

Gödl-Hanisch, Isabel

**Working Paper**

## Bank Concentration and Monetary Policy Pass-Through

CESifo Working Paper, No. 10378

**Provided in Cooperation with:**

Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

*Suggested Citation:* Gödl-Hanisch, Isabel (2023) : Bank Concentration and Monetary Policy Pass-Through, CESifo Working Paper, No. 10378, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:

<https://hdl.handle.net/10419/272022>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

# Bank Concentration and Monetary Policy Pass-Through

*Isabel Gödl-Hanisch*

## **Impressum:**

CESifo Working Papers

ISSN 2364-1428 (electronic version)

Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH

The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute

Poschingerstr. 5, 81679 Munich, Germany

Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email [office@cesifo.de](mailto:office@cesifo.de)

Editor: Clemens Fuest

<https://www.cesifo.org/en/wp>

An electronic version of the paper may be downloaded

- from the SSRN website: [www.SSRN.com](http://www.SSRN.com)
- from the RePEc website: [www.RePEc.org](http://www.RePEc.org)
- from the CESifo website: <https://www.cesifo.org/en/wp>

# Bank Concentration and Monetary Policy Pass-Through

## Abstract

This paper analyzes the implications of the gradual rise in bank concentration since the 1990s for the transmission of monetary policy. I use branch-level data on deposit and loan rates to evaluate the monetary policy pass-through conditional on the level of local bank concentration and bank capitalization. I find that banks operating in high-concentration markets and under-capitalized banks adjust short-term lending rates more. I then build a theoretical model with heterogeneous banks that rationalizes the empirical findings and explains the underlying mechanism. In the model, monopolistic competition in local deposit and loan markets, along with bank capital requirements, lead to frictions on the pass-through to the real economy. Counterfactual analyses highlight that the rise in bank concentration alters monetary policy pass-through by two channels: the market power and capital allocation channels. Both channels further strengthen monetary policy transmission to output and investment, amplify the credit cycle, and flatten the Phillips curve.

JEL-Codes: E440, E510, E520, G210.

Keywords: monetary transmission, bank heterogeneity, monopolistic competition, bank regulation.

*Isabel Gödl-Hanisch*  
*LMU Munich / Germany*  
*isabel.goedl-hanisch@econ.lmu.de*

This version: April, 2023. Click here for the latest version:

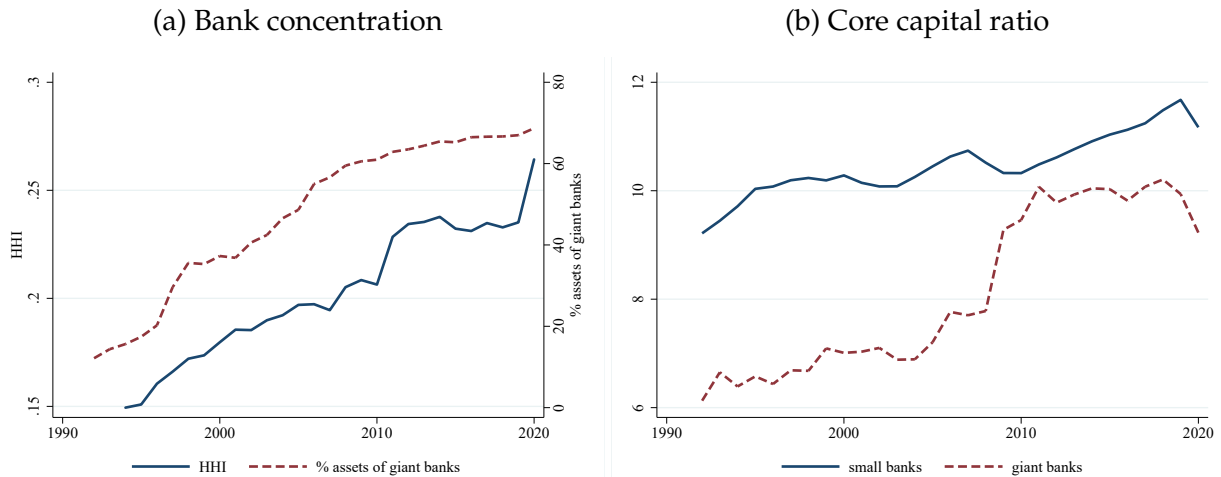
[https://www.dropbox.com/s/90gy2hmbknzwb2l/Paper\\_current\\_version.pdf?dl=0](https://www.dropbox.com/s/90gy2hmbknzwb2l/Paper_current_version.pdf?dl=0)

I am grateful to Rudiger Bachmann, Christiane Baumeister, Eric Sims, and Cynthia Wu for their guidance and support. I received helpful comments from seminar participants at the Federal Reserve Board, European Central Bank, 2020 CEBRA annual meeting, Banco Central de Chile, 15th Economics Graduate Student Conference at Washington University in St. Louis, 2nd WIMFEH Workshop, University of Wurzburg, 20th FDIC/JFSR Bank Research Conference, SNDE 2022, IE University, Bank of England, LMU Munich, Norges Bank, De Nederlandsche Bank, WU Vienna, Miami University, Federal Reserve Bank of Philadelphia, EEA Milano 2022, MEG 2022, SNDE Symposium for Young Researchers, 4th EMMMC, CEPR Workshop Empirical Monetary Economics 2022, CRC TR 224 Workshop 2022, EWMES 2022, and 1st X Amsterdam Macroeconomic Workshop. I thank the Notre Dame Department of Economics and Mendoza School of Business for purchasing the microdata for this project and the ISLA Graduate Student Research Award for financial support.

# 1 Introduction

Over the last two decades, the US banking sector has become increasingly concentrated, as relaxed banking regulation before the financial crisis and bank consolidation after the financial crisis significantly reduced the number of banks in many local banking markets.<sup>1</sup> In 1994, giant US banks, defined as banks with more than \$100 billion in assets, owned 16% of total commercial bank assets; that share increased to 69% by 2020. During the same time, the local Herfindahl-Hirschman Index (HHI) steadily grew from a moderate level of 0.15 in 1994 to a highly concentrated level of 0.26 in 2020, as shown in Panel (a) of Figure 1.<sup>2</sup> This paper studies the implications of the gradual rise in bank concentration over the last two decades for monetary policy transmission to the real economy.

Figure 1: The US banking sector over time



Notes: HHI is shown at the average county level and is weighted by total deposits, % assets of giant banks is the asset share of banks > \$100 billion in assets (in \$2018), and core capital ratio measures mean core capital over assets by group. Source: Federal Deposit Insurance Corporation.

In order to assess the role of bank concentration for monetary policy pass-through, it is crucial to consider the observed differences in retail rates and lending volumes within a given bank across regions (e.g., Wells Fargo branches in Cook County vs. St. Joseph County) as well as across bank institutions within a region (e.g., Wells Fargo vs. Bank of America branches in St. Joseph County). The variation in retail rates serves to shed light on how (i) the concentration of local markets and (ii) the size distribution of banks affect the aggregate transmission of monetary policy via two channels. The first channel is the

<sup>1</sup>For example, the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 permitted banks to open branches across states, and the Glass-Steagall Act's repeal in 1999 allowed commercial banks to offer both securities and insurance, providing incentives for scale economies (Corbae and D'Erasmus, 2020).

<sup>2</sup>Appendix A.1 decomposes national bank concentration growth and finds that within-county growth and rising concentration in counties with deposit inflows contribute significantly to the overall effect (Figure A.1).

*market power channel*: a higher concentration in local banking markets leads to a widening spread between the central bank's policy rate and the commercial banks' loan and deposit rates. This might be expected to weaken monetary policy transmission to deposit rates but strengthen the transmission to loan rates. The second channel is the *capital allocation channel*: a higher banking concentration implies that giant banks, which tend to have relatively low capital ratios, as shown in Panel (b) of Figure 1, handle an increasing share of total loans and deposits. This might be expected to amplify financial frictions arising from regulatory requirements on giant banks. In the past years, Basel III reforms had the effect of mitigating the decline in *aggregate* capitalization driven by increasing concentration. Overall, a rise in bank concentration may thus strengthen pass-through to lending rates via both channels but dampen the pass-through to deposit rates via the market power channel.

Previous research has shown that both bank market power (e.g., Drechsler et al., 2017; Scharfstein and Sunderam, 2016) and bank size and capitalization (e.g., Kashyap and Stein, 2000; Van den Heuvel, 2002) impact the effectiveness of monetary policy. Yet, these studies have analyzed these aspects in isolation. Little attention has been paid to the effects of the banking sector's composition on monetary policy pass-through and the relative importance of each channel. The contribution of this paper is to provide a unified framework for deposit and loan market power and allow for a role of bank financial regulation. This joint modeling approach allows me to rationalize variation across branches *within* a bank institution and across banks *within* a market. My results underscore the importance of compositional effects for the transmission of monetary policy and demonstrate that a partial analysis falls short of accounting for interaction effects and, therefore, overstates the effect of rising concentration on monetary transmission. Similarly, abstracting from loan market power would underrate the effect of rising concentration on monetary transmission to output and inflation by 43% and 34%, respectively.

This paper starts by building a simple model of heterogeneous monetary policy pass-through to retail rates inspired by the canonical Monti-Klein model (e.g., Monti 1972, Klein 1971). To micro-found the differences between branches of the same bank across locations and the differences across bank institutions in the same location, I combine two conventional building blocks. First, banks have market power in local deposit and loan markets. Second, banks face a capital requirement that imposes additional frictions on monetary policy pass-through. The theoretical model predicts that monetary policy pass-through to loan rates is an increasing function of local bank concentration, as the markup acts as a multiplier on the policy rate; whereas monetary policy pass-through to deposit rates is a decreasing function of local bank concentration, as the markdown

acts as a multiplier on the policy rate. The model also predicts that monetary policy pass-through to loan rates is a decreasing function of bank capitalization, as the capital constraint imposes an additional lending cost. Further, it suggests interaction effects between a bank's capitalization and market power: a greater impact of financial frictions in an environment with market power.

In the empirical part of the paper, I present novel facts on rate dispersion across space and time using US bank branch-level data from *RateWatch* from January 1998 to March 2019. I document substantial rate dispersion *within* banks and locations in line with the assumptions of the theoretical model. I then test the model's predictions by studying monetary policy pass-through to consumer retail rates. I define monetary policy pass-through as the extent to which loan and deposit rates respond to changes in the monetary policy rate. To control for potential endogeneity in monetary policy, I use monetary policy surprises from Nakamura and Steinsson (2018) as instruments for the policy rate. I exploit variation in local bank concentration and bank capitalization to assess the relative importance of the *market power channel* and the *capital allocation channel*. The empirical results confirm the model's predictions. While monetary policy pass-through to loan rates is higher for branches operating in high-concentration counties and for banks with low capital ratios, monetary policy pass-through to deposit rates is lower for branches operating in high-concentration counties and almost unaffected by the bank's capital ratio. Cross-sectional differences have significant implications for the pass-through to loan and deposit rates across regions: estimated county-level pass-through coefficients for loan and deposit rates across the US range from 0.29 to 2.21 and 0.46 to 0.82, respectively, implying substantial exposure to monetary policy in some counties but almost none in others.

To quantify the relative importance of the different frictions and perform counterfactual analyses, I embed the simple model into a dynamic New Keynesian model extending Gerali et al. (2010). In segmented markets, patient households provide deposits to the banking sector, while impatient households and entrepreneurs demand credit. The presence of financial frictions on the banking side impairs the intermediation of credit between the agents. The banking sector consists of heterogeneous bank headquarters facing size-dependent capital requirements and branches operating in spatially segmented markets with different market structures, varying by the degree of competition. More specifically, the elasticities of loan demand and deposit supply are heterogeneous and closely linked to the level of bank concentration in a market. The framework lends itself to conducting different counterfactuals: to what extent have sectoral shifts in the US banking sector, particularly in the composition of local markets and bank size distribution, as well as secular trends in markups and capital ratios, affected monetary transmission over time?

The counterfactual analyses show that increasing bank concentration from 1994 to 2019 amplified monetary policy pass-through to loan rates. In other words, loan rates and bank lending became more sensitive to monetary policy changes. Quantitatively, pass-through to loan rates increased by 50% but decreased for deposit rates by the same proportion. Decomposing the total pass-through change over time reveals that the *market power channel*, i.e., increasing markups and local market share changes, is the most significant contributor to the overall effect. The impacts of the *capital allocation channel*, i.e., rising capital requirements and giant banks' market share changes over time, is relatively small, as the effect of bank regulation plays a more tangential role relative to market power. However, additional significant interaction effects emerge as higher market power increases the response to changes in marginal costs and financial frictions, a part underestimated in a partial analysis and true for most of the literature. Disentangling the *market power channel* into the marginal contribution of deposit and loan market power indicates that market power in deposit markets is important, consistent with previous work (Drechsler et al., 2017); a new result is that market power in loan markets is quantitatively even more important for aggregate dynamics and monetary transmission to output, explaining 43% in differential transmission over time compared to 20% explained by deposit market power.

Further, rising bank concentration alters monetary policy transmission to the macroeconomy. It amplifies monetary transmission to output and investment but dampens its impact on inflation. Specifically, the output contraction becomes twice as large in the medium run, though inflation reacts by only  $\frac{1}{3}$  in a high vs. low bank concentration environment. The opposite effects on output and inflation imply a flatter empirical Phillips curve over time, consistent with recent US data.<sup>3</sup> There are two sets of factors at play. The slope of the Phillips curve depends on the level of resource costs from the banking sector, leading to a wealth effect. Rising bank concentration increases these costs and widens the gap between production and effective output, breaking the link between output and marginal costs. Labor supply frictions, specifically wage rigidity and habit formation, individually and jointly lead to a further decoupling of output, marginal costs, and inflation and flatten the Phillips curve over time. The extent of macroeconomic implications depends on whether the households and firms are financially constrained. Adding borrowing constraints à la Iacoviello (2005) to households and firms lowers their overall sensitivity to loan rates, and compositional shifts in the banking sector become less important.

---

<sup>3</sup>Hazell et al. (2022), Matheson and Stavrev (2013), Ball and Mazumder (2011), and Kuttner and Robinson (2010) deliver evidence and alternative explanations. Similarly, Gilchrist et al. (2017) connect the flattening of the Phillips curve to financial frictions but approach the topic from a different angle.



**Related Literature.** This paper relates to research explaining differences in monetary policy pass-through based on bank characteristics and local market conditions. Similar to the structural approach of Wang et al. (2022), I quantify the implications of several frictions for monetary policy pass-through, comparing the role of loan and deposit market power and capital constraints shown to be important by Kashyap and Stein (2000), Kishan and Opiela (2000), Altavilla et al. (2019), and Van den Heuvel (2002). I add to Wang et al. (2022)'s analysis of bank lending by looking at the cross-section of retail rates, taking into account that banks operate in local markets, and by offering micro-foundations for the various frictions at play.<sup>4</sup> Drechsler et al. (2017) establish that banks in highly concentrated markets have a lower pass-through to deposit rates. Similarly, Scharfstein and Sunderam (2016) analyze the pass-through of mortgage-backed securities (MBS) yields to mortgage refinancing and the role of bank concentration therein, finding that banks in high-concentration markets are less sensitive to changes in MBS yields. While this paper also focuses on mortgages, the emphasis lies on the pass-through of changes in the policy rate to short-term mortgage rates and the role of bank concentration. Another contribution is to connect the findings on local bank concentration and bank characteristics. On top of that, I control for endogenous changes in the policy rate as a regressor to rule out a potential response to credit conditions. Using local projections instead of panel techniques shows the pass-through dynamics and easily incorporates state dependencies (see, e.g., Ramey and Zubairy 2018). Similarly, but using different methods and models, Corbae and D'Erasmus (2020) and De Loecker et al. (2020) point to higher markups and concentration in the financial sector over time.

On the theoretical side, I build on the canonical studies by Monti (1972) and Klein (1971). In the same vein as Gerali et al. (2010) and Andres and Arce (2012), I model the banking sector with monopolistic competition, which assumes that deposits and loans are baskets of differentiated products with constant elasticity of substitution leading to a constant markup. Gerali et al. (2010) compare the transmission of shocks with and without financial frictions in the banking sector in a New Keynesian model, finding that bank capital requirements, imperfect competition, and sticky rates alter monetary policy transmission. I extend their framework to include heterogeneous bank headquarters and branches to compare the pass-through in different banking environments. In addition, this paper fits into the growing theoretical literature on the state dependency of monetary policy transmission. Amongst them, Brunnermeier and Koby (2018) demonstrate that an accommodative monetary policy shock reverses and becomes contractionary when the policy rate falls below a certain level. Likewise, Wang (2019) and Ulate (2021) study

---

<sup>4</sup>Most papers study the effect on total lending or imputed loan rates (e.g., Drechsler et al. 2021).

monetary policy transmission to deposit and loan rates, focusing on low and negative rates. In contrast, the contribution of this paper focuses on the cross-sectional pass-through of monetary policy to retail rates given different banking sector structures.

The remainder of this paper is structured as follows. Section 2 proposes a simple model of heterogeneous monetary policy pass-through. Section 3 describes the data set. Section 4 presents novel facts on deposit and loan rate pass-through. Section 5 outlines the richer quantitative model, performs counterfactual analyses, and decomposes the effect of rising concentration on monetary transmission and the Phillips curve. Section 6 concludes. Details and robustness checks are available in the appendices.

## 2 Simple Model of Heterogeneous Pass-Through

To provide intuition for the empirical section, I build a simple model of heterogeneous monetary policy pass-through to retail rates inspired by the canonical Monti–Klein model. The model rationalizes retail rate differences between *branches* of the same bank across locations and *bank institutions* within the same location. In short, the model provides three predictions for cross-sectional pass-through differences, *ceteris paribus*: (i) a higher pass-through to loan rates in high-concentration locations, (ii) a lower pass-through to deposit rates in high-concentration locations, and (iii) a higher pass-through for low capitalization banks. In addition, the model also suggests an interaction between the *market power channel* and *capital allocation channel*.

In the stylized model, banks are financial intermediaries and originate loans funded by deposits and bank capital in different locations indexed by  $c$ . Financial regulation requires banks to hold specific bank capital ratios. Assume that banks are exogenously endowed with heterogeneous bank capital, implying variation in bank lending and deposit holdings across banks due to size-dependent capital constraints. Banks operate under monopolistic competition, taking local market conditions into account, wherein market power could arise from spatial and product differentiation. Table 1 shows a bank’s balance sheet with loans  $L_i^c$  and reserves  $R_i^c$  as assets, and deposits  $D_i^c$  and bank capital  $K_i^c$  as liabilities.

Table 1: Bank  $i$ ’s balance sheet in location  $c$

Assets		Liabilities	
Loans	$L_i^c$	Deposits	$D_i^c$
Reserves	$R_i^c$	Bank capital	$K_i^c$

Each bank  $i$  in location  $c$  seeks to maximize profits,  $\Pi_i^c = r_i^{l,c} L(r_i^{l,c}) + r^f R_i^c - r_i^{d,c} D(r_i^{d,c})$ ,

subject to (i) a capital requirement,  $K_i^c \geq \nu_i L_i^c$ , governed by  $\nu_i$ , the minimum bank capital adequacy ratio; (ii) local loan demand,  $L(r_i^{l,c}) = \left(\frac{r_i^{l,c}}{\bar{r}^{l,c}}\right)^{-\epsilon^{l,c}} \bar{L}^c$ , depending on local elasticity,  $\epsilon^{l,c}$ , aggregate loan rate,  $\bar{r}^{l,c}$ , aggregate loan demand,  $\bar{L}^c$ , and offered loan rate,  $r_i^{l,c}$ ; <sup>5</sup> (iii) local deposit supply,  $D(r_i^{d,c}) = \left(\frac{r_i^{d,c}}{\bar{r}^{d,c}}\right)^{-\epsilon^{d,c}} \bar{D}^c$ , depending on local elasticity,  $\epsilon^{d,c}$ , aggregate deposit rate,  $\bar{r}^{d,c}$ , aggregate deposit supply,  $\bar{D}^c$ , and offered deposit rate,  $r_i^{d,c}$ ; and (iv) a balance sheet constraint,  $L_i^c + R_i^c = D_i^c + K_i^c$ . <sup>6</sup>

Solving the maximization problem and rewriting the first-order conditions yields the loan and deposit rate decision as a function of the local markup and markdown on bank  $i$ 's marginal cost and policy rate  $r^f$ , where  $\phi_i$  reflects the multiplier on the capital constraint:

$$r_i^{l,c} = \underbrace{\frac{\epsilon^{l,c}}{(\epsilon^{l,c} - 1)}}_{\text{markup}} \underbrace{(r^f + \nu_i \phi_i)}_{\text{marginal cost}}, \quad (1)$$

$$r_i^{d,c} = \underbrace{\frac{\epsilon^{d,c}}{(\epsilon^{d,c} - 1)}}_{\text{markdown}} r^f. \quad (2)$$

As shown in equation (1), marginal costs for bank lending are heterogeneous across banks due to differences in the capital requirement  $\nu_i$  interacting with  $\phi_i$ , the multiplier on the capital constraint. Lending is relatively more costly for constrained banks since an increase in  $\phi_i$  raises marginal costs and loan rates. Equation (2) indicates that the policy rate  $r^f$  solely influences deposit rates. The capital requirement does not have an effect. Further, loan and deposit rates depend on markups and markdowns, which vary across locations due to monopolistic competition in local markets. The markups and markdowns are functions of loan demand and deposit supply elasticities in location  $c$ . The lower the elasticity, the higher the markup and the lower the markdown, linked to high concentration. <sup>7</sup>

The total derivatives of the loan and deposit rate with respect to the policy rate,  $r^f$ , inform about monetary policy pass-through. <sup>8</sup>

<sup>5</sup>The CES demand setup is isomorphic to assuming heterogeneous borrowers with stochastic utility and type 1 extreme value distribution (Ulate, 2021).

<sup>6</sup>A further reserve requirement would impose additional frictions and affect loan and deposit rates. I abstract from reserve requirements, as those likely have not been binding in the last years, particularly since the Federal Reserve began to pay interest on reserves in 2008. In March 2020, the Federal Reserve eliminated reserve requirements. For details, see, e.g., the website of the Federal Reserve Board.

<sup>7</sup>For example, under Cournot competition, the demand elasticity, markup, and HHI are tightly connected.

