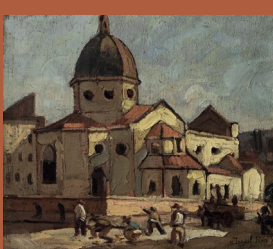
Heterogeneity in Macroeconomics: Implications for Monetary Policy

Sofía Bauducco
Andrés Fernández
Giovanni L. Violante
editors

Banco Central de Chile / Central Bank of Chile

Heterogeneity in Macroeconomics: Implications for Monetary Policy

There is important heterogeneity among households, firms, and banks; and the way shocks or policies affect these agents depends critically on that heterogeneity. There was a rapid surge in the awareness of academic researchers and policymakers of the nexus between heterogeneity and monetary policy, with the emergence of a new class of models subsequently known as HANK, an acronym for heterogeneous-agent new Keynesian. The HANK models combine two long-standing traditions of macroeconomic theory: (i) the new Keynesian approach to the study of business cycles and stabilization policies and (ii) the heterogeneous-agent incomplete market approach to the study of the wealth distribution and of those policies that offer social insurance, promote income mobility, and redistribute income across households. This volume focuses on the role of heterogeneity in macroeconomics and its implications for monetary policy in general and Chile in particular. Understanding the heterogeneous micro implications of a given macro aggregate shock can improve our knowledge of how the economy works and help us forecast its future evolution.



Trabajadores frente a iglesia
Israel Roa Villagra
Oil on canvas, 39 x 45.5 cm
Collection of the Central Bank of Chile

Heterogeneity in Macroeconomics: Implications for Monetary Policy

Sofía Bauducco
Andrés Fernández
Giovanni L. Violante
Editors

Central Bank of Chile / Banco Central de Chile

Series on Central Banking, Analysis,
and Economic Policies

The Book Series on "Central Banking, Analysis, and Economic Policies" of the Central Bank of Chile publishes new research on central banking and economics in general, with special emphasis on issues and fields that are relevant to economic policies in developing economies. The volumes are published in Spanish or English. Policy usefulness, high-quality research, and relevance to Chile and other economies are the main criteria for publishing books. Most research in this Series has been conducted in or sponsored by the Central Bank of Chile.

Book manuscripts are submitted to the Series editors for a review process with active participation by outside referees. The Series editors submit manuscripts for final approval to the Editorial Board of the Series and to the Board of the Central Bank of Chile. Publication is done in both paper and electronic format.

The views and conclusions presented in the book are exclusively those of the authors and do not necessarily reflect the position of the Central Bank of Chile or its Board members.

Heterogeneity in Macroeconomics: Implications for Monetary Policy

Sofía Bauducco
Andrés Fernández
Giovanni L. Violante
Editors

Central Bank of Chile / Banco Central de Chile

ISBN (print) 978-956-7421-74-9
ISBN (digital) 978-956-7421-73-2
Intellectual Property Registration 2024-A-10537

ISSN 0717-6686 (Series on Central Banking, Analysis, and Economic Policies)

Production Team

Editors:
Sofía Bauducco
Andrés Fernández
Giovanni L. Violante

Supervisor:
Giancarlo Acevedo
Pedro Schilling

Copy Editor:
María Marta Semberoiz

Designer:
Maru Mazzini

Printer:
Andros Impresores

Contributors

The articles in this volume are revised versions of the papers presented at the Twenty-fifth Annual Conference of the Central Bank of Chile on Heterogeneity in Macroeconomics: Implications for Monetary Policy, held in Santiago on 21-22 November 2022. The list of contributing authors and conference discussants follows.

Contributing Authors:

Sushant Acharya
Bank of Canada
Centre for Economic Policy Research

Felipe Alves
Bank of Canada

Adrien Auclert
Stanford University
Center for Economic and Policy Research
National Bureau of Economic Research

Sofía Bauducco
Central Bank of Chile

William Chen
Massachusetts Institute of Technology

Dean Corbae
University of Wisconsin-Madison
National Bureau of Economic Research

Marco Del Negro
Federal Reserve Bank of New York
Centre for Economic Policy Research

Pablo D'Erasmo
Federal Reserve Bank of Philadelphia

Keshav Dogra
Federal Reserve Bank of New York

Andrés Fernández
International Monetary Fund

Benjamin Garcia
Central Bank of Chile

Mario Giarda
Central Bank of Chile

Aidan Gleich
Federal Reserve Bank of New York

Shlok Goyal
Harvard University

Donggyu Lee
Federal Reserve Bank of New York

Carlos Lizama
Central Bank of Chile

Emiliano Luttini
The World Bank

Ethan Matlin
Harvard University

Alisdair McKay
Federal Reserve Bank of Minneapolis

Hugo Monnery
Harvard University

Ernesto Pastén
Central Bank of Chile

Matthew Rognlie
Northwestern University
National Bureau of Economic Research

Elisa Rubbo
University of Chicago

Reca Sarfati
Massachusetts Institute of Technology

Thomas J. Sargent
New York University
Sikata Sengupta
University of Pennsylvania

Ludwig Straub
Harvard University
Center for Economic and Policy Research
National Bureau of Economic Research

Giovanni L. Violante
Princeton University
Centre for Economic Policy Research
Institute for Fiscal Studies
National Bureau of Economic Research

Christian K. Wolf
Massachusetts Institute of Technology
National Bureau of Economic Research

Conference Discussants

Jordi Galí
CREI – Universitat Pompeu Fabra

Johathan Heathcote
Federal Reserve Bank of Minneapolis

Alexandre Janiak
Pontificia Universidad Católica de Chile

Markus Kirchner
Central Bank of Chile

David Moreno
Central Bank of Chile

Gastón Navarro
U.S. Federal Reserve Board

Pablo Ottonello
University of Michigan

Table of Contents

Heterogeneity in Macroeconomics: Implications for Monetary Policy An Overview

Sofía Bauducco
Central Bank of Chile

Andrés Fernández
International Monetary Fund

Giovanni L. Violante
Princeton University
Centre for Economic Policy Research
Institute for Fiscal Studies
National Bureau of Economic Research

This volume collects some of the papers presented at the XXV Annual Conference of the Central Bank of Chile, which took place in November 2022 in Santiago, Chile.[1] The theme of the conference was **Heterogeneity in Macroeconomics: Implications for Monetary Policy**. The main objective of this conference was to invite some of the most prominent macroeconomists working on models with salient heterogeneity among households, firms, and banks to present their research and discuss how this class of models can inform the design of monetary policy.

The rapid surge in interest, among academic researchers and policymakers, in the nexus between heterogeneity and monetary policy is associated with the emergence of a new class of models referred to as HANK, an acronym for Hetergoeneous-Agent New Keynesian. HANK models combine two long-standing traditions of macroeconomic theory: (i) the new Keynesian approach to the study of business cycles and stabilization policies and (ii) the heterogeneous-agent incomplete-market approach to the study of the wealth distribution and of those

1. The full program is available on the Central Bank's website.

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

policies that offer social insurance, promote income mobility, and redistribute across households.[2]

The production and monetary policy blocks of this model are exactly the same as in the representative-agent new Keynesian (RANK) model and, as in that framework, they are summarized by three aggregate equations: (i) the Phillips curve, which specifies a relation between inflation and output dynamics; (ii) the Taylor rule, which summarizes how the monetary authority operates its main policy instrument, the nominal interest rate; (iii) and the Fisher equation, which links the real interest rate, the policy rate, and expected inflation. The crucial innovation lies in replacing the representative consumer, and hence the aggregate Euler equation (or the IS curve), with the modern theory of consumption and saving. The starting point of this theory is that households differ ex ante because of innate heterogeneity, and ex post because of idiosyncratic income shocks; then, due to financial market imperfections, these differences transmit to consumption, saving, and welfare. In equilibrium, the absence of perfect risk-sharing yields a non-degenerate cross-sectional distribution of income, consumption, and wealth, as well as individual mobility dynamics across the distribution, both of which resemble their data counterparts. As the bulk of macroeconomics of the last four decades, this class of models is also deeply rooted in the tradition of dynamic stochastic general equilibrium and rational expectations. More recently, economists have extended the analysis of the relation between distributions and monetary policy beyond the household sector to firms as well.[3]

The reason why central banks became engaged in this research program is that HANK models subvert some of the classic tenets of their representative-agent complete-market counterpart. A number of new policy lessons have emerged: (i) the transmission mechanism of monetary policy is no longer centered on intertemporal substitution, the dominant channel in RANK, but it revolves around equilibrium effects operating via shifts in labor income and asset prices; (ii) the cyclicality of income inequality and that of uninsurable labor income risk—both absent from RANK models—can substantially amplify the propagation of aggregate shocks; (iii) monetary policy leaves significant fiscal footprints because of the failure of Ricardian equivalence, a fixture of RANK; (iv) redistributive and social insurance policies

2. We refer the reader to Mankiw and Romer (1991) and Heathcote and others (2009) for an overviews of these two approaches.

3. See Ottonello and Winberry (2020).

are also aggregate stabilization policies and vice versa, i.e., the stark dichotomy between stabilizing business cycles and addressing imperfect insurance and inequality—which was an integral part of Samuelson's neoclassical synthesis—vanishes.[4] For a more detailed discussion of these new ramifications of HANK models, we refer the reader to Violante (2022).

The papers at the conference echoed all these messages. T.J. Sargent gave the keynote address at the conference. His paper, entitled **HAOK and HANK Models**, is a comparison—rich with informative historical references to the founders of modern macroeconomics—between the Heterogeneous-Agent Old Keynesian (HAOK) framework and the Heterogeneous-Agent New Keynesian (HANK) framework.

The old Keynesian paradigm is built on the idea that the macroeconomy can be, at times, in an equilibrium characterized by underemployment of capital and labor because of nominal rigidities. Other times, though, it operates efficiently under full employment, and markets alone are successful in setting prices correctly and allocating resources. In light of this observation, John Maynard Keynes advocated (what Sargent calls) "light-handed" monetary-fiscal interventions during downturns in order to restore full utilization and promote aggregate efficiency.

Paul Samuelson called this theory-policy pair a "neoclassical synthesis." At the heart of this view, Sargent argues, there is the implicit belief in the existence of well-functioning state-contingent transfers through markets, families or social safety nets that effectively insure households against adverse idiosyncratic shocks, such as job losses. As a result, macroeconomists could focus their attention on aggregate business cycles, without any reference to distributional issues. This perspective is what justified James Tobin's definition of macroeconomics as "a field that attains workable approximations by ignoring the effects of distributions on aggregates."

M. Friedman, R. Lucas, E.C. Prescott, and of course T.J. Sargent himself, together with other practitioners of twentieth-century macroeconomics, embraced this view. Most notably, they identified severe logical inconsistencies in early attempts to estimate Keynesian models through simultaneous equations systems. Their research program led to the rational expectations revolution and the paradigm that, seen through the lens of a structural model, time series are

4. For a more detailed discussion of these new ramifications of HANK models, we refer the reader to Violante (2022).

equilibrium stochastic processes. The class of dynamic stochastic general equilibrium models which emerged from that collective intellectual effort and which today pervades macroeconomics lays out environments where "a theory and an econometrics fit together consistently." These models still relied on the idea that the complete-market assumption offers a good approximation to actual economies, and thus the household sector of the economy could be collapsed into a representative agent. In his Presidential Address to the American Economic Association entitled "Macroeconomic Priorities", Bob Lucas wrote that "for individual behavior and welfare, of course, heterogeneity is everything. [...] for determining the behavior of aggregates, [...] household heterogeneity just does not matter very much" (see Lucas, 2003).

Then along came HANK models, which put heterogeneity and market incompleteness front and center in the study of business cycles. As a result of this additional complexity, solving and estimating these models requires new tools, many of which are showcased in the papers presented at the conference. According to Sargent, however, "the HANK revolution is not about tools but about substance. HANK research undermines the neoclassical synthesis." Sargent refers to the dichotomy between stabilization and redistribution implicit in that approach. In HANK economies, instead, traditional stabilization instruments such as countercyclical spending or interest rate cuts are necessarily redistributive and alter the amount of insurance against idiosyncratic shocks.[5] Similarly, traditional redistributive or social insurance instruments—such as tax reforms or expansion of unemployment insurance benefits—necessarily induce aggregate fluctuations because of the heterogeneity of marginal propensity to consume across the population.[6]

Sargent concludes his paper by airing the concern that this new theory-policy package could undermine traditional mandates for monetary policies and provide ammunition to constituencies that want to assign to central banks goals that involve redistribution and reallocation, with the risk of losing sight of price stability. While we share these

5. Bhandari and others (2021) study optimal monetary policy in HANK models and conclude that most of the gains accrue through improved consumption insurance, not higher aggregate efficiency.

6. There are, however, knife-edge cases where HANK models preserve this stark distinction, as articulated by Werning (2015).

concerns, we remain optimistic that policymakers will make good use of this new class of models without misinterpreting their implications. For example, HANK models can be beneficial to help choose between two policy interventions that attain, approximately, the same main objective (e.g., a certain trade-off between inflation and output) with different distributional consequences.

Motivated by the significant surge in energy prices in 2021 and the ensuing debate around the appropriate monetary and fiscal policy responses, the chapter by Adrien Auclert, Hugo Monnery, Matthew Rognlie, and Ludwig Straub (AMRS hereafter) titled **Managing an Energy Shock: Fiscal and Monetary Policy** examines the macroeconomic impact of energy price shocks in advanced, energy-importing economies. Their paper builds a HANK model of a small open economy that imports energy, by adding an energy good to the framework developed in Auclert and others (2021). This model allows the authors to explore how high energy prices may impact consumer demand, a recurrent concern voiced by policymakers. Such demand channel had remained largely unexplored either because existing work focused on the supply-side effects of energy price shocks, abstracting from nominal rigidities, or included sticky prices in RANK models where the demand channel is quantitatively trivial.

AMRS show how, under a realistic calibration of substitution elasticities and marginal propensities to consume (MPCs), energy price shocks can impact GDP via their effect on aggregate demand. Their main analytical result is that, when monetary policy keeps the real interest rate constant, the negative real-income effect (consumers demand less of all goods) dominates the substitution effect (households spend more on domestically produced goods) and aggregate output falls. In addition, aggregate dynamics do not display price-wage spirals because the recession caused by the shock pushes labor demand and wages down, thus offsetting workers' desire for higher wages linked to their decline in purchasing power. The paper also studies alternative monetary policy responses to the shock and uncovers interesting monetary and fiscal policy spillovers across countries. For example, while small individual countries acting alone would be unable to influence global prices, coordinated monetary tightening among energy-importing countries can reduce global energy demand, leading to lower energy prices and imported inflation. Turning to fiscal policies, AMRS argue that energy price subsidies can shield domestic consumers, but tend to have negative spillovers on other economies because they sustain the rise in world energy prices.

This paper offers an example of how models with realistic distributions of marginal propensity to consume can lead to a propagation mechanism of shocks that differs significantly from the one arising in representative-agent models.[7]

Measuring the Redistributive Effects of Monetary Policy: An Application to the Chilean Economy, by Emiliano Luttini, Ernesto Pastén, and Elisa Rubbo, explores the impact of monetary policy shocks across workers who are heterogeneous in their ex-ante (demographic) characteristics, consume different bundles, and work in sectors which differ in their capital intensity and their degree of nominal rigidity. Their multisector model, based on Rubbo (2023) and calibrated to the Chilean economy, shows that the response of employment and income of older, high-income men is almost 8 times larger than that of middle-aged, middle-income men. The reason for this unequal effect of monetary policy shocks can be almost exclusively traced back to the fact that the former group tends to work in industries with more severe nominal rigidities. This result hinges on the specificities of the Chilean economy and on the particular shock analyzed and, therefore, cannot be easily generalized. Taken at face value, though, it suggests that demand shocks might have a stronger impact on groups of workers that are, arguably, less liquidity-constrained and have lower MPCs. This particular distribution of exposure across households dampens the aggregate impact of the shock relative to a representative-agent model. A follow-up question for future research is whether the allocation of labor to sectors that differ in their exposure to demand shocks is an outcome of the optimal behavior of maximizing agents with different ability to self-insure.

The transmission mechanisms of monetary policy in environments with salient heterogeneity, this time in terms of financial intermediaries, is the topic of the chapter by Dean Corbae and Pablo D'Erasmo. In **The Bank Lending Channel Across Time and Space**, the authors set up an oligopoly model of heterogeneous banks with endogenous entry and exit to rationalize stylized facts about the U.S. banking sector after the Riegle-Neal Act, which permitted banks to cross state lines. The policy reform increased bank concentration at the national and state levels, but led to more geographic diversification of local shocks. The authors use the model to study the effects of this change in regulation on the bank lending channel of monetary policy.

7. Kaplan and Violante (2018) discuss various notions of equivalence between equilibrium outcomes in RANK and HANK models.

One important mechanism that generates substantial geographical heterogeneity in their model is that tighter monetary policy influences the equilibrium composition of the banking industry at the state level through the extensive margin, i.e., entry and exit. The reason is that large banks are less sensitive than small ones to a rise in the cost of external funds in the model, consistently with the microdata.[8]

The chapter by Sushant Acharya and coauthors entitled **Estimating HANK for Central Banks** provides a first assessment of the out-of-sample forecasting performance of HANK models, an operational issue that is at center stage for central banks. The authors use the HANK model of Bayer et al. (2024) which features the same types of shocks and frictions as the benchmark representative-agent new Keynesian (RANK) model of Smets and Wouters (2007). The paper makes a methodological contribution by explaining why and how the use of the Sequential Monte Carlo method can yield considerable efficiency gains when estimating HANK models. These gains are instrumental for the task of performing an out-of-sample assessment of these models, as one must estimate them multiple times.

Their main result is that no consistent improvement is found in the out-of-sample forecasting ability of HANK models. In fact, while for some series such as inflation, the forecasting ability is similar to that of standard RANK models commonly used by central banks, for other series, notably consumption growth, the performance is worse. This finding is surprising because the consumption block of the HANK model is much richer and more sophisticated than its RANK counterpart. The authors conjecture that a possible cause is that many parameters in their HANK model, namely all those affecting the model's steady state, continue to be calibrated ex ante.

They conclude that these results should motivate researchers to explore further ways to enhance the quantitative and out-of-sample properties of this class of models, which is still unsatisfactory. This is a worthy effort for a central bank which wants to fully understand the transmission mechanism of monetary policy, and its redistribution implications.

The standard view of macroeconomic dynamics—rooted in the monumental work of Burns and Mitchell (1946)—is that aggregate time series can be decomposed into a long-run component (the trend) and an orthogonal short-run component (the business cycle), which fluctuates around the trend. Quantitative DSGE models used for

8. See Kashyap and Stein (2000).

research and policy analysis fit into this description and, consistently with this view, routinely assume that transitory shocks have no long-term effects on aggregates.

A more nuanced view of business cycles is that of "macroeconomic hysteresis".[9] According to this interpretation, there is no longer a clear-cut separation between cycle and trend, and transitory shocks have very persistent, even permanent, effects on the level of economic activity. In their chapter entitled **From Micro to Macro Hysteresis: Long-Run Effects of Monetary Policy**, Felipe Alves and Giovanni L. Violante explore this alternative view by developing a HANK model built on the micro evidence that job losses lead to persistently lower individual earnings through a combination of skill decay and abandonment of the labor force. They show that these labor market micro-level sources of negative hysteresis give rise to macroeconomic hysteresis in response to transitory negative aggregate demand shocks, modeled as monetary policy innovations. In the model, the strength of these effects increases as one moves down the wage distribution: a decade after the shock, the scarring of labor earnings for workers in the lowest skill quartile is almost ten times as large as the average scarring effect. Hysteresis, thus, operates disproportionately through the labor market trajectories of low-wage workers. Despite the long shadow cast on output, the shock generates only short-lived movements in inflation, which quickly returns to its target. The reason for these dynamics is the decline in labor productivity and labor force participation, which jointly generate inflationary pressures that offset the long-run deflationary pull coming from the persistent decline in output.

Overall, the paper demonstrates that, thanks to their ability to richly represent heterogeneous exposure to aggregate shocks, HANK models are a natural laboratory to explore macroeconomic hysteresis that arises through the aggregation of microeconomic behavior in the labor market or, possibly, other markets.

There is much academic and policy interest centered on the question of how fiscal policy can be used to manage an economy that is stuck at the zero lower bound (ZLB). A number of classic results derived from RANK models prove equivalence, in terms of aggregate outcomes, between standard monetary policy, i.e., adjusting the nominal rate in response to demand shocks, and certain fiscal tools such as time-

9. See Cerra and others (2023) for a survey.

varying consumption subsidies.[10] These equivalence results are very useful to design stabilization policies when the ZLB binds, because they imply that these unconventional fiscal interventions can almost perfectly substitute for the lack of a monetary lever. In their chapter entitled **On the Optimal Use of Fiscal Stimulus Payments at the Zero Lower Bound**, Alisdair McKay and Christian K. Wolf revisit this question from the perspective of a HANK model. They study the optimal policy response for a government that wants to stabilize inflation and output but also dislikes consumption inequality in excess of its steady-state level. Their key conclusion is that, for canonical ZLB-type shocks—like a tightening in borrowing constraints or a distributional shock concentrated on low-income households—the best alternative to classical unconstrained monetary policy is not consumption subsidies, but uniform transfer stimulus payments. The reason is that, beyond perfectly stabilizing aggregate output and inflation, they boost consumption of low-income households, directly counteracting the distributional incidence of the original business-cycle shock. As a result, stimulus payments do not just substitute for conventional monetary policy—they strictly improve upon it.

This paper is a stark example of how HANK models can be useful to policymakers in choosing among alternative policies that achieve very similar aggregate outcomes, but yield different distributional implications.

Another example of this logic can be found in **The Role of Progressivity on the Economic Impact of Fiscal Transfers: a HANK for Chile**, by Benjamin García, Mario Giarda, and Carlos Lizama. The authors start by documenting a strong non-Ricardian response of the Chilean economy to fiscal transfers. In addition, they find that more progressive fiscal transfers display significantly larger effects on consumption than less progressive ones. Motivated by these empirical results, they set up a HANK model with search and matching frictions and calibrate it to key moments of the Chilean economy, such as the fraction of hand-to-mouth households from household surveys and income dynamics from administrative data. The model is able to reproduce the main finding that fiscal transfers geared towards households with higher MPCs have a larger macroeconomic impact. Furthermore, they show that this impact is amplified if transfers are financed through debt instead of taxes. The authors conclude by speculating that the right combination of expansionary fiscal and

10. See Correia and others (2013).

contractionary monetary policy could entail significant redistribution without displaying large adverse aggregate effects.

Overall, the Central Bank of Chile conference showcased a rapidly evolving and vibrant field that is challenging some acquired wisdom, while remaining well anchored to the successful research program of dynamic stochastic general equilibrium models in macroeconomics. Its success will depend on two factors. First, the extent to which practitioners of these models will be able to make a convincing and robust case that a two-way feedback between inequality and the macroeconomy exists and is quantitatively important. Second, the development of better computational and econometric tools to solve globally and estimate stochastic models. Recent advances based on neural nets appear to offer a promising avenue.

References

Auclert, A., M. Rognlie, M. Souchier, and L. Straub. 2021. "Exchange Rates and Monetary Policy with Heterogeneous Agents: Sizing Up the Real Income Channel." Technical Report, National Bureau of Economic Research.

Bayer, C., B. Born, and R. Luetticke. 2024. "Shocks, Frictions, and Inequality in U.S. Business Cycles." *American Economic Review* 114(5): 1211–47.

Bhandari, A., D. Evans, M. Golosov, and T.J. Sargent. 2021. "Inequality, Business Cycles, and Monetary-Fiscal Policy." *Econometrica* 89(6): 2559–99.

Burns, A.F. and W.C. Mitchell. 1946. "Measuring Business Cycles." National Bureau of Economic Research.

Cerra, V., A. Fatás, and S.C. Saxena. 2023. "Hysteresis and Business Cycles." *Journal of Economic Literature* 61(1): 181–225.

Correia, I., E. Farhi, J.P. Nicolini, and P. Teles. 2013. "Unconventional Fiscal Policy at the Zero Bound." *American Economic Review* 103(4): 1172–211.

Heathcote, J., K. Storesletten, and G.L. Violante. 2009. "Quantitative Macroeconomics with Heterogeneous Households." *Annual Review of Economics* 1(1): 319–54.

