card Companies Screen for Behavioral Biases?" Working Paper 22360, National Bureau of Economic Research. 29

Sampat, B. and H. L. Williams (2019): "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome," *American Economic Review*, 109, 203–36. 99

Scandizzo, S. (2011): "A framework for the analysis of reputational risk," *The Journal of Operational Risk*, 6, 41–63. 71

Schankerman, M. (1998): "How Valuable is Patent Protection? Estimates by Technology Field," *The RAND Journal of Economics*, 29, 77–107. 129, 132

Schankerman, M. and A. Pakes (1986): "Estimates of the Value of Patent Rights in European Countries during the Post-1950 Period," *Economic Journal*, 96, 1052–1076. 110, 114

Schankerman, M. and F. Schuett (2022): "Patent Screening, Innovation, and Welfare," *The Review of Economic Studies*, 89, 2101–2148. 96, 123, 127, 128, 129, 132, 153, 155, 157

Scholz, F. and A. Zhu (2019): "package 'kSamples'," *CRAN Reference Manual*. 37

Scholz, F. W. and M. A. Stephens (1987): "K-Sample Anderson-Darling Tests," *Journal of the American Statistical Association*, 82, 918–924. 37

Sidak, Z., P. Sen, and J. Hajek (1999): *Theory of Rank Tests*, Elsevier. 37

Stango, V. (2002): "Pricing with Consumer Switching Costs: Evidence from the Credit Card Market," *The Journal of Industrial Economics*, 475–492. 29

Stango, V. and J. Zinman (2009): "Exponential Growth Bias and Household Finance," *The Journal of Finance*, 64, 2807–2849. 29

——— (2015): "Borrowing High versus Borrowing Higher: Price Dispersion and Shopping Behavior in the U.S. Credit Card Market," *The Review of Financial Studies*, 29, 979–1006. 29

Stewart, N. (2009): "The Cost of Anchoring on Credit-Card Minimum Repayments," *Psychological Science*, 20, 39–41. 29

Stiglitz, J. E. and A. Weiss (1981): "Credit Rationing in Markets with Imperfect Information," *The American Economic Review*, 71, 393–410. 15

The Economist (2015): "Time to fix patents," *The Economist Group*, August 8th-14th, 9. 91

Train, K. E. (2003): *Discrete Choice Methods with Simulation*, Cambridge University Press. 57

Wang, L. (2023): "Regulating Competing Payment Networks," *Unpublished Working Paper*. 29

Xifra, J. and E. Ordeix (2009): "Managing reputational risk in an economic downturn: The case of Banco Santander," *Public Relations Review*, 35, 353–360. 71

Yang, S., L. Markoczy, and M. Qi (2007): "Unrealistic optimism in consumer credit card adoption," *Journal of Economic Psychology*, 28, 170–185. 29