<sup>8</sup>While the IO-literature defines pass-through as the *percentage* change in prices resulting from a one *percentage* change in marginal costs, the interpretation of pass-through in this paper differs: pass-through reflects a *percentage point* change in retail rates resulting from a one *percentage point* change in the policy rate.

$$\frac{dr_i^{l,c}}{dr^f} = \underbrace{\frac{\epsilon^{l,c}}{(\epsilon^{l,c} - 1)}}_{\text{market power channel}} + \frac{\epsilon^{l,c}}{(\epsilon^{l,c} - 1)} \underbrace{\nu_i \frac{d\phi_i}{dr^f}}_{\text{capital allocation channel}} \quad (3)$$

$$\frac{dr^{d,c}}{dr^f} = \underbrace{\frac{\epsilon^{d,c}}{(\epsilon^{d,c} - 1)}}_{\text{market power channel}} \quad (4)$$

Equation (3) indicates that changes in the policy rate,  $r^f$ , affect loan rates by more in relatively less competitive locations. Intuitively, banks with high market power can easily pass changes in marginal costs to the consumer. Shifts in market structure thus affect loan rate pass-through directly: A lower elasticity of loan demand leads to higher markups and pass-through (i.e., the *market power channel*). Further, the magnitude of loan rate pass-through depends on the bank's capitalization and regulation. Hence, capital requirement shifts directly affect loan rate pass-through: Lower capitalization  $\nu_i$  leads to a higher pass-through (i.e., the *capital allocation channel*). The reason is that the multiplier on the constraint,  $\phi_i$ , declines in response to a monetary tightening as higher rates curb loan demand. Increased capitalization allows banks to benefit more from an easing constraint. Conversely, loan rates of more levered, less capitalized banks fluctuate more. Further, a non-negligible interaction effect results, as market power amplifies the *capital allocation channel*. In contrast, as shown in equation (4), deposit rate pass-through increases with competitiveness due to a declining markdown and is unaffected by the capital constraint. The extended model in Section 5 embeds this framework. Section 4 tests and quantifies the model's cross-sectional pass-through predictions:

1. Pass-through to loan rates increases with bank market power:  $\epsilon^{l,c} \downarrow \Rightarrow \frac{dr_i^{l,c}}{dr^f} \uparrow$ .
2. Pass-through to loan rates declines with bank capitalization:  $\nu_i \uparrow \Rightarrow \frac{dr_i^{l,c}}{dr^f} \downarrow$ .
3. Pass-through to deposit rates declines with bank market power:  $|\epsilon^{d,c}| \downarrow \Rightarrow \frac{dr^{d,c}}{dr^f} \downarrow$ .

### 3 Data Description

This paper combines multiple banking data sources, county-level and national macroeconomic data, and monetary policy surprises to study pass-through to loan and deposit rates. First, I use a panel of offered deposit and loan rates at a branch level for US commercial banks from January 1998 to March 2019, provided by *RateWatch*. The data provider regularly surveys 76,000 financial institution locations and collects quotes of deposit, mortgage,

and consumer loan rates. In this way, *RateWatch* serves as an advertisement and informational platform for consumers and business-to-business marketers, who expect the posted rates to be accurate and available.<sup>9</sup> The sampled loan rates provide information for the “best” borrowers, i.e., those with exceptional FICO scores defined by a default threshold of 740 (e.g., Bank of America or Chase), for a particular constant loan volume. In the case of mortgages, the volume is \$175,000. The data set includes fixed-rate and adjustable-rate (ARM) mortgages. For more information on the survey and a sample pricing sheet, see Appendix A.2. Second, using the branch identifier, the rate data is then merged with the FDIC’s Summary of Deposits, including annual county-level branch deposit holdings and historical ownership information. Third, the sample is combined with the Statistics on Depository Institutions (SDI), including bank balance sheet information, using the bank identifier.

I construct three key metrics to evaluate heterogeneous pass-through: (i) local bank concentration, (ii) bank-level capitalization, and (iii) a monetary policy measure.

**Measuring local bank concentration.** The canonical market concentration measure is the Herfindahl-Hirschman Index (HHI). The US Department of Justice’s antitrust division applies the measure to assess bank mergers. The HHI measures the sum of each bank institution’s squared market share by county for each point in time:<sup>10</sup>

$$HHI_{c,t} = \sum_{i=1}^I s_{i,c,t}^2 = s_{1,c,t}^2 + s_{2,c,t}^2 + \dots + s_{I,c,t}^2, \quad (5)$$

where  $s_{c,t,i}$  reflects bank  $i$ ’s market share in county  $c$  at time  $t$ . An HHI of 1 indicates a perfect monopoly, and  $1/N$  is an oligopoly with  $N$  equal-sized banks. The Department of Justice classifies a market with an HHI between 0.1 and 0.18 as “moderately concentrated” and above 0.18 as “highly concentrated,” according to the Federal Reserve Bank of St. Louis. In the baseline, I construct the HHI by county time and based on branch deposit holdings per county, similar to Drechsler et al. (2017). Figure 2 shows bank concentration across counties in the US in 2019. Considerable cross-sectional variation emerges among the HHIs ranging from 0.05 to 1, both across and within states. For example, Florida’s Leon County had an HHI of 0.1 in 2019, while surrounding counties Jefferson and Wakulla

<sup>9</sup>The *RateWatch* sample average follows the pattern of the aggregate time series of Freddie Mac closely. A small constant spread of 0.6 p.p. remains due to differences in points and other characteristics.

<sup>10</sup>A market is defined at the county level, consistent with the average distance to a lender of 1.25 miles in Canada (Allen et al., 2019) and Fannie Mae’s National Housing Survey (Q1 2019) documenting that 2 of 5 recent home buyers did not shop around for mortgage lenders. The results are robust to defining competition at the MSA level instead of the county level.

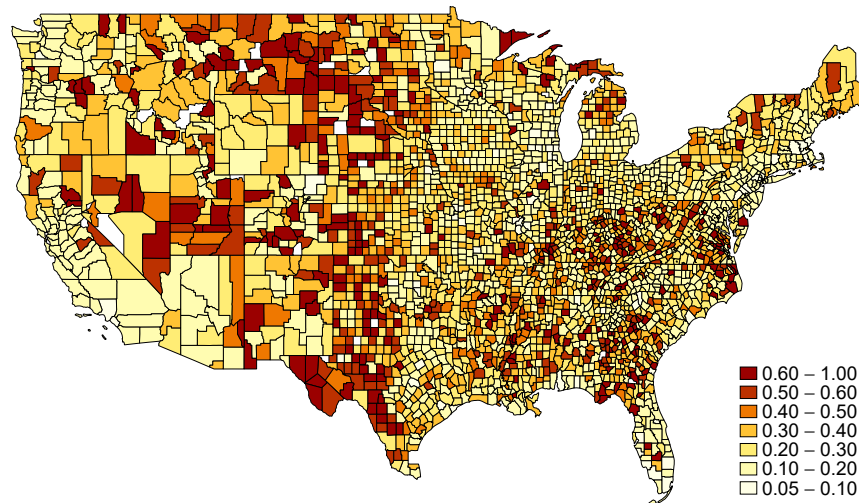
had HHIs of 0.66 and 0.44. Figure A.2 in Appendix A.1 provides evidence of shifts in local bank concentration between 1994 and 2019 and turns to the drivers behind the rise in aggregate concentration. A large proportion of counties, particularly those with large banking sectors (measured in terms of deposits), observed increasing concentration over time.

Instead of focusing on deposit market concentration, an alternative is to look directly at mortgage market concentration similar to Scharfstein and Sunderam (2016) using Home Mortgage Disclosure Act (HMDA) data. Unlike the deposit-based measure, the mortgage-based measure pertains to flows, resulting in higher volatility and less reasonable estimates at a granular level, but includes credit unions and non-bank lenders. Despite these fundamental differences, both concentration measures are highly correlated at the county level, as shown in Appendix A.3. Similarly, the trend in the rising market share of giant banks, measured in terms of deposits, loans, or mortgages, is remarkably similar (Figure A.4). Deposit market concentration can therefore provide a good proxy of loan market power. Appendix A.3 contrasts the empirical results for both measures and confirms the main results' robustness to the choice of concentration measure. A third alternative considers markups (e.g., De Loecker et al. 2020, Pasqualini 2021). Using a "poor man's estimate" of markups and markdowns instead of the county-level HHI confirms the results. The so-called poor man's markup and markdown estimates correspond to the ratio of the branch-level average loan rate over the policy rate and the policy rate over the branch-level deposit rate, taking out time trends.<sup>11</sup>

---

<sup>11</sup>More specifically, I regress the branch-level markup/markdown on time and branch-fixed effects and isolate the branch-fixed effect as an estimate for the markdown/markup. For this step, the sample is restricted to pre-2009, as markups would be overestimated at the zero lower bound.

Figure 2: Bank concentration by county in 2019



Notes: 2019 HHI by county based on deposit holdings. Source: FDIC Summary of Deposits.

**Measuring bank capitalization.** Since the financial crisis, risk-weighted measures have become an integral part of bank regulation. I measure bank capitalization as the tier 1 risk-based capital ratio, a key pillar of the Basel III requirements. This is also in line with the theoretical model.<sup>12</sup> Robustness checks using the *core-capital ratio*, the *total risk-based capital ratio* and the *equity capital to asset ratio* yield similar results (Appendix A.4) since all measures are strongly correlated at the bank level.

**Measuring monetary policy.** I measure policy changes using monetary policy surprises computed from financial market variable changes within 30 minutes around Federal Open Market Committee meetings using the approach by Nakamura and Steinsson (2018) extended up to 2019. The surprises correspond to the first principal component of high-frequency movements in federal funds and Eurodollar futures with one year or less maturity.<sup>13</sup> The policy indicator captures, therefore, a forward guidance component consistent with the short-term loan rate maturity. Other monetary policy surprises, Romer and Romer (2004)'s narrative monetary policy shocks, and raw changes in the federal funds rate confirm the results (Appendix A.5).

<sup>12</sup>For details, see Bank for International Settlements (BIS) on Basel III. Mortgage loans enter with a risk weight of 20-100% into risk-weighted asset calculation used by the BIS in practice for regulation.

<sup>13</sup>The principal component analysis includes five futures: (i) the current month, (ii) and three-month ahead federal funds, and the eurodollar at the horizons of (iii) two, (iv) three, and (v) four quarters.

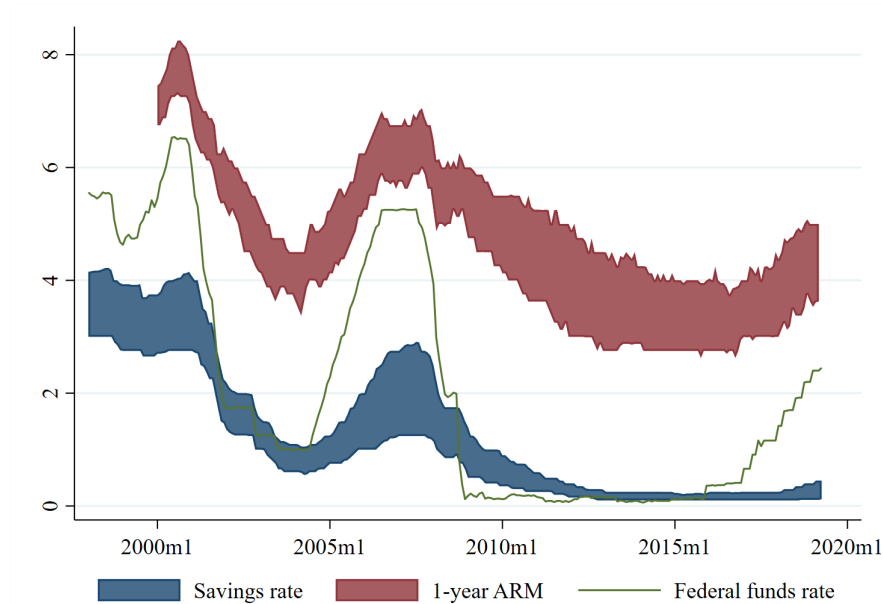
## 4 Empirical Findings

This section presents novel insights on loan and deposit rates across locations, bank institutions, and time. First, I document substantial heterogeneity in rate-setting behavior - across counties and bank institutions. Second, I study the role of local bank concentration and bank capitalization in the pass-through of monetary policy using state-dependent local projections for explaining time-varying cross-sectional dispersion.

### 4.1 Rate Dispersion across Space and Time

Figure 3 plots the interquartile range (IQR) of the deposit and loan rates across all surveyed branches, along with the federal funds rate.<sup>14</sup> A couple of facts summarize the evidence:

Figure 3: Deposit and loan rate IQR across bank branches



*Notes:* The shaded areas reflect the IQR of the 1-year adjustable mortgage rate and deposit rate for money market accounts with deposits of \$25,000 from January 1998 to March 2019. The solid line represents the federal funds rate. Source: RateWatch, Federal Reserve Economic Data.

**Dispersion within banks and locations.** Bank loan and deposit rates are dispersed in the cross-section, both across locations within a bank institution and across institutions within a given location. The IQR measures total dispersion between 50 and 200 basis points in the

<sup>14</sup>Appendix A.6 expands the analysis to a broader set of loan and deposit rates. The focus on short-term rates abstracts from term premium effects. The 30-year fixed rate's cross-sectional dispersion is relatively small. Banks typically do not keep these loans on their balance sheets, selling or securitizing them. The ARM share was above 50% before 2007, then declined. Source: CoreLogic.



cross-section varying over time. Assuming a \$175,000 mortgage, the discrepancy in loan rates results in an annual interest payment differential of \$600 to \$2,400. Similarly, LendingTree.com economists suggest consumers refinance their loans when the rate declines by about 50 basis points (see, e.g., MarketWatch) suggesting that the observed cross-sectional dispersion is of economic significance and importance to households.

Table 2 decomposes the average loan and deposit rate dispersion (i.e., IQR) into dispersion within locations and institutions. Focusing on loan rate dispersion in the upper part, within-location dispersion is higher than within-bank dispersion, at 0.97 versus 0.31, suggesting marginal costs play a more significant role than local concentration and the relevance of the bank’s cost structure. The average deposit rate dispersion shown in the bottom part is smaller, at 0.50 and 0.19, for within-location and within-bank, suggesting that costs and market power play a relatively more minor role. Telephone interviews with loan officers at large US banks (e.g., Chase and PNC) conducted in August 2019 shed light on the underlying reasons for different pricing strategies within banks: Institutions set rates strategically across locations depending on their local market share and origination costs.

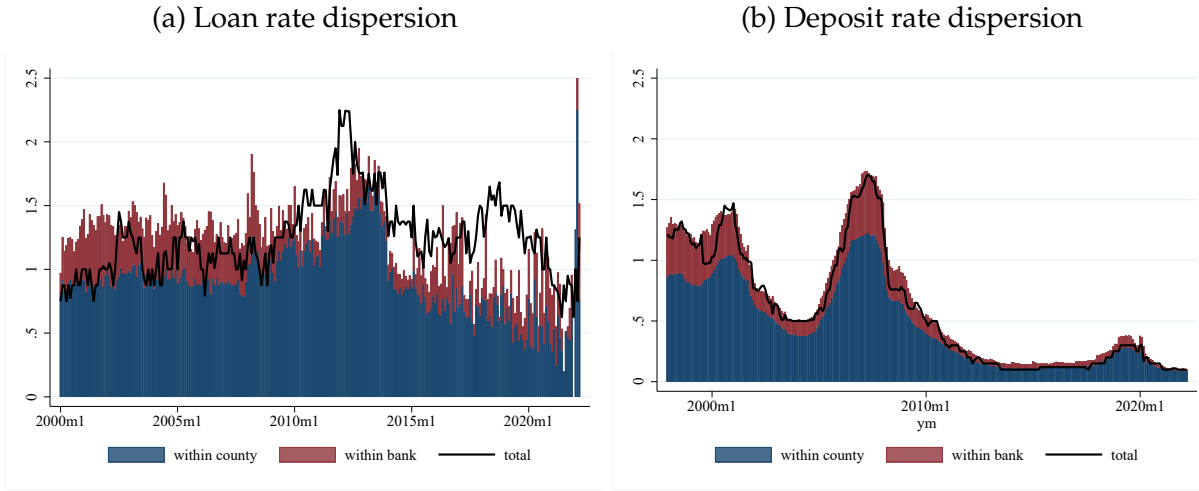
Table 2: Dispersion within-location and within-bank

	$mean(\overline{IQR}_t^{loc})$	$mean(\overline{IQR}_t^{bank})$
$r_t^l$	0.97	0.31
$r_t^d$	0.50	0.19

Notes:  $r_t^l$  reflects loan rate,  $r_t^d$  deposit rate,  $mean(\overline{IQR}_t^{loc})$  within-county average dispersion, and  $mean(\overline{IQR}_t^{bank})$  within-bank average dispersion. Loan rate: 1-year adjustable mortgage rate; deposit rate: money market accounts with deposits of \$25,000. Source: RateWatch.

To understand how dispersion evolves, Figure 4 decomposes loan and deposit rate dispersion into variations within-county and within-bank for each point in time. As seen in Table 2, within-county dispersion is much more significant for loan and deposit rates than within-bank across the entire period, reflected by the longer bars. Quantitatively, total and within-county loan rate dispersion exceeds the variation observed in deposit rates, especially during the low-interest rate environment.

Figure 4: Decomposition of rate dispersion



Notes: Decomposition of the total loan and deposit rate dispersion into within-county and within-bank dispersion. Loan rate: 1-year adjustable mortgage rate; deposit rate: money market accounts with deposits of \$25,000. Source: RateWatch.

**Countercyclical rate dispersion.** Loan and deposit rate dispersion varies with the federal funds rate. Loan rate dispersion tends to be high during times of low federal funds rates, and deposit rate dispersion tends to be high during times of high federal funds rates. Correspondingly, the correlation between loan rate dispersion and the federal funds rate is -0.55, and 0.89 for deposit rate dispersion, as shown in the right column of Table 3. Similarly, loan rate dispersion is 28 basis points higher for low rates, while deposit rate dispersion is 78 points higher for high rates. The negative correlation between loan rate dispersion and the federal funds rate suggests that banks' marginal costs are more heterogeneous during low versus high rate periods as capital requirements tighten.

Table 3: Dispersion for low and high federal funds rates

	$\rho(r_t^f, \overline{IQR}_t)$	$mean(\overline{IQR}_t   r_t^f < 2)$	$mean(\overline{IQR}_t   r_t^f \geq 2)$
$r_t^l$	-0.55	1.37	1.09
$r_t^d$	0.89	0.32	1.10

Notes:  $\rho$  reflects the correlation coefficient of dispersion and the federal funds rate,  $r_t^f$ ;  $mean$ , is the conditional mean of loan rate,  $r_t^l$ , and deposit rate,  $r_t^d$ , IQRs during low, ( $r_t^f < 2$ ), and high, ( $r_t^f \geq 2$ ), federal funds rate periods. Loan rate: 1-year adjustable mortgage rate; deposit rate: money market accounts with deposits of \$25,000. Source: RateWatch, Federal Reserve Economic Data.

## 4.2 Monetary Policy Pass-Through in the Cross Section

This section examines pass-through dynamics using local projection methods (Jordà, 2005), which provide a flexible framework and allow for heterogeneity and state dependency. The analysis focuses on the speed and extent of monetary policy pass-through, i.e., how fast and completely banks pass cost changes to consumers and how much it varies across locations and bank institutions.

The baseline model estimates the pass-through of monetary policy shocks to loan and deposit rates at each horizon,  $h \in [0, H]$ , by regressing branch  $i$ 's retail rate adjustment in location  $c$ ,  $r_{i,c,t+h} - r_{i,c,t-1}$ , on the monetary policy shock,  $s_t$ , interacted with the variable of interest,  $X_{i,c,t-1}$ , separately for loan and deposit rates:

$$r_{i,c,t+h} - r_{i,c,t-1} = \alpha_i^h + \beta^h s_t + \underbrace{\gamma^h s_t \times X_{i,c,t-1}}_{\substack{\text{local HHI or} \\ \text{bank capitalization}} + \theta^h X_{i,c,t-1} + \eta^h Z_{c,t} + \epsilon_{t+h,i,c} \quad (6)$$

where  $r_{i,c,t+h} - r_{i,c,t-1}$  reflects the loan or deposit rate change between  $t + h$  and  $t - 1$ . The regression is estimated for each horizon  $h$  and includes branch fixed effects,  $\alpha_i^h$ , and controls for national and local economic conditions,  $Z_{c,t}$ . The set of controls includes two lags of the county-level and national unemployment rate, real GDP growth, CPI inflation, county-level median debt-to-income ratio, county-level house price growth, lags of the dependent variable and monetary shock, and a dummy for the zero lower bound period. To address endogeneity concerns in the level of local HHI and bank capitalization, I use the lagged values of the interaction variables.<sup>15</sup> The main coefficient of interest in equation (6) is  $\gamma^h$ , the local HHI or bank capitalization's marginal effect on pass-through.  $\beta^h$  serves as a reference point to indicate average pass-through.<sup>16</sup> To facilitate the interpretation of the marginal effect of bank concentration and capitalization, the impulse responses are presented for high and low states, defined as two standard deviations above or below the mean of characteristic  $X_{i,c,t}$ . Accordingly, pass-through in the high and low state is calculated as  $\beta^h + \gamma^h (m^X \pm 2sd^X)$ . This representation simplifies interpretation but maintains a continuous interaction term. The monetary shock is scaled to increase the federal funds rate by one percentage point on impact.

**Local bank concentration.** Panels (a) and (b) of Figure 5 present impulse response functions for branch-level loan and deposit rates to a monetary shock at both a high

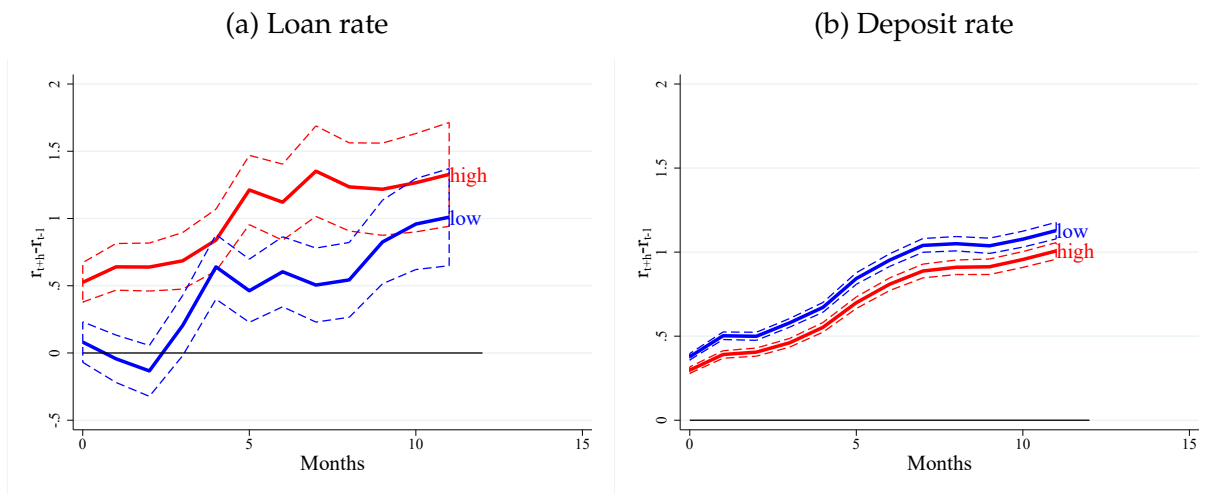
<sup>15</sup>To control for time trends in the underlying bank capital ratio variable, I use the deviation from the period average. The obtained estimate is then based on cross-sectional differences, not on time differences.