Kaplan, G. and G.L. Violante. 2018. "Microeconomic Heterogeneity and Macroeconomic Shocks." *Journal of Economic Perspectives* 32(3): 167–94.

Kashyap, A.K. and J.C. Stein. 2000. "What Do a Million Observations on Banks Say about the Transmission of Monetary Policy?" *American Economic Review* 90(3): 407–28.

Lucas, R.E. 2003. "Macroeconomic Priorities." *American Economic Review* 93(1): 1–14.

Mankiw, N.G. and D. Romer. 1991. New Keynesian Economics: Coordination Failures and Real Rigidities." Boston, MA: MIT Press.

Ottonello, P. and T. Winberry. 2020), "Financial Heterogeneity and the Investment Channel of Monetary Policy." *Econometrica* 88(6): 2473–502.

Rubbo, E. 2023. "Monetary Non-Neutrality in the Cross-Section." Technical Report.

Smets, F. and R Wouters. 2007. "Shocks and Frictions in U.S. Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97(3): 586–606.

Violante, G.L. 2022. "What Have We Learned from HANK Models, thus Far?" In *Beyond the Pandemic: the Future of Monetary Policy*. ECB Forum on Central Banking 2021, European Central Bank, Frankfurt, Germany.

Werning, I. 2015. "Incomplete Markets and Aggregate Demand." Technical Report, National Bureau of Economic Research.

HAOK and HANK Models

Thomas J. Sargent
New York University

Accounting for and managing heterogeneities in economic agents' preferences, information sets, and opportunities have always been central to macroeconomic theory. Long before macroeconomics existed as a distinct field, conflicts of interest preoccupied those who designed monetary-fiscal policies.[1] Section 1 describes heterogeneous agent old Keynesian (HAOK) models and the reasons why distinguished twentieth-century macroeconomists used them to analyze the consequences of alternative monetary and fiscal policies. Section 2 describes how informal NBER reference cycle models created by Burns and Mitchell (1946) and single-factor descriptive statistical models, like those sketched by Koopmans (1947) and formalized by Sargent and Sims (1977), framed evidence that motivated HAOK theorists. The goal of quantifying HAOK models motivated the construction of a statistical theory for estimating systems of vector difference equations. Section 3 recalls Kenneth Arrow's skepticism about the consistency of HAOK models with modern general equilibrium theory. Section 4 describes how authors of HANK models challenge key empirical motivations underlying HAOK models and how they subvert the logic underlying the light-handed monetary-fiscal policies affiliated with a neoclassical synthesis. Section 5 tells how functional autoregressions and related descriptive statistical models are being used to gather evidence that might discriminate between HAOK and HANK models. Section 6 concludes by offering opinions about how the HANK project creates promises and controversies.

I thank Tanvi Bansal and Dean Parker for helpful suggestions.

1. See appendix A for some nineteenth-century U.S. examples.

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

1. The Neoclassical Synthesis

The K in HAOK and HANK honors Sir John Maynard Keynes. It is useful to recall the sense in which he intended his *General Theory of Employment, Interest and Money*[2] to be precisely that—a general theory. Keynes wanted his theory to:

- explain equilibria with underemployed resources and excess supplies,
- reduce to "classical" (i.e., Walrasian) general equilibrium theory when resources are fully employed, and
- rationalize light-handed monetary-fiscal interventions that depend only on aggregate data.

Keynes wanted macroeconomic policies to promote aggregate efficiency by letting individuals' choices guide the allocation of resources. To accomplish this, he advocated:

- a price-level target,[3] and
- keeping two government budgets—a current account and a capital account:

- Always balancing the current-account budget.
- Not requiring period-by-period balancing of the capital budget but requiring only its present-value balance.
- Using countercyclical capital-account deficits, but not current-account deficits, to finance public works.[4]

Keynes's advocacy of these light-handed macroeconomic policies presumed the presence of a U.K. 1920s-style social safety net.

In a nutshell, Keynes advocated (i) achieving full employment by using well-timed public investment to sustain adequate demand and then (ii) relying on markets to set relative prices and allocations. Paul Samuelson called this theory-policy package a "neoclassical synthesis". Here is how Keynes described it:

> When 9,000,000 men are employed out of 10,000,000 willing and able to work, there is no evidence that the labour of these 9,000,000 men is misdirected. The complaint against the present system is not that these 9,000,000 men ought to be employed on different tasks, but that tasks should be available for the remaining 1,000,000 men. It is in determining the volume, not the direction, of actual employment that the existing system has broken down.[5]

2. Keynes (1936).

3. Keynes (1924, 1925) emphasized the priority of present value government budget balance as essential determinant of the price level.

4. Proposals to time public works to attenuate the business cycle were in the air in the 1920s. For example, see Foster and others (1928), and Foster and Catchings (1930).

5. See Keynes (1936, chapter 24).

A package of ideas that culminated in his neoclassical synthesis emerged gradually during the years from 1911 to 1931, when Keynes practiced what he later called "classical" macroeconomics. To follow his progress, read chapter 1 in *A Tract in Monetary Reform* (Keynes, 1924), where he analyzed how inflation disrupted (1) distributions of wealth and consumption among (a) investors, (b) the business class, and (c) earners as well as (2) production (i.e., the allocation of resources).[6] His analysis of those disruptions led Keynes to advocate price-level targeting:

> We leave Saving to the private investor, and we encourage him to place his savings mainly in titles to money. We leave responsibility for setting Production in motion to the business man, who is mainly influenced by the profits which he expects to accrue to himself in terms of money. Those who are not in favor of drastic changes in the existing organization of society believe that these arrangements, being in accord with human nature, have great advantages. But they cannot work properly if the money, which they assume as a stable measuring-rod, is undependable. Unemployment, the precarious life of the worker, the disappointment of expectation, the sudden loss of savings, the excessive windfalls to individuals—the speculator, the profiteer—all proceed, in large measure, from the instability of the standard of value.[7]

Keynes disapproved of episodes of redistributions via unforeseen inflations:

> There is no record of a prolonged war or a great social upheaval which has not been accompanied by a change in the legal tender, but an almost unbroken chronicle in every country which has a history, back to the earliest dawn of economic record, of a progressive deterioration in the real value of the successive legal tenders which have represented money.[8]

He regarded those past inflation-engineered redistributions as purposeful:

> Moreover, this progressive deterioration in the value of money through history is not an accident and has had behind it two great driving forces—the impecuniosity of Governments and the superior influence of the debtor class.

6. The way Keynes (1924, chapter 1) sorted through the effects of inflation on distribution and production reminds me of recent analyses of contending effects of alternative government policies in HANK models in terms of imputations of welfare consequences of alternative government policies that flow from (i) redistribution, (ii) insurance, and (iii) efficiency. See Bhandari and others (2023, 2021).

7. Keynes (1924).

8. Ibid.

> . . . the benefits of a depreciating currency are not restricted to the Government. Farmers and debtors and all persons liable to pay fixed money dues share in the advantage. As now in the persons of businessmen, so also in former ages these classes constituted the active and constructive elements in the economic scheme.[9]

Appendix A provides some U.S. historical examples of the episodes that Keynes probably had in mind.[10] The appendix describes some nineteenth-century controversies about how the U.S. federal government should use monetary-fiscal policy to redistribute wealth among nominal net creditors and debtors, controversies that recurred often from the founding of the U.S. republic until Keynes's time. Instances of the same controversies occurred in England, France, and other European countries in the eighteenth and nineteenth centuries. Keynes participated actively and passionately in widespread debates about similar issues that occurred in Europe after World War I. Keynes's response to these debates was to advocate separating a government's price-level goals from its concerns about redistribution:[11]

Keynes advocated targeting the price level.

> If we are to continue to draw the voluntary savings of the community into "investments," we must make it a prime object of deliberate State policy that the standard of value, in terms of which they are expressed, should be kept stable; adjusting in other ways (calculated to touch all forms of wealth equally and not concentrated on the relatively helpless "investors") the redistribution of the national wealth, if in the course of time, the laws of inheritance and the rate of accumulation have drained too great a proportion of the income of the active classes into the spending control of the inactive.[12]

Samuelson, Tobin, Friedman, Lucas, Prescott, and other creators and practitioners of twentieth-century macroeconomics accepted and implemented Keynes's neoclassical synthesis. But first they had to resolve the ambiguities and confusions inherent in Keynes's mostly literary (i.e., nonmathematical) style of analysis. A project to do that began with a string of contributions by Hicks (1937), Tinbergen (1939), Samuelson (1939), Modigliani (1944), and Tobin (1955). They translated and transformed Keynes's analysis into a "general equilibrium" system of n equations in n unknowns having

9. Ibid.

10. Brunnermeier and others (2023) document how German monetary policy during the 1922–1923 hyperinflation purposefully benefitted some citizens at the expense of others. See Newcomb (1865) for a related analysis and criticism of U.S. monetary policy during the 1861–1865 Civil War.

11. Also see Keynes (1931a).

12. See Keynes (1924).

a neat partition into n endogenous variables and several exogenous variables representing monetary and fiscal policy actions. Solutions of those equations could be used to analyze alternative settings of the government's monetary and fiscal actions. To perform the types of statistical implementation and verification of Keynes's general theory that Tinbergen sought, it was necessary to have in hand a specific "n-equations-in-n-unknowns" system of this kind. All of these early works accepted Keynes's reasoning in terms of broad macroeconomic aggregates—employment, interest, and money—because the Great Depression of the 1930s convinced them that understanding and attenuating adverse fluctuations in those aggregates were scientific problems of pressing moral importance.

Meanwhile, little impressed or influenced by Keynes's theorizing but vitally interested in business cycles, for many years Wesley C. Mitchell and Arthur Burns and their teammates at the National Bureau of Economic Research had patiently interrogated many "witnesses" to U.S. business cycles by assembling and studying time series of a diverse collection of quantities and prices, a long line of work that culminated in Burns and Mitchell (1946). From an immense dataset, they extracted a U.S. business cycle by using a home-made data-reduction technique. To summarize their dataset, they constructed a nine-part "reference cycle" onto which they "projected" each of their many time series. From their inductive approach, they organized evidence that, even to economists having more taste and patience for economic theory than Burns and Mitchell did, seemed to justify a constructing macroeconomic theory.

Although Burns and Mitchell (1946) and Tinbergen (1939) used very different methods, both were interested in the same data that somehow "nature" had generated through one process. Both sought to learn about that process by enlisting what modern statisticians call an "inductive bias" or "statistical prior". Indeed, both hypothesized a single-dimensional aggregate. Filling in technical details required to justify and extend the analytical approach of either Burns and Mitchell (1946) or Tinbergen (1939) would require talent and time. Thus, the statistical theory appropriate for estimating parameters of a system of n equations in n unknowns —required to complete Tinbergen's project—had not yet been created. Connections between such a statistical theory and the sorts of statistics that Burns and Mitchell (1946) had assembled were unknown.[13] In the next section, we briefly describe early efforts to learn these connections.

13. King and Plosser (1994) connect the two approaches.

2. Two Types of Statistical Model

In the tradition of Koopmans (1950), I define a statistical model as a probability distribution $f(y \mid \theta)$ of a random vector y indexed by parameters $\theta \in \Theta$. The set Θ describes a manifold of statistical models. In economics and other sciences too, statistical models come, or pretend to come, in two types—descriptive and structural.

- Parameters θ_{descr} of a descriptive model are data summarizers like regression coefficients and entries of covariance matrices of shock vectors. These parameters are not directly interpreted as preference or technology parameters of an economic theory. Instead, they are dimension-reducers, i.e., data-compression devices.
- Some or all of the parameters θ_{struct} of a structural model pin down preferences, technologies, endowments, information structures, surprises that instigate "mistakes of foresight", and so on. These parameters are objects in which economic theories are cast.

Descriptive models are designed to detect patterns and assemble interesting "facts", but not to explain them. Structural models are designed to explain them in terms of the parameters that quantify determinants of demands and supplies. Both types of model play important roles in macroeconomics. The purposes of a descriptive model are dimension reduction, data compression, and pattern recognition. The purpose of a structural model is to uncover invariants that can support theoretical analysis of historically unprecedented policy interventions.

Koopmans and his colleagues at the Cowles Commission initiated a research program that would connect the two types of statistical models. Koopmans (1947) wanted to construct a mapping $\theta_{descr} = F(\theta_{struct})$ so that he could study how to invert it and recover $\theta_{struct} = F^{-1}(\theta_{descr})$. Koopmans (1949, 1950) advocated "structural" Keynesian econometric models that could be used to recommend

aggregate demand management policies that would implement the neoclassical synthesis. [14, 15]

Mid-twentieth-century theorists and econometricians who were inspired by the noble goal of understanding and moderating business cycles and preventing a recurrence of the geopolitical disaster that was the Great Depression of the early 1930s, introduced a distinction between descriptive and structural statistical models that pervades applied econometrics to this day.[16] Leading theorists and econometricians repaired loose ends left by Keynes by representing his ideas as n equations in n unknowns that formed vector stochastic difference equations that could be matched to data. In the five years after WWII, parallel efforts by raw empiricists Burns and Mitchell at the National Bureau of Economic Research and theorist-econometricians at the Cowles Commission, first at the University of Chicago and then at Yale, came to fruition. A memorable debate pitted Koopmans against Burns and Mitchell and posed enduring issues. Koopmans was remarkably even-handed in setting forth and refining a case for using Burns and Mitchell's approach before delineating its limitations:

> When Tycho Brahe and Johannes Kepler engaged in the systematic labor of measuring the positions of the planets, and charting their orbits, they started with conceptions and models of the planetary system which later proved incorrect in some aspects, irrelevant in others. Tycho always, and Kepler initially, believed in uniform circular motion as the natural basic principle underlying the course of celestial bodies. Tycho's main contribution was a systematic accumulation of careful measurements. Kepler's outstanding success was due to a willingness to strike out for new models and

14. Koopmans (1949, 1950) usually started with a structural model with parameters θ_{struct} and then deduced an associated "reduced form" descriptive model with parameters $\theta_{descr} = G(\theta_{struct})$. A major theme of Hansen and Sargent (2013) was to pursue this approach by characterizing the mapping from a structural dynamic model that takes the form of a linear hidden Markov model to an associated vector autoregression that characterizes its likelihood function and that represents its reduced form. Unfortunately, today, the expression "reduced form" is too often used, not in its original Cowles Commission sense, but in the corrupted sense of "incompletely articulated descriptive model".

15. Koopmans prefigures what we now call "indirect inference" as perfected by Gallant and Tauchen (1996). For Gallant and Tauchen, an auxiliary model is a descriptive statistical model that (1) is a likelihood function that describes data well, and (2) can be computed and maximized easily. It is a good idea to estimate a structural model by using score functions of an auxiliary model to generate an appropriate generalized method of moments (GMM) criterion.

16. It pervades "machine learning" as well.

> hypotheses if such were needed to account for the observations obtained. He was able to find simple empirical "laws" which were in accord with past observations and permitted the prediction of future observations. This achievement was a triumph for the approach in which large scale gathering, sifting, and scrutinizing of facts precedes, or proceeds independently of, the formulation of theories and their testing by further facts.
>
> . . . in due course, the theorist Newton was inspired to formulate the fundamental laws of attraction of matter, which contain the empirical regularities of planetary motion discovered by Kepler as direct and natural consequences. The terms "empirical regularities" and "fundamental laws" are used suggestively to describe the "Kepler stage" and the "Newton stage" of the development of celestial mechanics. It is not easy to specify precisely what is the difference between the two stages. Newton's law of gravitation can also be looked upon as describing an empirical regularity in the behavior of matter. The conviction that this "law" is in some sense more fundamental, and thus constitutes progress over the Kepler stage, is due, I believe, to its being at once more elementary and more general. It is more elementary in that a simple property of mere matter is postulated. As a result, it is more general in that it applies to all matter, whether assembled in planets, comets, sun or stars, or in terrestrial objects—thus explaining a much wider range of phenomena.[17]
>
> . . . even for the purpose of systematic and large-scale observation of such a many-sided phenomenon, theoretical preconceptions about its nature cannot be dispensed with . . .[18]

As a sympathetic and constructive critic of Burns and Mitchell's reference-cycle technique, Koopmans indicated how it could be formalized as a single-factor dynamic version of a factor-analytic model of the type that psychologists had used to summarize student test scores as an intelligence quotient.[19]

> The notion of a reference cycle itself implies the assumption of an essentially one-dimensional basic pattern of cyclical fluctuation, a background pattern around which the movements of individual variables are arranged in a manner dependent on their specific nature as well as on accidental circumstances. (There is a similarity here with Spearman's psychological hypothesis of a single mental factor common to all abilities.) This "one-dimensional" hypothesis may be a good first approximation, in the same sense in which the

17. See Koopmans (1947), page 161.

18. Ibid, page 163.

19. Lovie and Lovie (1993) describe the origins and early applications of factor analysis.

> assumption of circular motion provides a good first approximation to the orbits of the planets. It must be regarded, however, as an assumption of the "Kepler stage", based on observation of many series without reference to the underlying economic behavior of individuals. It is in this sense, I believe that the authors refer (page 3) to their definition of business cycles as "a tool of research, similar to many definitions used by observational sciences and, like its analogues, subject to revision or abandonment if not borne out by observation." I believe that the authors would not object to the addition: "or by the logical consequences of observations of a wider range of phenomena."[20]

Thus, Koopmans indicated that some of Burns and Mitchell's data summaries could be organized and sharpened in terms of a factor analytic model, a suggestion that Geweke (1977), Sargent and Sims (1977), and Geweke and Singleton (1981), and others would eventually pursue.

Although Koopmans (1947) had regarded *Measuring Business Cycles* by Burns and Mitchell (1946) as an extensive pattern-recognition and data-reduction exercise that fell short of formally producing a descriptive statistical model, even without such a formalization, Burns and Mitchell's concept of a one-dimensional "reference cycle" influenced leading macroeconomic model builders. I audited Robert E. Lucas's Economics 331 PhD first-year macro class at the University of Chicago in the winter quarter of 1977. Lucas devoted several lectures to describing Burns and Mitchell's procedures for constructing reference cycles through a process of taking moving averages, removing trends, and applying subjective judgments. Using Brock and Mirman (1972) as a benchmark model, Lucas took Burns and Mitchell's single-factor "all business cycles are alike" finding as his starting point. Then he set out to explain "real" and "nominal" outcomes in terms of preferences and constraints facing households, firms, and governments. From Burns and Mitchell's diagrams and other sources, Lucas inferred that, while a one-factor model could approximate quantities well, it seemed that another factor was needed to account for nominal prices. Additional tentative support for Lucas's inferences emerged from Sargent and Sims (1977).

In summary, two interrelated ideas guided authors of HAOK models: (1) an empirical judgment that "all business cycles are similar" captured by Burns and Mitchell's application of their reference-cycle procedure to many U.S. time series, and (2) Keynes's neoclassical synthesis that

20. See Koopmans (1947), page 165.

justified James Tobin's definition of macroeconomics as "*a field that ignores distribution effects*". While many leading U.S. economists after World War II endorsed this approach, not everyone did.

3. Arrow's Challenge

When he reviewed the collected works of Paul Samuelson (1966), Kenneth Arrow called the neoclassical synthesis a scandal:

> . . . Samuelson has not addressed himself to one of the major scandals of current price theory, the relation between microeconomics and macroeconomics. Neoclassical macroeconomic equilibrium with fully flexible prices presents a beautiful picture of the mutual articulations of a complex structure, full employment being one of its major elements. What is the relation between this world and either the real world with its recurrent tendencies to unemployment of labor, and indeed of capital goods, or the Keynesian world of an underemployment equilibrium?[21]

Arrow asserted that:

> If the neoclassical model with full price flexibility were sufficiently unrealistic that stable unemployment equilibrium be possible, then in all likelihood the bulk of the theorems derived by Samuelson, myself, and everyone else from the neoclassical assumptions are also counterfactual. The problem is not resolved by what Samuelson has called "the neoclassical synthesis," in which it is held that achievement of full employment requires Keynesian intervention, but that neoclassical theory is valid when full employment is reached.[22]

Elaborating, Arrow wrote:

> The Samuelson-Keynes view of the world is that full employment is a valid proposition in K(g) only for special values of g, whereas full employment holds in W(g) for all g. If g* is such that full employment holds in K(g*), can it be true that theorems valid in W(g*) are also valid in K(g*)? Obviously, it is not true that the two systems respond similarly to changes in g, since full employment remains valid in one but not in the other.[23]

It is natural to expect that Arrow's criticisms would be taken to heart especially by rational expectations macroeconomists like Lucas and Prescott, who were eager to bring lessons from Arrow's and Debreu's analysis of general models into macroeconomics.[24]

21. Arrow (1967), page 734.
22. Ibid, page 735.
23. Ibid, page 735.
24. See Prescott and Lucas (1972).

Lucas (1987) addressed some of Arrow's doubts, though at the end of the day, Lucas embraced the neoclassical synthesis. Manuelli and Sargent (1988) discussed some of the steps that Lucas took to separate redistribution and insurance from the determinants of aggregate outcomes.

After criticizing the theoretical foundations of the neoclassical synthesis, Arrow commended statistical findings that had modified recent refinements of macroeconomic theories:

> The major developments, the development of more subtle theories of the consumption function and the distributed-lag theories of investment, have been closely associated with econometric investigation.[25]

In section 2, we described empirical findings that fortified a HAOK modeling tradition that embraced a neoclassical synthesis. In section 5, we'll describe how more recent investigations bear on the HANK project.

4. HANK Models

Although a neoclassical synthesis dominated quantitative macroeconomics for many decades, heterogeneous agent models were always present and taken seriously as early as the multiple-class models of Kalecki (2016), that emphasized heterogeneous marginal propensities to consume and their implications for fiscal policy. Indeed, important components of Friedman (1956) were his empirical and theoretical analyses of differences in marginal propensities to consume across classes of consumers who faced stochastic processes of nonfinancial income with different mixtures of permanent and temporary components. Furthermore, a substantial body of work by macroeconomists occupying the last third of Ljungqvist and Sargent (2018) applied recursive contracts to analyze how to arrange social insurance in the presence of information and enforcement difficulties.[26]

25. Arrow (1967), page 733.

26. Interesting examples of such work are Pavoni and Violante (2007) and Pavoni and others (2016), who analyze optimal arrangements for inducing welfare recipients to enter gainful employment. They do "recursive mechanism design", also known as "dynamic programming squared", in which history dependent allocations are represented recursively by using agents' continuation values as state variables in a planner's value function. Thus, Pavoni and others (2016) deploy ". . . *several policy instruments (e.g., job-search, assisted search, mandated work) the principal can use, in combination with welfare benefits, in order to minimize the costs of delivering promised utility to the agent. The generosity of the program and the skill level of the unemployed agent determine the optimal policy instrument to be implemented.*"

Nevertheless, n-equations-in-n-unknowns quantitative models of macroeconomic equilibrium continued to be cast in terms of macroeconomic aggregates (i.e., cross-section averages).[27] Macroeconomists refined how to acknowledge heterogeneity but still preserve a macroeconomic analysis cast solely in terms of aggregates. Prominent examples include Lucas (1982, 1987, 2003).[28] Thus, recall how Lucas (1982) carefully arranged a complete set of state-contingent contracts and an initial distribution of wealth across countries to prevent the distribution of wealth across countries from affecting prices and aggregate quantities. Lucas (1987, 2003) assumed a complete set of state-contingent contracts, an effective social safety net, and a monetary-fiscal policy that eliminated avoidable adverse fluctuations. I read Lucas as estimating the residual gains to aggregate efficiency that remained possible beyond those that had been achieved by Volcker and Greenspan. His finding that they were small induced Lucas to advocate focusing research and policy improvements on secular growth, rather than on further attenuating business cycles. In similar ways, creators of representative-agent New Keynesian (RANK) models that swept into central banks and macro textbooks in the 1990s also pushed heterogeneity into the background to justify casting their n-equations-in-n-unknowns models in terms of aggregates.[29]

Then along came HANK models.

HANK models are part of a broad project to put heterogeneity front and center in macroeconomics. They substantially increase the dimension n in n-equation-in-n-unknown models by including higher moments of cross sections of wealth and income components as determinants of cross-section means. Dynamic programming, dynamic programming squared (i.e., recursive contracts), vector autoregressions, and structural macroeconometrics are HANK modelers' hammers and saws. The HANK revolution is not about tools but about substance. HANK research undermines the neoclassical synthesis in several ways. First, it contributes descriptive statistical models.[30] These models detect relations among the higher moments

27. Edward Prescott urged his students and everyone else who would listen to say "aggregate economics", not "macroeconomics".

28. Prescott (2005, 2006a, 2006b) used distinct theories of aggregation to construct an aggregate labor supply curve, one based on Rogerson employment lotteries, the other based on incomplete markets, self-insurance, and time-averaging. He switched from one to the other in between the two published versions of his Nobel lecture.

29. For many RANK models, $n = 3$.

30. For example, see Guvenen and others (2014, 2021), and Heathcote and others (2023).

and the means of cross sections of incomes and wealth means. They indicate that current values of higher moments contain information about future cross-section averages. Second, it has invented structural HANK models[31] that undermine the HAOK prescription from Keynes that macroeconomic policy should be light-handed and separate from policies that redistribute income and wealth. Furthermore, HANK modelers would replace a low-inflation mandate (or a low-inflation plus low-unemployment mandate) for a central bank and focus instead on other outcomes.

Thus, Bhandari and others (2021) apply recursive contracts analysis to an ex-ante heterogeneous agent HANK model. They compare outcomes and policies under optimal history-dependent policies with those recommended by ordinary Taylor rule and interpret differences in terms of motivations of a Ramsey planner. Responses of optimal policies to aggregate shocks differ qualitatively from what they would be in a corresponding representative agent economy. They are an order of magnitude larger. An ordinary Taylor rule is strongly dominated. A motive to provide insurance that arises from heterogeneity and incomplete markets outweighs price stabilization motives that ordinarily rule in a representative-agent New Keynesian model. To understand sources of welfare gains relative to an ordinary Taylor rule, they use a decomposition of those gains proposed by Bhandari and others (2023) into parts attributable to insurance, redistribution, and aggregate efficiency. They find that an insurance component is positive and greater than 100 percent, that a redistribution component is small, and that an aggregate efficiency component is negative. They summarize their results as follows:

> . . . essentially all the welfare gains from optimal HANK policies arise from the additional insurance that they provide. Provision of insurance comes at the cost of sacrificing price stability, which creates deadweight losses and lowers total aggregate resources available for consumption. This explains why the aggregate efficiency component is negative.

31. For example, see Kaplan and others (2018), and Kaplan and Violante (2018).

5. Functional Autoregressions and HANK

The HANK modeling project fosters both descriptive and structural statistical models. In terms of descriptive models, new tools—or extensions of old ones—are being applied to revisit Burns and Mitchell's (1946) characterization of business cycles with NBER reference cycles and with dynamic versions of the Spearman single-factor models mentioned by Koopmans (1947). This work is directed at reexamining and refining the single-factor characterization of macro time series that originally buttressed the neoclassical synthesis. Here I briefly describe a useful tool for constructing descriptive models of cross-section dynamics that extends the vector autoregression technology that for 45 years macroeconomists have deployed to construct descriptive models of macroeconomic variables. Its purpose is to construct an autoregression for a stochastic process of cross-section densities $p_t(x)$, $t \in T$, where T is the set of integers. Density $p_t(x)$ has dimension infinity. It is convenient to work with log densities $\ell_t = \log p_t(x)$ and to fit a VAR for an $\ell_t(x)$ process.To approximate an infinite dimensional VAR, one estimates a finite K-dimensional VAR for coefficients of K-basis functions for a cross-section density. Thus, let a first-order functional VAR be

$$\ell_{t+1}(x) = \int B(x,\tilde{x})\,\ell_t(\tilde{x})\,d\tilde{x} + u_{t+1}(x)$$

or

$$\ell_{t+1} = B\ell_t + u_{t+1}, u_{t+1} \perp \ell_t$$

Make an approximation

$$\ell_t(x) \approx \left[\xi_1(x),\ldots,\xi_K(x)\right]\cdot\left[\alpha_{1t}\ldots\alpha_{Kt}\right],$$

where the basis functions $\xi_i(x)$ might be sieves or functional principal components. Run a first-order VAR on the basis coefficients

$$\alpha_{t+1} = A\alpha_t + u_{\alpha,t+1}, u_{\alpha,t+1} \perp \alpha_t.$$

Then back out approximate log cross-section densities $\ell_t(x)$.

Time series macro-econometricians at the University of Indiana have fit functional VARs to interesting cross-section log densities. They have fit functional VARs as ingredients of both descriptive

and structural statistical models. To acknowledge the prevalence of stochastic geometric growth in state-of-the-art ways, Chang and others (2019) describe how to incorporate cointegration and additive functionals in the spirit of Hansen (2012). Liu and Plagborg-Møller (2021) estimate a heterogenous-agent structural model. Chang and others (2022a) formulate a functional VAR for aggregates and a cross-section consumption density as a hidden Markov model. More Indiana macro is on the way in a work-in-progress paper by Chang and others (2022b).

Findings of these papers bear on the plausibility and promise of the HANK project. I'll confine myself here to a few remarks about Chang and others (2022a). After they fit a descriptive functional VAR as a hidden Markov model, in the process displaying high technical virtuosity, they offer an informative discussion of mappings $\theta_{struct} = F^{-1}(\theta_{descr})$ for some HANK models simulated under some interesting scenarios. Their findings are bound to be controversial because their descriptive model detects limited dynamic influences that pass from higher cross-section moments to cross-section averages. This seems to be a discouraging finding for the HANK project. But I hesitate to conclude that, because maybe the findings describe outcomes after prevailing social safety-net and aggregate demand management policies have generated effective "off-equilibrium" feedbacks from cross-section dynamics to aggregates, while observed equilibrium paths conceal those feedbacks. This interpretation is a counterpart to my earlier interpretation of costs of business cycles quantified by Lucas (2003).

6. Concluding Remarks

The HANK project is promising and provocative. It is being pursued by technically able researchers who are full of ideas and analytical powers, and who thoroughly know the HAOK and real business cycle models that they want to improve.[32] Their HANK project has an electric charge and is bound to be controversial because it challenges the neoclassical synthesis and a widely believed prescription for separating macro policy design from policies to redistribute income and wealth. Because they undermine single and dual mandates for monetary policies, HANK research is bound to attract attention from constituencies that today want to assign goals to central banks that involve redistribution and reallocation. Some of these goals are so

32. Most of them are diplomats, so they'd say "improve" instead of "replace".

foreign to what Keynes (1924, 1936) advocated that perhaps we should remove the K from HANK.

The descriptive modeling branch of the HANK research project brings new interest to tools, both old and new. An old tool whose promise was long neglected or unrealized was invented by Koopman (1931). He constructed an operator that, by measuring appropriate functions of the state (some eigenfunctions), maps a lower-order nonlinear dynamic system into a higher-order linear system. In doing so, the Koopman operator makes the optimal linear control theory that has long been a mainstay of rational expectations econometrics[33] applicable to an interesting class of nonlinear models. It also brings links to functional autoregressions, in particular to some recent applications of machine learning to fluid dynamics in the form of dynamic mode decompositions, called DMD. DMD can be a fast way of estimating a first-order functional VAR by applying a singular value decomposition (SVD) to a tall-skinny data matrix X.[34]

33. See the introduction to Lucas and Sargent (1981).
34. See Tu and others (2014), and Brunton and Kutz (2022).

REFERENCES

Arrow, K.J. 1967. "Samuelson Collected." *Journal of Political Economy* 75(5): 730–37.

Bhandari, A., D. Evans, M. Golosov, and T.J. Sargent. 2021. "Inequality, Business Cycles, and Monetary-Fiscal Policy." *Econometrica* 89(6): 2559–99.

Bhandari, A., D. Evans, M. Golosov, and T.J. Sargent. 2023. "Efficiency, Insurance, and Redistribution Effects of Government Policies." Technical Report, Working Paper.

Brock, W.A. and L.J. Mirman. 1972. "Optimal Economic Growth and Uncertainty: The Discounted Case." *Journal of Economic Theory* 4(3): 479–513.

Brunnermeier, M.K., S.A. Correia, S. Luck, E. Verner, and T. Zimmermann. 2023. "The Debt-Inflation Channel of the German Hyperinflation." Working Paper No. 31298, National Bureau of Economic Research.

Brunton, S.L. and J.N. Kutz. 2022. *Data-driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press.

Burns, A.F. and W.C. Mitchell. 1946. "Measuring Business Cycles." National Bureau of Economic Research.

Chang, M., X. Chen, and F. Schorfheide. 2022a. "Heterogeneity and Aggregate Fluctuations." Technical Report, University of Pennsylvania and Yale University.

Chang, Y., B. Hu, and J.Y. Park. 2019. "Econometric Analysis of Functional Dynamics in the Presence of Persistence. Technical Report, Department of Economics, Indiana University.

Chang, Y., S. Kim, and J.Y. Park. 2022b. "How Do Macroaggregates and Income Distribution Interact?" Technical Report, Department of Economics Indiana University.

Edwards, S. 2018. *American Default: The Untold Story of FDR, the Supreme Court, and the Battle over Gold*. Princeton University Press.

Fisher, I. 1933. "The Debt-deflation Theory of Great Depressions." *Econometrica: Journal of the Econometric Society* 337–57.

Foster, W.T. and W. Catchings. 1930. "Mr. Hoover's Road to Prosperity." *Review of Reviews* 81: 50–52.

Foster, W.T. and W. Catchings. 1928. *Road to Plenty*. Boston, MA: Houghton Mifflin Company.

Friedman, M. 1956. *A Theory of the Consumption Function*. Princeton, NJ: Princeton University Press.

Gallant, A.R. and G. Tauchen. 1996. "Which Moments to Match?" *Econometric Theory* 12(4): 657–81.

Geweke, J.F. 1977. "The Dynamic Factor Analysis of Economic Time Series." In *Latent Variables in Socio-economic Models*, edited by D. Aigner and A. Goldberger. New York, NY: North Holland.

Geweke, J.F. and K.J. Singleton. 1981. "Maximum Likelihood 'Confirmatory' Factor Analysis of Economic Time Series. *International Economic Review* 37–54.

Guvenen, F., S. Ozkan, and J. Song. 2014. "The Nature of Countercyclical Income Risk." *Journal of Political Economy* 122(3): 621–60.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song. 2021. "What Do Data on Millions of U.S. Workers Reveal About Lifecycle Earnings Dynamics?" *Econometrica* 89(5): 2303–39.

Hall, G.J. and T.J. Sargent. 2014. "Fiscal Discriminations in Three Wars." *Journal of Monetary Economics* 61: 148–66.

Hall, G.J. and T.J. Sargent. 2021. "Debt and Taxes in Eight U.S. Wars and Two Insurrections." *The Handbook of Historical Economics*: 825–80.

Hansen, L.P. and T.J. Sargent. 2013. *Recursive Models of Dynamic Linear Economies*. The Gorman Lectures in Economics Series. Princeton, NJ: Princeton University Press.

Hansen, L.P. 2012. "Dynamic Valuation Decomposition within Stochastic Economies." *Econometrica* 80(3): 911–67.

Heathcote, J., F. Perri, G.L. Violante, and L. Zhang. 2023. "More Unequal We Stand?" Inequality Dynamics in the United States 1967–2021. *Review of Economic Dynamics* 50: 235–66.

Hicks, J.R. 1937. "Mr. Keynes and the 'Classics': A Suggested Interpretation." *Econometrica: Journal of the Econometric Society*: 147–59.

Kalecki, M. 2016. *Studies in the Theory of Business Cycles*: 1933-1939. Oxfordshire, U.K.: Routledge.

Kaplan, G. and G.L. Violante. 2018. "Microeconomic Heterogeneity and Macroeconomic Shocks." *Journal of Economic Perspectives* 32(3): 167–94.

Kaplan, G., B. Moll, and G.L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

Keynes, J.M. 1924. *A Tract on Monetary Reform*. New York, NY: Harcourt, Brace, and Company.

Keynes, J.M. 1925. *The United States and Gold. In European Currency and Finance*, edited by J.P. Young. Washington, DC: Government Printing Office.

Keynes, J.M. 1930. *A Treatise on Money: Pure Theory of Money*, vol. I. London, UK: Macmillan.

Keynes, J.M. 1931a. *Essays in Persuasion. An Open Letter to the French Minister of Finance* (1926). Edinburgh, U.K.: R.& R. Clark, Limited.

Keynes, J.M. 1931b. *A Treatise on Money*, vol. 2.

Keynes, J.M. 1936. *The General Theory of Employment, Interest and Money*. London, U.K.: Macmillan.

King, R.G. and C.I. Plosser. 1994. "Real Business Cycles and the Test of the Adelmans." *Journal of Monetary Economics* 33(2): 405–38.

Koopman, B.O. 1931. "Hamiltonian Systems and Transformation in Hilbert Space." *Proceedings of the National Academy of Sciences* 17(5): 315–18.

Koopmans, T.C. 1947. "Measurement without Theory." *Review of Economics and Statistics* 29(3): 161–72.

Koopmans, T.C. 1949. "The Econometric Approach to Business Fluctuations." *American Economic Review* 39(3): 64–72.

Koopmans, T.C. 1950. *Statistical Inference in Dynamic Economic Models*. New York, NY: Wiley.

Liu, L. and M. Plagborg-Møller. 2021. "Full-information Estimation of Heterogeneous Agent Models Using Macro and Micro data. CAEPR Working Paper Series 2021-001.

Ljungqvist, L. and T.J. Sargent. 2018. *Recursive Macroeconomic Theory*. Cambridge, MA: MIT Press, 4th ed.

Lovie, A.D. and P. Lovie. 1993. "Charles Spearman, Cyril Burt, and the Origins of Factor Analysis." *Journal of the History of the Behavioral Sciences* 29(4): 308–21.

Lucas, R.E., Jr. 1982. Interest rates and currency prices in a two-country world. Journal *of Monetary Economics* 10 (3): 335–59.

Lucas, R.E., Jr. 1987. *Models of business cycles*. Oxford and New York: Basil Blackwell.

Lucas, R.E., Jr. 2003. "Macroeconomic Priorities." *American Economic Review* 93(1): 1–14.

Lucas, R.E., Jr. and T.J. Sargent. 1981. *Rational Expectations and Econometric Practice*. University of Minnesota Press.

Manuelli, R. and T.J. Sargent. 1988. "Models of Business Cycles: A Review Essay." *Journal of Monetary Economics* 22(3): 523–42.

Modigliani, F. 1944. "Liquidity Preference and the Theory of Interest and Money." *Econometrica, Journal of the Econometric Society* 45–88.

Newcomb, S. 1865. *A Critical Examination of our Financial Policy during the Southern Rebellion*. New York, NY: D. Appleton & Co.

Pavoni, N. and G.L. Violante. 2007. "Optimal Welfare-to-work Programs." *Review of Economic Studies* 74(1): 283–318.

Pavoni, N., O. Setty, and G. Violante. 2016. "The Design of 'Soft' Welfare-to-work Programs." *Review of Economic Dynamics* 20: 160–80.

Prescott, E.C. 2005. "The Transformation of Macroeconomic Policy and Research." In *Les Prix Nobel 2004*, 370-395. Stockholm: Almqvist & Wiksell International.

Prescott, E.C. 2006a. "Comment." In *NBER Macroeconomics Annual 2006*, edited by D. Acemoglu, K. Rogoff, and M. Woodford. Cambridge, MA: MIT Press.

Prescott, E.C. 2006b. Nobel Lecture: "The Transformation of Macroeconomic Policy and Research." *Journal of Political Economy* 114(2): 203–35.

Prescott, E.C. and R.E. Lucas, Jr. 1972. "A Note on Price Systems in Infinite Dimensional Space." *International Economic Review* 416–22.

Samuelson, P.A. 1939. "Interactions between the Multiplier Analysis and the Principle of Acceleration." *Review of Economics and Statistics* 21(2): 75–8.

Samuelson, P.A. 1966. *The Collected Scientific Papers of Paul A. Samuelson*, vol. 1,2,3. Cambridge, MA: MIT press.

Sargent, T.J. 2012. "Nobel Lecture: United States Then, Europe Now." *Journal of Political Economy* 120 (1): 1–40.

Sargent, T.J. and C.A. Sims. 1977. "Business Cycle Modeling without Pretending to Have Too Much a priori Economic Theory. In *New Methods in Business Cycle Research*, edited by C.A. Sims. Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Tinbergen, J. 1939. *Business Cycles in the United States of America, 1919-1932*. League of Nations.

Tobin, J. 1955. "A Dynamic Aggregative Model." *Journal of Political Economy* 63(2): 103–15.

Tu, J.H., C.W. Rowley, D.M. Luchtenburg, S.L. Brunton, and J.N. Kutz. 2014. "On Dynamic Mode Decomposition: Theory and Applications." *Journal of Computational Dynamics* 1(2): 391–421.

Appendix A. Keynes as a Historian and Prognosticator

I describe some of the monetary-fiscal policy controversies that Keynes had in mind when, in the passage cited in section 1, he said that "*There is no record of a prolonged war or a great social upheaval which has not been accompanied by a change in the legal tender, but an almost unbroken chronicle in every country which has a history, back to the earliest dawn of economic record, of a progressive deterioration in the real value of the successive legal tenders which have represented money*.[35] While section A.1 indicates that Keynes's "unbroken chronicle" characterization doesn't describe nineteenth-century U.S. outcomes well, it does capture how contending interests sought to turn Federal monetary policy decisions to their advantage. Section A.2 then documents how twentieth-century U.S. outcomes confirmed Keynes's pessimism about "*progressive deterioration in the real value of the successive legal tenders which have represented money*".

A.1 Nineteenth-Century U.S. Episodes

I confine this subsection to controversies that raged during the U.S. Civil War (1861-1865) and the 15 years that followed its end. Monetary-fiscal policies that contributed to outcomes during those years were influenced by statesmen's memories and understandings of earlier wars that had unleashed similar forces. Thus, rehearsals for those Civil War monetary-fiscal controversies occurred during and following the U.S. War for Independence from 1776 to 1783 and again during and following the U.S. War of 1812.[36] After glancing at some of the nineteenth-century outcomes, I'll turn briefly to some U.S. data from the twentieth century. All of these episodes illustrate how the issues and forces described by Keynes had preoccupied U.S. monetary-fiscal policymakers and their constituencies. I'll reproduce graphs of U.S. price levels and ex-post returns on Federal public-debt data assembled by George Hall of Brandeis University.[37]

35. See the chapters on historical evidence in Keynes (1930, 1931b).

36. The War of 1812 outcome pattern reversed one that characterized the U.S. War of Independence and its aftermath, a consequence of deliberate policy choices described by Hall and Sargent (2014) and Sargent (2012).

37. For many more details see Hall and Sargent (2021) and Hall and Sargent (2014).

Figure A.1 Log Price Level

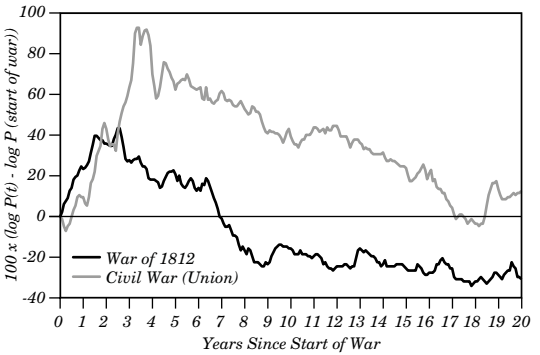


Figure A.2 Cumulative Real Returns

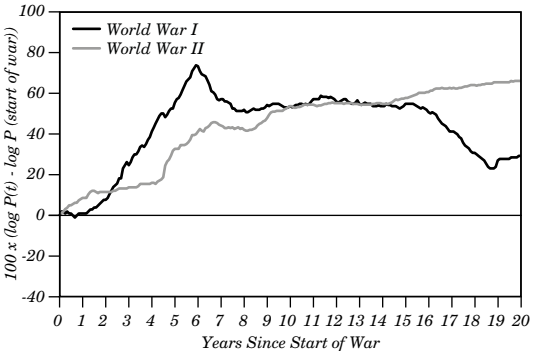


Figure 1 shows the logarithm of the U.S. price levels during and after two big nineteenth-century U.S. wars—the War of 1812 and the Civil War. Figure 2 shows cumulative returns on a representative portfolio of U.S. federal debt during and after those two wars. I'll focus on the Civil War. In 1862, the Union (northern) government left the gold standard and issued an inconvertible paper currency called greenbacks that it made a legal tender for most, but not all, debts, both public and private. By 1864, the greenback had depreciated to about 40 gold cents per greenback dollar, the gold-greenback exchange rate moving with outcomes of battles between Union and Confederate forces. The war ended in April 1865 with gold at 60 cents per greenback dollar. The

price level was denominated in greenbacks; its movements mirrored those of the gold-greenback exchange rate. Our graphs show how the price level rose during the war and how federal creditors received low returns during the war but high returns afterward. This pattern echoed the U.S. experience during the War of 1812. [38]

From 1865 until 1879 and beyond, controversy swirled about whether to make the greenback convertible into gold, and at what exchange rate. It was especially heated from 1865 until March 1869, when Ulysses S. Grant was inaugurated as President.[39] Congress had left ambiguous whether it intended the face value of important classes of bonds (the famous 5-20s) to be paid in greenbacks or gold. Many private bonds had been denominated in greenbacks, including many railroad bonds. Advocates for creditors contended with advocates for debtors, provoking debates cutting across both major political parties and regions. The following words from two of the highest authorities are examples of the contending positions. As an advocate of "rescheduling" (i.e., partial default) we cite President Andrew Johnson, in his Fourth Annual Message of December 9, 1868:

> There seems to be a general concurrence as to the propriety and justness of a reduction in the present rate of interest . . . The lessons of the past admonish the lender that it is not well to be over-anxious in exacting from the borrower rigid compliance with the letter of the bond.

Against President Johnson and most of the Democratic party, the Republican party advocated honoring all public debts, as stated in plank 3 of their Republican Party Platform (1868):

> We denounce all forms of repudiation as a national crime; and national honor requires the payment of the public indebtedness in the utmost good faith to all creditors at home and abroad, not only according to the letter but the spirit of the laws under which it was contracted.

Republican candidate General Ulysses S. Grant won the 1868 election. At his first Inaugural Address, on 4 March 1869, he said:

38. It also echoed experience in England during and after the wars with France from 1797 to 1815. It differed from the U.S. experience during and after the U.S. War of independence in ways that persuaded policymakers during the War of 1812 to do things differently. See Hall and Sargent (2014).

39. Newcomb (1865) criticized Union monetary policy for provoking adverse redistributions consequent on its making inconvertible greenbacks a legal tender. His book is remarkable in a number of ways, one being how far he gets deploying the labor theory of value, another being an information-theoretic analysis of optimal taxation in which ingredients of Ramsey and Mirrlees theories are both present.

> A great debt has been contracted in securing to us and our posterity the Union. The payment of this, principal and interest, as well as the return to a specie basis as soon as it can be accomplished without material detriment to the debtor class or to the country at large, must be provided for. To protect the national honor, every dollar of Government indebtedness should be paid in gold, unless otherwise expressly stipulated in the contract. Let it be understood that no repudiator of one farthing of our public debt will be trusted in public place, and it will go far toward strengthening a credit which ought to be the best in the world and will ultimately enable us to replace the debt with bonds bearing less interest than we now pay.

The Republicans delivered on Grant's promise in a process full of improvisations and postponements that unfolded during and after the two Grant administrations (1869–1877). The U.S. Treasury made greenbacks convertible at par into gold starting on 1 January 1879.[40]

A.2 Twentieth-Century U.S. Outcomes

The preceding graphs and quotes provide examples of some of the same disputes about manipulating the price level to redistribute wealth among creditors and debtors that concerned Keynes (1924). In those nineteenth-century U.S. episodes, a coalition that did not want to use the price level to redistribute wealth from nominal creditors to nominal debtors had prevailed. Those nineteenth-century episodes are exceptions to Keynes's characterization of secular debasement of legal tenders as an "*unbroken chronicle in every country which has a history*". Economic historians have presented many more such exceptions in the nineteenth and earlier centuries. But outcomes in the twentieth century differed from the nineteenth century. Figures 3 and 4, respectively, show the log of price level and cumulative real returns on the U.S. Federal debt from the beginnings of World Wars I and II.