<sup>16</sup>Adding time dummy variables yields qualitatively similar results but provides no benchmark.

and low bank concentration level. High-concentration branches adjust loan rates more in response to the shock than low-concentration branches, by about 50 basis points on impact and consistently over ensuing months. In the low-concentration region, pass-through is incomplete, i.e., less than one after 12 months. The findings are in line with the predictions from the heterogeneous pass-through model in Section 2. Banks operating in high-concentration markets serve customers with relatively low demand elasticity and exhibit high market power, leading to higher markups and pass-through. The divergence of loan rates across branches in response to a monetary shock also manifests in a widening dispersion during policy changes in Figure 3.

By contrast, high-concentration branches adjust deposit rates less in response to the shock than low-concentration branches by about 10 basis points on impact and over ensuing months. The finding on the deposit side is consistent with Drechsler et al. (2017): high-concentration branches increase deposit spreads in response to a change in the federal funds rate.

Figure 5: Impulse responses of rates by local bank concentration



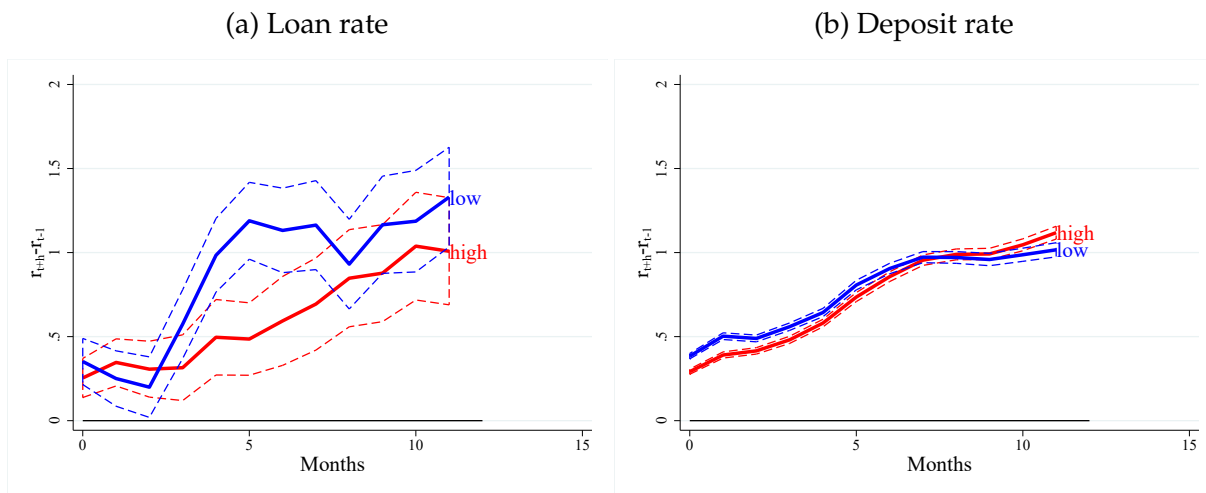
Notes: Impulse response functions of the 1-year adjustable mortgage rate and deposit rate for money market accounts with deposits of \$25,000 to a monetary policy shock at both high and low local bank concentrations, calculated as  $\beta^h + \gamma^h (m^{HHI} \pm 2sd^{HHI})$ . Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

The general pattern holds across monetary policy shocks and is robust to using raw changes in the federal funds rate (Figures A.10 and A.11 in Appendix A.5). If there is any difference, the pattern is more pronounced for loan rate pass-through using monetary policy shocks than for changes in the federal funds rate (similar to Bluedorn et al., 2017). Similarly, using the mortgage market concentration measure yields similar results for loan rate pass-through (Appendix A.3).

**Bank capitalization.** Banks with low capitalization and those with relatively illiquid balance sheets respond more to monetary policy (e.g., Kashyap and Stein, 2000). Banks with a relatively low bank capital ratio will adjust loan rates more to changes in funding costs and benefit less from a looser capital constraint.

Panels (a) and (b) of Figure 6 plot the branch-level loan and deposit rate impulse response functions to a monetary shock for low and high bank capital ratios. Low-capitalized banks adjust loan rates by more, in line with the simple model. However, bank capitalization seems to play a lesser role than concentration, particularly so for deposit rates; there is a smaller difference in impulse responses, i.e., the coefficient for capitalization interaction term  $\hat{\gamma}^h$  is an order of magnitude smaller for deposit versus loan rates. The temporary divergence of loan rates across banks in response to a monetary shock also explains a widening dispersion during monetary policy changes in Figure 3.

Figure 6: Impulse responses of rates by bank capitalization



*Notes:* Impulse response functions of the 1-year adjustable mortgage rate and deposit rate for money market accounts with deposits of \$25,000 to a monetary policy shock at both high and low bank capitalization, calculated as  $\beta^h + \gamma^h$  ( $m\% \pm 2sd^h$ ). Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

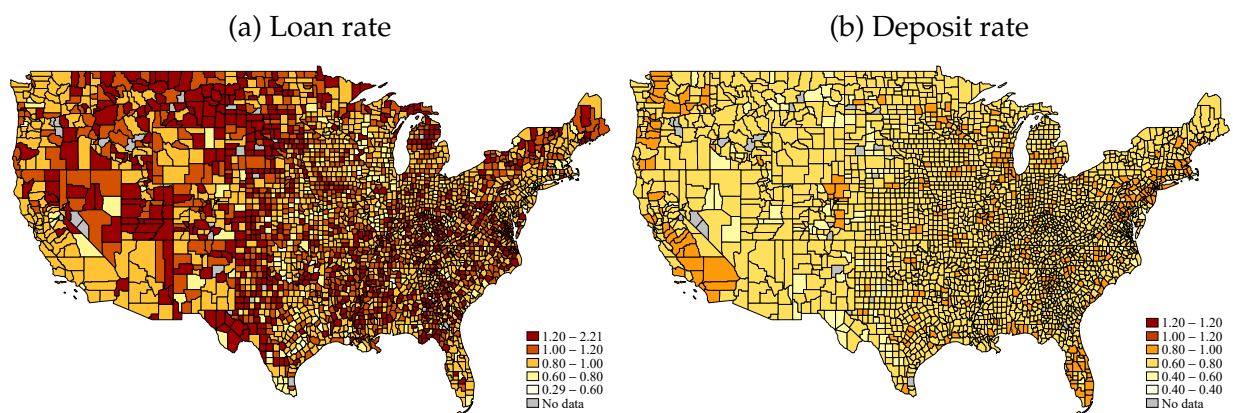
The general pattern holds across monetary policy shocks and using changes in the federal funds rate (Figures A.12 and A.13 in Appendix A.5). For capitalization, the other shocks reveal significant differences for loan rate pass-through in banks with low and high capitalization. Further, the results are robust to alternative capitalization measures (Appendix A.4).

**Heterogeneous pass-through across US counties.** What are the implications of differences in local market structures for pass-through at the regional level? How large are

the differences across counties in the United States? County-level pass-through,  $PT_{c,t}^h$ , depends on two factors:<sup>17</sup> the level of local bank concentration,  $HHI_{c,t-1}$ , and the average capitalization in a county,  $\bar{m}(\%)_{t-1}$ , and is calculated as:  $\hat{\beta}^h + \hat{\gamma}_1^h HHI_{c,t-1} + \hat{\gamma}_2^h \bar{m}(\%)_{t-1}$ .

There are large pass-through differences for both deposit and loan rates. Panels (a) and (b) of Figure 7 present the pass-through estimates for loan and deposit rates across US counties in 2019 at a horizon of six months,  $h = 6$ . The loan rate pass-through ranges from 0.29 to 2.21, implying that monetary policy reaches borrowers across counties unevenly. While the central bank’s policy strongly affects some counties, particularly in the center of the US, it leaves other parts mostly unaffected. Similarly, deposit rate pass-through ranges from 0.46 to 0.82, implying that monetary policy affects savers differently.

Figure 7: Estimated monetary policy pass-through after six months across US counties



Notes: Estimated monetary policy pass-through to the 1-year adjustable mortgage rate and deposit rate for money market accounts with deposits of \$25,000 after six months,  $h = 6$ , for 2019 by US county, calculated as  $\hat{\beta}^h + \hat{\gamma}_1^h HHI_{c,t-1} + \hat{\gamma}_2^h \bar{m}(\%)_{t-1}$ .  $HHI_{c,t-1}$  reflects the lagged county-level HHI and  $\bar{m}(\%)_{t-1}$  the weighted mean of the lagged capital ratio by county.

Figure A.15 in Appendix A.7 presents the maps for the estimated pass-through to loan and deposit rates in 1995. Comparing 1995 to 2019 indicates a higher estimated pass-through to loan rates in many counties in 2019, reflected by more red and dark-red counties. The higher local loan rate pass-through suggests that aggregate pass-through has increased. A back-of-the-envelope calculation states that the aggregate loan rate pass-through rose from 0.87 to 1.09 between 1995 and 2019, accounting for changes in local market concentration and bank capitalization and weighting counties by deposit holdings. In contrast, the aggregate deposit rate pass-through decreased from 0.80 to 0.75. Similar

<sup>17</sup>This exercise focuses on the role of bank concentration and capitalization, abstracting from other factors such as demographics, socioeconomic factors, ARM share that additionally might affect pass-through heterogeneity across counties.

evidence provides Appendix A.8 documenting an increase in pass-through to loan rates and lending over time using aggregate data. Section 5.2 further considers changes in markups and capital ratios over time for assessing changes in pass-through.

## 5 Quantitative Model

This section introduces a dynamic stochastic general equilibrium model to quantify the relative importance of market power and capital requirements for monetary policy pass-through and assess how rising bank concentration affects monetary transmission to the real economy using counterfactual analyses. The model features segmented financial markets, where patient households provide deposits to the banking sector, and impatient households and entrepreneurs demand credit for investment in housing and capital. A monetary authority sets the policy rate via a Taylor rule. Extending the simple model in Section 2, in which banks operate in an environment with monopolistic competition and bank capital regulation, the quantitative model additionally considers that changing the amount of lending and deposit holdings is costly. The remaining standard New Keynesian building blocks follow Gerali et al. (2010). The model details beyond the banking sector and calibration are relegated to Appendices B.1, B.2, and B.3.

### 5.1 The Banking Sector

The banking sector is divided into three parts: (i) representative wholesale management units (comparable to bank headquarters) with *heterogeneous bank capital requirements*, (ii) a continuum of retail deposit branches in *different locations* operating under monopolistic competition, and (iii) a continuum of loan branches in *different locations* operating under monopolistic competition.<sup>18</sup>

#### 5.1.1 Wholesale Unit

Wholesale units manage funds between retail deposit and loan branches and are subject to a size-dependent, bank-specific bank capital requirement. Each wholesale unit  $i$  finances loans to households,  $L_{i,c,t}^H$ , and entrepreneurs,  $L_{i,c,t}^{bE}$  in different locations  $c$ . Hence, total bank lending,  $L_{i,t}$ , equals  $L_{i,t} = L_{i,t}^{bH} + L_{i,t}^{bE}$ , where  $L_{i,t}^{bH}$  and  $L_{i,t}^{bE}$  equal the sum of lending across all locations  $c$ :  $L_{i,t}^{bH} = \sum_{c=1}^C L_{i,c,t}^{bH}$  and  $L_{i,t}^{bE} = \sum_{c=1}^C L_{i,c,t}^{bE}$ . The wholesale unit obtains funds from deposit branches,  $D_{i,t}$ , and holds bank capital,  $K_{i,t}$ .  $D_{i,t}$  equals the sum of

<sup>18</sup>The banking sector expands on Gerali et al. (2010) featuring a representative wholesale management unit and one market for loans and deposits.

deposit across all locations  $c$ :  $D_{i,t} = \sum_{c=1}^C D_{i,c,t}$ . In order to cover incidental management costs, the wholesale unit retains the previous period's profits. Bank capital,  $K_{i,t}$ , evolves according to equation (7):

$$\pi_t K_{i,t} = (1 - \delta^b) K_{i,t-1} + \Pi_{i,t-1}^b, \quad (7)$$

where  $\Pi_{i,t-1}^b$  reflects retained profits,  $\delta^b$  the marginal management costs for bank capital, and  $\pi_t$  the inflation rate. Any deviation from the required bank capital,  $\nu_i$ , is modeled with a quadratic cost function,  $\mathbb{A}_K \left( \frac{K_{i,t}}{L_{i,t}} \right) = \frac{\kappa_K}{2} \left( \frac{K_{i,t}}{L_{i,t}} - \nu_i \right)^2$ , governed by cost parameter  $\kappa_K$ .<sup>19</sup>  $\nu_i$  differs by wholesale unit  $i$ , implying heterogeneous costs for banks due to the bank regulations.

The wholesale unit makes profits from providing wholesale funding to its retail loan branches,  $L_{i,t}$ , at the wholesale funding rate,  $R_{i,t}^b$ , minus expenses paid to deposit branches,  $D_{i,t}$ , at the wholesale lending rate,  $R_{i,t}^d$ . The wholesale lending rate,  $R_{i,t}^d$ , equals the central bank policy rate,  $r_t^f$ , in equilibrium since the wholesale bank could always obtain marginal funds at the central bank at the price of the policy rate. The wholesale unit discounts future profits with the stochastic discount factor of the patient household,  $\Lambda_{0,t}^P$ , and maximizes:

$$\max_{L_{i,t}, D_{i,t}} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[ R_{i,t}^b L_{i,t} - R_{i,t}^d D_{i,t} - \mathbb{A}_K \left( \frac{K_{i,t}}{L_{i,t}} \right) K_{i,t} \right], \quad (8)$$

subject to the wholesale unit's balance sheet constraint:

$$L_{i,t} = D_{i,t} + K_{i,t}. \quad (9)$$

Solving the wholesale unit's maximization problem and rewriting the first-order condition yields the wholesale funding rate as a function of bank capital ratio,  $\nu_i$ , and policy rate,  $r_t^f$ :

$$R_{i,t}^b = r_t^f - \kappa_K \left( \frac{K_{i,t}}{L_{i,t}} - \nu_i \right) \left( \frac{K_{i,t}}{L_{i,t}} \right)^2. \quad (10)$$

Outside the steady state, the loan rate depends inversely on the bank capitalization, similar to the simple model in Section 2. This relation results from the negative cost term in parentheses during a monetary easing: A decline in the policy rate expands bank lending,  $L_{i,t}$ , by more than bank capital,  $K_{i,t}$ . The more so, the higher the cost parameter,  $\kappa_K$ , and steady-state bank capital ratio,  $\nu_i$ .

<sup>19</sup>Instead of explicitly modeling the capital constraint, the quadratic cost function avoids any non-linearities while otherwise similar (e.g., Gerali et al., 2010, Brunnermeier and Koby, 2018). An extension could consider an asymmetric quadratic cost function with high deviation costs for low capitalization only.



### 5.1.2 Retail Deposit Branches

In each location  $c$ , retail deposit branches collect deposits from patient households and store these at the wholesale unit at the policy rate  $r_t^f$ .<sup>20</sup> The deposit branches earn a positive spread on the deposit rate due to monopolistic deposit market competition in each local market  $c$ . Deposit branches incur adjustment costs from changing deposits, as attracting new customers requires additional processing and advertising. Flannery (1982) regards deposits as “quasi-fixed” inputs. Hence, keeping a constant deposit funding stock may explain why deposit rates exceeded the federal funds rate for some periods in Figure 3. Adjustment costs,  $\mathbb{A}_D$ , are proportional to aggregate deposit expenses,  $\bar{r}_{c,t}^d \bar{D}_{c,t}$ , expressed as deviations from the steady-state deposit level,  $D_{c,ss}$ , and take the form:  $\frac{\kappa_d}{2} \left( \frac{D(r_{c,t}^d)}{D(r_{c,ss}^d)} - 1 \right)^2$ , governed by cost parameter  $\kappa_d$ . Each deposit branch maximizes the sum of future profits discounted by the patient household’s stochastic discount factor  $\Lambda_{0,t}^P$ :

$$\max_{r_{c,t}^d} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[ r_t^f D(r_{c,t}^d) - r_{c,t}^d D(r_{c,t}^d) - \mathbb{A}_D (D(r_{c,t}^d)) \bar{r}_{c,t}^d \bar{D}_{c,t} \right], \quad (11)$$

subject to the local deposit supply function:

$$D(r_{c,t}^d) = \left( \frac{r_{c,t}^d}{\bar{r}_{c,t}^d} \right)^{-\epsilon^{d,c}} \bar{D}_{c,t}, \quad (12)$$

where  $\bar{r}_{c,t}^d$  and  $\bar{D}_{c,t}$  reflect the aggregate deposit rate and deposits in location  $c$ . The local deposit supply elasticity,  $\epsilon^{d,c}$ , depends on the local market structure and differs by location  $c$ . After imposing symmetry, ( $D_{c,t} = \bar{D}_{c,t}$ ,  $r_{c,t}^d = \bar{r}_{c,t}^d$ ), the deposit branch’s optimality condition is:

$$-\epsilon^{d,c} \frac{r_t^f}{r_{c,t}^d} + (\epsilon^{d,c} - 1) + \epsilon^{d,c} \kappa_d \left( \frac{D_{c,t}}{D_{c,ss}} - 1 \right) \frac{D_{c,t}}{D_{c,ss}} = 0 \quad (13)$$

The branch determines the deposit rate based on (i) deposit supply elasticity,  $\epsilon^{c,d}$ , (ii) the policy rate,  $r_t^f$ , and (iii) deviation from the steady-state deposit level. Accordingly, cross-sectional heterogeneity may emerge in deposit rates due to differences in deposit supply elasticity  $\epsilon^{c,d}$ , as shown in the simple model, and adjustment costs,  $\kappa_d$ , or the steady-state deposit level (i.e., branch size).

<sup>20</sup>The deposit branch’s optimization decision is thus independent of its affiliation to wholesale unit  $i$ . For better visualization and without loss of generality, the subscript  $i$  is omitted in the deposit branch description.

### 5.1.3 Retail Loan Branches

In each location  $c$ , retail loan branches of type  $\tau$ , with  $\tau \in \{bH, bE\}$ , finance loans to impatient households,  $L_{i,c,t}^{bH}$  or entrepreneurs,  $L_{i,c,t}^{bE}$ . The retail loan branches belong to headquarters  $i$  and obtain funding at the headquarters-specific wholesale funding rate,  $R_{i,t}^b$ , respectively. Similar to the retail deposit branches, retail loan branches earn a positive spread due to monopolistic loan market competition. Each loan branch incurs costs from adjusting lending,  $\mathbb{A}_l$ , as expanding the loan portfolio increases processing costs and requires additional staff. Adjustment costs are proportional to aggregate loan returns,  $\bar{r}_{c,t}^\tau \bar{L}_{c,t}^\tau$ , defined in terms of deviations from the steady-state loan level,  $L_{i,c,ss}^\tau$ , and take the form:  $\frac{\kappa_\tau}{2} \left( \frac{L_{i,c,t}^\tau}{L_{i,c,ss}^\tau} - 1 \right)^2 \quad \forall \tau \in \{bH, bE\}$ , governed by cost parameter  $\kappa_\tau$ . Each loan branch belonging to headquarters  $i$  in location  $c$  maximizes the sum of future profits discounted by the patient household's stochastic discount factor  $\Lambda_{0,t}^P$ :

$$\max_{r_{i,c,t}^\tau} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[ r_{i,c,t}^\tau L^\tau(r_{i,c,t}^\tau) - R_{i,t}^b L^\tau(r_{i,c,t}^\tau) - \mathbb{A}_\tau (L^\tau(r_{i,c,t}^\tau)) \bar{r}_{c,t}^\tau \bar{L}_{c,t}^\tau \right] \quad (14)$$

subject to the local loan demand function:

$$L^\tau(r_{i,c,t}^\tau) = \left( \frac{r_{i,c,t}^\tau}{\bar{r}_{c,t}^\tau} \right)^{-\epsilon^{\tau,c}} \bar{L}_{c,t}^\tau \quad \forall \tau \in \{bH, bE\} \quad (15)$$

where  $\bar{r}_{c,t}^\tau$  and  $\bar{L}_{c,t}^\tau$  reflect the aggregate loan rate and loans of type  $\tau$  in location  $c$ . The local loan demand elasticity,  $\epsilon^{\tau,c}$ , depends on the local market structure and differs by location  $c$ . After imposing symmetry, the loan branch's optimality condition is:

$$-(\epsilon^{\tau,c} - 1) + \epsilon^{\tau,c} \frac{R_{i,t}^b}{r_{i,c,t}^\tau} + \epsilon^{\tau,c} \kappa_\tau \left( \frac{L_{i,c,t}^\tau}{L_{i,c,ss}^\tau} - 1 \right) \frac{L_{i,c,t}^\tau}{L_{i,c,ss}^\tau} = 0 \quad \forall \tau \in \{bH, bE\} \quad (16)$$

The loan rate decision is determined by: (i) loan demand elasticity,  $\epsilon^{\tau,c}$ , (ii) wholesale funding rate,  $R_{i,t}^b$ , and (iii) loan portfolio changes. Hence, heterogeneity in monetary policy pass-through to retail rates can be explained by differences in market power,  $\epsilon^{\tau,c}$ , adjustment costs,  $\kappa_\tau$ , steady-state loans volumes (i.e., branch size), and bank capital constraint determinants,  $\nu_i$  and  $\kappa_K$ .

### 5.1.4 Aggregation

In the aggregate, lending to households and entrepreneurs in the economy equals the sum of individual lending by all bank headquarters  $i$ :  $L_t^\tau = \sum_{i=1}^I L_{i,c,t}^\tau \quad \forall \tau \in \{bH, bE\}$ .

Similarly, aggregate deposits in the economy equal the sum of individual deposit holdings by all bank headquarters  $i$ :  $D_t = \sum_{i=1}^I D_{i,t}$ . The loan and deposit rates are weighted averages by bank market share  $\alpha^b$  and local market size  $\alpha^m$ :

$$r_t^j = \sum_{c=1}^C \sum_{i=1}^I \alpha^b \alpha^m r_{i,c,t}^j \quad \forall j \in \{d, bH, bE\},$$

where  $\alpha^b = \frac{\omega_i}{\omega} \quad \forall \omega \in \{D, L^{bH}, L^{bE}\}$  and  $\alpha^m = \frac{\omega_c}{\omega} \quad \forall \omega \in \{D, L^{bH}, L^{bE}\}$  and are initially taken as exogenous.

To capture bank heterogeneity in a tractable framework, assume two types along each dimension: regional and giant banks, denoted by the superscripts  $r$  and  $g$ , paired with a continuum of branches in low- and high-concentration markets, denoted  $l$  and  $h$ . The approach yields four types of bank branches: (i) regional banks in low-concentration markets, (ii) regional banks in high-concentration markets, (iii) giant banks in low-concentration markets, and (iv) giant banks in high-concentration markets. Correspondingly, there is a share of branches operating in high-concentration markets,  $\alpha^m$ , and giant banks,  $\alpha^b$ . Table 4 presents the derived branch-specific loan and deposit rates depending on local market structure,  $\epsilon^{s,c} \quad \forall c \in \{l, h\}$ ,  $s \in \{d, bH, bE\}$ , and headquarters-specific marginal costs,  $R_{j,t} \quad \forall j \in \{r, g\}$ , abstracting from adjustment costs.