40. It remained there until 1933. Proposals to redistribute via inflation resurfaced often after 1879.

Figure A.3 Log Price Level

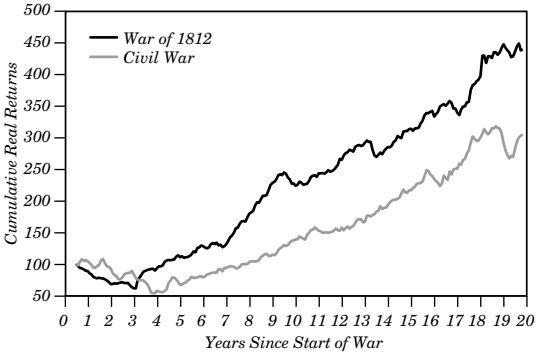
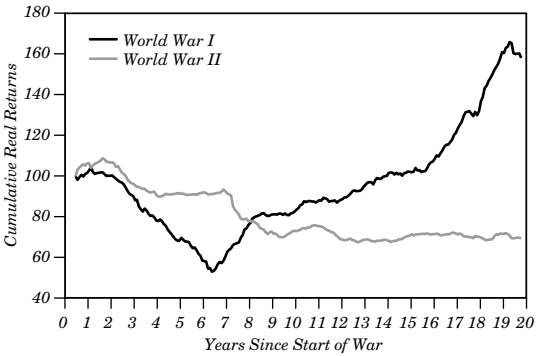


Figure A.4 Cumulative Real Returns



Price levels rose persistently after the starts of both world wars. The Great Depression from 1929 until the end of our graph rise after WWI temporarily reversed the rise. That reversal, and the redistributions to nominal creditors from nominal debtors that accompanied it, had concerned Keynes (1924) as well as Fisher (1933). Those concerns inspired monetary-fiscal policies of President Franklin Delano Roosevelt, which were explicitly designed to redistribute from nominal creditors to nominal debtors.[41]

41. See Edwards (2018).

Managing an Energy Shock: Fiscal and Monetary Policy

Adrien Auclert
Stanford University
Center for Economic and Policy Research
National Bureau of Economic Research

Hugo Monnery
Harvard University

Matthew Rognlie
Northwestern University
National Bureau of Economic Research

Ludwig Straub
Harvard University
Center for Economic and Policy Research
National Bureau of Economic Research

In recent years, advanced economies have faced a large increase in the price of energy.[1] Prices for natural gas, crude oil, and electricity began to rise in 2021, then surged after the Russian invasion of Ukraine in February 2022 and, while they have fallen somewhat since, their future path remains uncertain. This sudden increase has

Prepared for the Proceedings of the XXV Annual Conference of the Central Bank of Chile, held in November 2022. We are grateful for helpful comments and suggestions from Jenny Chan, Sebastián Edwards, Jordi Galí, Peter Ganong, Pierre-Olivier Gourinchas, Jonathan Heathcote, Oleg Itskhoki, Enisse Kharroubi, Robert Kollmann, Moritz Lenel, Ben Moll, Dmitry Mukhin, Iván Werning, as well as from seminar participants at the 2023 AEA Meetings, Bank of Canada, Bank of England, Central Bank of Chile, Chicago Fed, EABCN Monetary Policy conference, Hamburg, Hoover, IMF, National Bank of Belgium, San Francisco Fed, Sciences Po, 2023 SED Conference, Yale, and the University of Pennsylvania. This research is supported by the National Science Foundation grant awards SES-1851717 and SES2042691 as well as the Domenic and Molly Ferrante Award.

1. See figure 1a.

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

led to debate about the appropriate response of monetary and fiscal policy—especially in Europe, where much energy is imported.

A key concern for policymakers has been the likely adverse impact of high energy prices on consumer demand. For instance, ECB chief economist Phillip Lane has argued that:[2]

> In addition to the direct and indirect impact of a surge in energy prices on inflation, it is necessary to recognize the adverse income and wealth effects of rising energy import prices on aggregate demand. Since the euro area is a large-scale net energy importer, an increase in the relative price of energy [implies] a net outward income transfer to the countries supplying energy to the euro area, [...] an adverse terms of trade movement, and a decline in real incomes, [...] with knock-on effects for consumption behavior.

This concern for knock-on effects on consumption motivated numerous fiscal packages, including direct transfers to households, VAT cuts, and other price regulations aimed at cushioning the impact of energy prices on real incomes.[3] Yet, in spite of a large literature on the macroeconomic effects of energy price shocks, standard theoretical models do not feature a direct link between high energy prices and aggregate demand.

Papers that study the supply-side effect of energy price shocks, such as Baqaee and Farhi (2019), Baqaee and Farhi (2022), and Bachmann and others (2022), find that rises in energy prices have a very limited effect on GDP, given realistic substitution elasticities. Since these papers abstract from nominal rigidities, they do not feature an aggregate demand channel. Yet, concerns about depressed aggregate demand appear to be well founded. For instance, the European GDP performance has been lackluster, at least compared to the United States,[4] with consumption playing a significant role in accounting for this difference. Moreover, research has found that the marginal propensity to consume (MPC) out of energy price increases is quite large.[5]

Papers that do feature an aggregate demand channel, such as New Keynesian models with oil, usually feature households that have a very low MPC out of energy, either because they use complete markets

2. See Inflation Diagnostics at the blog in the European Central Bank site, 25 November 2022.

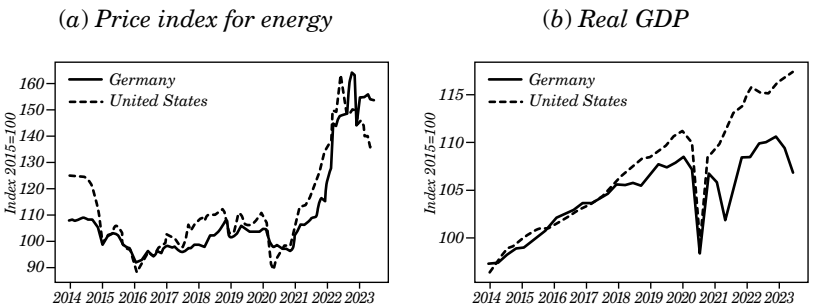
3. See Ari and others (2022) and Sgaravatti and others (2023).

4. See Figure 1b.

5. See Gelman and others (2023).

to insure against changes in oil prices,[6] or because their permanent-income behavior leads them to smooth the effect of any price change on their consumption.[7] In these models, oil price shocks can cause a recession, but only because of the endogenous response of monetary policy to the inflation caused by the shock, rather than the direct effect of the shock on household real incomes and spending.[8,9] Yet it is this direct effect that seems to concern policymakers. Further, tightening of monetary policy in the euro area has lagged behind the United States, so that it is difficult to argue that the difference in figure 1b can be accounted for by more restrictive monetary policy in Germany.[10]

Figure 1. Energy Price Index and Real GDP in Germany vs. the United States



Source: Authors' calculations.
(a): Energy CPI in the U.S. (FRED: CPIENGSL) and energy HICP for Germany (Eurostat: EI_CPHI_M:CP-HIE).
(b): GDP in the U.S. (FRED:GDPC1) and in Germany (EEurostat:NAMQ_10_GDP:B1G). All indexed to 100 in 2015.

6. See Blanchard and Galí (2007a), and Soto and Medina (2005).

7. See Bodenstein and others (2011).

8. See Bernanke and others (1997), Leduc and Sill (2004), and Bodenstein and others (2013).

9. For empirical evidence that oil shocks can be expansionary at the ZLB, see Miyamoto and others (2023).

10. Instead, this differential performance of Germany relative to the U.S. is consistent with Phillip Lane's concerns about depressed aggregate demand, together with his observation that "the energy-related terms of trade sharply differentiates the current euro area and U.S. situations, since the U.S. is broadly balanced in its energy trade due to its large-scale domestic production of energy." (Inflation Diagnostics, cited above.)

This paper studies the macroeconomic effects of energy price shocks in a heterogeneous-agent New Keynesian model of a small open economy that imports energy, by adding an energy good to the model of Auclert and others (2021a). We show that, when MPCs are realistically large and the elasticity of substitution between energy and domestic goods is realistically low, this model does feature a direct link between high energy prices and aggregate demand: increases in energy prices depress real incomes and cause a recession, even if the central bank does not tighten monetary policy. We use our model as a laboratory to study potential monetary and fiscal policy responses to an energy shock, including their distributional effects.

To isolate the direct channel from energy price increases to aggregate demand, we begin by studying the case where monetary policy keeps the real interest rate constant in the face of energy shocks. We show analytically that the effect on aggregate GDP depends on a race between two effects: first, a substitution effect (when foreign energy is more expensive, consumers consume more domestically produced goods), which raises GDP and is governed by a certain elasticity of substitution χ, and second, a real-income effect (with real incomes depressed, consumers consume less of all goods), which lowers GDP and is governed by MPCs. Under a realistic calibration of substitution elasticities and MPCs, the second effect dominates, and energy price shocks cause a domestic contraction. This result contrasts with the predictions of a complete-market representative-agent model à la Blanchard and Galí (2007a) where, under this monetary policy, the substitution effect is the only effect, and the shock unambiguously causes an expansion; and also with the predictions of a representative-agent incomplete-market (RA-IM) model à la Bodenstein and others (2011), where the shock causes an expansion that is not offset by a real-income effect unless the shock is very persistent.

We then turn to the effect of the oil shock on price and wage inflation. Motivated by recent concerns about wage-price spirals in advanced economies, we ask whether the energy price shock can cause such a spiral, with nominal wages rising to catch up to nominal prices.[11] Under a standard parameterization of the wage Phillips curve, we find that, in fact, the answer is no: while the decline in purchasing power does lead households to desire higher wages, the recession caused by the shock makes them ask for lower wages, and the second force always dominates. However, we find that, when combining

11. See Blanchard (1986), Lorenzoni and Werning (2023b,a).

nominal rigidities with real-wage rigidities as in Blanchard and Galí (2007b), a spiral can occur: both wages and prices can rise after the energy price shock. Even in this case, the rise in nominal wages does not mitigate the real-wage decline caused by the shock: instead, the rise in nominal prices always outpaces the rise in nominal wages.

Next, we study alternative monetary policy responses to the shock. The natural reaction of an inflation-targeting central bank to an inflationary shock is to raise interest rates to limit inflation, even if that means a weakening of economic activity. Our model suggests an important caveat of such a policy: a shock that is caused by rising energy prices at the world level is hard to counteract with contractionary monetary policy by an individual energy importer, as the effect on world energy prices is bound to be limited. The only remaining way to affect domestic energy prices is via an exchange rate appreciation, but the effects of monetary policy on exchange rates are likely too weak to materially affect inflation.[12]

Tightening domestic monetary policy does tame domestic energy demand. This suggests that monetary policy has positive externalities on other countries. Indeed, we find that when all energy importers in our model coordinate and tighten monetary policy together, there is a material reduction in world energy prices and domestic energy inflation. In other words, in the wake of an energy price shock, monetary policy among energy-importing countries suffers from a free-rider problem: each central bank may find it individually optimal to keep a loose stance, while all central banks hiking together could materially limit world energy inflation.

We then turn to fiscal policy. We study three types of fiscal measures: energy price subsidies; untargeted lump-sum transfers; and targeted lump-sum transfers, proportional to households' exposure to the energy shock. All policies are deficit-financed and ultimately repaid by raising income taxes. As with monetary policy, we first study these policies when used by an individual energy-importing country in isolation, and then we consider externalities across countries.

We show that, when used by an individual country, fiscal policy can curtail the negative GDP effects of the energy shock. This is easiest to do by using energy subsidies. When households are insulated from higher energy prices, there is no real-wage loss and no associated

12. A back-of-the-envelope calculation, using the uncovered interest-rate parity condition, shows that monetary tightening of 1pp. for one year only causes the nominal exchange rate to appreciate by one percent.

reduction in aggregate demand. Instead, by moving the shock from private balance sheets to its own balance sheet, the government is able to smooth out the impact of the shock over time. Transfers are also able to mitigate the effects of the shock, albeit somewhat less effectively. They mostly support consumer spending and hence aggregate demand. Inflation is higher when transfers are being used, as wage inflation increases with higher aggregate demand. All three kinds of fiscal policy reduce consumption inequality—a measure of welfare inequality—in response to the shock.

In contrast to these domestic benefits, we find that fiscal policy imposes strongly negative externalities on other countries. This is most salient for energy price subsidies. Since these subsidies limit incentives to substitute away from energy, world energy prices increase in response. The policy of any individual country only causes a small increase in world prices, but when all energy importers employ price subsidies, world energy demand becomes almost price inelastic, requiring a sharp rise in prices to clear the world energy market. This makes subsidies largely self-defeating: they are unable to effectively insulate countries from the shock and cause such a burden on government balance sheets that even a smoothed tax plan significantly deepens the recession. Transfers also cause negative externalities on other energy importers, albeit to a lesser extent.

In summary, our paper suggests that any individual country's monetary tightening is costly and of limited use in fighting inflation after an energy price shock; but that it comes with positive externalities on other energy importers. Inversely, fiscal policy can be very powerful in cushioning the effects of energy price shocks but tends to have negative externalities on other countries. In light of these results, a promising combination of monetary and fiscal policy could be one that focuses on aggressive, coordinated monetary tightening, combined with fiscal relief targeted to the poor—crucially avoiding energy price subsidies.

Our paper is one of the first to analyze an import price shock in an open-economy New Keynesian macro model with household heterogeneity. As such, it relates to an emerging literature that brings household heterogeneity à la Bewley (1977)-Aiyagari (1994) into small open-economy New Keynesian models à la Galí and Monacelli (2005), which has focused on different kinds of shocks.[13] The paper builds in

13. See the early work of de Ferra and others (2020), as well as Guo and others (2023), Oskolkov (2023), Zhou (2022), and Aggarwal and others (2023), among others.

particular on Auclert and others (2021a), who studied exchange rate shocks. Import price shocks are different: for instance, as in the earlier paper, we derive an equivalence between representative-agent (RA) and heterogeneous-agent (HA) economies, but here this equivalence occurs for a parameterization with unitary elasticities and is therefore more closely related to Cole and Obstfeld (1991)'s seminal paper.

Several papers study supply shocks, e.g., to energy, in closed-economy New Keynesian models with household heterogeneity. Guerrieri and others (2022) emphasize how incomplete markets among households can lead to negative demand spillovers from adverse supply shocks. Känzig (2023) studies the macroeconomic effects of carbon pricing in a closed-economy setup with tractable heterogeneity à la Bilbiie (2021) and Bilbiie and others (2022). Pieroni (2023) analyzes the effects of an energy shock in a full-blown closed-economy heterogeneous-agent New Keynesian model à la Kaplan and others (2018) and Auclert and others (2023).[14] Absent monetary tightening, aggregate demand for labor is a lot more likely to increase in a closed-economy setting, even with heterogeneity, since higher energy prices increase real incomes in such a setting.

An established literature exists around the propagation of oil price shocks in open-economy representative-agent models. A vexing question in this literature has been why oil price shocks empirically have such large negative effects on GDP.[15] Rotemberg and Woodford (1996) argued that this is caused by endogenously increasing markups. Bernanke and others (1997) argued that it is mostly contemporaneous monetary tightening. Blanchard and Galí (2007a) substantiate this point by using a model with real-wage rigidities. In the model, the real interest rate required to stabilize nominal-wage inflation rises sharply in response to an oil shock, inducing a strong recession when inflation is stabilized. Bodenstein and others (2011) present a two-country representative-agent model with incomplete markets. They do find wealth effects on consumer spending to matter, under the assumption of nearly permanent shocks. However, even with monetary tightening, hours increase in their baseline simulation in response to a negative oil shock.[16] Our paper shows that, once one allows for

14. Kuhn and others (2021) analyzes an energy shock in a similar model, but with flexible prices.

15. See Hamilton (1983), Barsky and Kilian (2004), Kilian (2009), Baumeister and Hamilton (2019), and Känzig (2021) for empirical evidence on the macroeconomic effects of oil price shocks.

16. See their figure 8.

household heterogeneity, even temporary energy shocks can lead to significant contractions in real GDP.

Our results on policy spillovers are reminiscent of the literature on currency wars and competitive easing.[17] This literature points out that monetary easing hurts other countries at the zero lower bound, stimulating the domestic economy at the expense of others. Our results emphasize that there is a related spillover via the world energy market since monetary easing boosts world energy demand, which hurts other energy importers. In Fornaro and Romei (2022), monetary policy does not internalize its impact on the world supply of tradable goods. Fiscal policy externalities have also previously been analyzed in Gourinchas and others (2021), Aggarwal and others (2023), and Devereux and others (2023), though not with regard to energy-related policies or spillovers via energy prices.

Finally, the recent surge in energy prices has led to many papers studying their implications for current policy. Lorenzoni and Werning (2023b), Blanchard and Bernanke (2023), and Gagliardone and Gertler (2023) find that energy prices can explain recent inflation developments. Kharroubi and Smets (2023) study their implications for the natural rate of interest when energy demand is non-homothetic. Closest to us, Chan, Diz, and Kanngiesser (2022), and Langot and others (2023) study the effects on aggregate demand in an open-economy heterogeneous-agent New Keynesian setting. Chan and others (2022) restrict heterogeneity by studying a two-agent model and are able to derive implications for optimal policy. Langot and others (2023) conduct a policy analysis for France, backing out the shocks that rationalize the data and then using the model for policy counterfactuals.

1. Model

Our model builds on the open-economy heterogeneous-agent New Keynesian model in Auclert and others (2021a), extended to study energy shocks.[18] This extension allows for an energy good, a small continuum of energy importers, and a real-wage stabilization motive. We focus on the effects of energy price shocks on the demand side of the

17. See Caballero and others (2021).

18. The Auclert and others (2021a) model itself is a combination of the canonical Galí and Monacelli (2005) model with the closed-economy heterogeneous-agent framework in Auclert and others (2023).

economy, initially leaving the supply side intact. We argue in section 2.4 that energy entering the supply side causes very similar behavior.

1.1 Model Setup

Time is discrete and the horizon is infinite. We consider a nested small open-economy environment. The world consists of a mass-one two-dimensional continuum of countries, e.g. $[0,1]^2$, of which a one-dimensional subset of length 1, e.g. $\{0\} \times [0,1]$, labels all energy-importing countries. We make the simplifying assumptions that these countries are the sole purchasers and consumers of energy in the world and that energy is supplied entirely by the rest of the world.

We first focus on one representative energy-importing country, 'home', and then turn to the set of energy-importing countries as a whole to explore coordinated policy responses. We denote variables corresponding to the entire world economy with a star superscript.

We consider perfect-foresight impulse responses to shocks starting from a steady state without aggregate uncertainty ("MIT shocks"). We use the sequence-space Jacobian method from Auclert and others (2021b) and linearize with respect to these shocks. By certainty equivalence, these impulse responses are the same as those from the model with aggregate risk.

There are three goods in the economy. The 'home' good, H, is domestically produced and can be exported. The 'energy' good, E, and 'foreign' good, F, are produced abroad and imported.

Domestic households. The economy is populated by a unit mass of households. Each household is subject to idiosyncratic income risk, driven by productivity shocks e_{it}, which follow a first-order Markov chain with mean $\mathbb{E}e_{it} = 1$. Households can invest their assets in a domestic mutual fund, but cannot insure their idiosyncratic risk. A household with asset position a and productivity level e at time t optimally chooses its consumption c and saving a' by solving the dynamic programming problem

$$\begin{aligned} V_t(a,e) = \max_{c,a'} & U(c,N_t) + \beta \mathbb{E}_t\left[V_{t+1}(a',e')\right] \\ \text{s.t.} \quad & c + a' = (1+r_t)a + eZ_t\,. \\ & a' \geq \underline{a} \end{aligned} \tag{1}$$

Here r_t denotes the ex-post mutual-fund return in units of the consumer price index (CPI) P_t; W_t is the nominal wage; N_t denotes

labor supplied by households, determined by union demand as specified below; Z_t is aggregate labor income,

$$Z_t \equiv \frac{W_t}{P_t} N_t, \tag{2}$$

and $\underline{a} \leq 0$ parametrizes the borrowing constraint agents face. The utility function, which is common across households, is separable and takes the form

$$U(c, N_t) = u(c) - v(N_t),$$

where

$$u(c) = \frac{c^{1-\sigma}}{1-\sigma'}, \; v(N) = v_\varphi \frac{N^{1+\varphi}}{1+\varphi}.$$

The parameter $\sigma > 0$ is the inverse elasticity of intertemporal substitution, and $\varphi > 0$ is the inverse Frisch elasticity of labor supply. $\nu_\varphi > 0$ is a normalization constant.

The household's consumer basket, c, is formed by a constant-elasticity-of-substitution (CES) combination of energy consumption c_E and non-energy consumption c_{HF}, where the non-energy bundle results from a CES combination of home consumption c_H, and foreign consumption c_F,

$$c = \left[\alpha_E^{1/\eta_E} c_E^{(\eta_E - 1)/\eta_E} + (1-\alpha_E)^{1/\eta_E} \, c_{HF}^{(\eta_E - 1)/\eta_E} \right]^{\eta_E/(\eta_E - 1)} \tag{3}$$

$$c_{HF} = \left[\alpha_F^{\frac{1}{\eta}} c_F^{\frac{\eta-1}{\eta}} + (1-\alpha_F)^{\frac{1}{\eta}} c_H^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}.$$

Here $\eta > 0$ is the elasticity of substitution between home and foreign goods, and $\eta_E > 0$ is the elasticity of substitution between energy and non-energy goods. The CPI for these preferences is

$$P = \left[\alpha_E P_E^{1-\eta_E} + (1-\alpha_E) P_{HF}^{1-\eta_E} \right]^{\frac{1}{1-\eta_E}} \tag{4}$$

$$P_{HF} = \left[\alpha_F P_F^{1-\eta} + (1-\alpha_F) P_H^{1-\eta} \right]^{\frac{1}{1-\eta}}.$$

Here, P_{Et} and P_{Ft} are the nominal price of energy and foreign goods, respectively, in domestic currency units, and P_{Ht} is the price of domestic goods.

Households differ in their level of spending but have the same consumer basket and price index. Defining $\alpha \equiv \alpha_E + (1-\alpha_E)\alpha_F$, by standard two-step budgeting arguments, a household in state (a,e), with consumption $c_t(a,e)$, splits its purchases between energy, foreign, and home goods according to

$$c_{Et}(a,e) = \alpha_E \left(\frac{P_E}{P}\right)^{-\eta_E} c_t(a,e), \tag{5}$$

$$c_{Ft}(a,e) = (1-\alpha_E)\alpha_F \left(\frac{P_F}{P_{HF}}\right)^{-\eta} \left(\frac{P_{HF}}{P}\right)^{-\eta_E} c_t(a,e), \tag{6}$$

$$c_{Ht}(a,e) = (1-\alpha)\left(\frac{P_H}{P_{HF}}\right)^{-\eta} \left(\frac{P_{HF}}{P}\right)^{-\eta_E} c_t(a,e). \tag{7}$$

Foreign households. Foreign households in other energy-importing countries face the same problem as domestic households. Households in the rest of the world, which fully account for the demand for home exports, face an almost identical problem, except that they do not consume energy. These households consume an exogenous and constant quantity C^* of worldwide goods, and spread their own consumption of foreign goods across all foreign countries, with an elasticity of substitution across countries of $\gamma > 0$. Denoting by P^*_{Ht} the foreign-currency price of domestically produced goods, export demand for home goods is given by

$$C^*_{Ht} = \alpha^* \left(\frac{P^*_{Ht}}{P^*}\right)^{-\gamma} C^*. \tag{8}$$

We assume that the law of one price holds for home goods, so that P^*_{Ht} is equal to the cost $P_{Ht}/\mathcal{E}_t$ of a domestic good in foreign-currency units:

$$P^*_{Ht} = \frac{P_{Ht}}{\mathcal{E}_t}, \tag{9}$$

where $\mathcal{E}_t$ is the nominal exchange rate. With this convention, an increase in $\mathcal{E}_t$ indicates a nominal depreciation.

Monetary policy abroad keeps the price of foreign goods in foreign currency constant, $P^*_{Ht} = P^*_t = 1$. The world nominal interest rate, i^*, is constant.

Production of home goods. We allow for energy to be used as an input in production, though our main results concern the version of the model in which labor is the only input.[19] Output is produced from domestic intermediates and imported energy. The intermediate inputs to be used in home goods production are produced by a continuum of monopolistically competitive firms each using the technology

$$Y_t = A_N N_t, \tag{10}$$

where N_t is labor, and A_N is the constant level of TFP. Let ϵ denote the elasticity of substitution between intermediates. We assume that prices are fully flexible so that the price of labor for production is set at a constant markup μ over nominal marginal costs,

$$P^I_t = \mu \frac{W_t}{A_N}$$

where $\mu = \epsilon/(\epsilon - 1)$. Total real dividends generated by domestic firms are then equal to

$$D_t = \frac{P^I_t Y_t - W_t N_t}{P_t}. \tag{11}$$

Firms have a unit mass of shares outstanding, with end-of-period price j_t.

Home goods are produced competitively from domestic intermediates and energy with the constant returns to scale production function,

$$\bar{Y}_t = \left[(1 - \xi_E)^{\frac{1}{v}} Y_t^{\frac{v-1}{v}} + \xi_E^{\frac{1}{v}} E_t^{\frac{v-1}{v}} \right]^{\frac{v}{v-1}}, \tag{12}$$

19. This is mostly for simplicity. See section 2.4 for an argument that an economy with energy in the production function behaves very similar to one with energy in consumption.

where E_t is energy used in production (the ξ_E= 0 case corresponds to the case without energy in production). The price is then set equal to the marginal cost

$$P_{Ht} = \left[(1-\xi_E) \left(\mu \frac{W_t}{A_N} \right)^{1-v} + \xi_E P_{Et}^{1-v} \right]^{\frac{1}{1-v}}. \tag{13}$$

Real GDP is always equal to Y_t in this economy.

Energy suppliers. Energy is supplied to the energy-importing countries by a measure one of price-taking firms, which are owned by foreign agents. These energy suppliers each have a claim to a source of energy that by default costlessly generates $\overline{E}_t$ in each period t. A firm i can pull supply forward by a single period by extracting additional energy today, at some cost, leaving less energy to be costlessly extracted tomorrow. Similarly it can delay extraction, facing a symmetric cost. Call the 'inventory', $I_{i\,t}^E$, of energy the cumulative shortfall of extraction relative to the default path $\{\overline{E}_t\}$. So

$$I_{i,t+1}^E = I_{i,t}^E + \left(\overline{E}_t - E_{it} \right).$$

Then the amount of energy that can be costlessly extracted by firm i at t is then $I_{i,t}^E + \overline{E}_t$. The value of an energy supplier is the present discounted value of their dividends

$$\sum_{j=0}^{\infty} \left(\frac{1}{1+r^*} \right)^j \left[P_{E,t+j}^* E_{i,t+j} - \mathcal{C}\left(E_{i,t+j} - \overline{E}_{t+j} - I_{i,t+j}^E \right) \right],$$

where the adjustment cost paid is

$$\mathcal{C}\left(E_{i,t} - \overline{E}_t - I_{i,t}^E \right) = \frac{\Gamma}{2} \left(E_{i,t} - \overline{E}_t - I_{i,t}^E \right)^2.$$

Then the energy 'inventory' carried over from period t to $t+1$ is

$$I_{i,t+1}^E = \frac{\left(\frac{1}{1+r^*} \right) P_{E,t+1}^* - P_{E,t}^*}{\Gamma}.$$

Financial sector. We assume frictionless capital flows across countries. At home, an unconstrained, risk-neutral mutual-fund

issues claims to households, with aggregate real value A_t at the end of period t. The mutual fund may invest in nominal bonds and firms, both at home and abroad. Its objective is to maximize the (expected) real rate of return on its liabilities r_{t+1}. In equilibrium, this implies that expected returns on all these assets are equal.

Equating returns from the nominal bonds, we get the standard uncovered interest parity (UIP) condition,

$$1 + i_t = \left(1 + i_t^*\right) \frac{\mathcal{E}_{t+1}}{\mathcal{E}_t}. \tag{14}$$

Define the ex-ante real interest rate as

$$1 + r_t^{\text{ante}} \equiv \left(1 + i_t\right) \frac{P_t}{P_{t+1}} \tag{15}$$

and define the real exchange rate as

$$Q_t \equiv \frac{\mathcal{E}_t}{P_t}. \tag{16}$$

We can combine (14), (15), and (16) to obtain a real version of the UIP condition

$$1 + r_t^{\text{ante}} = \left(1 + i^*\right) \frac{Q_{t+1}}{Q_t}. \tag{17}$$

Since the ex-ante returns are equated, the initial mutual-fund portfolio is indeterminate, and the ex-post return for all dates $t \geq 1$ is independent of the portfolio, $r_{t+1} = r_t^{\text{ante}}$. To determine r_0, we assume that coming into date 0, the mutual fund holds the entire stock of the home goods firms. So we can write

$$1 + r_{t+1} = \frac{j_{t+1} + D_{t+1}}{j_t},$$

where the end-of-period share price of domestic firms is the present discounted value of dividends,

$$j_t = \frac{D_{t+1} + j_{t+1}}{1 + r_t^{\text{ante}}}. \tag{18}$$

We define the net foreign-asset position to be the difference between the value of assets accumulated domestically, A_t, and the total value of assets in net supply domestically, i.e.,

$$\text{nfa}_t \equiv A_t - j_t. \tag{19}$$

Unions. We assume a formulation for sticky wages with heterogeneous households, similar to Auclert and others (2023). A union employs all households for an equal number of hours N_t and is in charge of setting nominal wages by maximizing the welfare of the average household. Relative to the Phillips curve in Auclert and others (2023), we assume here that the union puts an extra weight on stabilizing real wages relative to the steady-state real wage, incorporating the ideas of Blanchard and Galí (2007b). We show in appendix A.1 that this problem leads to the wage Phillips curve

$$\pi_{wt} = \kappa_w \left(\frac{v'(N_t)/u'(C_t)}{\frac{1}{\mu_w}(W_t/P_t)^{1+\zeta_{BG}}} - 1 \right) + \beta \pi_{wt+1}, \tag{20}$$

where π_{wt} denotes nominal-wage inflation,

$$\pi_{wt} \equiv \frac{W_t}{W_{t-1}} - 1.$$

Here, $\zeta_{BG} \geq 0$ is the parameter characterizing the extent of the real-wage stabilization motive. When $\zeta_{BG} = 0$, the wage Phillips curve has the standard form,[20] with wage inflation rising when the marginal rate of substitution (numerator) exceeds the marked-down after-tax real wage, now or in the future.[21] If we derive this equation from a Calvo specification where the probability of keeping the wage fixed is θ_w, then $\kappa_w = \frac{(1-\beta\theta_w)(1-\theta_w)}{\theta_w}$. When $\zeta_{BG} > 0$, unions are averse to departures of real wages from their steady-state value.

20. See Erceg and others (2000).

21. In Auclert and others (2023)'s formulation of the union problem, the consumption level that enters the Phillips curve in (20) is equal to a consumption aggregator $\overline{C_t} \equiv (u')^{-1}\left(\mathbb{E}\left[e_{it}u'(c_{it})\right]\right)$ that takes into account inequality in labor earnings. Here we opt for the simpler formulation in (20), because it helps streamline some of our analytical results.

Monetary policy. The monetary authority sets the nominal interest rate according to a monetary rule. For the analytical results that we develop in the paper, our baseline is a specification in which monetary policy holds the real interest rate constant,

$$i_t = r_{ss} + \pi_{t+1} + \epsilon_t. \tag{21}$$

This is a CPI-based Taylor rule with a coefficient of 1 on expected inflation. This monetary rule achieves a middle ground between standard CPI-based Taylor rules with responsiveness larger than 1 , and zero-lower-bound specifications with a fixed nominal interest rate, and is widely used in the literature as a device to partial out the effects of monetary policy in the study of the effects of shocks to aggregate demand.[22] In the context of energy price shocks, rule (21) can be thought of as a 'neutral' monetary policy stance, in which monetary policy hikes nominal interest rates just enough to keep up with inflation. We consider alternative monetary rules in section 3.

Equilibrium. We are now ready to define two different notions of equilibrium. We define an (uncoordinated) small open-economy (SOE) equilibrium as follows.

Definition. Given sequences of foreign energy price shocks $\{P_{Et}^*\}$ and monetary shocks $\{\epsilon_t\}$, an initial wealth distribution $\mathcal{D}_0(a,e)$, and an initial portfolio allocation for the mutual fund, a SOE equilibrium is a path of policies $\{c_{Ht}(a,e), c_{Ft}(a,e), c_{Et}(a,e), c_t(a,e), a_{t+1}(a,e)\}$ for households, distributions $\mathcal{D}_t(a,e)$, prices $\{\mathcal{E}_t, Q_t, P_t, P_{Ht}, P_{Ft}, P_{Et}, W_t, p_t, i_t, r_t, r_t^p\}$, and aggregate quantities $\{C_t, C_{Ht}, C_{Ft}, C_{Et}, Y_t, \overline{Y}_t, A_t, D_t, \text{nfa}_t,\}$, such that all agents optimize, firms optimize, and the domestic goods market clears:

$$C_{Ht} + C_{Ht}^* = \overline{Y}_t, \tag{22}$$

where $C_{Ht} \equiv \sum_e \pi_e \int c_{Ht}(a,e) \mathcal{D}_t(a,e)$ denotes aggregate consumption of home goods, and C_t, C_{Ft}, C_{Et}, A_t, are defined similarly. We focus on equilibria in which the long-run exchange rate returns to its steady-state level, $Q_\infty = Q_{ss}$.

We also consider (coordinated) world equilibria, in which total energy demand must be met by total energy supply.

Definition. A coordinated equilibrium is an uncoordinated equilibrium in which the path of world energy prices $\{P_{Et}^*\}$ is chosen such that energy demand C_{Et} equals energy supply in each period t.

22. See Woodford (2011), McKay and others (2016), Auclert and others (2023).

Further equilibrium objects. In equilibrium, the current account identity holds:

$$\text{nfa}_t = NX_t + \left(1 + r_{t-1}^{\text{ante}}\right)\text{nfa}_{t-1} + \left(r_t - r_{t-1}^{\text{ante}}\right)A_{t-1} - \left(r_t^H - r_{t-1}^{\text{ante}}\right)j_{t-1}, \tag{23}$$

where $NX_t \equiv \mathcal{E}_t \frac{P_{Ht}^*}{P_t} C_{Ht}^* - \mathcal{E}_t \frac{P_{Ft}^*}{P_t} C_{Ft} - \mathcal{E}_t \frac{P_{Et}^*}{P_t} C_{Et}$ is the value of net exports in units of the CPI. The last two terms capture a balance of valuation effects. r_t^H is the ex-post return on the home-good-producing firms. These valuation terms are zero for all $t \geq 1$.

We consider a steady state with no inflation and no initial gross positions across borders. That is, the domestic mutual fund owns all stocks issued by home-good-producing firms and the net foreign-asset position is zero.[23] We normalize foreign demand such that $\alpha^* = \alpha + \frac{\xi_E}{1 - \zeta_E}$. Then, we can normalize prices to 1 in this steady state, implying that $P_{\text{Hss}}, P_{Fss}, P_{Ess}, P_{ss}, P_{Hss}^*, \mathcal{E}_{ss}, Q_{ss}$ are all equal to 1. Moreover, we normalize domestic GDP Y_{ss} as well as consumption C_{ss} and C^* to 1, implying output $\overline{Y}_{ss} = \frac{1}{1 - \xi_E}$.

Following the same arguments as in Auclert and others (2021a) the unique $Q_\infty = 1$ steady state, to which the economy returns after transitory shocks, also has no net foreign-asset position and $C_\infty = Y_\infty = 1$. Hence, our heterogeneous-agent model is stationary without the need for a debt-elastic interest rate, as in Schmitt-Grohé and Uribe (2003) or the large literature that followed.

Complete-market representative-agent model ("RA model"). We also consider the canonical representative-agent model of Galí and Monacelli (2005), in which there are complete markets across households and across countries. Following the same arguments as in Auclert and others (2021a), in that model, the consumption behavior of the representative domestic household is described by the Backus-Smith condition

$$Q_t C_t^{-\sigma} = C_{ss}^{-\sigma}. \tag{24}$$

Calibration. We calibrate the model at a quarterly frequency. Table 1 summarizes our calibration parameters, which are aimed

23. Note that the steady-state value of the importing firms is zero.

at capturing a large European energy-importing country. We follow the calibration in Auclert and others (2021a). We assume discount factor heterogeneity in order to match aggregate wealth. We consider permanent heterogeneity, with a three-point distribution at $\left\{\beta - \frac{\Delta}{2}, \beta, \beta + \frac{\Delta}{2}\right\}$ and a third of agents in each. We set β to achieve an annualized real interest rate of $r = 4.0\%$ in steady state. We set the initial steady-state net foreign-asset position to 0, with all mutual-fund assets invested in domestic stocks. We consider standard values of $\sigma^{-1} = 1$ for the elasticity of intertemporal substitution, and $\varphi^{-1} = 0.5$ for the Frisch elasticity of labor supply.

We target an import-to-GDP ratio of 30 percent.[24] So we set α_F to achieve $\alpha = 0.3$. We set the energy share, α_E, at four percent of GDP.[25] As in Bachmann and others (2022), we consider a low elasticity of substitution between energy and non-energy goods equal to 0.1. We set the elasticity of substitution between home and foreign goods, η, equal to that between varieties of foreign goods, γ. We set these such that χ, defined in (30), equals 0.3. We do not explicitly model delayed substitution, but we focus our analysis on the short run and so choose low elasticities in line with Boehm and others (2023). We set the real-wage stabilization parameter to $\zeta_{BG} = 5$.[26] We set θ_w so that peak nominal-wage inflation matches the EA-19 peak of 3.9 percent.

24. In 2021, imports to GDP across the five largest European energy-importing countries were as follows: U.K. 28%, Italy 30%, France 32%, Spain 33%, Germany 42%. Overall, our economies are slightly less open than in Galí and Monacelli (2005), where $\alpha = 0.4$.

25. We take data on complete energy balances from Eurostat and consider the EU-27 in 2021. We measure energy consumption by gross available energy (GAE), which combines production, net imports, and rundown of stocks. We use the TTF price for natural gas, the Brent crude-oil price for oil and petroleum products, and IHS Northwest European coal prices for solid fossil fuels. Together, GAE for these three fuels makes up 2.9% of EU-27 GDP. In common energy units, they account for 69% of total GAE and over 95% of energy imports. A simple extrapolation to the remaining energy sources would yield $\alpha_E \approx 2.94\%/0.69 = 4.3\%$.

Also in common energy units, 41% of GAE is domestically produced. In value weighted terms, the 2021 figure is likely lower since oil and gas (both largely imported) prices were already rising.

We price the remaining fuels—the largest two being nuclear and renewables—at the (unweighted) mean of the three known prices. This gives an energy share of 4.1% of which 35% is domestically produced. In most of section 2, we will assume this is entirely imported, as this simplifies the analytic results. However, we additionally consider the case where some energy is produced domestically, and this is the case we use in our quantitative model.

26. If we eliminate the nominal-wage rigidity in our model, our assumption of $\zeta_{BG} = 5$ lies squarely between the two values in Blanchard and Galí (2007b), 1.5 and 9. We show this in appendix A.2.

Auclert and others (2021a) argue that the implied θ_F estimated for Italy and the U.K. are 0.94 and 1.00, respectively, although lower in other cases. We set $\theta_F = 0.9$. Finally, we set $\theta_E = 0.65$, making the passthrough on impact around 40 percent.

For the energy shock itself, we let P_E^* follow an AR(1), with persistence giving a half-life of 16 quarters and with an initial impact of 100 percent.

1.2 Intertemporal MPCs

An important part of our analysis is to analyze household spending behavior in energy-importing countries. To do so, we summarize aggregate consumption behavior in terms of a function $\mathcal{C}_t$ that maps sequences of ex-ante real interest rates $\left\{r_s^{\text{ante}}\right\}$ and real aggregate income $\{P_{Hs}/P_s \cdot Y_s\}$ into the sequence of aggregate consumption $\{C_t\}$. We describe this function for the case where energy only appears in consumption, $\zeta_E = 0$. The map works in two steps:

Table 1. Model Calibration

Parameter	*Benchmark model*	*Parameter*	*Benchmark model*
σ	1	r	0.01
φ	2	β	0.95
η_E	0.1	s.s. nfa	0
η	0.51	ζ_{BG}	5
γ	0.51	θ_w	0.938
α_E	0.04	θ_E	0.65
α_F	0.27	θ_F	0.9
μ	1.03	ρ_e	0.96

Source: Authors' calculations.

First, it maps ex-ante interest rates and real income into ex-post returns $\{r_s\}$. For all $s > 0$, this map is simply given by $r_s = r_{s-1}^{\text{ante}}$. For $s = 0$, r_0 picks up a valuation effect, and is determined by

$$1 + r_0 = \frac{D_0 + j_0}{j_{sS}},$$

with $D_t = \left(1 - \frac{1}{\mu}\right)\frac{P_{Ht}}{P_t} Y_t$ and j_t given by (18).

Second, it maps ex-post returns $\{r_s\}$ and real income $\{P_{Hs}/P_s \cdot Y_s\}$ into consumption. This works because the only two endogenous aggregates in (1) are ex-post returns and aggregate labor income $Z_t = \frac{1}{\mu}\frac{P_{Ht}}{P_t} Y_t$. Once the paths of these two aggregates are determined, all consumption and saving policies $c_t(a,e)$, $a_t(a,e)$ and the evolution of the distribution $\Psi_t(a,e)$ (assuming the initial distribution is at the steady state) can be solved for, so aggregate consumption can be written as

$$C_t = \int c_t(a,e)\, d\Psi_t(a,e) = \mathcal{C}_t\left(\left\{r_s^{\text{ante}}, \frac{P_{Hs}}{P_s} \cdot Y_s\right\}_{s=0}^{\infty}\right).$$

Finally, since we initially focus on an economy in which ex-ante real interest rates are kept constant, we will write consumption simply as a function of aggregate real income,

$$C_t = \mathcal{C}_t\left(\left\{P_{Hs} / P_s \cdot Y_s\right\}_{s=0}^{\infty}\right). \tag{25}$$

Intuitively, $\mathcal{C}_t$ captures spending behavior in response to arbitrary paths of aggregate real income. Aggregate real income here affects spending in two ways. First, it reprices outstanding assets, as dividends are a given fraction of aggregate real income; and the associated capital gains lead to a spending response of households. Second, it increases aggregate labor income, which again results in a spending response.

As in previous work, e.g., Auclert and others (2023), we linearize (25) around the steady state and express changes in spending over time, stacked as the vector $d\mathbf{C} \equiv \left(dC_0, dC_1, \ldots\right)$, as a function of changes in real income $d\left(\frac{\mathbf{P}_H}{\mathbf{P}}\mathbf{Y}\right) \equiv \left(d\left(\frac{P_{H0}}{P_0} Y_0\right), d\left(\frac{P_{H1}}{P_1} Y_1\right), \ldots\right)$,

$$d\mathbf{C} = \mathbf{M} \cdot d\left(\frac{\mathbf{P}_H}{\mathbf{P}} \mathbf{Y} \right). \tag{26}$$

Here, $\mathbf{M}$ is the sequence-space Jacobian of $\mathcal{C}_t$ defined as the collection of partial derivatives

$$M_{ts} \equiv \frac{\partial \mathcal{C}_t}{\partial \left(P_{Hs} / P_s \cdot Y_s \right)}$$

around the steady state. We call the entries of $\mathbf{M}$ intertemporal marginal propensities to consume (iMPCs). iMPCs are a richer set of moments than standard MPCs, in that they capture both the entire dynamic response of consumption to unanticipated (aggregate) income changes—the entries in the first column ($M_{\cdot,0}$) of $\mathbf{M}$—as well as the entire dynamic response of consumption to anticipated income changes—the entries in column s,($M_{\cdot,s}$), for an anticipated income change at date $s > 0$.

2. Energy Price Shocks and Heterogeneity

We begin by studying the response of one individual energy importer to a (first-order) shock to the world price of energy P^*_{Et}, denoted by dP^*_{Et}. We assume that the shock is AR(1), that is,

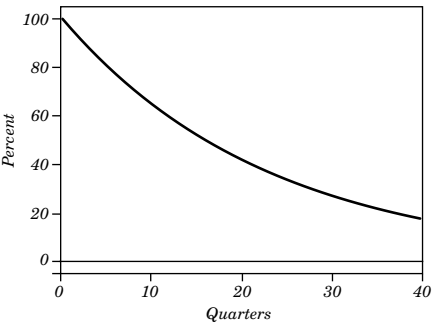
$$dP^*_{Et} = dP^*_{E0} \cdot \rho_e^t,$$

where $\rho_e \in (0,1)$ is the persistence of the shock. We choose a baseline persistence of $\rho_e = 0.96$ and normalize the shock such that $P^*_{E0} = 1$. The shock path is shown in figure 2. As described above, we assume that, for now, the ex-ante real interest rate is kept constant by monetary policy. We study alternative monetary policy rules in section 3 below. Up until section 2.4 below, we do not consider energy usage in production and keep $\zeta_E = 0$.

Our analysis is centered around the home goods market clearing condition (22). After substituting in the demands (7)-(8) and the price-setting condition for PCP (9), we can write this condition as

$$Y_t = (1-\alpha)\left(\frac{P_{Ht}}{P_{HFt}} \right)^{-\eta} \left(\frac{P_{HFt}}{P_t} \right)^{-\eta_E} C_t + \alpha \left(\frac{P_{Ht}}{\mathcal{E}_t} \right)^{-\gamma} C^*. \tag{27}$$

Figure 2. The Energy Price Shock



Source: Authors' calculations.
Note: AR (1) shock to P^*_{Et} with persistence 0.96. This represents a doubling of energy prices on impact, with a half-life of four years.

Aggregate demand for home goods, the right-hand side of (27), is influenced by the shock either due to changing relative prices $\frac{P_{Ht}}{P_{HFt}}, \frac{P_{HFt}}{P_t}, \frac{P_{Ht}}{\mathcal{E}_t}$, or due to changing domestic spending C_t. We next explore how a representative-agent model behaves in response to the shock; then we will compare that to a heterogeneous-agent model.

2.1 Representative Agent

In the complete-market representative-agent model, aggregate consumption remains constant, $C_t = C_{ss}$. This is easiest to see by combining the Backus-Smith condition (24) with the real UIP condition (17). Since ex-ante real interest rates are kept constant, the real exchange rate is constant as well, $Q_t = Q_{ss}$, and so is consumption. With this, we can characterize equilibrium output and consumption as follows.

Proposition 1. *In the complete-market representative-agent model with real interest rate rule (21), the linearized deviations from steady-state consumption over output, $dC_t = (C_t - C_{ss})/Y_{ss}$ and output $dY_t = (Y_t - Y_{ss})/Y_{ss}$ in response to shocks to the world energy price $dP^*_{Et} = (P^*_{Et} - P^*_{E,ss})/P^*_{E,ss}$ are given by*

$$dC_t = 0 \tag{28}$$

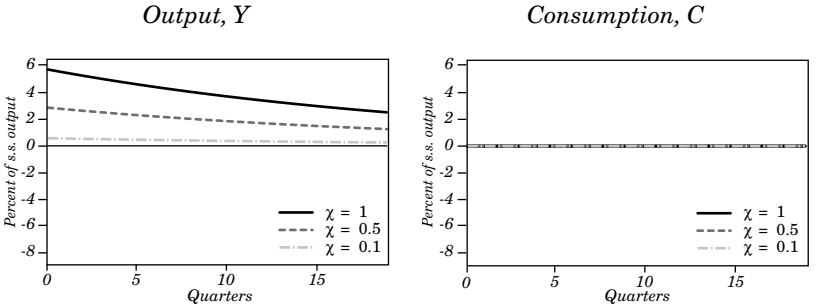
$$dY_t = \frac{\alpha_E}{1-\alpha} \cdot \chi \cdot dP^*_{Et}, \tag{29}$$

where χ *is a weighted average elasticity of substitution,*

$$\chi \equiv (1-\alpha)\left(\alpha_F \eta + (1-\alpha_F)\eta_E\right) + \alpha\gamma. \tag{30}$$

Proposition 1 shows that the output response in the RA economy is proportional to the energy price shock. Its scale is determined by two factors: the share of energy in consumption, α_E, relative to home consumption,$1 - \alpha$, and an appropriately weighted average of the elasticities of substitution in the economy, χ. Crucially, the output response (29) is always positive in response to a positive energy price shock. This can be explained by consumers substituting away from imported energy towards domestically produced goods, thus causing a boom in economic activity in the domestic economy. In fact, as consumer spending remains constant, the entire output response is driven by expenditure switching. We plot impulse responses in figure 3 for various substitution elasticities χ.