Table 4: Heterogeneous bank rates across bank types and markets

		Bank types		
		Regional	Giant	Share
Local market concentration	Low	$r_{l,r,t}^\tau = \frac{\epsilon^{\tau,l}}{\epsilon^{\tau,l}-1} R_{r,t}$	$r_{l,g,t}^\tau = \frac{\epsilon^{\tau,l}}{\epsilon^{\tau,l}-1} R_{g,t}$	$\alpha^m$
		$r_{l,t}^d = \frac{\epsilon^{d,l}}{\epsilon^{d,l}-1} r_t^f$	$r_{l,t}^d = \frac{\epsilon^{d,l}}{\epsilon^{d,l}-1} r_t^f$	
	High	$r_{h,r,t}^\tau = \frac{\epsilon^{\tau,h}}{\epsilon^{\tau,h}-1} R_{r,t}$	$r_{h,g,t}^l = \frac{\epsilon^{\tau,h}}{\epsilon^{\tau,h}-1} R_{g,t}$	$(1 - \alpha^m)$
		$r_{h,t}^d = \frac{\epsilon^{d,h}}{\epsilon^{d,h}-1} r_t^f$	$r_{h,t}^d = \frac{\epsilon^{d,h}}{\epsilon^{d,h}-1} r_t^f$	
Share		$\alpha^b$	$(1 - \alpha^b)$	

*Notes:* Branch-level loan rate of type  $\tau \in \{bH, bE\}$  depends on local market structure,  $\epsilon^{\tau,c} \quad \forall c \in \{l, h\}$ , and headquarters-specific marginal costs,  $R_{j,t} \quad \forall j \in \{r, g\}$ . Branch-level deposit rate depends on local market structure,  $\epsilon^{d,c} \quad \forall c \in \{l, h\}$ .  $(1 - \alpha^m)$  refers to high-concentration; and  $(1 - \alpha^b)$  giant banks' share.

Taking into account that deposit rates do not differ across bank headquarters, aggregate

retail rates and deposits and loans to households and entrepreneurs simplify to:

$$D_t = D_{h,r,t} + D_{l,r,t} + D_{h,g,t} + D_{l,g,t}$$

$$L_t^\tau = L_{h,r,t}^\tau + L_{l,r,t}^\tau + L_{h,g,t}^\tau + L_{l,g,t}^\tau \quad \forall \tau \in \{bH, bE\}$$

$$r_t^d = \alpha^m r_{l,t}^d + (1 - \alpha^m) r_{h,t}^d$$

$$r_t^\tau = \alpha^b \alpha^m r_{l,r,t}^\tau + (1 - \alpha^b) \alpha^m r_{l,g,t}^\tau + \alpha^b (1 - \alpha^m) r_{h,r,t}^\tau + (1 - \alpha^b) (1 - \alpha^m) r_{h,g,t}^\tau \quad \forall \tau \in \{bH, bE\}$$

## 5.2 Quantitative Assessment of the Rise in Bank Concentration

This section uses counterfactual analyses to quantify the implications of rising bank concentration for monetary policy pass-through, distinguishing between the *market power channel*, changes in the underlying *market environment*, and the *capital allocation channel*, shifts in the composition of the *banking sector*. The counterfactual analyses contrast monetary policy pass-through and transmission in an environment featuring a relatively low- and high-concentrated banking sector, calibrated to the US banking sector for 1994 and 2019. Specifically, I consider differences along (i) the *extensive margin*, i.e., the share of high-concentration markets,  $(1 - \alpha^m)$ , and giant banks,  $(1 - \alpha^b)$ , in line with US trends presented in Appendix B.4, and (ii) the *intensive margin*, trends in markups,  $\epsilon$ , and bank capital ratios,  $\nu$ , over time.

Table 5: Calibration of banking sector parameters

	Parameter	$\alpha^m$	$\alpha^b$	$\epsilon^d$	$\epsilon^{bH, bE}$	$\nu$
1994	Bank/Branch I	0.7	0.9	-2.60	2.51	0.09
	Bank/Branch II	0.3	0.1	-1.03	2.05	0.06
2019	Bank/Branch I	0.4	0.4	-0.99	1.68	0.12
	Bank/Branch II	0.6	0.6	-0.32	1.46	0.09

*Notes:* The row Branch/Bank I (Bank/Branch II) presents the calibration of  $\epsilon^d$ ,  $\epsilon^{bH}$ ,  $\epsilon^{bE}$  and  $\nu$  for the low-concentration market and regional bank (high-concentration market and giant bank) by period, 1994 and 2019.  $\alpha^m$  and  $\alpha^b$  reflect the share of low-concentration markets and regional banks, respectively.

Table 5 presents the calibration details for the counterfactual analyses. I calibrate low-concentration markets' share,  $\alpha^m$ , regional banks' share,  $\alpha^b$ , deposit supply elasticity,  $\epsilon^d$ , elasticities of loan demand from households,  $\epsilon^{bH}$ , and entrepreneurs,  $\epsilon^{bE}$ , and bank capital requirement,  $\nu$ , separately for low- and high-concentration markets and regional and giant banks, as well as two periods, 1994 and 2019.<sup>21</sup> I calibrate low-concentration markets' share,

<sup>21</sup>The 1994 and 2019 calibrations rely on bank data for the periods 2000-2008 and 2009-2019.

$\alpha^m$ , and deposit supply elasticity,  $\epsilon^{d,c}$ , and loan demand elasticity,  $\epsilon^{j,c} \forall j \in \{bH, bE\}$  for market  $c \in \{l, h\}$ .  $\alpha^m$  is derived from the county-level HHI distribution across time.  $\epsilon^{j,c} \forall j \in \{d, bH, bE\}$  is inferred from bank-level interest income and expense data and calibrated to the average cross-sectional, asset-weighted markups/markdowns and dispersion.<sup>22</sup> Giant banks are defined as those above \$100 billion in assets (in \$2018).<sup>23</sup> I calculate giant banks' share,  $(1 - \alpha^b)$ , and the annual weighted group means of the bank capital ratio,  $\nu$ , separately for giant and regional banks, defined as those with assets below \$100 billion. The adjustment cost parameter  $\kappa_\tau$  for  $\tau \in \{d, bH, bE\}$  is the same across specifications and calibrated to consistent results with the empirical part.

Comparing the parameter values for 1994 and 2019 shows an increased share of high-concentrated markets  $(1 - \alpha^m)$  from 0.3 to 0.6. At the same time, the share of giant banks  $(1 - \alpha^b)$  rose from 0.1 to 0.6. Similarly, the elasticities of loan demand and deposit supply  $\epsilon$  have decreased (in absolute value), implying an increase in markups and markdowns for both low and high-concentration branches. Likewise, capital ratios  $\nu$  have each increased for regional and giant banks.

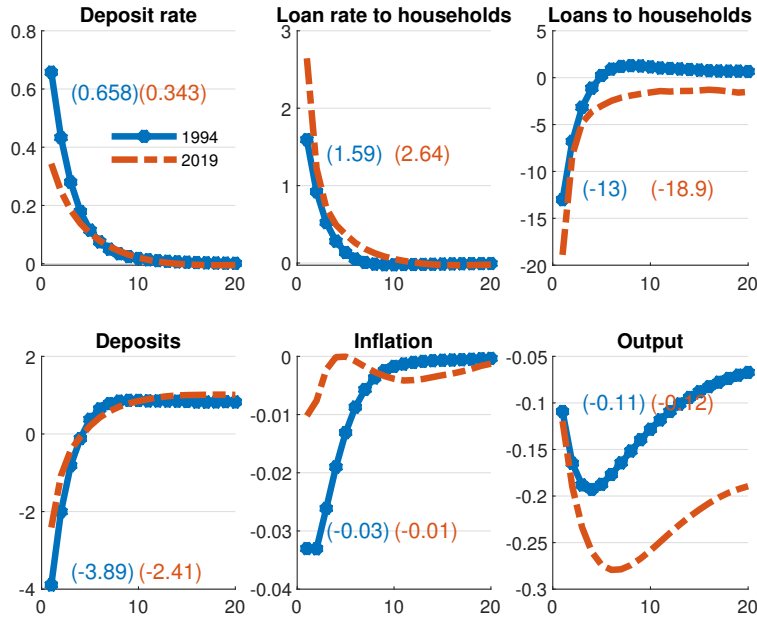
### 5.2.1 Total Effect of the Rise in Bank Concentration

Figure 8 shows impulse response functions to a monetary tightening under the low- and high-concentration banking sector calibration, labeled 1994 and 2019, respectively. Loan rate pass-through has increased over time, while deposit rate pass-through has declined. The strengthened loan rate pass-through is in line with increasing markups over time and a larger share of giant banks, whereas the dampened deposit rate pass-through goes back to the presence of higher markdowns. Similarly, loans to households declined by more but deposits by less in response to a monetary policy shock. Focusing on macroeconomic variables, transmission to output amplifies, i.e., output contracts by more in the medium run, but the effect on inflation dampens. Quantitatively, output declines by 9% more and inflation by 67% less on impact.

<sup>22</sup>The markup/markdown,  $m^j$ , of each bank  $j$  is calculated as the average markup over the federal funds rate excluding periods when the federal funds rate is below 1%, as markups/markdowns below are abnormally high/low and bias results. The implied  $\epsilon^j$  is inferred from the steady-state relationship between retail and policy rates and calculated as  $\epsilon^j = \frac{m^j}{m^j - 1}$ . The calibration of three parameters,  $\alpha^m$ ,  $\epsilon^{j,l}$ , and  $\epsilon^{j,h}$ , based on aggregate mean and standard deviation leaves one degree of freedom. I select  $\alpha^m$  to target an HHI threshold to minimize distance across moments: unconditional asset-weighted group means and dispersion and distance between model and data group means.

<sup>23</sup>Classification and cutoff follow the definition of the Federal Reserve Board of Governors for large financial institutions.

Figure 8: Impulse responses to a monetary tightening: low vs. high bank concentration



Notes: Impulse response functions to a positive monetary shock in an environment with low concentration banking sector, labeled 1994 (solid blue line), and high concentration banking sector, labeled 2019 (dashed red line). The calibration considers differences in  $(1 - \alpha^m)$ ,  $(1 - \alpha^b)$ ,  $\epsilon$ , and  $\nu$ . The impact effect is displayed in parentheses.

As banks adjust loan rates more in response to policy rate changes, borrowers also respond by demanding less credit, leading to a more significant contraction in credit. Consequently, firms and households invest less in capital and housing, which also causes a more significant output contraction. In contrast, banks adjust saving rates by less, leading to a smaller outflow of deposits. As saving becomes less attractive, households decide to consume more, counteracting the dampened demand, but also suffer from adverse income effects.

Adding borrowing constraints à la Iacoviello (2005) to households and firms lowers their sensitivity to loan rates, and compositional shifts in the banking sector become less important for macroeconomic aggregates, as shown in Figure B.2 in Appendix B.5. While the pass-through to loan and deposit rates in both banking sector environments is similar to Figure 8, the impact on aggregate lending and saving is almost muted in an environment with financial frictions on the borrower's side. Similar to before, the results still point towards a more dampened response to inflation in an environment with high bank concentration but no impact on the response of output.

## 5.2.2 Decomposition of the Rise in Bank Concentration

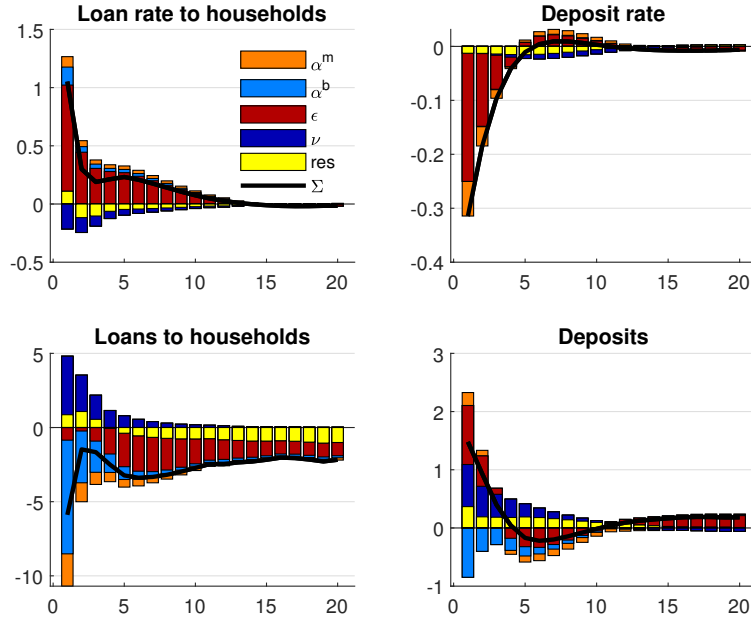
This section decomposes the total effect of rising bank concentration on monetary policy pass-through into five components and compares their relative contribution. The unified framework also allows for assessing interaction effects between the different components, particularly the *market power* and *capital allocation channels*. As summarized in equation (17), the total effect,  $\Sigma$ , considers the extensive and intensive margins. In particular, the total effect includes changes in: (i) share of low-concentration markets  $\alpha^m$ ; (ii) share of regional banks  $\alpha^b$ ; (iii) loan demand and deposit supply elasticities, i.e., markups and markdowns  $\epsilon$ ; (iv) bank capital ratio  $\nu$ ; and (v) an interaction effect  $res$ .

$$\Delta_{t+h}^{\Sigma} = \underbrace{\Delta_{t+h}^{\alpha^m}}_{\% \text{ high-concentration markets}} + \underbrace{\Delta_{t+h}^{\alpha^b}}_{\% \text{ regional banks}} + \underbrace{\Delta_{t+h}^{\epsilon}}_{\text{markup}} + \underbrace{\Delta_{t+h}^{\nu}}_{\text{bank capital ratio}} + \underbrace{res_{t+h}}_{\text{interaction}} \quad \forall h \in [0, H], \quad (17)$$

where  $\Delta_{t+h}^j \quad \forall j \in \{\Sigma, \alpha^m, \alpha^b, \epsilon, \nu, res\}$  reflects the difference between the impulse response functions of each variable from 2019 and 1994 under calibration  $j$ , calculated as  $\Delta_{t+h}^j = IRF_{t+h}^{j,2019} - IRF_{t+h}^{j,1994}$  for each horizon. More specifically, to assess the marginal contribution of one component, I marginally change its parameter value, e.g.,  $\alpha^b$ , while keeping the other parameters constant. The interaction effect  $res_{t+h}$  equals the residual.

Figure 9 decomposes the total change in monetary policy pass-through for the aggregate deposit rate, loan rate, household loans, and deposits. Increasing markups,  $\epsilon$ , primarily drive the total increase in loan rate pass-through. Compositional shifts in  $\alpha^m$  and  $\alpha^b$ , and the interaction effect  $res$  further strengthen pass-through. In contrast, increases in bank capital ratios,  $\nu$ , dampen pass-through and counteract the other forces. Pass-through to the deposit rate declined due to increasing markdowns from shifts along the intensive and extensive margins (i.e., increases in  $\alpha^m$  and decreases in  $|\epsilon|$ ), whereas capital ratios do not alter deposit rate dynamics, as shown already by equation (2) in Section 2.

Figure 9: Decomposing the change in monetary pass-through to rates and bank aggregates



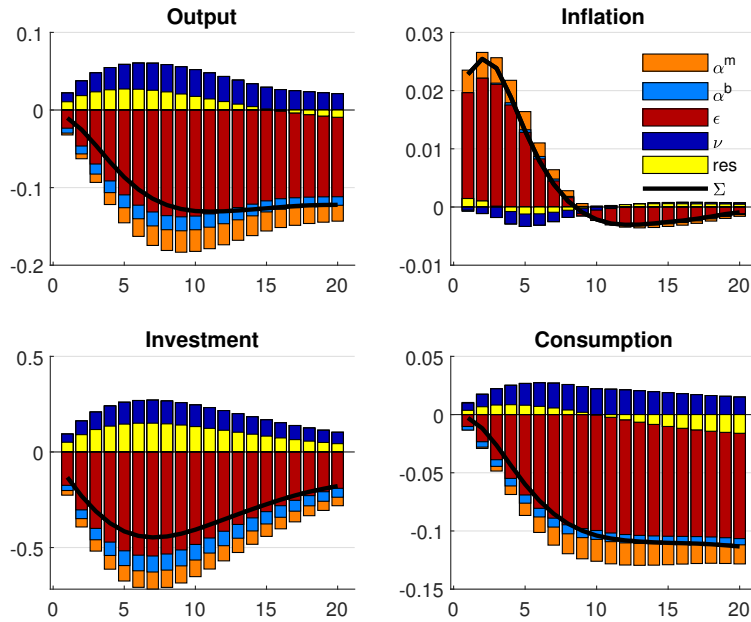
*Notes:* Differences in impulse response functions to a monetary shock between high and low bank concentration environments. Decomposition of the total effect  $\Sigma$  into five components: changes in the share of low-concentration markets  $\alpha^m$ , the share of regional banks  $\alpha^b$ , the elasticity of loan demand and deposit supply  $\epsilon$ , the bank capital ratio  $\nu$ , and an interaction effect  $res$ . The x-axis represents the horizon.

Aggregate loans and deposits present a near mirror image of the aggregate loan and deposit rate. The decrease in transmission is due to compositional effects but also markups  $\epsilon$  and interaction effects play a substantial role. Rising markups interact with financial frictions and lead to a more substantial decline in lending, which a partial analysis would fail to capture.

Figure 10 presents the decomposition for macroeconomic variables. Total monetary policy transmission to output, investment, and consumption strengthened in 2019; that is, those variables declined more in response to a positive shock, also reflected by the negative difference. The amplification results primarily from rising markups  $\epsilon$  explaining approx. 60% on impact. The rise in bank capital ratios  $\nu$  and interaction effect  $res$  counteracted the amplification. Hence, a partial analysis leaving out interaction effect  $res$  would overstate the total effects by 26% on impact. In contrast to output, monetary policy transmission to inflation is more muted, indicating that rising bank concentration has opposite implications for the transmission to prices and output. Similarly, the dampened transmission to inflation is attributed to rising markups and shifts towards concentrated markets. Overall, the importance of markup shifts indicates that secular trends outweigh composition effects.



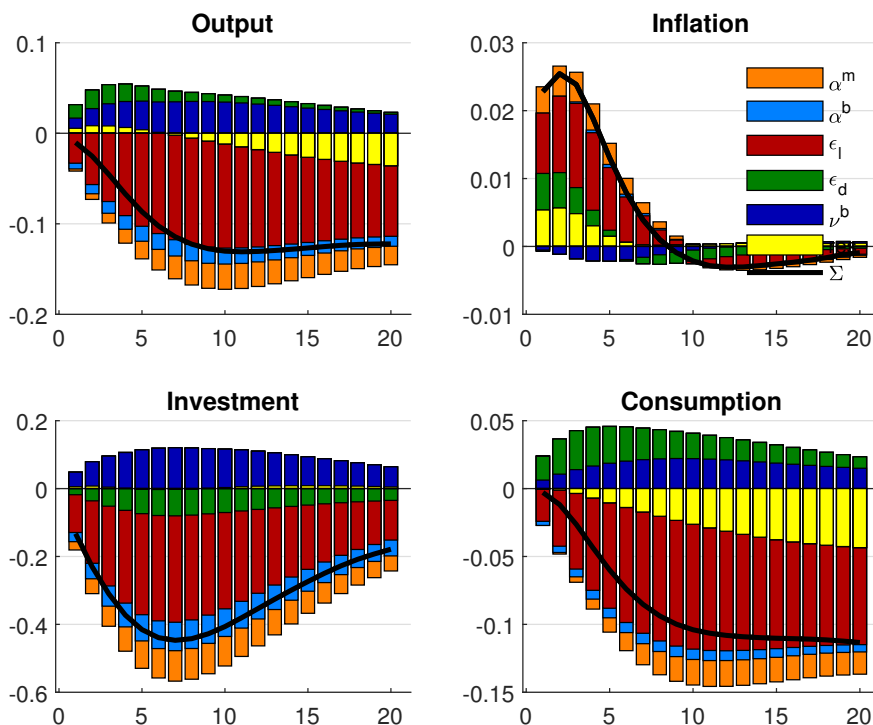
Figure 10: Decomposing the change in monetary policy transmission to the macroeconomy



*Notes:* Differences in impulse response functions to a monetary shock between high and low bank concentration environments. Total effect  $\Sigma$  is decomposed into five components: changes in the share of low-concentration markets  $\alpha^m$ , the share of regional banks  $\alpha^b$ , the loan demand and deposit supply elasticity  $\epsilon$ , the bank capital ratio  $\nu$ , and an interaction effect  $res$ . The x-axis represents the horizon.

**Relevance of deposit vs. loan market power.** Existing literature has focused on the importance of deposit market power. For instance, Drechsler et al. (2017)'s deposits channel of monetary policy shows that higher deposit market concentration, i.e., higher deposit market power, indirectly affects lending. That is, a sluggish pass-through to deposit rates with increasing spreads in response to monetary tightening leads to an outflow of deposits, which ultimately causes a contraction in lending. In other words, even abstracting from market power on loans, there are spillovers on lending from market power on deposit markets. Considering deposit and loan market power in a unified framework allows for assessing the relative importance of either. For this, Figure 11 decompose the *market power channels* contribution into two parts: (i) shifts in markdowns  $\epsilon_d$  and (ii) shifts in markups  $\epsilon_l$ .

Figure 11: Decomposing the change in monetary transmission: loan vs. deposit market power



*Notes:* Differences in impulse response functions to a monetary shock between high and low bank concentration environments. Total effect  $\Sigma$  is decomposed into six components: changes in the share of low-concentration markets  $\alpha^m$ , the share of regional banks  $\alpha^b$ , the loan demand elasticity  $\epsilon^l$ , the deposit supply elasticity  $\epsilon^d$ , the bank capital ratio  $\nu$ , and an interaction effect *res.* The x-axis represents the horizon.

Comparing the marginal contribution of loan market power  $\epsilon^l$  to the deposit market power  $\epsilon^d$  indicates loan market power is more relevant for the change in transmission to output as reflected by larger shares. While loan market power leads to a more significant contraction in output, deposit market power dampens the effect of a monetary tightening on output. Quantitatively, loan market power explains 43% of the differential impact effect on output, whereas deposit market power explains only 20%. In contrast, the change in transmission to inflation is more evenly distributed (34% vs. 22%) due to higher loan and deposit market power, respectively; both sides lead to a dampening response. Interestingly, and in line with previous findings, loan market power plays a significant role in investment by affecting the decision variable, the lending rate. In contrast, deposit market power does not affect it. For consumption, the opposite argument holds: the contraction of consumption dampens in the presence of higher deposit market power; however, it amplifies in the presence of higher loan market power. This finding is also related to previous work by Wong (2019) and Di Maggio et al. (2017) documenting a higher interest rate sensitivity and marginal propensity to consume of homeowners and

low-wealth households.

The results are not due to differences in the relative size of the markdowns and markups. In fact, markups and markdowns on both sides are roughly equal, as shown in Table 5. In relative terms, savers were slightly more elastic in 1994. However, borrowers were slightly more elastic in 2019, implying that savers eventually became relatively more inelastic, entailing a dampened response in the patient household’s consumption exceeding the amplification of the impatient household’s consumption.