Figure 3. Output and Consumption Responses to an Energy Price Shock in the RA Model



Source: Authors' calculations.
Note: Impulse responses in the representative-agent model to the energy price shock P^*_{Et} displayed in figure 2. χ is the average substitution elasticity between energy and domestically produced goods. It is defined in (30).

Proposition 1 should not be interpreted as saying that there can never be a bust after an energy price shock in RA models, though. Instead, when there is a bust,[27] it has to be because of monetary tightening in response to the shock, in the sense of a rising real interest rate, rather than the shock itself. In terms of the textbook three-equation New Keynesian model,[28] proposition 1 implies that a suitable interpretation of an energy shock in an RA model is one of a cost-push shock, paired with a positive aggregate-demand shock.

Going forward, it will be convenient to express impulse responses as vectors, just like in (26). With this notation, (28)–(29) become $d\mathbf{C} = 0$ and $d\mathbf{Y} = \frac{\alpha_E}{1-\alpha} \cdot \chi \cdot d\mathbf{P}_E^*$.

2.2 Heterogeneous Agents

In light of our discussion in section 1.2, one way to explain the RA result is to point out that, with complete markets across countries, an RA model essentially behaves like a model with zero iMPCs, $\mathbf{M}^{RA} = 0$. In other words, the complete-market RA model features no real-income effect on consumption.[29] This is the key difference from our heterogeneous-agent economy, where we find the following result for output and consumption.

Proposition 2. *With a real interest rate rule and a matrix of intertemporal MPCs* $\boldsymbol{M}$*, the impulse responses of consumption and output following an energy price shock are given by*

$$d\mathbf{C} = \underbrace{-\frac{\alpha_E}{1-\alpha}\mathbf{M} \cdot d\mathbf{P}_E^*}_{\text{Real income-channel}} + \underbrace{\mathbf{M} \cdot d\mathbf{Y}}_{\text{Multiplier}} \tag{31}$$

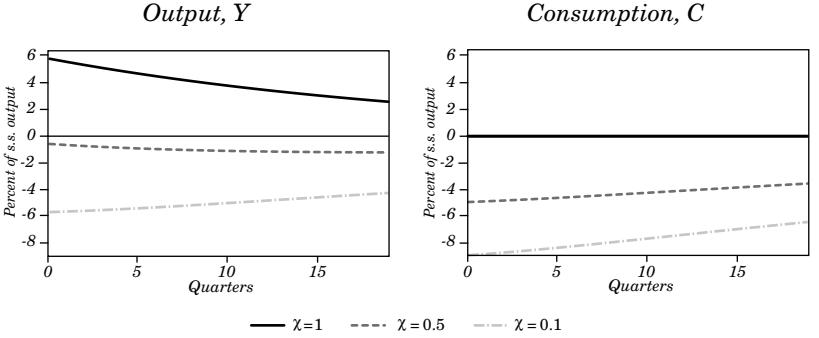
$$d\mathbf{Y} = \underbrace{\frac{\alpha_E}{1-\alpha}\chi d\mathbf{P}_E^*}_{\text{Exp. switching channel}} - \underbrace{\alpha_E \mathbf{M} \cdot d\mathbf{P}_E^*}_{\text{Real income-channel}} + \underbrace{(1-\alpha)\mathbf{M} \cdot d\mathbf{Y}}_{\text{Multiplier}}. \tag{32}$$

27. See Bodenstein and others (2011).

28. See Galí (2008).

29. We analyze a RA-IM model in section 2.4 and show that it implies quantitatively very small real-income effects.

Figure 4. Output and Consumption Responses to an Energy Price Shock in the HA Model



Source: Authors' calculations.
Note: Impulse responses in the representative-agent model to the energy price shock P_{Et}^* displayed in figure 2. χ is the average substitution elasticity between energy and domestically produced goods. It is defined in (30).

Proposition 2 shows that the impulse responses of consumption and output now also depend on the matrix of intertemporal MPCs $\mathbf{M}$. Equation (31) finds that there are two ways in which real income $\frac{P_{Ht}}{P_t} Y_t$, and hence consumption $d\mathbf{C}$, are affected by an energy shock $d\mathbf{P}_E^*$. First, increased energy prices increase the CPI P_t relative to the price of home goods P_{Ht}. This reduces real income all else equal, leading agents to cut consumption by $\mathbf{M} \times \frac{\alpha_E}{1-\alpha} d\mathbf{P}_E^*$. We refer to this as the real-income channel of energy price shocks. Second, the energy price shock will, indirectly, also affect the path of output $d\mathbf{Y}$, which also enters real income and changes consumption by $\mathbf{M} \times d\mathbf{Y}$. This is a standard (Keynesian) multiplier effect.

Linearizing goods market clearing (27) and substituting in (31), we obtain equation (32), whose form is like that of a standard Keynesian cross, where the relevant multiplier is the product of MPCs $\mathbf{M}$ by the degree of home bias $(1-\alpha)$. Including expenditure switching, there are altogether three distinct channels that jointly determine the output response to any given shock. The next proposition derives the general solution to (32).

Proposition 3. *Assuming* $\boldsymbol{M} \geq 0$*, the equilibrium output response is unique and given by*

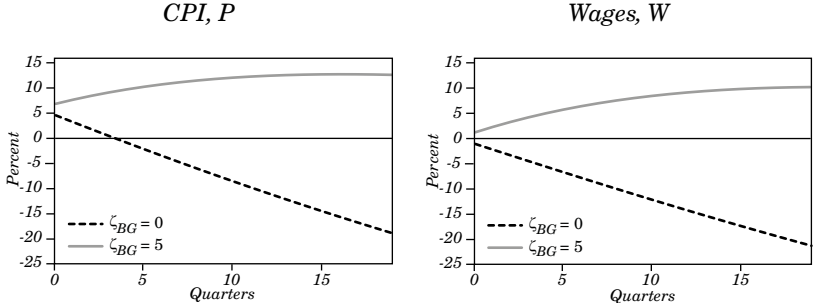
$$d\mathbf{Y} = \frac{\alpha_E}{1-\alpha} \chi d\mathbf{P}_E^* + \alpha_E (\chi - 1) \mathbf{M} \cdot (\mathbf{I} - (1-\alpha)\mathbf{M})^{-1} d\mathbf{P}_E^*. \tag{33}$$

In particular, if $\chi = 1$*, all aggregate quantities and prices are the same as in the RA model, including* $d\boldsymbol{Y} = d\boldsymbol{Y}^{RA}$*. Moreover, provided that* $\boldsymbol{M} > 0$*, for an energy shock* $d\mathbf{P}_E^* \geq 0$*, we have*

$$d\mathbf{Y} \lesseqgtr d\mathbf{Y}^{RA} \text{ and } d\mathbf{C} \lesseqgtr 0 \Leftrightarrow \chi \lesseqgtr 1.$$

Proposition 3 solves the Keynesian cross fixed point in (32) for $d\boldsymbol{Y}$. Similar to Auclert and others (2021a), it establishes a formal neutrality result for $\chi = 1$, showing that the RA and HA models have identical implications for aggregate quantities and prices.[30] When the substitution elasticity lies below one ($\chi < 1$), however, the output response in the HA model is more muted relative to the RA model. The intuition for this result is that when $\chi = 1$, the real-income and multiplier channels in (32) exactly offset each other, and $d\boldsymbol{Y}$ is entirely driven by expenditure switching, as in the RA model. Reducing χ below 1 leads to a smaller expenditure switching channel, and hence also a smaller multiplier effect, making the HA output response fall below RA.

Figure 5. Wage-price Spiral with Real-Wage Stabilization Motive



Source: Authors' calculations.
Note: Impulse responses in the heterogeneous-agent model to the energy price shock P_{Et}^* displayed in figure 2. ζ_{BG} is the weight on the Blanchard and Galí (2007b) real-wage stabilization motive.

30. One important difference from Auclert and others (2021a), however, is that in (30), $\chi = 1$ is implied by all primitive elasticities being unity, as in Cole and Obstfeld (1991), whereas in Auclert and others (2021a), $\chi = 1$ requires primitive elasticities below unity.

We illustrate proposition 3 in figure 4, plotting the output and consumption responses to the energy shock for various choices of χ. While the responses are identical to those for the RA model (figure 3) when $\chi = 1$, output turns negative for modest substitution elasticities around $\chi \approx 0.5$. With realistic energy substitution elasticities of around $\chi = 0.1$, the shock causes a sizable contraction.

2.3 Wage-Price Spirals

Our result in proposition 3 characterizes the quantity response to the energy shock. What about prices and wages?

A useful starting point is the real wage $w_t \equiv W_t / P_t$. Given flexible prices, we can write

$$d \log w_t = d\left(\frac{P_{Ht}}{P_t} \right) = -\frac{\alpha_E}{1-\alpha} dP^*_{Et}. \tag{34}$$

The real wage is directly determined by the shock, independent of the nominal-wage Phillips curve. Given the responses of the real wage, output (or, equivalently, hours), and consumption, the nominal-wage Phillips curve (20) then pins down the behavior of nominal wages and, by (34), the behavior of the price level. This separation, which allows us to first solve the "real economy" including real wages, before solving for nominal objects, is a useful consequence of the combination of a real interest rate monetary policy rule, sticky nominal wages, and flexible prices.[31]

Figure 5 plots prices and wages as implied by the nominal-wage Phillips curve (20) without the real-wage stabilization motive (dashed line) and with the real-wage stabilization (solid line). Without the real-wage stabilization motive, an initial jump up in the price level is actually followed by a sustained decline in prices, even below their original level. This is because wages start declining as households' consumption and hours fall with the shock, raising their willingness to work. With the real-wage stabilization motive, unions attempt to raise nominal wages to counteract declining real wages.

31. See Auclert et al. (2023), Auclert and others (2021a), Aggarwal and others (2023) for recent applications of this idea. We have found in Auclert and others (2021a) that the main results in this environment are robust to alternative monetary policy rules and sticky prices in addition to sticky wages.

Interestingly, our economy is one in which the real-wage stabilization motive is entirely self-defeating and does not succeed in pushing up real wages (34). Higher average nominal wages W_t lead to higher domestic prices P_{Ht}, a higher price index P_t, and ultimately a depreciated exchange rate $\mathcal{E}_t$. The depreciated exchange rate $\mathcal{E}_t$ leads to higher import prices, so that altogether, the entire CPI bundle becomes more expensive, in line with the increases in W_t.[32] A wage-price spiral emerges.

Going forward, we work with the model that features a wage-price spiral.

2.4 Extensions

We consider six extensions to our analysis of the baseline HA model.

Large shocks. Our analysis has assumed small, first-order shocks thus far. The energy shocks we are seeing in the world in 2022 seem anything but first order, however. Figure 6 compares a nonlinear MIT shock with a first-order one. We see that our model does not imply a hugely nonlinear impulse response.

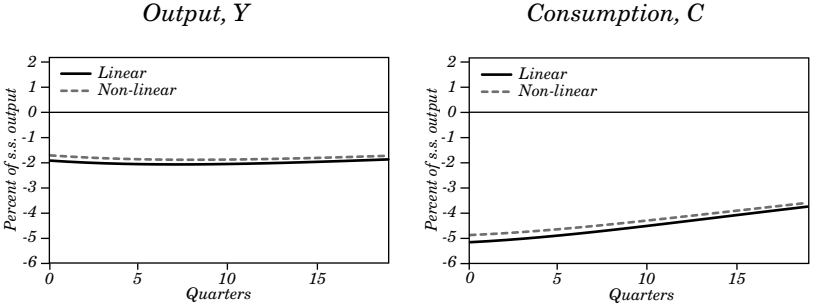
Representative-agent model with incomplete markets across countries. Our RA model benchmark assumes complete markets across countries. A natural question is what happens in a RA-IM model across countries. Figure 7 redoes figure 3 but with incomplete markets. Comparing the figures, we see that incomplete markets do not change the response by a significant amount. The main reason for this is that rather than $\mathbf{M}^{\text{RA}} = 0$, the RA-IM model has positive, but very small intertemporal MPCs.

With very persistent shocks, the effective MPC rises in the RA model with incomplete markets. However, as we show in figure 8, this model struggles to generate substantial contractionary effects without very long-lived shocks.

Two-agent model. A natural next extension is to compare our HA model with a model with simplified heterogeneity with just two types, à la Campbell and Mankiw (1989), Galí and others (2007), and Bilbiie (2008). We make such a comparison in appendix C.

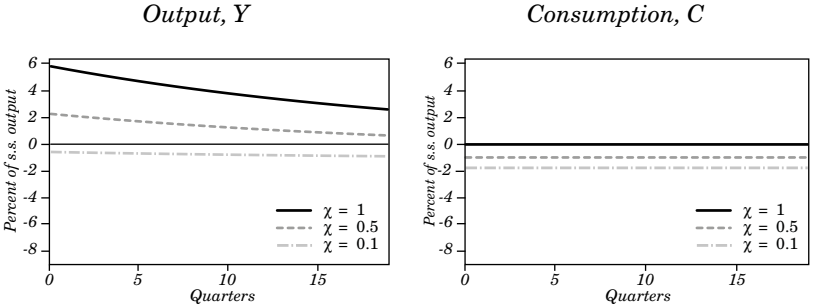
32. See appendix D.2.

Figure 6. First-Order vs. Higher-Order MIT Shocks



Source: Authors' calculations.
Note: Impulse responses in the heterogeneous-agent model to the energy price shock P_{Et}^{*} displayed in figure 2. The figure compares the first-order impulse response with the nonlinear "MIT shock" (perfect-foresight) solution.

Figure 7. Output and Consumption Responses to an Energy Price Shock in the RA Model with Incomplete Markets



Source: Authors' calculations.
Note: Impulse responses in a representative-agent model with incomplete markets to the energy price shock P_{Et}^{*} displayed in figure 2. χ is the average substitution elasticity between energy and domestically produced goods. It is defined in (30).

Energy in production. One natural question is whether the response in our RA model of GDP and consumption would look different if energy were used in production rather than consumption. The answer is no.

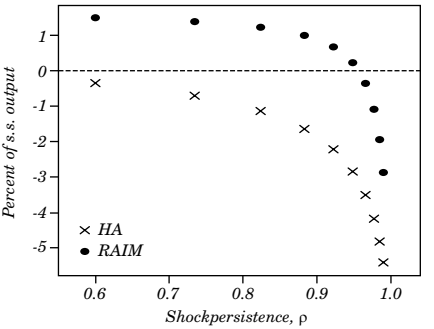
Proposition 4. *In the economy in which energy enters production but not consumption,* $\xi_E > 0$ *and* $\alpha_E = 0$*, the response of GDP is given by*

$$d\mathbf{Y} = \underbrace{\frac{\xi_E}{1-\xi_E} v d\mathbf{P}_E^*}_{\text{Exp. switching channel}} \underbrace{-(1-\alpha_F)\xi_E \mathbf{M} \cdot d\mathbf{P}_E^*}_{\text{Real-income channel}} + \underbrace{(1-\xi_E)(1-\alpha_F)\mathbf{M} \cdot d\mathbf{Y}}_{\text{Multiplier}}. \quad (35)$$

In particular, when setting ξ_E, α_F, *and* ν *in the "energy in production model" to be equal to* $(1-\alpha_E)\alpha_F$, $\frac{\alpha_E}{1-(1-\alpha_E)\alpha_F}$, *and* χ *in the "energy in consumption" model, the GDP response* $d\boldsymbol{Y}$ *to an arbitrary* $d\boldsymbol{P}_E^*$ *shock with energy in production is exactly the same as the GDP response with energy in consumption shown in proposition 3.*

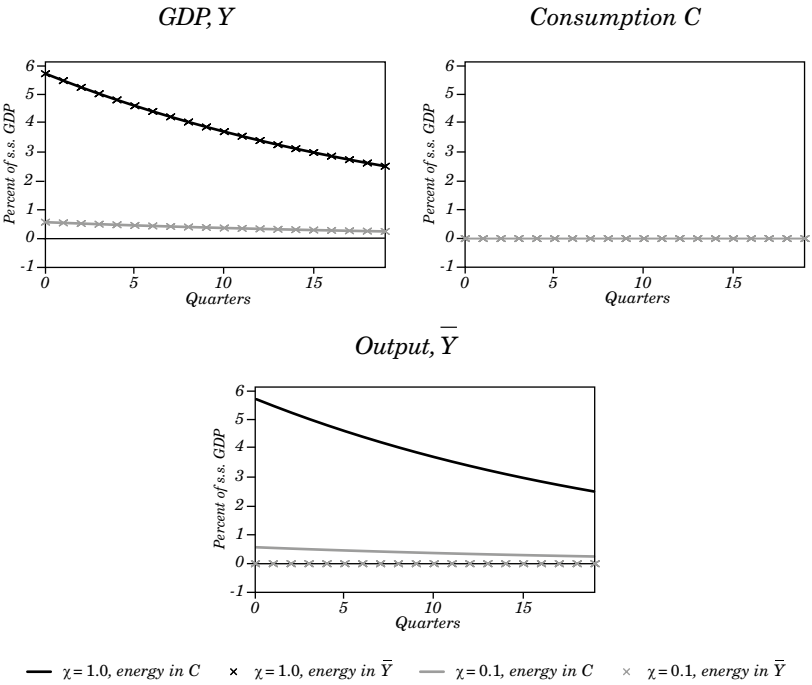
Figure 9 illustrates the proposition. Where before it was households that switched their expenditure from imported energy to domestically produced goods, it is now firms that make the same substitution. Under the condition stated in proposition 4, the response of GDP will be identical. The condition is intuitive: It simply ensures that the effective spending shares on the three goods, H, F, E, by domestic households are the same in the two models.

Figure 8. Date-0 Output Response to an Energy Price Shock in the RA-IM and HA Models



Source: Authors' calculations.
Note: Impact response of output in a representative-agent model with incomplete markets and in a heterogeneous-agent model to the energy price shock P_{Et}^* displayed in figure 2. Here we set $\chi = 0.3$ as in our baseline calibration.

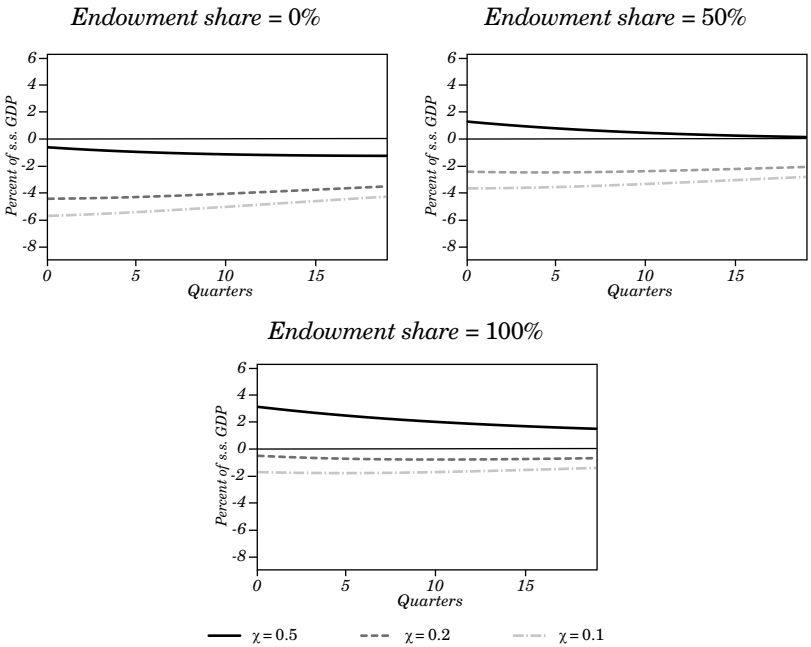
Figure 9. Energy in Consumption versus Production in the RA Model



Source: Authors' calculations.
Note: Impulse responses in a representative-agent model. "Energy in C" refers to energy directly entering the household's consumption bundle. "Energy in $\overline{Y}$'" indicates that energy is instead used in production of the home good. χ is the average substitution elasticity between energy and domestically produced goods in the "energy in C" case. It is defined in (30).

Endowment of energy. In our baseline model, energy-importing countries do not produce any energy themselves. Here we allow for energy to be produced at home. This energy is produced and sold by energy suppliers, exactly as described above. These firms are entirely owned by domestic households, and they sell energy at the global price, P^*_{Et}. In figure 10, we vary the endowment of energy between zero and the level of total energy consumption. Increasing the energy share mitigates the hit to employment and home production, $\overline{Y}$. However, even with a 100 percent energy share, if χ is low enough, we still see a decline in $\overline{Y}$ as the shock redistributes towards lower MPC agents.

Figure 10. Response of Home Production to an Energy Shock in the HA Model with Energy Endowments



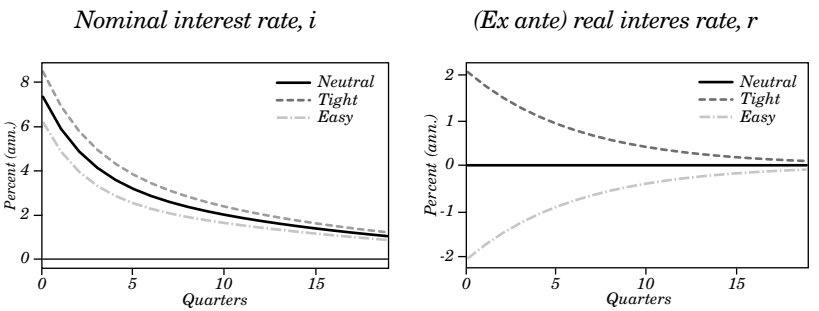
Source: Authors' calculations.
Note: Impulse responses for $\overline{Y}$—production of the home good—in the heterogeneous-agent model to the energy price shock P_{Et}^{*} displayed in figure 2. Under the baseline (endowment share = 0%), no energy is produced domestically, and all energy for consumption is imported. We also show the results when domestic energy production is equal to 50% and 100% of domestic energy consumption, respectively.

Markup shocks. In appendix D.3, we show that, under a real rate rule, modeling the energy shock as a markup shock fails to generate a decline in output. Under a Taylor rule, the markup shock generates a notably smaller recession. This suggests that an energy price shock is a more difficult problem for monetary policy than a standard cost-push shock.

3. Monetary Policy Response

Our analysis so far has concentrated on a specific monetary policy rule, namely one that achieves a stable real interest rate path. A natural question is then to what extent a more active monetary policy stance can meaningfully bring down inflation or mitigate the recession.

Figure 11. Monetary Policy Scenarios in Response to the Energy Price Shock



Source: Authors' calculations.
Note: This figure shows three scenarios for the monetary policy response to the energy price shock. The solid line represents a monetary response that keeps the real interest rate constant. The dashed line represents a monetary response that raises the on-impact real interest rate by 2 percentage points (annual), and then follows an AR(1) trajectory back to the original real rate (persistence =0.85). The dot-dashed line does the opposite.

In this section, we will compare three monetary policy responses to the shock: the neutral stance we have analyzed before, as well as an 'easy' and a 'tight' alternative response. We parameterize those alternatives as AR(1) paths for real interest rates that either start at plus or minus two percentage points (annualized). The shock as well as the induced nominal interest rate paths can be seen in figure 11.

One issue with our baseline model that can be seen in section 2.3 is that prices jump by a significant margin at date 0, which implies an unreasonably large inflation response on impact. To solve this issue, we first introduce slow passthrough of world prices into consumer prices and then study the effects of monetary policy.

3.1 The Quantitative Model

Slow passthrough. We allow for a slow passthrough of import prices of both F and E goods into consumer prices.[33] This implies that local currency prices for E and F, denoted P_{ET} and P_{FT}, are no longer simply equal to converted world prices $\mathcal{E}_t P^*_{Et}$ and $\mathcal{E}_t P^*_{Ft}$.

33. Since there is immediate passthrough of the exchange rate to export prices but slow passthrough to import prices, this is analogous to what the U.S. experiences in the "Dollar Currency Pricing" paradigm (DCP). We think of this as reasonable to model Europe, with many imports and exports goods priced in euros.

There is a continuum of monopolistically competitive firms that import the foreign good. Each importer produces their variety of the foreign imports at unit real cost $\frac{\mathcal{E}_t P^*_{Ft}}{P_t}$. The importing firms are also subject to a Calvo friction, and can only adjust their price each period with probability $1 - \theta$. The foreign imports are combined by a competitive sector by using CES aggregation. We focus on the case where these imports are highly substitutable, with the steady-state gross markup going to 1, and generating the foreign good Phillips curve

$$\pi_{F,t} = \kappa_F \left[\frac{\mathcal{E}_t P^*_{Ft}}{P_{F,t}} - 1 \right] + \frac{1}{1 + r_{t+1}} \pi_{F,t+1},$$

where $\kappa_F = \frac{(1-\theta_F)\left(1 - \frac{\theta_F}{1+r_{SS}}\right)}{\theta_F}$ and r_{SS} denote the steady-state interest rate. The foreign good importers pay out total dividends

$$D_{Ft} = \left(\frac{P_{Ft} - \mathcal{E}_t P^*_{Ft}}{P_t} \right) C_{Ft}.$$

The energy good is imported in the same manner. The equations governing energy price inflation π_{Et} and dividends of energy firms D_{Et} are the direct analog of those for π_{Ft} and D_{Ft}. A high κ_E corresponds to the case where world energy price or exchange rate changes rapidly pass through to domestic energy prices.

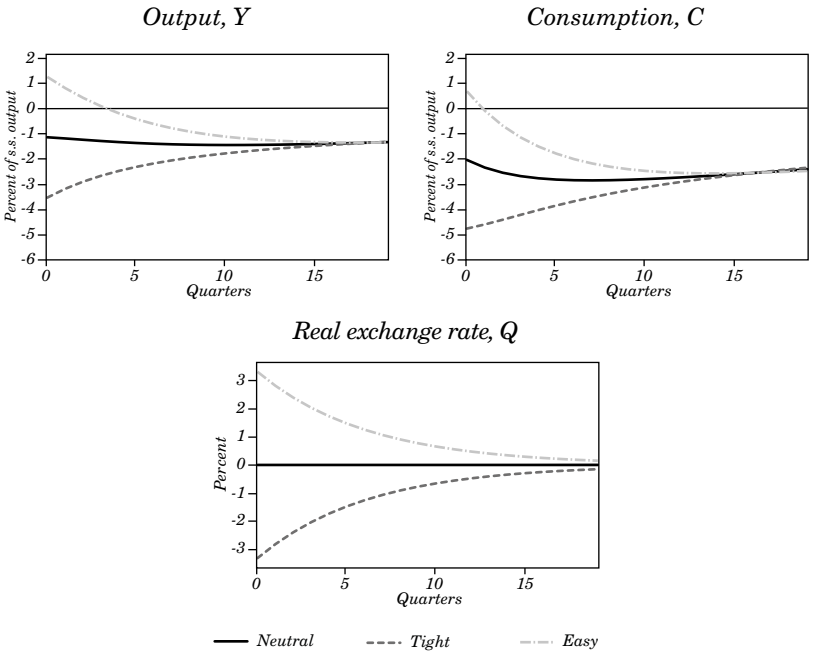
In order not to distort the steady state of the model with the introduction of a slow passthrough, we assume that importers of E and F goods are owned by foreigners. This changes our expression of net exports in section 1 to

$$NX_t \equiv \frac{\mathcal{E}_t P^*_{Ht}}{P_t} C^*_{Ht} - \mathcal{E}_t \frac{P_{Ft}}{P_t} C_{Ft} - \mathcal{E}_t \frac{P_{Et}}{P_t} C_{Et}.$$

All other equilibrium conditions are left untouched by this addition.

Domestic energy production. Another feature we include in our numerical model is an energy endowment, as discussed in section 2.4. Introducing an energy endowment makes the response to the energy price shock less contractionary and more inflationary in our model. It also emphasizes the importance of heterogeneous agents—as we allow for domestic energy production, the RA-IM is increasingly unable to generate a sizable recession in response to the shock.

Figure 12. Effect of Monetary Policy on Output and Consumption



Source: Authors' calculations.
Note: This figure shows the output and consumption responses to an energy price shock across the three monetary policy scenarios detailed in figure 11.

We retain the share of energy consumption in GDP at $\alpha_E = 0.04$, but now suppose that a third of this is domestically produced.[34]

3.2. Effects of Monetary Policy on Output and Inflation

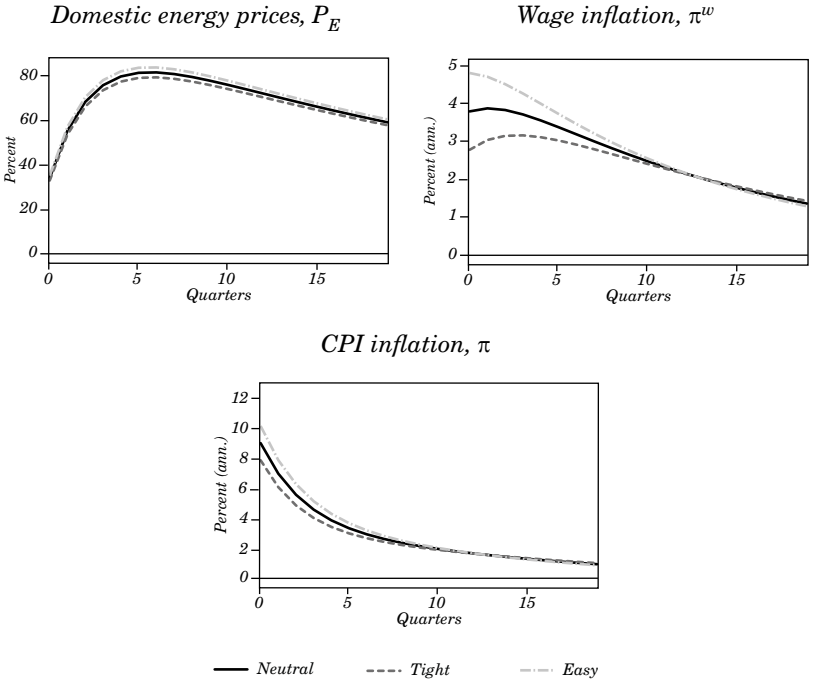
Figure 12 shows the effects of the two alternative monetary policy responses on output and consumption. As one would expect, monetary easing ameliorates the recession induced by the energy shock, while monetary tightening deepens the recession. There is a small reversal a few quarters out, as tighter monetary policy actually aids the recovery. This emerges as households see higher interest rates as an incentive to save more and improve their balance sheet position,

34. See footnote 13 for details.

thus increasing their ability to spend later. This effect also appeared in Auclert and others (2021a) and does not occur in standard closed-economy heterogeneous-agent environments.

We plot the response of inflation and domestic energy prices to the alternative monetary policy responses in figure 13. We see that wage inflation reacts significantly to changes in monetary policy, but since domestic energy prices move very little, it is very hard to reduce CPI inflation in a meaningful way given the large initial increase in inflation. This is largely coming from the fact that the shock to CPI inflation is large, and monetary policy primarily affects inflation via wage inflation, which is relatively sticky. Crucially, any small energy importer's monetary policy is unable to affect world energy prices, which implies that it cannot move the price that lies at the origin of the shock at all. We return to this point below, in section 5.

Figure 13. Effect of Monetary Policy on Inflation



Source: Authors' calculations.
Note: This figure shows the price and wage inflation responses to an energy price shock across the three monetary policy scenarios detailed in figure 11.

3.3. Effectiveness of Monetary Policy by Source of the Shock

In this section, we explore how this imported inflationary shock can be more difficult for monetary policy than a domestic inflationary shock. To do so, we ask what decline in output would be required to achieve zero inflation in the presence of downward nominal-wage rigidity. We show the results in figure 14. With the energy price shock we have considered throughout, monetary policy stabilizes the CPI by raising rates to (1) appreciate the currency, lowering P_E and P_F, and (2) contract output, lowering W and so P_H. With downward nominal-wage rigidity, the second channel is shut down, and the central bank must cause a bigger recession to sufficiently appreciate the currency. We contrast this with a "domestic shock" that generates the same path for CPI. In this case, wages pull up the CPI, and so the downward nominal-wage rigidity does not bind. As such, monetary policy is more effective in fighting domestically generated inflation.

4. Fiscal Policy Response

An important component of the actual policy response to the energy shocks in 2022 and 2023 has been fiscal support programs. We now consider the effects of three such policies. To introduce them, we first extend the model to allow for a government. We keep a slow passthrough and the energy endowment, which we introduced above in section 3.1.

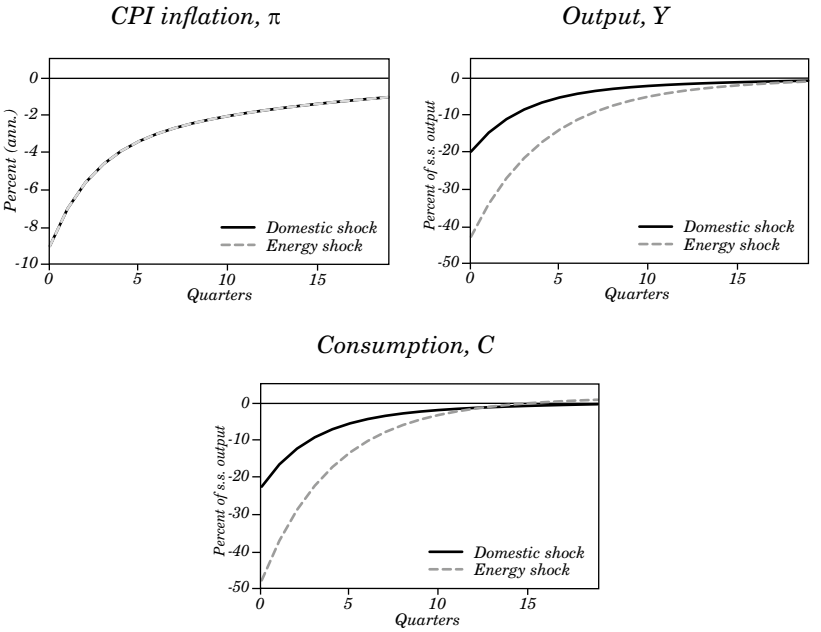
4.1 Government

The government runs three possible programs: it can subsidize energy domestically, and it can send targeted or untargeted transfers to households. It finances those programs with deficits initially, which are ultimately repaid with labor income taxes.

Energy subsidies. The government may subsidize the real energy price that households face

$$\frac{P_{Et}^{hh}}{P_t} = \left(1 - \tau^E\right)\frac{P_{Et}}{P_t} + \tau^E \frac{P_{E,ss}}{P_{ss}}.$$

Figure 14. Different Inflation-Output Tradeoffs for Foreign and Domestic Shocks



Source: Authors' calculations.
Note: This plot shows the change in inflation, output, and consumption required to offset the degree of inflation generated by the energy price shock, given two different sources of the shock, and in the presence of downward nominal-wage rigidity.

Here, P_{Et}^{hh} denotes the nominal price paid by households after the subsidy. Before the subsidy, the price is still denoted by P_{Et}. It is important to subsidize real energy prices such that permanent shifts in the price level as a result of the shock do not lead to permanent subsidies.

Targeted transfers. The government may make targeted transfers to households, indexed to their counterfactual level of energy consumption absent the shock. Under a targeted transfer, household i in idiosyncratic state (a,e) with counterfactual energy consumption $c_{i,ss}^{E} \equiv c_{E,ss}(a,e)$ receives a real transfer $T_{i,t}$ that insures a fixed proportion ins^{E} of the net increase in energy costs,

$$T_{i,t} = \text{ins}^{E} \cdot c_{i,ss}^{E} \cdot \left(\frac{P_{Et}}{P_t} - \frac{P_{E,ss}}{P_{ss}} \right).$$

Untargeted transfers. The government may also make an untargeted (real) transfer, by giving all households an equal amount, T_t^{unt}. The level of T_t^{unt} is set so that the total subsidy is the same as in the targeted case.

Labor income taxes. The proportional labor income tax rate is denoted by τ_t^L. We henceforth take Z_t to denote after-tax labor income. Replacing (2), Z_t is now given by

$$Z_t = \left(1-\tau_t^L\right)\frac{W_t}{P_t}N_t,$$

and the wage Phillips curve is now based on the after-tax wage $\left(1-\tau_t^L\right)W_t / P_t$,

$$\pi_{wt} = \kappa_w\left(\frac{v'\left(N_t\right)/u'\left(C_t\right)}{\frac{1}{\mu_w}\left[\left(1-\tau_t^L\right)W_t/P_t\right]\cdot\left(W_t/P_t\right)^{\zeta_{BG}}}-1\right)+\beta\pi_{wt+1}.$$

Government budget constraint. The government issues real bonds B_t to satisfy the government budget constraint

$$B_t = \left(1+r_{t-1}^{\text{ante}}\right)B_{t-1}+\tau^E\left(\frac{P_{Et}}{P_t}-\frac{P_{E,ss}}{P_{ss}}\right)C_{Et}+\text{ins}^E\left(\frac{P_{Et}}{P_t}-\frac{P_{E,ss}}{P_{ss}}\right)C_{E,ss}+T_t^{\text{unt}}-\tau_t^L\frac{W_t}{P_t}N_t.$$

The rate of income tax is proportional to the level of debt

$$\tau_t^L = \psi_B\left(B_{t-1}-B_{ss}\right),$$

where $\psi_B > 0$ parameterizes the speed with which debt is brought back to the steady state. The net foreign-asset position is now given by

$$\text{nfa}_t \equiv A_t - j_t - B_t$$

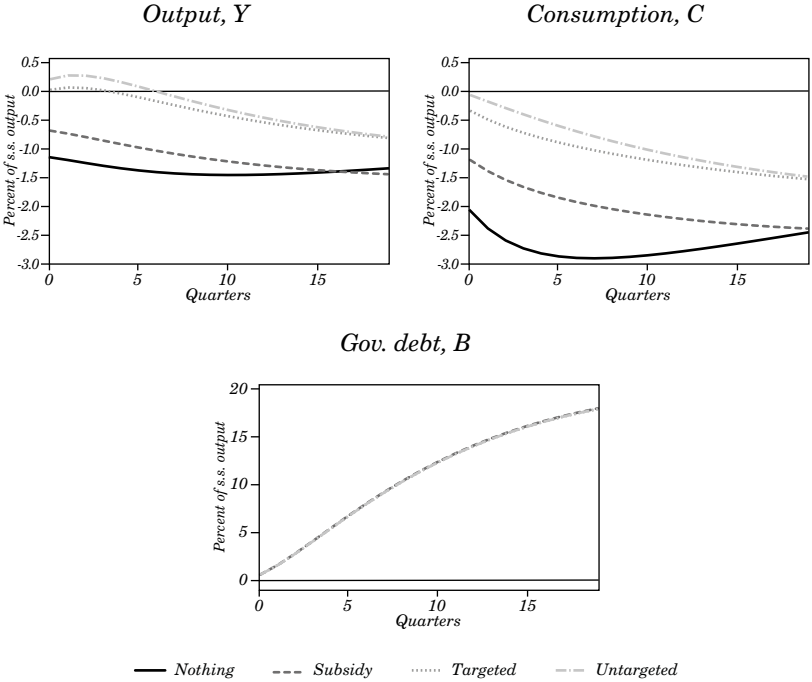
rather than (19).

Calibration. In order to keep the policies comparable, we set $\tau^E = \text{ins}^E$. We then set the untargeted transfer path to match the overall (ex-post) transfer in the targeted case. We explore the case of a 50 percent subsidy of deviations from the steady-state price, $\tau^E = 0.5$. We set $\psi_B = 0.04$. In the absence of government spending, this implies a half-life of government debt of just under six years.

4.2 Effects of Fiscal Policy on Output and Inflation

Figure 15 shows the effects of the three types of fiscal policies on output and consumption. It is clear that all three policies are able to significantly limit the real economic fallout of the energy shock. Both output and consumption are considerably higher under the policies. There is a very limited reversal 15–20 quarters out, which is due to labor income taxes being raised to bring down the additional debt that has been accumulated. We show in appendix D.5 that, if a government has less fiscal space and is therefore forced to run a balanced budget, the three policies are significantly less effective.

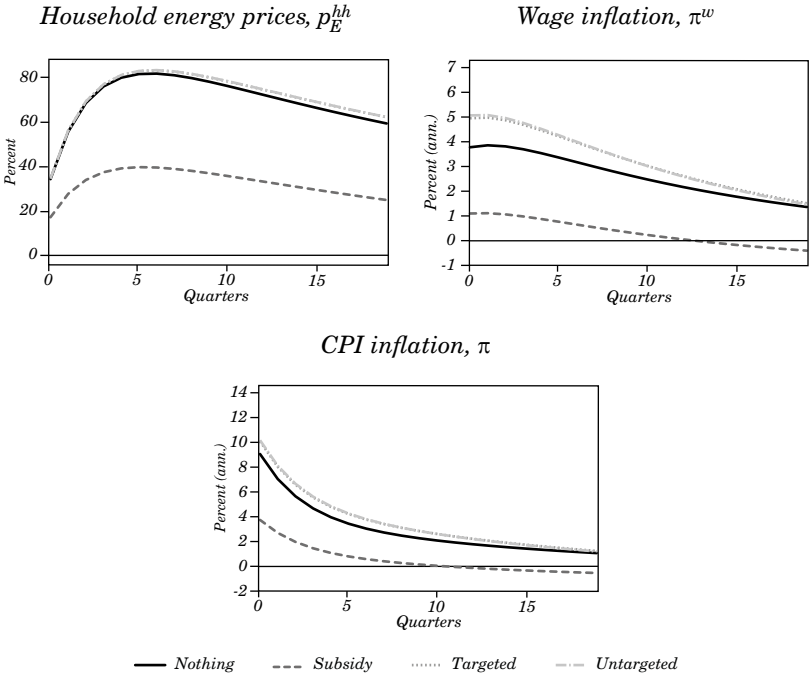
Figure 15. Effect of Fiscal Policy on Output and Consumption



Source: Authors' calculations.

Note: This figure compares the output and consumption responses to an energy price shock under no fiscal policy with the three fiscal policy programs explained in section 4.1. All policies are financed by a deficit initially and slowly paid for via increased proportional labor income taxes.

Figure 16. Effect of Fiscal Policy on Inflation

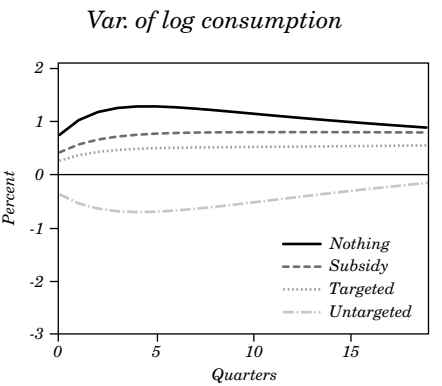


Source: Authors' calculations.
Note: This figure compares the wage and price inflation responses to an energy price shock under no fiscal policy with the three fiscal policy programs explained in section 4.1. All policies are financed by a deficit initially and slowly paid for via increased proportional labor income taxes.

Where the three types of policies differ more is in their predictions for inflation.[35] Targeted and untargeted transfers cause a significant uptick in CPI inflation, largely driven by a strong increase in wage inflation. This is to be expected, as deficit-financed transfers raise aggregated demand and stimulate the economy when MPCs are sizable.[36] Subsidies, on the other hand, are able to tame inflationary pressures in the economy to a large extent. By construction, energy prices faced by households come way down; this puts less pressure on real wages and therefore lessens the desire of unions to call for strong nominal-wage increases; and ultimately CPI inflation only mildly overshoots its target.

35. See figure 16.
36. See Farhi and Werning (2016), Auclert and others (2023).

Figure 17. Fiscal Policy and Inequality after an Energy Shock



Source: Authors' calculations.
Note: This figure compares the inequality response to an energy price shock under no fiscal policy with the three fiscal policy programs explained in section 4.1. Since we have three household types (indexed g), the variance of log consumption at date t is computed as $\mathbb{E}_G$ [Var[log (c_{it}) | $i \in g$]]-Var$_G$ [$\mathbb{E}$[log (c_{it}) | $i \in g$]].

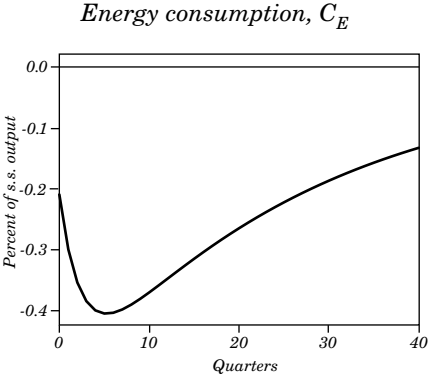
At the country level, therefore, energy subsidies appear to be a silver bullet: they tackle the shock at its root by bringing down energy prices and therefore reduce the recessionary and inflationary forces in the economy. We return to this logic below, in section 5.

Effects on inequality. Our heterogeneous-agent model enables us to also study predictions on inequality across households, as in the work of Pieroni (2023) and Kuhn and others (2021). Figure 17 shows the evolution of the variance of log differences in consumption across households, var$_{(a,e)}$ (log $c_t(a,e)$ – log $c_{ss}(a,e)$). We see that inequality rises due to the shock itself (solid line), but is significantly reduced by fiscal policy.

5. Role of Policy Coordination

So far we have limited our attention to an individual energy importer. Yet, all energy importers in our model face a similar situation and are likely to consider policy responses. In this section, we study the cross-border spillovers of fiscal and monetary policies implied by our model. To do so, we focus on a given energy importer and compare the macroeconomic effects of policies if the country is the only one engaging in the policy ('uncoordinated') to a situation in which all energy-importing countries engage in the same policy ('coordinated').

Figure 18. The Energy Supply Shock

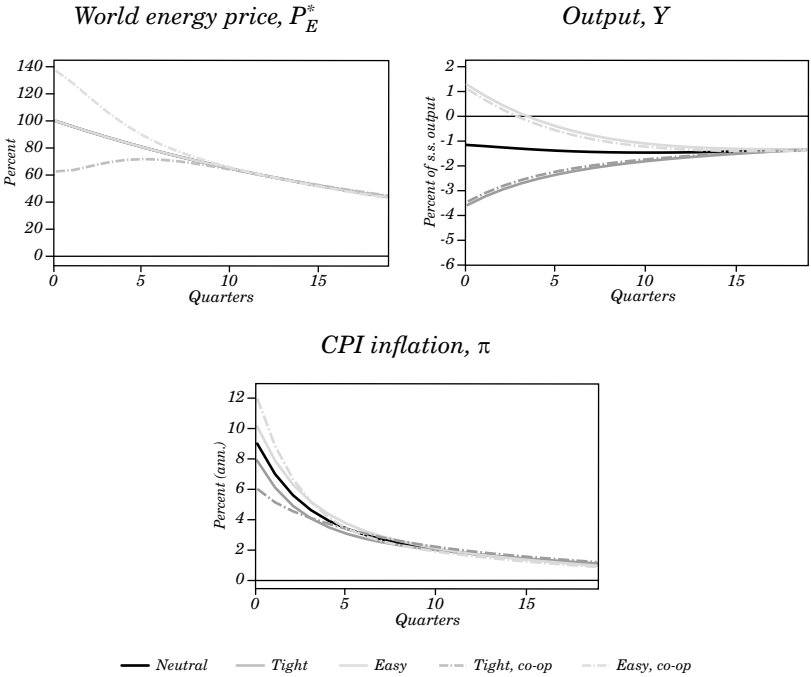


Source: Authors' calculations.
Note: Shock path is chosen such that, if all countries follow a neutral monetary policy and have no fiscal response, world energy prices P_{Et}^{*} endogenously follow the AR(1) process shown in figure 2.

We study coordinated policies by analyzing the world equilibrium, as defined in section 1, in which energy prices are endogenous. We choose the path of the energy supply shock $\overline{E}_t$ to be such that when all countries follow a neutral monetary policy with no fiscal response, energy prices endogenously follow the same AR(1) path that we analyze in the single-country equilibrium (figure 2). This makes the coordinated world equilibrium comparable to the uncoordinated single-country equilibrium. We show the energy supply shock that we arrive at in figure 18.

Coordinated monetary policy. Figure 19 compares uncoordinated with coordinated monetary policy. The key reason why coordinated monetary policy operates differently from uncoordinated policy is that coordinated policy is able to affect world energy prices. For example, coordinated tightening reduces world energy prices in the model by around 35 percentage points on impact. Even though passthrough to consumer prices is slow, the reduction in world energy prices brings down CPI inflation by more than twice as much on impact. The associated output cost of tightening is also mitigated when all energy importers hike in a coordinated fashion, as real wages now fall by less. This discussion suggests that there are positive externalities from monetary tightening across energy importers, in the sense that one central bank's tightening marginally reduces world energy prices for other countries.