### 5.2.3 Implications for the Phillips Curve

To examine the impact on the slope of the Phillips curve, I derive a simplified version of the model’s log-linearized Phillips curve, expressing changes in current inflation,  $\tilde{\pi}_t$ , in terms of changes in output,  $\tilde{y}_t$ , and expected future inflation,  $\mathbb{E}_t \tilde{\pi}_{t+1}$ , starting from equation (M.26) in Appendix B.2 and abstracting from indexation:

$$\tilde{\pi}_t = \Phi \tilde{y}_t + \beta^P \mathbb{E}_t \tilde{\pi}_{t+1}, \quad (18)$$

where  $\Phi$  summarizes the coefficients on output, including the Rotemberg price adjustment,  $\kappa_p$ , and elasticity of substitution across goods,  $\epsilon^y$ , parameters.

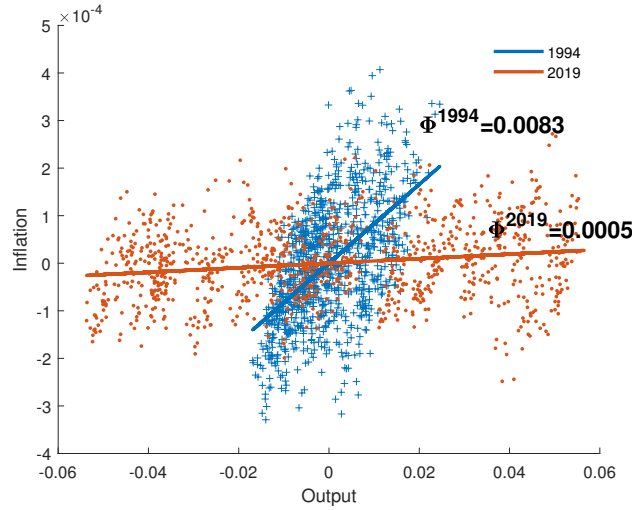
Figure 12 shows the inflation-output relationship based on simulated data for the 1994 and 2019 model calibrations.<sup>24</sup> The results use the monetary shock as the only source of stochastic uncertainty but remain robust for any other demand shock (e.g., preference shock). A comparison of the two calibrations’ estimated slope indicates the Phillips curve flattens over time, consistent with empirical evidence (e.g., Hazell et al., 2022). My calibration suggests a decline by a factor of 16.6, aligning with the estimates in the literature.<sup>25</sup> The flattening is though more attenuated in an environment with borrowing constraints à la Iacoviello (2005) but remains at approximately 20% (Figure B.3 in Appendix B.5).

What is the mechanism behind the flattening of the Phillips curve? The result relies upon two sets of factors. First, output and the slope of the Phillips curve depend on the level of resource costs responsible for a wealth effect. In simple words, the financial sector absorbs part of the resources in the economy, such as operating and bank management

<sup>24</sup>To control for changes in inflation expectations, the y-axis shows:  $\pi_t - \beta \mathbb{E}_t \pi_{t+1}$ . Alternatively, controlling directly for inflation expectations in the regression yields similar results. The simulation is based on 1,000 periods and includes 10,000 initial burn-in periods.

<sup>25</sup>Hazell et al. (2022) find that the Phillips curve flattens by a factor of 2 to 100, depending on model specification, using US state-level variation in inflation and unemployment for 1978-1990 and 1991-2018.

Figure 12: Phillips curves: relation between inflation and output



Notes: Simulated data for output and inflation based on low and high bank concentration environment, labeled 1994 and 2019, respectively. Data expressed in terms of deviations from the steady-state level (unconditional mean).

costs (reflected by  $\delta^b$  in equation (7)), and creates a dead-weight loss.<sup>26</sup> With rising bank concentration and higher bank management costs, “effective” output (i.e., output net of adjustment and management costs) becomes more volatile and disentangles from production. Second, the slope of the Phillips curve depends on the level of frictions affecting labor supply. Wage rigidities and habit formation interact with the wealth channel, further breaking the link between output, marginal costs, and inflation; see also Appendix B.6 for more details.

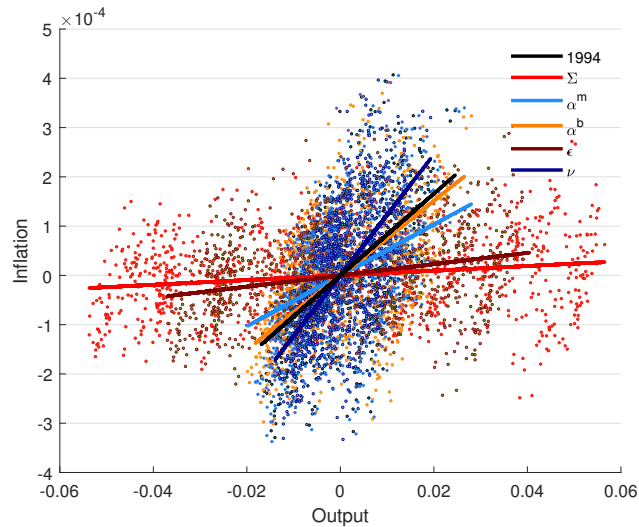
Wider wedges and markups affect the stochastic discount factor of savers and borrowers and the inter-temporal consumption and labor supply decision. While a higher loan rate and lower stochastic discount factor of borrowers lead to a stronger response today, lower saving rates and a higher stochastic discount factor of savers make future consumption more valuable and lead to a weaker response today. Overall the first effect dominates, and net consumption contracts by more. The Phillips curve result is robust to redefining output net of resource costs, but the flattening decreases. Similarly, the result does not go away by eliminating labor market frictions. However, the effect disappears by redefining output and eliminating all labor supply frictions. I regard the results on the flattening of the Phillips curve as complementary to existing explanations pointing to changes in the conduct of monetary policy and inflation expectations (e.g., Carlstrom et al., 2009), less frequent price adjustments (e.g., Kuttner and Robinson, 2010), and higher

<sup>26</sup>Redistributing the management costs to the patient household as a transfer still yields qualitatively similar results and a flattening of the Phillips curve over time.

worker bargaining power (e.g., Ng et al., 2018).

What are the relative importance of the *market power* and *capital allocation channels* for the Phillips curve flattening? I analyze the marginal impact of structural changes in (i) low-concentration markets' share,  $\alpha^m$ , (ii) regional banks' share,  $\alpha^b$ , (iii) loan demand and deposit supply elasticity,  $\epsilon$ , and (iv) bank capital ratio,  $\nu$ . Figure 13 contrasts the estimated Phillips curves for each specification with the 1994 baseline. Rising markups  $\epsilon$  are the main driver. Although changes along the extensive margin, market shares of regional banks  $\alpha^b$  and low-concentration markets  $\alpha^m$  shift the Phillips curve in the same direction, their effects are relatively small. An increase in bank capital ratios  $\nu$  leads to a steeper curve, slightly counteracting the other forces. The findings are consistent with the decomposition in Section 5.2.2 and confirm the relevance of the *market power channel*. What is driving the *market power channel*: loan or deposit market power? Looking at Figure 11 shows that the friction on the lending side – increasing wedges and higher loan rates – due to higher bank concentration matter for the flattening of the Phillips curve. In contrast, deposit market power instead leads to a steepening curve, as foreshadowed by Figure 11.

Figure 13: Phillips curves based on different banking sector calibrations



Notes: Simulated data for output and inflation based on different banking sector calibrations. 1994 reflects the baseline calibration.  $\Sigma$  considers all structural changes, including changes in regional banks' share  $\alpha^b$ , low-concentration markets' share  $\alpha^m$ , demand elasticity  $\epsilon$ , and bank capital ratio  $\nu$ . Data are expressed in terms of deviations from the steady-state level (i.e., unconditional mean).

## 6 Conclusion

This paper examines how the banking sector's structure affects monetary policy pass-through at a disaggregated level. The variation in branch-level retail rates sheds light

on how the composition of local markets and the size distribution of banks affect the aggregate transmission of monetary policy via two channels: a *market power channel*, with higher concentration in local banking markets increasing loan rates and markups, and a *capital allocation channel*, with higher banking concentration implying a lower aggregate banking sector capitalization and amplifying financial frictions due to regulation.

I deliver empirical evidence for heterogeneous monetary policy pass-through to loan and deposit rates in the cross-section and over time and incorporate the novel insights into a quantitative model. I explain the cross-sectional heterogeneity via differences in market power across locations and marginal costs across banks stemming from bank capital ratios. Counterfactual analyses in a New Keynesian model with heterogeneous bank branches and banks calibrated to 1994 and 2019 reveal a strengthened monetary policy pass-through to loan rates and amplified the credit cycle due to the rise in bank concentration. This strengthening in monetary policy pass-through is due to both an increase in market power and shifts in the banks' size distribution. The rise in bank concentration amplifies monetary policy transmission to output and investment but dampens its impact on inflation. The opposite effects imply a flattening of the Phillips curve over time.

Rising bank concentration has important implications for monetary policy transmission and effectiveness. For the conduct of optimal monetary policy, both market power and capitalization of banks should be considered individually and jointly. Monetary policy became more effective in stimulating output over time. In other words, the central bank needs to adjust the policy rate by less to achieve a similar effect on output, though more to stimulate inflation. However, higher bank concentration is not necessarily desirable from a welfare perspective. Increasing wedges distort the banking sector and are sub-optimal from an efficiency point of view. In addition, such an assessment should include financial stability concerns. The framework allows also us to discuss the implications of financial regulation and macro-prudential policies on monetary policy transmission. Higher bank capital requirements lower pass-through to loan rates and monetary transmission, pointing towards significant interactions of macro-prudential and monetary policy.

Further, the results point towards heterogeneity at a disaggregated level relevant to policy design. Future work could expand the model to heterogeneous banks of more than two types and locations and closely study distributional effects across US counties and disparities. Similarly, an alternative model framework could consider a full HANK setup, such as in Kaplan et al. (2018), to study the implications for inequality at a more granular level.

## References

- Allen, J., Clark, R., and Houde, J.-F. (2019). Search frictions and market power in negotiated-price markets. *Journal of Political Economy*, 127(4):1550–1598.
- Altavilla, C., Canova, F., and Ciccarelli, M. (2019). Mending the broken link: Heterogeneous bank lending rates and monetary policy pass-through. *Journal of Monetary Economics*.
- Andres, J. and Arce, O. (2012). Banking competition, housing prices and macroeconomic stability. *The Economic Journal*, 122(565):1346–1372.
- Ball, L. M. and Mazumder, S. (2011). Inflation dynamics and the great recession. National Bureau of Economic Research (NBER) Working Paper 17044.
- Bluedorn, J. C., Bowdler, C., and Koch, C. (2017). Heterogeneous bank lending responses to monetary policy: New evidence from a real-time identification. *International Journal of Central Banking*, 47:95–149.
- Brennecke, C., Jacewitz, S., and Pogach, J. (2021). Shared destinies? Small banks and small business consolidation. *Small Banks and Small Business Consolidation. Federal Reserve Bank of Kansas City Working Paper*, (21-19).
- Brunnermeier, M. K. and Koby, Y. (2018). The reversal interest rate. National Bureau of Economic Research (NBER) Working Paper 25406.
- Carlstrom, C. T., Fuerst, T. S., and Paustian, M. (2009). Monetary policy shocks, Choleski identification, and DNK models. *Journal of Monetary Economics*, 56(7):1014–1021.
- Corbae, D. and D’Erasmus, P. (2020). Rising bank concentration. *Journal of Economic Dynamics and Control*, 115:103877.
- De Loecker, J., Eeckhout, J., and Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, 135(2):561–644.
- Di Maggio, M., Kermani, A., Keys, B. J., Piskorski, T., Ramcharan, R., Seru, A., and Yao, V. (2017). Interest rate pass-through: Mortgage rates, household consumption, and voluntary deleveraging. *American Economic Review*, 107(11):3550–3588.
- Drechsler, I., Savov, A., and Schnabl, P. (2017). The deposits channel of monetary policy. *Quarterly Journal of Economics*, 132(4):1819–1876.

- Drechsler, I., Savov, A., and Schnabl, P. (2021). Banking on deposits: Maturity transformation without interest rate risk. *The Journal of Finance*, 76(3):1091–1143.
- Flannery, M. J. (1982). Retail bank deposits as quasi-fixed factors of production. *The American Economic Review*, 72(3):527–536.
- Gerali, A., Neri, S., Sessa, L., and Signoretti, F. M. (2010). Credit and banking in a DSGE Model of the euro area. *Journal of Money, Credit and Banking*, 42:107–141.
- Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.
- Gilchrist, S., Schoenle, R., Sim, J., and Zakrajšek, E. (2017). Inflation dynamics during the financial crisis. *American Economic Review*, 107(3):785–823.
- Grant, C. (2007). Estimating credit constraints among US households. *Oxford Economic Papers*, 59(4):583–605.
- Hazell, J., Herreno, J., Nakamura, E., and Steinsson, J. (2022). The slope of the phillips curve: evidence from us states. *The Quarterly Journal of Economics*, 137(3):1299–1344.
- Iacoviello, M. (2005). House prices, borrowing constraints, and monetary policy in the business cycle. *American Economic Review*, 95(3):739–764.
- Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182.
- Kaplan, G., Moll, B., and Violante, G. L. (2018). Monetary policy according to hank. *American Economic Review*, 108(3):697–743.
- Kashyap, A. K. and Stein, J. C. (2000). What do a million observations on banks say about the transmission of monetary policy? *American Economic Review*, 90(3):407–428.
- Kishan, R. P. and Opiela, T. P. (2000). Bank size, bank capital, and the bank lending channel. *Journal of Money, Credit, and Banking*, 32(1):121.
- Klein, M. A. (1971). A theory of the banking firm. *Journal of Money, Credit and Banking*, 3(2):205–218.
- Kuttner, K. and Robinson, T. (2010). Understanding the flattening Phillips curve. *The North American Journal of Economics and Finance*, 21(2):110–125.



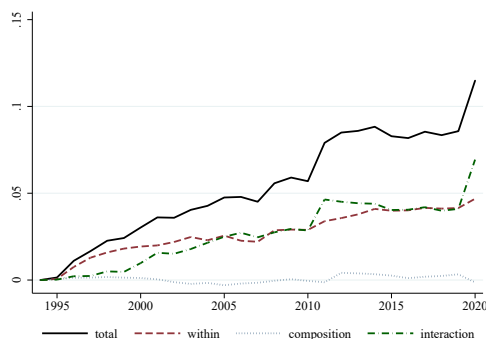
- Matheson, T. and Stavrev, E. (2013). The great recession and the inflation puzzle. *Economics Letters*, 120(3):468–472.
- Meyer, A. P. (2018). Market concentration and its impact on community banks. *St. Louis Fed or Federal Reserve System*.
- Monti, M. (1972). *Deposit, credit and interest rate determination under alternative bank objective function*. North-Holland/American Elsevier.
- Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: the information effect. *Quarterly Journal of Economics*, 133(3):1283–1330.
- Ng, M., Wessel, D., and Sheiner, L. (2018). The Hutchins Center explains: The Phillips curve. *Brookings Up Front*.
- Pasqualini, A. (2021). Markups, markdowns and bankruptcy in the banking industry.
- Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: evidence from us historical data. *Journal of Political Economy*, 126(2):850–901.
- Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.
- Scharfstein, D. and Sunderam, A. (2016). Market power in mortgage lending and the transmission of monetary policy. *Working Paper*.
- Ulate, M. (2021). Going negative at the zero lower bound: The effects of negative nominal interest rates. *American Economic Review*, 111(1):1–40.
- Van den Heuvel, S. J. (2002). The bank capital channel of monetary policy. *The Wharton School, University of Pennsylvania, mimeo*, pages 2013–14.
- Wang, O. (2019). Banks, low interest rates, and monetary policy transmission. Working paper, MIT.
- Wang, Y., Whited, T. M., Wu, Y., and Xiao, K. (2022). Bank market power and monetary policy transmission: Evidence from a structural estimation. *The Journal of Finance*, 77(4):2093–2141.
- Wong, A. (2019). Refinancing and the transmission of monetary policy to consumption. *Unpublished manuscript*.

# A Additional Empirical Results

## A.1 Decomposition of the Rise in US Bank Concentration

To what extent is the rise in bank concentration a general trend across US counties or due to compositional effects? The increase can be decomposed into three parts: (i) changes in concentrated counties' relative market size, (ii) changes in within-county bank concentration, and (iii) interaction effects.<sup>27</sup> Figure A.1 shows that the main drivers are increases within county and the interaction effect, contributing 0.05 and 0.07, respectively, to the total increase of 0.11 from 1994 to 2020.

Figure A.1: Decomposition of the rise in US bank concentration



Notes: Decomposition of national HHI growth from Figure 1(a) in: (i) changes in the market share of high-concentration counties, (ii) changes in concentration within county, and (iii) interaction effects.

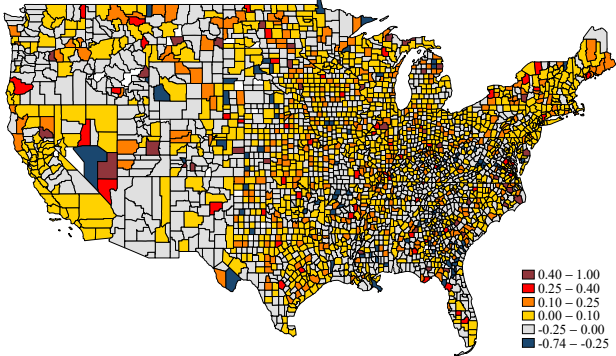
Figure A.2 examines county-level bank concentration changes between 1994 and 2019. While many rural counties observed a decrease in concentration, urban counties with a relatively larger banking sector in terms of deposit holdings observed an increase in bank concentration. The increase in concentration in large counties contributes significantly to the national pattern, reflected in Figure 1(a) by the interaction effect. These shifts are partly due to banks allocating assets to headquarters in big cities as New York City and Salt Lake City. Hence, for the early period, the unweighted and weighted average of local HHIs may lead to different conclusions but indicate both increases in concentration afterward (Brennecke et al., 2021; Meyer, 2018).

<sup>27</sup>Decomposition of the cumulative growth in national HHI relative to 1994:

$$HHI_t - HHI_{1994} = \sum_c \left\{ \underbrace{d_{1994}^c (HHI_t^c - HHI_{1994}^c)}_{\text{within}} + \underbrace{HHI_{1994} (d_t^c - d_{1994}^c)}_{\text{composition}} + \underbrace{(d_t^c - d_{1994}^c) (HHI_t^c - HHI_{1994}^c)}_{\text{interaction}} \right\},$$

where  $HHI_t^c$  and  $d_t^c$  are the HHI and deposit market share of county  $c$ .

Figure A.2: Change in bank concentration between 1994 and 2019 by county



Notes: Changes in HHIs between 1994 and 2019:  $HHI_c^{2019} - HHI_c^{1994}$ . Source: FDIC Summary of Deposits.

## A.2 Survey Instrument

Figure A.3 presents an extract of the *RateWatch* survey instrument. Every month the survey is sent out to branch loan officers to collect information on prices for financial advisors and conduct competitor analyses. The focus lies on loan rate quotes to the “best” customers, i.e., clients with excellent credit scores. To obtain standardized loan rates across branches and time, *RateWatch* specifically asks for offered rates with close to zero fees and points and a constant loan amount, e.g., a 30-year mortgage rate with a loan amount of \$175,000.

Figure A.3: Survey instrument

Institution Name:  
 Account Number:  
 Contact:  
 Today's Date:             
 Current Prime Rate:           

**RATEWATCH**  
 Accurate Financial Data Since 1989  
 RATEWATCH PHONE 800.348.1831

Send to: [submitrates@rate-watch.com](mailto:submitrates@rate-watch.com)

*Mortgages: Please list in-house rates first. If N/A then 2nd market rates. If not offered then N/A the category. Need as close to zero point/fees @ 60 day lock period. Purchase, single family owner occupied.*

1 YEAR ARM @ 175K LOAN		
AMOUNT	FIXED RATE	COMMENTS
RATE		
APR		
DISCOUNT POINTS		
DOWN PAYMENT TO AVOID PMI		
CAPS		
MAX AMORTIZATION TERM		
ORIGINATION FEES		

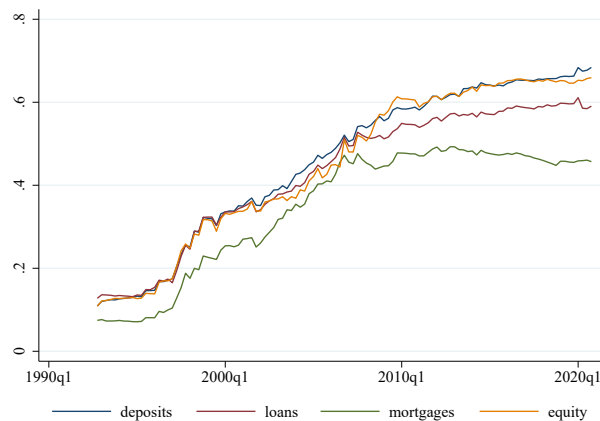
3 YEAR ARM @ 175K LOAN		
AMOUNT	FIXED RATE	COMMENTS
RATE		
APR		
DISCOUNT POINTS		
DOWN PAYMENT TO AVOID PMI		
CAPS		
MAX AMORTIZATION TERM		
ORIGINATION FEES		

Notes: An extract of the survey instrument *RateWatch* sends out to bank branches. Source: RateWatch.

### A.3 Alternative Concentration Measures

The baseline concentration measure of this paper is based on deposit market shares, as there is no information on other balance sheet items at the branch level. At the bank level, deposit, loan, and mortgage holdings are highly correlated. Figure A.4 shows that all concentration measures have similarly increased over the past two decades. Hence, branch-level deposits can serve as a proxy for other balance sheet items at the branch level.

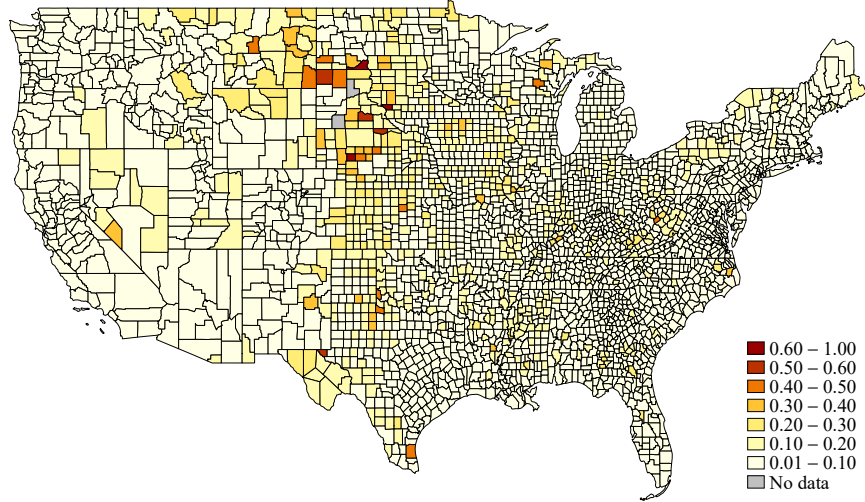
Figure A.4: Market share of giant banks



*Notes:* Market share of giant banks based on deposits, loans, real estate loans, and equity. Source: Federal Deposit Insurance Corporation.

In addition, the Home Mortgage Disclosure Act (HMDA) provides data on mortgage originations at a county level, a close substitute relying on flows instead of stocks. Following Scharfstein and Sunderam (2016), mortgage market concentration is based on each institution's share of mortgage originations per county/year. In contrast to the FDIC data, HMDA includes non-bank lenders and credit unions. Figure A.5 shows a map of the mortgage market concentration across counties in 2019. Overall, the mortgage market concentration is much lower than the deposit market concentration, reflected by lighter coloring. Similar to the deposit market concentration shown in Figure 2, high mortgage market concentration is predominantly in the Midwest and center of the United States. Table A.1 quantifies the relationship between the concentration measures for different periods. The correlation is relatively strong, consistently about 0.42 across time.

Figure A.5: Mortgage origination concentration by county based on HMDA data



Notes: County-level HHIs based on mortgage origination in 2019. Source: Home Mortgage Disclosure Act.

Table A.1: Correlation of county-level deposit and mortgage market concentration

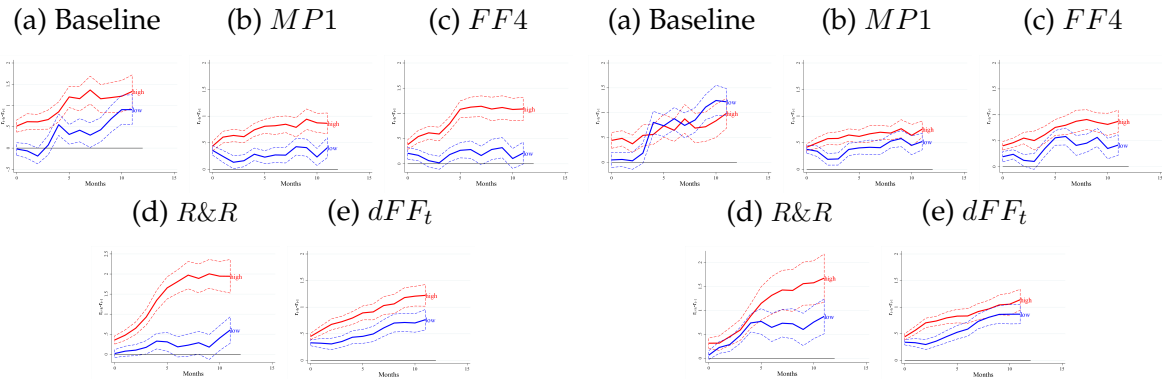
	2000-2019	2000-2008	2009-2019
$\rho(HHI^{dep}, HHI^{mortg})$	0.42	0.42	0.43

Notes:  $\rho$  reflects the correlation coefficient of the county-level HHIs based on deposits and mortgage originations. Source: FDIC Summary of Deposits, Home Mortgage Disclosure Act.

**HMDA Mortgage Market Concentration** I next contrast the impact of deposit and mortgage market concentration on monetary policy pass-through to loan rates. Panels (a) to (e) in Figures A.6 and A.9 present the conditional loan rate responses to different monetary shocks: (i) Nakamura and Steinsson (2018) surprises, (ii) current month's ( $MP1$ ) and (iii) three-month ahead future rate ( $FF4$ ) surprises, (iv) Romer and Romer (2004) narrative shocks ( $R\&R$ ), and (v) raw changes in the federal funds rate ( $dFF_t$ ), revealing that both concentration measures correlate positively with pass-through. Bank branches operating in markets with high deposit and mortgage concentration adjust loan rates faster to a monetary shock.

Figure A.6: Deposit market concentration

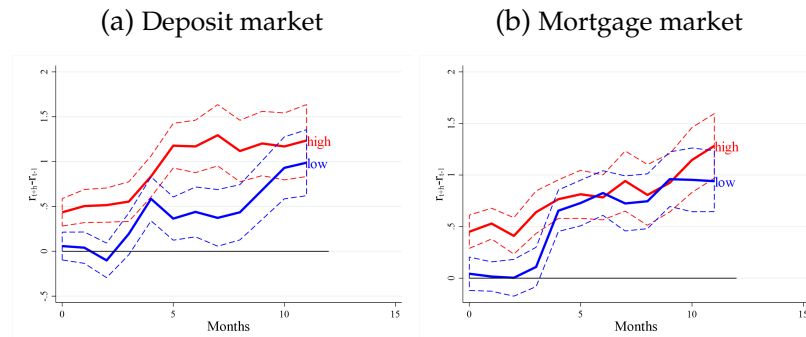
Figure A.7: Mortgage market concentration



Notes: Impulse responses of 1-year hybrid ARM rates to a monetary shock at both high and low concentration levels. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

A horse race between the two, estimating a regression model with both concentration measures jointly, indicates that both remain important independently (Figure A.8).

Figure A.8: Including both concentration measures simultaneously



Notes: Impulse responses of 1-year hybrid ARM rates to a monetary shock at both high and low concentration levels. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

## A.4 Alternative Bank Capitalization Measures

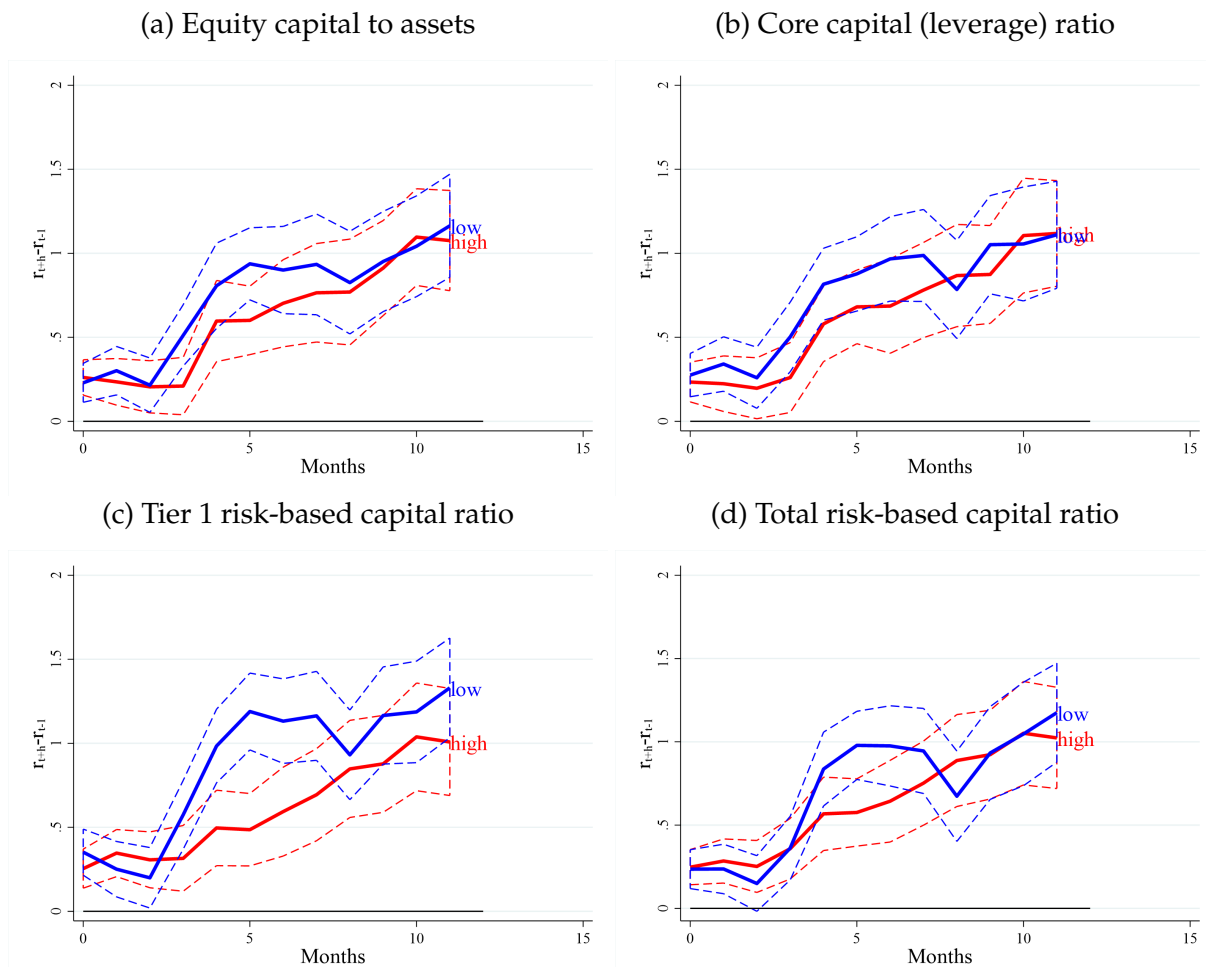
Different bank capitalization measures are strongly correlated (Table A.2). Figure A.9 considers the equity capital to assets ratio, the core capital, tier 1 risk-based capital, and total risk-based capital ratios. The results hold across measures.

Table A.2: Correlation coefficient of capitalization measures

	Eq. cap. to assets	Core cap. (leverage) ratio	Tier 1 risk-based cap. ratio	Total risk-based cap. ratio	Common eq. tier 1 cap. ratio
Eq. cap. to assets	1				
Core cap. (leverage) ratio	0.93	1			
Tier 1 risk-based cap. ratio	0.65	0.71	1		
Total risk-based cap. ratio	0.65	0.71	1	1	
Common eq. tier 1 cap. ratio	0.65	0.71	1	1	1

Notes: Correlation coefficient of equity capital to assets (*equ*), core capital (leverage) ratio (*rbc1aa*), tier 1 risk-based capital ratio (*rbc1rwa*), total risk-based capital ratio, (*rbcrrwa*), and common equity tier 1 cap. ratio (*rbct1cer*) (available 2015-). FDIC variable name in parentheses. Source: FDIC.

Figure A.9: Different bank capitalization measures



Notes: Impulse responses of 1-year hybrid ARM rates to a monetary shock at both high and low bank capitalization levels. Horizon is in months, and standard errors are clustered at the county level (90% CI).



## A.5 Alternative Monetary Policy Measures

Appendix A.5 presents the results estimating equation (6) with alternative monetary policy measures: surprises in the current month's future rate ( $MP1$ ), three-month ahead future rate ( $FF4$ ),<sup>28</sup> Romer and Romer (2004) narrative monetary shocks ( $R\&R$ ),<sup>29</sup> and raw changes in the federal funds rate ( $dFF_t$ ).

Figure A.10: Loan rate - concentration

Figure A.11: Deposit rate - concentration

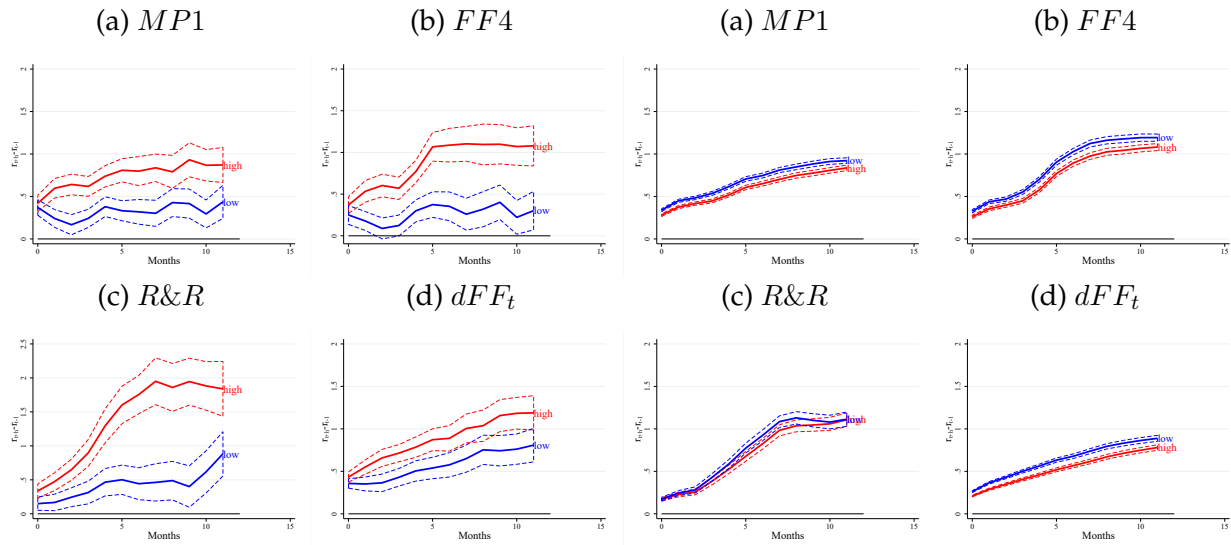
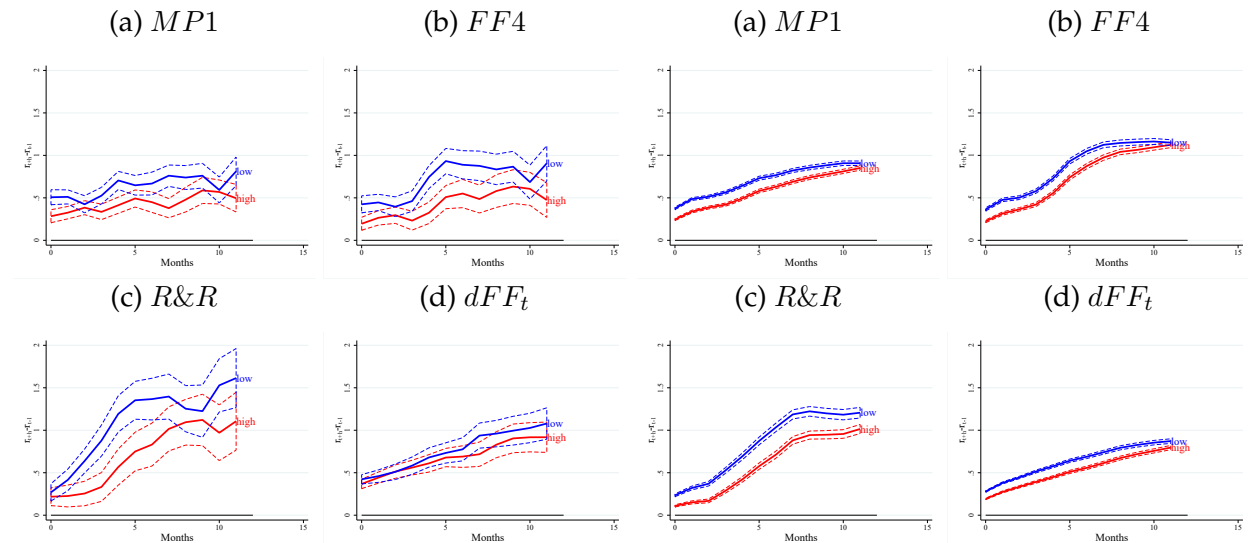


Figure A.12: Loan rate - capitalization

Figure A.13: Deposit rate - capitalization

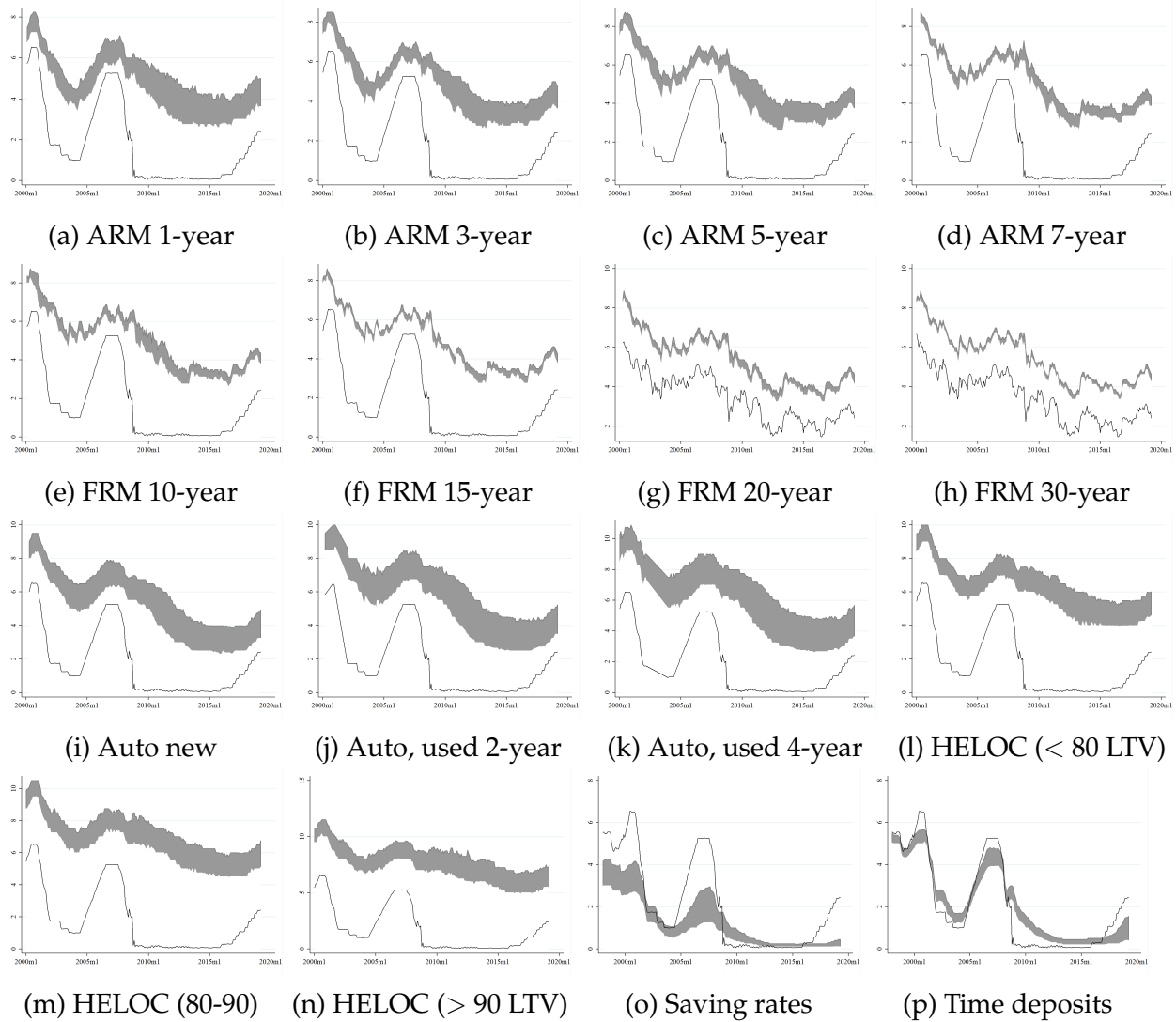


Notes: Impulse response functions of the 1-year hybrid ARM rate and deposit rate to a monetary policy shock at both high and low local bank concentrations in Figures A.10 and A.11 and at both high and low bank capitalization in Figures A.12 and A.13. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

<sup>28</sup>Gertler and Karadi (2015) use the three-month ahead future surprise to identify monetary shocks.

<sup>29</sup>Thank you to Johannes Wieland for providing an updated narrative shocks series for 2000-2007.

## A.6 Dispersion and Spread Over Time

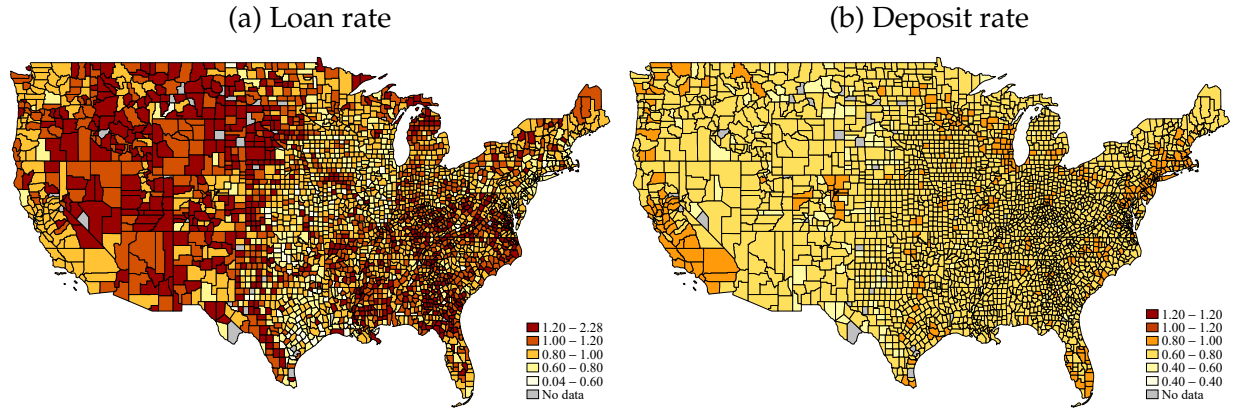


*Notes:* IQR of branch-level deposit and loan rates. ARM denotes adjustable rate mortgage, and FRM is a fixed-rate mortgage with a loan amount of \$ 175,000 and a maturity of 30 years. HELOC is a home equity line of credit with varying loan-to-value (LTV) ratios. Auto loan rates vary by car age (36 months contracts).

## A.7 Heterogeneous Pass-Through Across US Counties

Figure A.15 shows the estimated pass-through after six months,  $h = 6$ , for loan and deposit rates across US counties in 1995. County-level pass-through,  $PT_{c,t}^h$ , depends on two factors: the level of local bank concentration,  $HHI_{c,t-1}$ , and the average capitalization in a county,  $\bar{m}(\%)_{t-1}$ ,<sup>30</sup> and is calculated as:  $\hat{\beta}^h + \hat{\gamma}_1^h HHI_{c,t-1} + \hat{\gamma}_2^h \bar{m}(\%)_{t-1}$ .

Figure A.15: Estimated monetary policy pass-through after six months across US counties



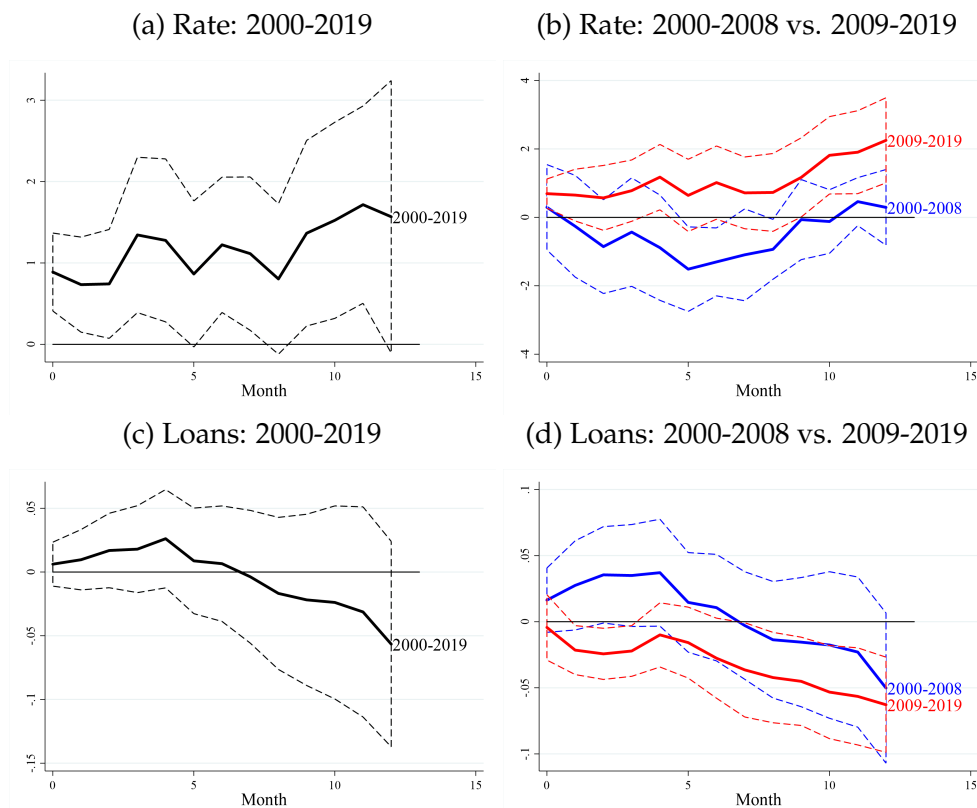
Notes: Estimated monetary policy pass-through to loan and deposit rates after six months,  $h = 6$ , for 1995 by US county, calculated as  $\hat{\beta}^h + \hat{\gamma}_1^h HHI_{c,t-1} + \hat{\gamma}_2^h \bar{m}(\%)_{t-1}$ .  $HHI_{c,t-1}$  reflects the lagged county-level HHI and  $\bar{m}(\%)_{t-1}$  the weighted mean of the lagged capital ratio by county.

<sup>30</sup>The average county-level capitalization is the bank headquarters capital ratio weighted by the deposit share in a county.

## A.8 Increasing Aggregate Pass-Through Over Time

Figure A.16 presents responses of the aggregate mortgage rate and real estate loans to Nakamura and Steinsson (2018) monetary surprises for the entire and a center-split sample.<sup>31</sup> The aggregate pass-through to mortgage rates is higher in the second period (red) than in the first period (blue). Similarly, real estate lending declined more sharply in the second period. The increasing aggregate pass-through between 2000-2008 vs. 2009-2019 confirms the expected increase due to compositional shifts in the banking sector and bank concentration over the same time.

Figure A.16: Impulse responses of aggregate mortgage rates and real estate lending by period



*Notes:* Impulse response functions of the aggregate mortgage rate and log real estate loans from commercial banks to a monetary policy shock. Horizon is in months. 90% confidence intervals.

<sup>31</sup>I control for aggregate national economic and financial conditions by interacting these with the monetary policy surprises.

## B Model Details

### B.1 Full Model

Next to the afore-described financial intermediaries, the full model includes two types of households, entrepreneurs, labor packers and unions, capital and final goods producers, and a monetary authority following Gerali et al. (2010). The baseline environment deviates from Gerali et al. (2010) in two ways: the impatient household and entrepreneur do not face a credit constraint, and the only source of uncertainty is a monetary shock.

#### B.1.1 Patient and Impatient Households

There is a unit mass of patient and impatient households, each denoted by  $i$ . In the baseline model, both types of households differ only in terms of their subjective discount factor  $\beta^\chi$ , with  $\chi \in \{P, I\}$ , where  $\beta^P > \beta^I$ .<sup>32</sup> Otherwise, the household preferences are the same. Both types consume, work, and own a housing stock, which is in aggregate in fixed supply.<sup>33</sup> Each household  $i$  of type  $\chi \in \{P, I\}$  maximizes expected utility:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \beta^{\chi,t} \left[ (1 - a^\chi) \log (c_t^\chi(i) - a^\chi c_{t-1}^\chi) + \epsilon^h \log h_t^\chi(i) - \frac{n_t^\chi(i)^{1+\phi}}{1 + \phi} \right],$$

depending on current consumption,  $c_t^\chi(i)$ , past aggregate consumption,  $c_{t-1}^\chi$ , housing stock,  $h_t^\chi(i)$ , and, individual labor supplied,  $n_t^\chi(i)$ .  $a^\chi$  governs the degree of external, group-specific habit formation.<sup>34</sup>  $\phi$  measures disutility of labor. The utility of housing follows a log form governed by  $\epsilon^h$ . The budget constraints differ across households, as the patient household provides deposits to the banking system, and the impatient household demands loans from the banking system.

The patient household's budget constraint is as follows:

$$c_t^P(i) + q_t^h (h_t^P(i) - h_{t-1}^P(i)) + d_t^P(i) \leq w_t^P n_t^P(i) + (1 + r_{t-1}^d) \frac{d_{t-1}^P(i)}{\pi_t} + \tau_t^P(i),$$

where  $d_t^P(i)$  is the patient household's deposit holding earning with gross interest income  $1 + r_{t-1}^d d_{t-1}^P(i)/\pi_t$ ,  $w_t^P$ , real wage,  $q_t^h$ , price of housing, and  $\tau_t^P(i)$  includes transfers from final goods producer and labor union, as these belong to the patient household.<sup>35</sup>

<sup>32</sup>The model extension with financial constraints adds a borrowing constraint to the impatient household.

<sup>33</sup>The housing market market-clearing condition is:  $\bar{h} = h_t^P + h_t^I$ , with constant housing supply,  $\bar{h}$ .

<sup>34</sup>Setting  $a^\chi$  to 0 nests the case without habit. Multiplying by  $(1 - a^\chi)$  cancels out steady-state distortions.

<sup>35</sup>The bank does not pay a dividend and retains profits for next period's bank capital.

The impatient household's budget constraint is as follows:

$$c_t^I(i) + q_t^h (h_t^I(i) - h_{t-1}^I(i)) + b_{t-1}^I(i) (1 + r_{t-1}^H) / \pi_t \leq w_t^I n_t^I(i) + b_t^I(i),$$

where  $b_t^I(i)$  reflects impatient household's outstanding debt with gross interest expenses  $1 + r_{t-1}^H b_{t-1}^I(i) / \pi_t$ , and,  $w_t^I$ , impatient household's real wage.

### B.1.2 Entrepreneurs

A unit mass of entrepreneurs  $i$  produces a homogeneous intermediate good using two inputs: capital,  $k_t^E$ , purchased from capital-good producers, and hired labor input from the patient,  $n_t^P$ , and impatient household,  $n_t^I$ . Similar to the households, the entrepreneur's utility depends on current individual consumption,  $c_t^E(i)$ , and lagged aggregate consumption,  $c_{t-1}^E$ , governed by  $a^E$ . The entrepreneur maximizes expected utility:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \beta_E^t \log (c_t^E(i) - a^E c_{t-1}^E),$$

subject to entrepreneur's budget constraint:

$$c_t^E(i) + w_t^I n_t^I(i) + w_t^P n_t^P(i) + \frac{1 + r_{t-1}^E}{\pi_t} L_{t-1}^E(i) + q_t^k k_t^E(i) + v(u_t(i)) k_{t-1}^E(i) \leq \frac{y_t^E(i)}{x_t} + L_t^E(i) + (1 - \delta) q_t^k k_t^E(i),$$

where  $b_t^E(i)$  is the entrepreneur's outstanding debt with gross interest expenses  $1 + r_{t-1}^E b_{t-1}^E(i) / \pi_t$ ,  $q_t^k$ , the price of physical capital,  $\delta$ , the depreciation rate,  $v(u_t(i))$ , capital utilization costs,  $w_t^I n_t^I(i)$  and  $w_t^P n_t^P(i)$ , the wage bill for hiring labor from impatient and patient households,  $x_t$ , the price markup, and,  $y_t^E(i)$ , the produced wholesale good. The production function follows:

$$y_t^E(i) = [u_t(i) k_{t-1}^E(i)]^\alpha [n_t^E(i)]^{1-\alpha} = [u_t(i) k_{t-1}^E(i)]^\alpha \left[ (n_t^P(i))^\mu (n_t^I(i))^{(1-\mu)} \right]^{1-\alpha}.$$

The labor input from the two types of households is combined to aggregate labor input,  $n_t^E(i) = (n_t^P(i))^\mu (n_t^I(i))^{(1-\mu)}$ , with  $\mu$  governing the patient household's labor income share.

### B.1.3 Labor Packers and Labor Unions

Perfectly competitive labor packers bundle differentiated labor inputs  $m$  using a CES aggregator and sell the homogenized bundle to the labor union. The labor union then provides the homogenized labor bundle to the entrepreneur as input. There exist two unions  $\chi$  for each type of labor input  $m$ , with  $\chi \in \{I, P\}$  for the impatient and patient household. Each labor union sets a nominal wage,  $W_t^\chi$ , subject to the entrepreneur's downward-sloping labor demand, and Rotemberg adjustment costs,  $\kappa_w$ . To cover for adjustment costs, the union charges a lump-sum fee and maximizes:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \beta_u^t \left\{ \Lambda_t^\chi(i, m) \left[ \frac{W_t^\chi(m)}{P_t} n_t^\chi(i, m) - \frac{\kappa_w}{2} \left( \frac{W_t^\chi(m)}{W_{t-1}^\chi(m)} - \pi_{t-1}^{\iota_w} \right)^2 \frac{W_t^\chi}{P_t} \right] - \frac{n_t^\chi(i, m)^{1+\phi}}{1+\phi} \right\},$$

subject to labor demand  $n_t^\chi(i, m) = \left( \frac{W_t^\chi(m)}{W_t^\chi} \right)^{-\epsilon^n} n_t^\chi$ , where  $\epsilon^n$  measures the degree substitutability. The labor union discounts future income with stochastic discount factor,  $\Lambda_t^\chi(i, m)$ , of the respective household. Adjustment costs incur relative to a weighted average of steady-state,  $\pi^{1-\iota_w}$ , and lagged inflation,  $\pi_{t-1}^{\iota_w}$ , with weight  $\iota_w$  on lagged inflation.

In the symmetric equilibrium, labor supply of household with type  $\chi$  is:

$$\kappa_w \left( \pi_t^{w,\chi} - \pi^{\iota_w, t-1} \pi^{1-\iota_w} \right) \pi_t^{w,\chi} = \beta^\chi \mathbb{E}_t \left[ \frac{\Lambda_{t+1}^\chi}{\Lambda_t^\chi} \kappa_w \left( \pi_{t+1}^{w,\chi} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w} \right) \right] + (1 - \epsilon^n) n_t^\chi + \frac{\epsilon^n n_t^{\chi, 1+\phi}}{w_t^\chi \Lambda_t^\chi},$$

where nominal wage inflation is defined as  $\pi_t^{w,\chi} = \frac{W_t^\chi}{W_{t-1}^\chi}$  and the real wage as  $w_t^\chi = \frac{W_t^\chi}{P_t}$ .

### B.1.4 Capital and Final Goods Producers

The capital good producer operates under perfect competition and purchases last period's depreciated physical capital stock,  $(1 - \delta^k) k_{t-1}$ , at a price  $q_t^k$  from the entrepreneur, and  $i_t$  units of the final good from retailers at a price  $P_t$ . The capital good producer converts the two input goods into new physical capital subject to quadratic investment adjustment costs, governed by cost parameter  $\kappa_i$ . It sells new capital back to entrepreneurs at the same price  $q_t^k$ . The capital good producer's objective is to maximize the sum of expected future

profits discounted by the entrepreneur's stochastic discount factor,  $\Lambda_{0,t}^E$ :

$$\mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^E (q_t^k [k_t - (1 - \delta^k)k_t] - i_t)$$

subject to the evolution of capital:

$$k_t = (1 - \delta^k)k_{t-1} + \left[ 1 - \frac{\kappa_i}{2} \left( \frac{i_t}{i_{t-1}} - 1 \right)^2 \right] i_t.$$

The final good firms operate under monopolistic competition. Each final good firm  $j$  buys intermediate goods from entrepreneurs at wholesale price,  $P_t^W$ , differentiates goods at no cost, and sells them to customers as a final good. Retail prices are sticky and indexed to an average of past and steady-state price inflation with weight  $\iota_p$  on past inflation. The firm incurs Rotemberg adjustment costs,  $\kappa_p$ , for changing prices beyond indexation. The final price,  $P_t(j)$ , is chosen to maximize profits:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[ P_t(j)y_t(j) - P_t^W y_t(j) - \frac{\kappa_p}{2} \left( \frac{P_t(j)}{P_{t-1}} - \pi_{t-1}^{\iota_p} \pi^{1-\iota_p} \right)^2 P_t y_t \right],$$

subject to final good demand of good  $j$  with demand price elasticity  $\epsilon^y$ :

$$y_t(j) = \left( \frac{P_t(j)}{P_t} \right)^{-\epsilon^y} y_t.$$

### B.1.5 Monetary Policy and Market Clearing

The central bank follows a standard Taylor rule:

$$(1 + r_t^f) = (1 + r^f)^{(1-\phi_R)} (1 + r_{t-1}^f)^{\phi_R} \left( \frac{\pi_t}{\pi} \right)^{\phi_\pi (1-\phi_R)} \left( \frac{y_t}{y_{t-1}} \right)^{\phi_y (1-\phi_R)} \varepsilon_t^R,$$

where  $\phi_R$  reflects the weight on the lagged policy rate,  $\phi_\pi$  and  $\phi_y$ , the responsiveness to inflation and output growth, and  $\varepsilon_t^R$  an i.i.d. monetary shock with standard deviation  $\sigma_R$ .

The goods market market-clearing condition is:

$$y_t = c_t^E + c_t^P + c_t^I + q_t^k [k_t - (1 - \delta) k_{t-1}] + k_{t-1} \phi(u_t) + \delta^K \frac{K_{t-1}^K}{\pi_t} + Adj_t.$$

where  $Adj_t$  combines all adjustment costs (prices, wages, and banks).



## B.2 Equilibrium Equations

$$c_t^I + q_t^h (h_t^I - h_{t-1}^I) + (1 + r_{t-1}^{BH}) \frac{L_{t-1}^I}{\pi_t} = w_t^I n_t^I + L_t^I \quad (\text{M.1})$$

$$\frac{(1 - a^I)}{c_t^I - a^I c_{t-1}^I} = \lambda_t^I \quad (\text{M.2})$$

$$\lambda_t^I q_t^h = \frac{\epsilon^h}{h_t^I} + \beta^I \mathbb{E}_t [\lambda_{t+1}^I q_{t+1}^h] \quad (\text{M.3})$$

$$\lambda_t^I = \beta^I \mathbb{E}_t \lambda_{t+1}^I \frac{(1 + r_t^{BH})}{\pi_{t+1}} \quad (\text{M.4})$$

$$\kappa_w \left( \pi_t^{w,I} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w} \right) \pi_t^{w,I} = \beta^I \mathbb{E}_t \frac{\lambda_{t+1}^I}{\lambda_t^I} \kappa_w \left( \pi_{t+1}^{w,I} - \pi_t^{\iota_w} \pi^{1-\iota_w} \right) \frac{\left( \pi_{t+1}^{w,I} \right)^2}{\pi_{t+1}} + (1 - \epsilon^n) n_t^I + \frac{\epsilon^n (n_t^I)^{1+\phi}}{w_t^{w,I} \lambda_t^I} \quad (\text{M.5})$$

$$\pi_t^{w,I} = \frac{w_t^{w,I}}{w_{t-1}^{w,I}} \pi_t \quad (\text{M.6})$$

$$c_t^P + q_t^h (h_t^P - h_{t-1}^P) + D_t = w_t^P n_t^P + (1 + r_{t-1}^d) \frac{D_{t-1}}{\pi_t} + \tau_t^P \quad (\text{M.7})$$

$$\frac{(1 - a^P)}{c_t^P - a^P c_{t-1}^P} = \lambda_t^P \quad (\text{M.8})$$

$$\lambda_t^P q_t^h = \frac{\epsilon^h}{h_t^P} + \beta^P \mathbb{E}_t \lambda_{t+1}^P q_{t+1}^h \quad (\text{M.9})$$

$$\lambda_t^P = \beta^P \mathbb{E}_t \lambda_{t+1}^P \frac{(1 + r_t^d)}{\pi_{t+1}} \quad (\text{M.10})$$

$$\kappa_w \left( \pi_t^{w,P} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w} \right) \pi_t^{w,P} = \beta^P \mathbb{E}_t \frac{\lambda_{t+1}^P}{\lambda_t^P} \kappa_w \left( \pi_{t+1}^{w,P} - \pi_t^{\iota_w} \pi^{1-\iota_w} \right) \frac{\left( \pi_{t+1}^{w,P} \right)^2}{\pi_{t+1}} + (1 - \epsilon^n) n_t^P + \frac{\epsilon^n (n_t^P)^{1+\phi}}{w_t^{w,P} \lambda_t^P} \quad (\text{M.11})$$

$$\pi_t^{w,P} = \frac{w_t^{w,P}}{w_{t-1}^{w,P}} \pi_t \quad (\text{M.12})$$

$$c_t^E + w_t^P n_t^P + w_t^I n_t^I + (1 + r_{t-1}^E) L_{t-1}^E / \pi_t + q_t^k k_t^E + v(u_t) k_{t-1}^E = \frac{y_t^E}{x_t} + L_t^E + q_t^k (1 - \delta) k_{t-1}^E \quad (\text{M.13})$$

$$v(u_t) = \zeta_1 (u_t - 1) + \zeta_2 (u_t - 1)^2 \quad (\text{M.14})$$

$$r_t^k = \zeta_1 + \zeta_2 (u_t - 1) \quad (\text{M.15})$$

$$\frac{(1 - a^E)}{c_t^E - a^E c_{t-1}^E} = \lambda_t^E \quad (\text{M.16})$$

$$\lambda_t^E = \beta^E \mathbb{E}_t \left[ \lambda_{t+1}^E \frac{(1 + r_t^E)}{\pi_{t+1}} \right] \quad (\text{M.17})$$

$$\lambda_t^E q_t^k = \beta^E \mathbb{E}_t \left\{ \lambda_{t+1}^E \left[ r_{t+1}^k u_{t+1} + q_{t+1}^k (1 - \delta) - \left( \zeta_1 (u_{t+1} - 1) + \frac{\zeta_2}{2} (u_{t+1} - 1)^2 \right) \right] \right\} \quad (\text{M.18})$$

$$y_t^E = [u_t k_{t-1}^E]^\alpha \left[ (n_t^P)^\mu (n_t^I)^{(1-\mu)} \right]^{1-\alpha} \quad (\text{M.19})$$

$$w_t^P = \mu (1 - \alpha) \frac{y_t^E}{n_t^P} \frac{1}{x_t} \quad (\text{M.20})$$

$$w_t^I = (1 - \mu) (1 - \alpha) \frac{y_t^E}{n_t^I} \frac{1}{x_t} \quad (\text{M.21})$$

$$r_t^k = \alpha [u_t k_{t-1}^E]^{\alpha-1} \left[ (n_t^P)^\mu (n_t^I)^{(1-\mu)} \right]^{1-\alpha} \quad (\text{M.22})$$

$$k_t = (1 - \delta) k_{t-1} + \left[ 1 - \frac{\kappa_i}{2} \left( \frac{i_t}{i_{t-1}} - 1 \right)^2 \right] i_t \quad (\text{M.23})$$

$$1 = q_t^k \left[ 1 - \frac{\kappa_i}{2} \left( \frac{i_t}{i_{t-1}} - 1 \right)^2 - \kappa_i \left( \frac{i_t}{i_{t-1}} - 1 \right) \frac{i_t}{i_{t-1}} \right] + \beta^E \mathbb{E}_t \frac{\lambda_{t+1}^E}{\lambda_t^E} q_{t+1}^k \kappa_i \left( \frac{i_{t+1}}{i_t} - 1 \right) \left( \frac{i_{t+1}}{i_t} \right)^2 \quad (\text{M.24})$$

$$\Pi_t^r = y_t \left( 1 - \frac{1}{x_t} \right) - \frac{\kappa_p}{2} (\pi_t - \pi_{t-1}^{\iota_p} \pi^{1-\iota_p})^2 \quad (\text{M.25})$$

$$0 = 1 - \epsilon^y + \frac{\epsilon^y}{x_t} - \kappa_p (\pi_t - \pi_{t-1}^{\iota_p} \pi^{1-\iota_p}) \pi_t + \beta^P \mathbb{E}_t \left[ \frac{\lambda_{t+1}^P}{\lambda_t^P} \kappa_p (\pi_{t+1} - \pi_t^{\iota_p} \pi^{1-\iota_p}) \pi_{t+1} \frac{y_{t+1}}{y_t} \right] \quad (\text{M.26})$$

$$L_{r,t} = D_{r,t} + K_{r,t} \quad (\text{M.27})$$

$$L_{g,t} = D_{g,t} + K_{g,t} \quad (\text{M.28})$$

$$\pi_t K_{r,t} = (1 - \delta^{b,r}) K_{r,t-1} + \Pi_{r,t-1}^b \quad (\text{M.29})$$

$$\pi_t K_{g,t} = (1 - \delta^{b,g}) K_{g,t-1} + \Pi_{g,t-1}^b \quad (\text{M.30})$$

$$\left( R_{r,t}^b - r_t^f \right) = -\kappa_K \left( \frac{K_{r,t}}{L_{r,t}} - \nu \right) \left( \frac{K_{r,t}}{L_{r,t}} \right)^2 \quad (\text{M.31})$$

$$\left( R_{g,t}^b - r_t^f \right) = -\kappa_K \left( \frac{K_{g,t}}{L_{g,t}} - \nu \right) \left( \frac{K_{g,t}}{L_{g,t}} \right)^2 \quad (\text{M.32})$$

$$\begin{aligned} \Pi_{g,t}^b = & r_{l,g,t}^{bH} L_{l,g,t}^{bH} + r_{l,g,t}^{bE} L_{l,g,t}^{bE} + r_{h,g,t}^{bH} L_{h,g,t}^{bH} + r_{h,g,t}^{bE} L_{h,g,t}^{bE} - r_{l,g,t}^d D_{l,g,t} - r_{h,g,t}^d D_{h,g,t} - \frac{\kappa_K}{2} \left( \frac{K_{g,t}}{L_{g,t}} - \nu_g \right)^2 K_{g,t} - \\ & \frac{\kappa_d}{2} \left( \frac{D_{l,g,t}}{D_{l,g,ss}} - 1 \right)^2 r_{l,g,t}^d D_{l,g,t} - \frac{\kappa_d}{2} \left( \frac{D_{h,g,t}}{D_{h,g,ss}} - 1 \right)^2 r_{h,g,t}^d D_{h,g,t} - \frac{\kappa_{bH}}{2} \left( \frac{L_{l,g,t}^{bH}}{L_{l,g,ss}^{bH}} - 1 \right)^2 r_{l,g,t}^{bH} L_{l,g,t}^{bH} - \\ & \frac{\kappa_{bH}}{2} \left( \frac{L_{h,g,t}^{bH}}{L_{h,g,ss}^{bH}} - 1 \right)^2 r_{h,g,t}^{bH} L_{h,g,t}^{bH} - \frac{\kappa_{bE}}{2} \left( \frac{L_{l,g,t}^{bE}}{L_{l,g,ss}^{bE}} - 1 \right)^2 r_{l,g,t}^{bE} L_{l,g,t}^{bE} - \frac{\kappa_{bE}}{2} \left( \frac{L_{h,g,t}^{bE}}{L_{h,g,ss}^{bE}} - 1 \right)^2 r_{h,g,t}^{bE} L_{h,g,t}^{bE} \end{aligned} \quad (\text{M.33})$$

$$\begin{aligned} \Pi_{r,t}^b = & r_{l,r,t}^{bH} L_{l,r,t}^{bH} + r_{l,r,t}^{bE} L_{l,r,t}^{bE} + r_{h,r,t}^{bH} L_{h,r,t}^{bH} + r_{h,r,t}^{bE} L_{h,r,t}^{bE} - r_{l,r,t}^d D_{l,r,t} - r_{h,r,t}^d D_{h,r,t} - \frac{\kappa_K}{2} \left( \frac{K_{r,t}}{L_{r,t}} - \nu_r \right)^2 K_{r,t} - \\ & \frac{\kappa_d}{2} \left( \frac{D_{l,r,t}}{D_{l,r,ss}} - 1 \right)^2 r_{l,r,t}^d D_{l,r,t} - \frac{\kappa_d}{2} \left( \frac{D_{h,r,t}}{D_{h,r,ss}} - 1 \right)^2 r_{h,r,t}^d D_{h,r,t} - \frac{\kappa_{bH}}{2} \left( \frac{L_{l,r,t}^{bH}}{L_{l,r,ss}^{bH}} - 1 \right)^2 r_{l,r,t}^{bH} L_{l,r,t}^{bH} - \\ & \frac{\kappa_{bH}}{2} \left( \frac{L_{h,r,t}^{bH}}{L_{h,r,ss}^{bH}} - 1 \right)^2 r_{h,r,t}^{bH} L_{h,r,t}^{bH} - \frac{\kappa_{bE}}{2} \left( \frac{L_{l,r,t}^{bE}}{L_{l,r,ss}^{bE}} - 1 \right)^2 r_{l,r,t}^{bE} L_{l,r,t}^{bE} - \frac{\kappa_{bE}}{2} \left( \frac{L_{h,r,t}^{bE}}{L_{h,r,ss}^{bE}} - 1 \right)^2 r_{h,r,t}^{bE} L_{h,r,t}^{bE} \end{aligned} \quad (\text{M.34})$$

$$\left( \epsilon^{d,l} - 1 \right) - \epsilon^{d,l} \frac{r_t^f}{r_{l,g,t}^d} + \epsilon^{d,l} \kappa_d \left( \frac{D_{l,g,t}}{D_{l,g,ss}} - 1 \right) \frac{D_{l,g,t}}{D_{l,g,ss}} = 0 \quad (\text{M.35})$$

$$- \left( \epsilon^{bH,l} - 1 \right) + \frac{\epsilon^{bH,l} R_{g,t}^b}{r_{l,g,t}^{bH}} + \epsilon^{bH,l} \kappa_{bH} \left( \frac{L_{l,g,t}^{bH}}{L_{l,g,ss}^{bH}} - 1 \right) \frac{L_{l,g,t}^{bH}}{L_{l,g,ss}^{bH}} = 0 \quad (\text{M.36})$$

$$- \left( \epsilon^{bE,l} - 1 \right) + \frac{\epsilon^{bE,l} R_{g,t}^b}{r_{l,g,t}^{bE}} + \epsilon^{bE,l} \kappa_{bE} \left( \frac{L_{l,g,t}^{bE}}{L_{l,g,ss}^{bE}} - 1 \right) \frac{L_{l,g,t}^{bE}}{L_{l,g,ss}^{bE}} = 0 \quad (\text{M.37})$$

$$(\epsilon^{d,l} - 1) - \epsilon^{d,l} \frac{r_t^f}{r_{l,t}^d} + \epsilon^{d,r,l} \kappa_d \left( \frac{D_{l,r,t}}{D_{l,r,ss}} - 1 \right) \frac{D_{l,r,t}}{D_{l,r,ss}} = 0 \quad (\text{M.38})$$

$$-(\epsilon^{bH,l} - 1) + \frac{\epsilon^{bH,l} R_{r,t}^b}{r_{l,r,t}^{bH}} + \epsilon^{bH,l} \kappa_{bH} \left( \frac{L_{l,t}^{bH}}{L_{l,r,ss}^{bH}} - 1 \right) \frac{L_{l,r,t}^{bH}}{L_{l,r,ss}^{bH}} = 0 \quad (\text{M.39})$$

$$-(\epsilon^{bE,l} - 1) + \frac{\epsilon^{bE,l} R_{r,t}^b}{r_{l,r,t}^{bE}} + \epsilon^{bE,l} \kappa_{bE} \left( \frac{L_{l,t}^{bE}}{L_{l,r,ss}^{bE}} - 1 \right) \frac{L_{l,r,t}^{bE}}{L_{l,r,ss}^{bE}} = 0 \quad (\text{M.40})$$

$$(\epsilon^{d,h} - 1) - \epsilon^{d,h} \frac{r_t^f}{r_{h,g,t}^d} + \epsilon^{d,h} \kappa_d \left( \frac{D_{h,g,t}}{D_{h,g,ss}} - 1 \right) \frac{D_{h,g,t}}{D_{h,g,ss}} = 0 \quad (\text{M.41})$$

$$-(\epsilon^{bH,h} - 1) + \frac{\epsilon^{bH,h} R_{g,t}^b}{r_{h,g,t}^{bH}} + \epsilon^{bH,h} \kappa_{bH} \left( \frac{L_{h,t}^{bH}}{L_{h,g,ss}^{bH}} - 1 \right) \frac{L_{h,g,t}^{bH}}{L_{h,g,ss}^{bH}} = 0 \quad (\text{M.42})$$

$$-(\epsilon^{bE,h} - 1) + \frac{\epsilon^{bE,h} R_{g,t}^b}{r_{h,g,t}^{bE}} + \epsilon^{bE,h} \kappa_{bE} \left( \frac{L_{h,t}^{bE}}{L_{h,g,ss}^{bE}} - 1 \right) \frac{L_{h,g,t}^{bE}}{L_{h,g,ss}^{bE}} = 0 \quad (\text{M.43})$$

$$(\epsilon^{d,h} - 1) - \epsilon^{d,h} \frac{r_t^f}{r_{h,r,t}^d} + \epsilon^{d,h} \kappa_d \left( \frac{D_{h,r,t}}{D_{h,r,ss}} - 1 \right) \frac{D_{h,r,t}}{D_{h,r,ss}} = 0 \quad (\text{M.44})$$

$$-(\epsilon^{bH,h} - 1) + \frac{\epsilon^{bH,h} R_{r,t}^b}{r_{h,r,t}^{bH}} + \epsilon^{bH,h} \kappa_{bH} \left( \frac{L_{h,t}^{bH}}{L_{h,r,ss}^{bH}} - 1 \right) \frac{L_{h,r,t}^{bH}}{L_{h,r,ss}^{bH}} = 0 \quad (\text{M.45})$$

$$-(\epsilon^{bE,h} - 1) + \frac{\epsilon^{bE,h} R_{r,t}^b}{r_{h,r,t}^{bE}} + \epsilon^{bE,h} \kappa_{bE} \left( \frac{L_{h,t}^{bE}}{L_{h,r,ss}^{bE}} - 1 \right) \frac{L_{h,r,t}^{bE}}{L_{h,r,ss}^{bE}} = 0 \quad (\text{M.46})$$

$$D_t = D_{l,g,t} + D_{l,r,t} + D_{h,g,t} + D_{h,r,t} \quad (\text{M.47})$$

$$L_t^{bH} = L_{l,g,t}^{bH} + L_{l,r,t}^{bH} + L_{h,g,t}^{bH} + L_{h,r,t}^{bH} \quad (\text{M.48})$$

$$L_t^{bE} = L_{l,g,t}^{bE} + L_{l,r,t}^{bE} + L_{h,g,t}^{bE} + L_{h,r,t}^{bE} \quad (\text{M.49})$$

$$r_t^d = \alpha^m r_{l,t}^d + (1 - \alpha^m) r_{h,t}^d \quad (\text{M.50})$$

$$r_t^{bH} = \alpha^b \alpha^m r_{l,r,t}^{bH} + (1 - \alpha^b) \alpha^m r_{l,g,t}^{bH} + \alpha^b (1 - \alpha^m) \quad (\text{M.51})$$

$$r_t^{bE} = \alpha^b \alpha^m r_{l,r,t}^{bE} + (1 - \alpha^b) \alpha^m r_{l,g,t}^{bE} + \alpha^b (1 - \alpha^m) \quad (\text{M.52})$$

$$\alpha^b = \frac{L_{l,r,ss}^{bE} + L_{h,r,ss}^{bE}}{L_t^{bE}} \quad (\text{M.53})$$

$$\alpha^m = \frac{L_{l,g,ss}^{bE} + L_{l,r,ss}^{bE}}{L_t^{bE}} \quad (\text{M.54})$$

$$(1 + r_t^f) = (1 + r^f)^{(1-\phi_R)} (1 + r_{t-1}^f)^{\phi_R} \left( \frac{\pi_t}{\pi} \right)^{\phi_\pi (1-\phi_R)} \left( \frac{y_t}{y_{t-1}} \right)^{\phi_y (1-\phi_R)} \varepsilon_t^R \quad (\text{M.55})$$

$$y_t = c_t^E + c_t^P + c_t^I + q_t^k [k_t - (1 - \delta) k_{t-1}] + k_{t-1} \phi(u_t) + \delta^K \frac{K_{t-1}^K}{\pi_t} + Adj_t \quad (\text{M.56})$$

$$\bar{h} = h_t^P + h_t^I \quad (\text{M.57})$$

$$n_t = n_t^I + n_t^P \quad (\text{M.58})$$

$$Y_t = c_t^E + c_t^P + c_t^I + i_t \quad (\text{M.59})$$

$$c_t = c_t^E + c_t^P + c_t^I \quad (\text{M.60})$$

$$L_t = L_t^{bE} + L_t^{bE} \quad (\text{M.61})$$

$$L_t^{bH} = L_t^I \quad (\text{M.62})$$

$$L_t^{bE} = L_t^E \quad (\text{M.63})$$

### B.3 Calibration of Baseline Model

Table B.1: Calibration of model parameters following Gerali et al. (2010)

Parameter	Description	Value
$\kappa^{Kb}$	Adjustment costs of bank capital ratio	11.49
$\delta^b$	Management cost of bank	0.1049 <sup>a</sup>
$\beta^P$	Discount factor of patient household	0.9943
$\beta^{I,E}$	Discount factor of impatient household and entrepreneur	0.975 <sup>b</sup>
$\phi$	Inverse of Frisch elasticity of labor supply	1
$\epsilon^h$	Housing preference	0.2
$a^{P,I,E}$	Habit consumption	0.86
$\epsilon^{m,I}$	Steady-state LTV-ratio for impatient households	0.7 <sup>c</sup>
$\alpha$	Output elasticity with respect to capital	0.25
$\mu$	Labor cost share of patient households costs	0.8
$\zeta_1$	Adjustment costs for capacity utilization	0.0478
$\zeta_2$	Adjustment costs for capacity utilization	0.00478
$\epsilon^{m,E}$	Steady-state LTV-ratio for entrepreneur	0.35 <sup>c</sup>
$\kappa_w$	Adjustment costs of wages	99.9
$\iota_w$	Indexation of wage inflation to past wage inflation	0.28
$\epsilon^n$	Steady-state labor market markup	5
$\delta$	Depreciation rate of physical capital	0.025
$\kappa_i$	Adjustment costs of investment	10.18
$\kappa_p$	Adjustment costs of good prices	28.65
$\iota_p$	Indexation of price inflation to past price inflation	0.16
$\epsilon^y$	Steady-state goods market markup	6
$\phi_R$	Taylor rule smoothing parameter	0.77
$\phi_\pi$	Taylor rule response to inflation	1.98 <sup>d</sup>
$\phi_x$	Taylor rule response to output	0.35
$\sigma_r$	Standard deviation of monetary shock	0.002

<sup>a</sup>  $\delta^b$  varies with  $\epsilon^d, \epsilon^{bH}, \epsilon^{bE}, \nu$  to satisfy in the steady state  $\delta^b = \Pi^b/K$ .

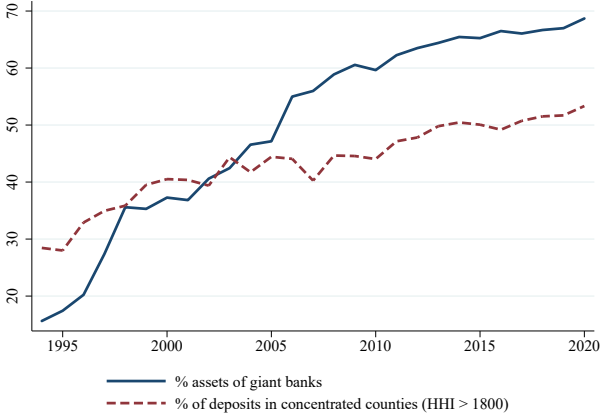
<sup>b</sup> In the baseline model without borrowing constraints  $\beta^{I,E}$  depends on  $\beta^P, \epsilon^d, \epsilon^{bH}, \epsilon^{bE}$ .

<sup>c</sup> Only used in the model with borrowing constraints.

<sup>d</sup> In Section 5.2, the coefficient on inflation is higher (2.9) to avoid indeterminacy issues.

# B.4 Calibration of Heterogeneous Bank Model

Figure B.1: Share of high-concentration markets and giant banks over time



Notes: The deposit-weighted market share of high-concentration markets from 1994 to 2019. The cutoff for high-concentration counties is 1800, following the Department of Justice’s classification defining markets with an HHI above 1800 points as highly concentrated. The share of assets held by banks with more than \$100 billion in assets (in \$2018). Source: Federal Deposit Insurance Corporation, Department of Justice.

## B.5 Extension Borrowing Constraints on the Household and Firm Side

This section examines the role of bank concentration for monetary policy pass-through in an environment where households and firms face financial frictions. Financial frictions are an important factor – with about 31% of households in the US being borrowing constrained (Grant, 2007). An LTV-ratio restricts most mortgage and investment loans. In the case of mortgages, the maximum loan volume corresponds to a fraction of the housing value. In this extension, the impatient household faces a borrowing constraint à la Iacoviello (2005) and the entrepreneur a borrowing constraint connected to the physical capital, shown in equations (M.64) and (M.65). The impatient household’s borrowing amount,  $(1 + r_t^{bH}) L_t^I$ , is limited by a maximum LTV-ratio,  $\epsilon^{m,I}$ , tied to the housing stock,  $h_t^I$ , times the expected future house price,  $\mathbb{E}_t q_{t+1}^h$ , and expected future inflation,  $\mathbb{E}_t \pi_{t+1}$ . Similarly, the entrepreneur’s borrowing amount,  $(1 + r_t^{bE}) L_t^E$ , is restricted by a maximum LTV-ratio,  $\epsilon^{m,E}$ , times the depreciated capital stock,  $(1 - \delta) k_t^E$ , the expected price of capital,  $\mathbb{E}_t q_{t+1}^k$ , and the expected future inflation rate,  $\mathbb{E}_t \pi_{t+1}$ .

$$(1 + r_t^{bH}) b_t^I \leq \epsilon^{m,I} \mathbb{E}_t [q_{t+1}^h h_t^I \pi_{t+1}] \quad (\text{M.64})$$

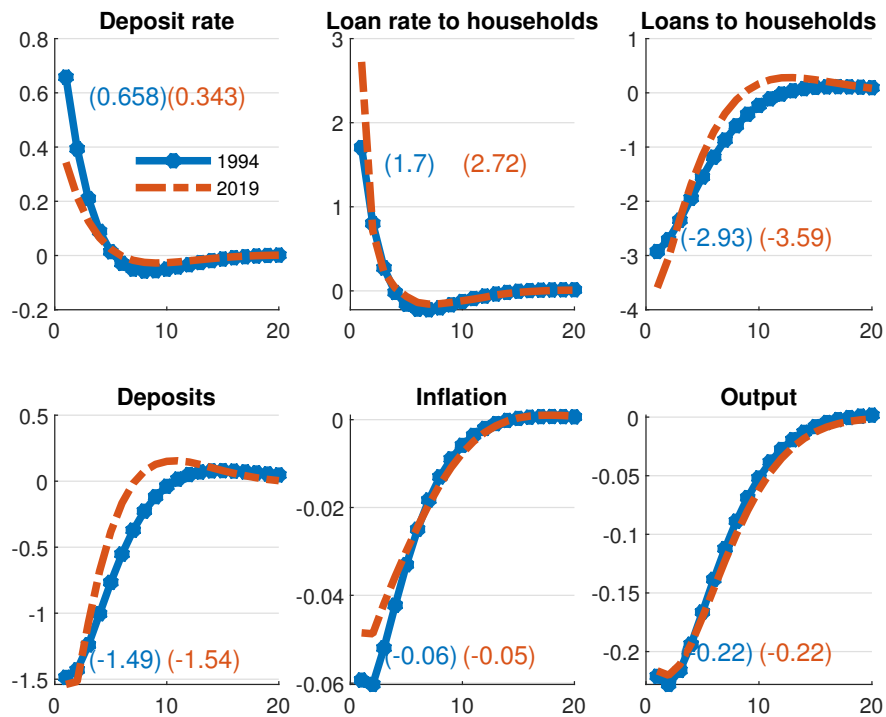
$$(1 + r_t^{bE}) b_t^E \leq \epsilon^{m,E} \mathbb{E}_t [(1 - \delta) q_{t+1}^k k_t^E \pi_{t+1}] \quad (\text{M.65})$$

This modification leads to a financial accelerator effect: a monetary tightening leads to a more severe economic downturn (i.a., lower inflation, output, and asset prices) as collateral constraints tighten and loan demand declines independently of higher interest costs. Consequently, this decreases the agent’s interest-rate sensitivity, i.e., making the agents less sensitive to changes in the loan rate.

Figure B.2 compares the impulse response functions of deposit and loan rate, deposits, household loans, output, and inflation to a monetary shock in a banking environment of 1994 and 2019. The impulse response functions of the loan and deposit rates are qualitatively similar to Figure 8. Therefore, adding borrowing constraints does not significantly alter the pass-through to interest rates. However, there are different effects on the credit cycle. Loans and deposits are more responsive in 2019 versus 1994, though the difference is smaller than seen in the unconstrained model. Further, the effect on inflation is more muted in 2019, similar to the unconstrained model, but the difference is smaller. The response of output is unaltered from bank concentration in this environment. However, the muted effect on inflation still implies a flatter observed Phillips curve over time shown in Figure B.3.

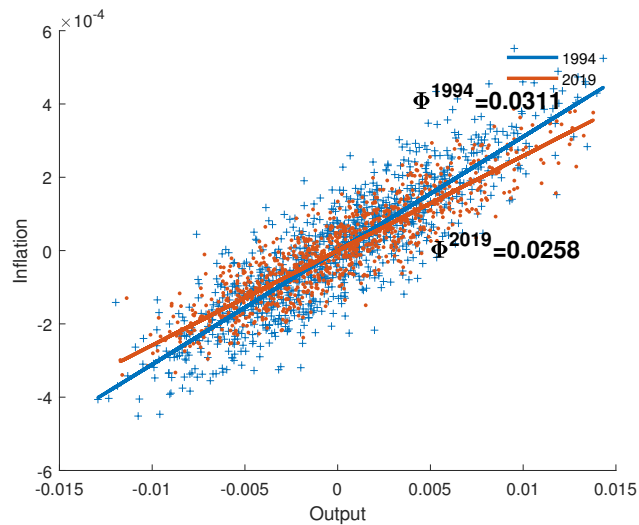


Figure B.2: Impulse responses to a monetary tightening varying  $\alpha^b$ ,  $\alpha^m$ ,  $\epsilon$ , and  $\nu$



Notes: Impulse responses to a positive monetary shock in 1994 (2019) in solid blue with asterisks (red-dashed). The difference between 1994 and 2019 are shifts in  $\alpha^b$ ,  $\alpha^m$ ,  $\epsilon$ , and  $\nu$ . The impact effect is displayed in parentheses.

Figure B.3: Phillips curves: relation between inflation and output

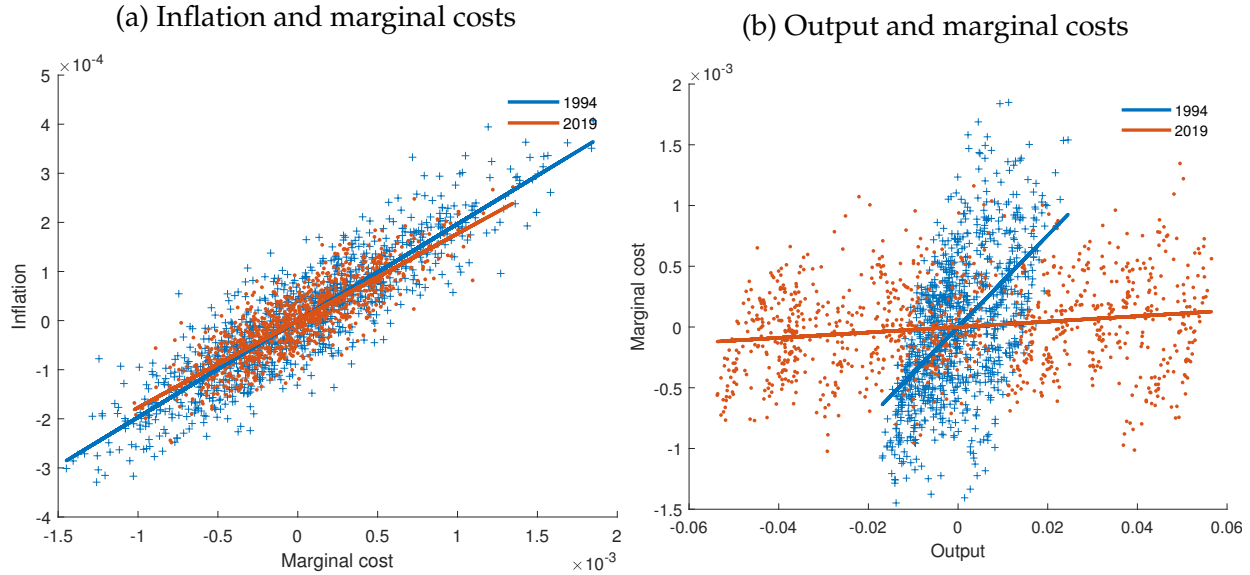


Notes: Simulated data for output and inflation based on banking sector calibration to 1994 and 2019. Data expressed in terms of deviations from steady state (unconditional mean).

## B.6 Disentangling the Flattening of the Phillips Curve

This section aims to shed light on the flattening of the Phillips curve. I decompose the total shift in the Phillips curve into three components: inflation, output, and marginal costs. Figure B.4 reveals that the break between the relationship of output and inflation is primarily due to a break between marginal costs and output and not between inflation and marginal costs.

Figure B.4: Relation between inflation, marginal costs and output



Notes: Simulated data for output, marginal costs, and inflation based on banking sector calibration to 1994 and 2019. Data expressed in terms of deviations from steady state (unconditional mean).