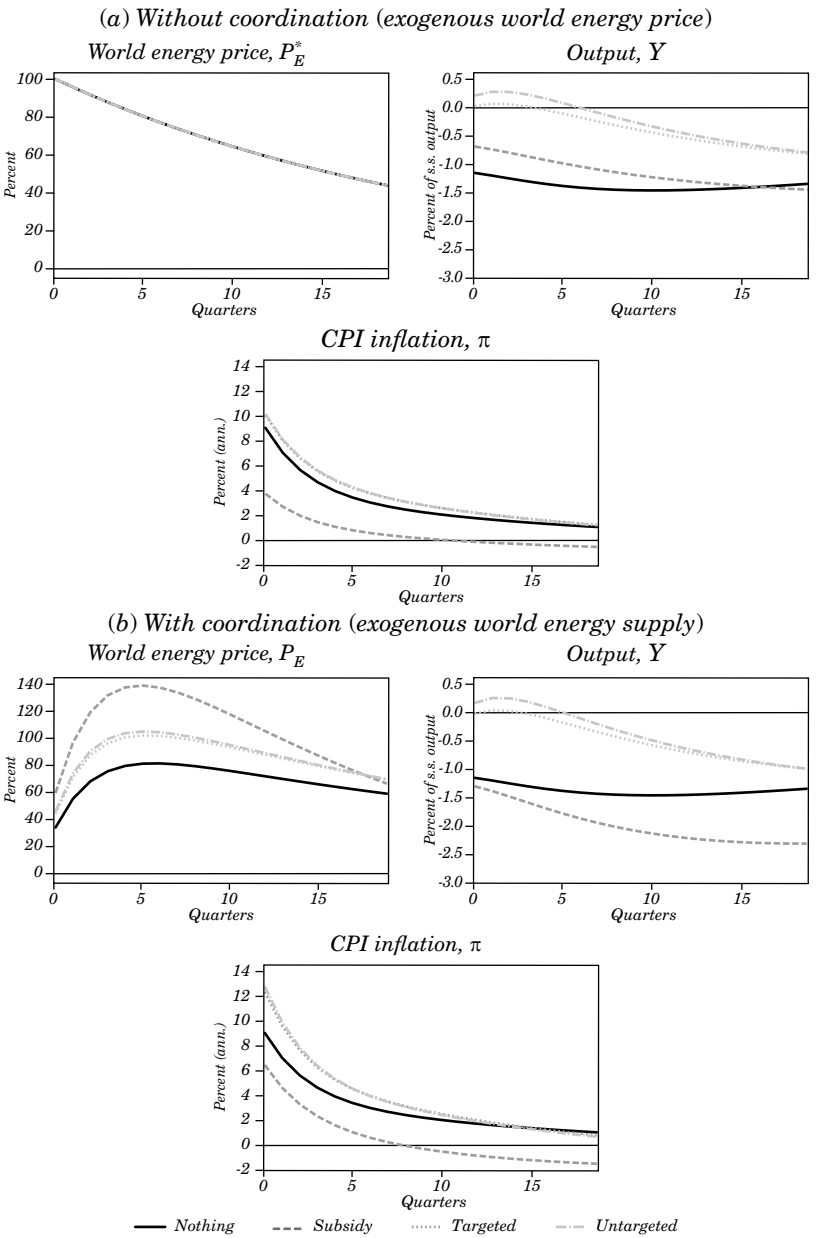
Figure 19. Coordinated vs. Uncoordinated Monetary Policy



Source: Authors' calculations.
Note: This figure compares the output and inflation responses to an energy price shock across the three monetary policy scenarios detailed in figure 11. Solid lines simulate the case when only a single economy engages in the monetary policy scenarios. Dot-dashed lines simulate the case when all economies use the same monetary policy.

Coordinated fiscal policy. Figure 20 compares uncoordinated with coordinated fiscal policy. Overall, the picture that emerges is one of negative externalities. Targeted and untargeted transfers lead to an even greater uptick in inflation in the coordinated world equilibrium. And, most importantly, energy subsidies lead to a large endogenous spike in world energy prices. This spike limits the insulating role of energy subsidies, with CPI inflation rising to similar levels as without energy subsidies. The recession actually worsens in a world with coordinated energy subsidies, as governments need significant increases in labor income taxes to stem the fiscal cost of sustaining the energy subsidies.

Figure 20. Coordinated vs. Uncoordinated Fiscal Policy



Source: Authors' calculations.

Note: This figure compares the output and inflation responses to an energy price shock across the fiscal policy scenarios detailed in section 4.1 when (a) a single economy carries out the policy and (b) all economies use the same fiscal policy.

Empirical evaluation of spillover channel. In this section, we empirically explore the effect of monetary policy shocks on the trade balance to verify our spillover channel is present in the data. We use the shocks constructed by Romer and Romer (2004) on their original sample (1969.3–1996.12). This exercise is therefore in a U.S. context, but we use it to confirm our channel is present and calibrated reasonably. To obtain impulse responses, we use a Jordà (2005) projection. We collect quarterly data on exports, imports, net exports, and output, which we interpolate to monthly frequency. We then run a Jordá projection, which for a generic outcome Y_t reads

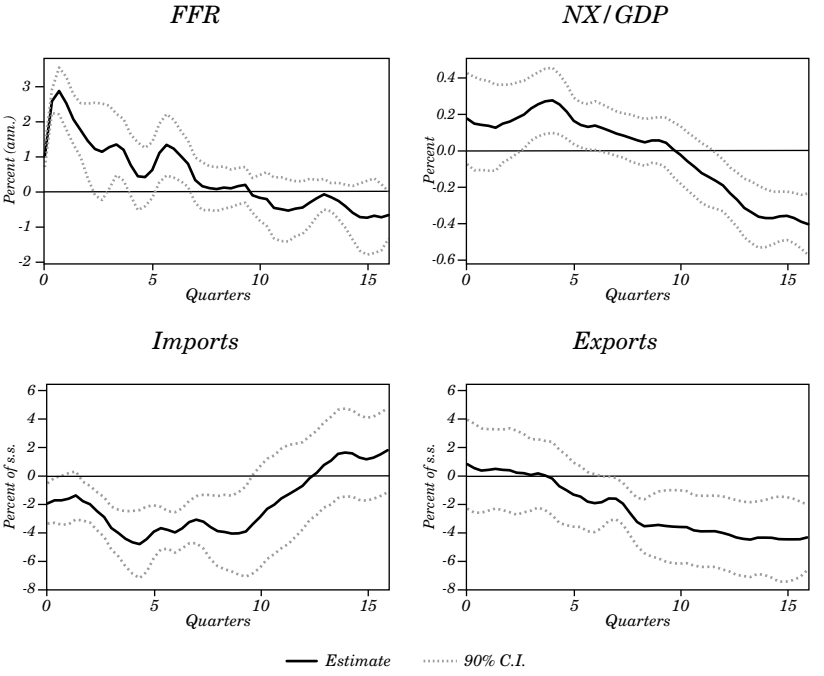
$$Y_{t+h} = J_h^Y \epsilon_t^m + \beta_h^Y X_t + \zeta_{t,h}^Y$$

separately for horizons $h = 1,\ldots, T$ up to $T = 48$ months, where ϵ_t^m is the Romer-Romer series, and $\zeta_{t,h}^Y$ is a regression error term. To control for the potential endogeneity of ϵ_t^m in practice, we include in X_t the set of controls that Ramey (2016) uses in her specification for figure 2, panel B: lags of industrial production, unemployment, the CPI, and a commodity price index. We compute the standard deviation of $\hat{J}_h^Y$ using a Newey and West (1987) correction for the autocorrelation in $\zeta_{t,h}^Y$.

The solid lines in figure 21 display the impulse responses, with the dotted lines indicating confidence intervals. We see that in response to a one percentage point increase in the federal funds rate, net exports rise by around 0.2 percent of GDP. While in the long run, we appear to get the decline suggested by the expenditure switching channel, the short run appears to be dominated by a fall in imports consistent with a decline in domestic real income and low elasticities of substitution. Our model is targeted to the short run, and indeed the average change in net exports to GDP in the first six quarters after such a shock is 0.19 in both our model and the estimated impulse-response functions (IRFs).[37]

37. To compute this, we aggregate the nominal interest rate IRF to quarterly frequency, and feed this shock into our model.

Figure 21. Trade Balance Response to a Monetary Policy Shock

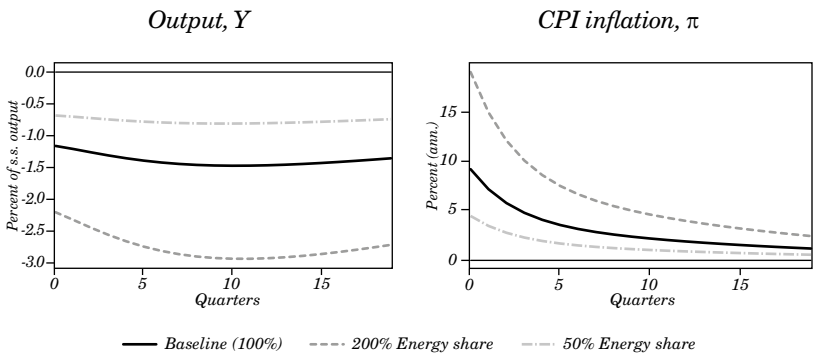


Source: Authors' calculations.
Note: This figure shows our estimated set of impulse responses to an identified Romer and Romer (2004) monetary policy shock (solid black line), with 90% confidence intervals (dotted gray lines).

6. State Dependence

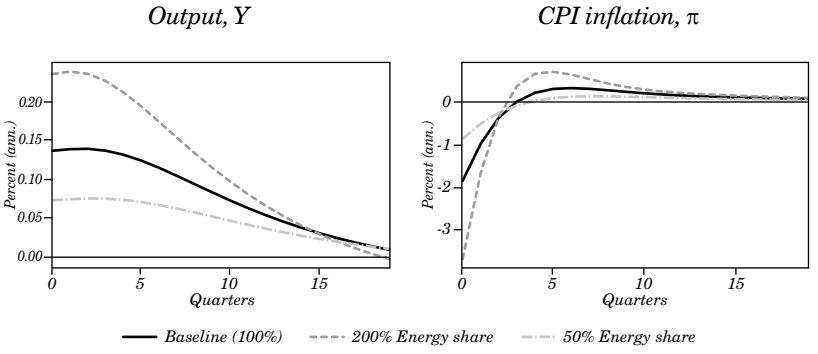
An important question is whether we should expect the mechanisms documented in this paper to always be present, or whether they depend on the presence of certain prerequisites. We now show that a crucial determinant of the presence of our mechanisms is the share of energy in an economy. To do so, we vary the share of energy in consumption between our baseline choice and double as well as half its value, i.e., $\alpha_E^{\text{high}} = 2\alpha_E$ and $\alpha_E^{\text{low}} = \frac{1}{2}\alpha_E$. We leave the rest of the calibration entirely the same, including the assumption that one third of energy is being produced by the small open economy itself.

Figure 22. Responses to an Energy Price Shock for Different Initial Energy-in-GDP Shares



Source: Authors' calculations.
Note: This figure compares the output and inflation responses to an energy price shock for different values of the energy-to-GDP ratio, α_E.

Figure 23. Responses to a Coordinated Monetary Policy Shock for Different Initial Energy-in-GDP Shares



Source: Authors' calculations.
Note: This figure compares the output and inflation responses for different values of the energy-to-GDP ratio, α_E. The shock is the world energy price path induced by all other energy-importing countries enacting the monetary policy tightening detailed in figure 11.

Figure 22 shows the responses of output and inflation to the energy shock across the three values of α_E. We clearly see that higher values of α_E leave an economy much more exposed to the energy shock. The responses are not entirely scaled versions of each other, as the average elasticity χ falls with a higher energy share, amplifying the effect of the shock.

Figure 23 highlights that the magnitude of the spillover effect of monetary policy is also state dependent and increases in the size of the energy share α_E. This suggests that, when examining the policies discussed above, the additional spillover channel of coordinated monetary tightening will play a particularly important role following a large, positive energy price shock.

7. Conclusion

We study the macroeconomic effects of energy price shocks in energy-importing economies using a heterogeneous-agent New Keynesian model. When MPCs are realistically large and the elasticity of substitution between energy and domestic goods is realistically low, there is a direct link between high energy prices and aggregate demand: increases in energy prices depress real incomes and cause a recession, even if the central bank does not tighten monetary policy. When nominal- and real-wage rigidities are both present, imported energy inflation can spill over to wage inflation through a wage-price spiral; this, however, does not mitigate the decline in real wages. Our model constitutes a useful framework to evaluate monetary and fiscal policy responses to energy price shocks.

We find that monetary tightening has a limited effect on imported inflation when done in isolation, but can be powerful when done in coordination with other energy importers by lowering world energy demand. Fiscal policy, especially energy price subsidies, can isolate individual energy importers from the shock, but it raises world energy demand and prices, imposing large negative externalities on other economies.

References

Aggarwal, R., A. Auclert, M. Rognlie, and L. Straub. 2023. "Excess Savings and Twin Deficits: The Transmission of Fiscal Stimulus in Open Economies." *NBER Macroeconomics Annual* 37: 325–412.

Aiyagari, S.R. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *Quarterly Journal of Economics* 109(3): 659–84.

Ari, A., N. Arregui, S. Black, O. Celasun, D. Iakova, A. Mineshima, V. Mylonas, I. Parry, I. Teodoru, and K. Zhunussova. 2022. "Surging Energy Prices in Europe in the Aftermath of the War: How to Support the Vulnerable and Speed up the Transition Away from Fossil Fuels." IMF Working Papers No. 152.

Auclert, A., M. Rognlie, M. Souchier, and L. Straub. 2021a. "Exchange Rates and Monetary Policy with Heterogeneous Agents: Sizing up the Real Income Channel." NBER Working Papers No. 28872.

Auclert, A., B. Bardóczy, M. Rognlie, and L. Straub. 2021 b. "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models." *Econometrica* 89(5): 2375–408.

Auclert, A., M. Rognlie, and L. Straub. 2023. "The Intertemporal Keynesian Cross." NBER Working Papers No. 25020.

Bachmann, R., D. Baqaee, C. Bayer, M. Kuhn, A. Löschel, B. Moll, A. Peichl, K. Pittel, and M. Schularick. 2022. "What If? The Economic Effects for Germany of a Stop of Energy Imports from Russia." ECONtribute Policy Brief No. 28/2022.

Baqaee, D. and E. Farhi. 2019. "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem." *Econometrica* 87(4): 1155–203.

Baqaee, D. and E. Farhi. 2022. "Networks, Barriers, and Trade." NBER Working Papers No. 26108.

Barsky, R.B. and L. Kilian. 2004. "Oil and the Macroeconomy Since the 1970s." *Journal of Economic Perspectives* 18(4): 115–34.

Baumeister, C. and J.D. Hamilton. "Structural Interpretation of Vector Autoregressions with Incomplete Identification: Revisiting the Role of Oil Supply and Demand Shocks." *American Economic Review* 109(5): 1873–910.

Bernanke, B.S., M. Gertler, and M. Watson. 1997. "Systematic Monetary Policy and the Effects of Oil Price Shocks." *Brookings Papers on Economic Activity* (1): 91–157.

Bewley, T. 1977. "The Permanent Income Hypothesis: A Theoretical Formulation." *Journal of Economic Theory* 16(2): 252–92.

Bilbiie, F.O. 2008. "Limited Asset Markets Participation, Monetary Policy and (inverted) Aggregate Demand Logic." *Journal of Economic Theory* 140(1): 162–96.

Bilbiie, F.O. 2021. "Monetary Policy and Heterogeneity: An Analytical Framework." Manuscript.

Bilbiie, F.O., D.R. Känzig, and Paolo Surico. 2022. "Capital and Income Inequality: An Aggregate-Demand Complementarity." *Journal of Monetary Economics* 126(C): 154–169.

Blanchard, O. and B.S. Bernanke. 2023. "What Caused the U.S. Pandemic-Era Inflation?" NBER Working Papers No. 31417.

Blanchard, O. and J. Galí. 2007a. "The Macroeconomic Effects of Oil Price Shocks: Why Are the 2000s so Different from the 1970s?" In *Dimensions of Monetary Policy*, edited by J. Galí and M. Gertler. Chicago, IL: International University of Chicago Press.

Blanchard, O. and J. Galí. 2007b. "Real Wage Rigidities and the New Keynesian Model." *Journal of Money, Credit and Banking* 39(s1): 35–65.

Blanchard, O.J. 1986. "The Wage Price Spiral." *Quarterly Journal of Economics* 101(3): 543–65.

Bodenstein, M., C.J. Erceg, and L. Guerrieri. 2011. "Oil Shocks and External Adjustment." *Journal of International Economics* 83(2): 168–84.

Bodenstein, M., L. Guerrieri, and C.J. Gust. 2013. "Oil Shocks and the Zero Bound on Nominal Interest Rates." *Journal of International Money and Finance* 32(1): 941–67.

Boehm, Ch.E., A.A. Levchenko, and N. Pandalai-Nayar. 2023. "The Long and Short (Run) of Trade Elasticities." *American Economic Review* 113(4): 861–905.

Caballero, R.J., E. Farhi, and P.O. Gourinchas. 2021. "Global Imbalances and Policy Wars at the Zero Lower Bound." *Review of Economic Studies* 88(6): 2570–621.

Campbell, J.Y. and N.G. Mankiw. 1989. "Consumption, Income, and Interest Rates: Reinterpreting the Time Series Evidence." NBER Macroeconomics Annual 4: 185–216.

Chan, J., S. Diz, and D. Kanngiesser. 2022. "Energy Prices and Household Heterogeneity: Monetary Policy in a Gas-TANK." Working Papers No. 4255158, Social Science Research Network.

Cole, H.L. and M. Obstfeld. 1991. "Commodity Trade and International Risk Sharing: How Much Do Financial Markets Matter?" *Journal of Monetary Economics* 28(1): 3–24.

De Ferra, S., K. Mitman, and F. Romei. 2020. "Household Heterogeneity and the Transmission of Foreign Shocks." *Journal of International Economics* 124(C): 1–18.

Devereux, M.B., K. Gente, and C. Yu. 2023. "Production Networks And International Fiscal Spillovers." *The Economic Journal* 133(653): 1871–900.

Erceg, C.J., D.W. Henderson, and A.T. Levin. 2000. "Optimal Monetary Policy with Staggered Wage and Price Contracts." *Journal of Monetary Economics* 46(2): 281–313.

Farhi, E. and I. Werning. "Chapter 31 - Fiscal Multipliers: Liquidity Traps and Currency Unions." In *Handbook of Macroeconomics*, vol. 2, edited by H. Uhlig and J.B. Taylor.

Fornaro, L. and F. Romei. 2022. "Monetary Policy During Unbalanced Global Recoveries." Working Papers No. 4026877, Social Science Research Network.

Gagliardone, L. and M. Gertler. 2023. "Oil Prices, Monetary Policy, and Inflation Surges." NBER Working Papers No. 31263.

Galí, J. 2008. *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton, NJ: Princeton University Press.

Galí, J. and T. Monacelli. 2005. "Monetary Policy and Exchange Rate Volatility in a Small Open Economy." *Review of Economic Studies* 72 (3): 707–34.

Galí, J., J.D. López-Salido, and J. Vallés. 2007. "Understanding the Effects of Government Spending on Consumption." *Journal of the European Economic Association* 5(1): 227–70.

Gelman, M., Y. Gorodnichenko, S. Kariv, D. Koustas, M.D. Shapiro, D. Silverman, and S. Tadelis. 2023. "The Response of Consumer Spending to Changes in Gasoline Prices." *American Economic Journal: Macroeconomics* 15(2): 129–60.

Gourinchas, P.O., S. Kalemli-Özcan, V. Penciakova, and N. Sander. 2021. "Fiscal Policy in the Age of COVID: Does it 'Get in all of the Cracks?'" NBER Working Papers No. 29293.

Guerrieri, V., G. Lorenzoni, L. Straub, and I. Werning. 2022. "Macroeconomic Implications of COVID-19: Can Negative Supply Shocks Cause Demand Shortages?" *American Economic Review* 112(5): 1437–74.

Guo, X., P. Ottonello, and D.J. Pérez. 2023. "Monetary Policy and Redistribution in Open Economies." *Journal of Political Economy Macroeconomics* 1(1): 191–241.

Hamilton, J.D. 1983. "Oil and the Macroeconomy since World War II." *Journal of Political Economy* 91(2): 228–48.

Jordà, Ò. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review* 95(1): 161–82.

Känzig, D.R. 2021. "The Macroeconomic Effects of Oil Supply News: Evidence from OPEC Announcements." *American Economic Review* 111(4): 1092–125.

Känzig, D.R.. "The Unequal Economic Consequences of Carbon Pricing." NBEF Working Papers No. 31221.

Kaplan, G., B. Moll, and G.L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

Kharroubi, E. and F. Smets. 2023. "Energy Shocks, the Natural Rate and Fiscal Policy." Manuscript.

Kilian, L.2009. "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market." *American Economic Review* 99(3): 1053–69.

Kuhn, F., M. Kehrig, and N.L. Ziebarth. 2021. "Welfare Effects of Gas Price Fluctuations." Manuscript.

Langot, F., S. Malmberg, F. Tripier, and J.O. Hairault. 2023. "The Macroeconomic and Redistributive Effects of Shielding Consumers from Rising Energy Prices: the French Experiment." Manuscript, CEPREMAP.

Leduc, S. and K. Sill. 2004. "A Quantitative Analysis of Oil-Price Shocks, Systematic Monetary Policy, and Economic Downturns." *Journal of Monetary Economics* 51(4): 781–808.

Lorenzoni, G. and I. Werning. 2023. "Inflation is Conflict." NBER Working Papers No. 31099.

Lorenzoni, G. and I. Werning. 2023. "Wage Price Spirals." Manuscript.

McKay, A., E. Nakamura, and J. Steinsson. 2016. "The Power of Forward Guidance Revisited." *American Economic Review* 106 (10): 3133–158.

Miyamoto, W., T.L. Nguyen, and D. Sergeyev. 2023. "How Oil Shocks Propagate: Evidence on the Monetary Policy Channel." Manuscript.

Newey, W.K. and K.D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55(3): 703–08.

Oskolkov, A. 2023. "Exchange Rate Policy and Heterogeneity in Small Open Economies." *Journal of International Economics* 142: 103750.

Pieroni, V. 2023. "Energy Shortages and Aggregate Demand: Output Loss and Unequal Burden from HANK." *European Economic Review* 154: 104428.

Ramey, V.A. 2016. "Chapter 2 - Macroeconomic Shocks and Their Propagation." In *Handbook of Macroeconomics*, vol. 2, Elsevier, edited by J.B. Taylor and H. Uhlig.

Romer, C.D., and D.H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *American Economic Review* 94(4): 1055–84.

Rotemberg, J.J., and M. Woodford. 1996. "Imperfect Competition and the Effects of Energy Price Increases on Economic Activity." *Journal of Money, Credit and Banking* 28 (4): 549–77.

Schmitt-Grohé, S. and M. Uribe. 2003. "Closing Small Open Economy Models." *Journal of International Economics* 61(1): 163–85.

Sgaravatti, G., S. Tagliapietra, C. Trasi, and G. Zachmann. 2023. National Fiscal Policy Responses to the Energy Crisis. See Bruegel website.

Soto, C. and J.P. Medina. 2005. "Oil Shocks and Monetary Policy in an Estimated DSGE Model for a Small Open Economy." Central Bank of Chile Working Papers No. 353.

Woodford, M. 2011. "Simple Analytics of the Government Expenditure Multiplier." *American Economic Journal: Macroeconomics* 3(1): 1–35.

Zhou, H. 2022. "Open Economy, Redistribution, and the Aggregate Impact of External Shocks." Manuscript.

APPENDICES

Appendix A. Model Details

A.1 Derivation of the Wage Phillips Curve

In this section, we derive the wage Phillips curve with the real-wage stabilization motive. At time t, union k sets its wage W_{kt} to maximize the utility of its average worker,

$$\sum_{\tau\geq 0}\beta^{t+\tau}\left(u\left(C_{t+\tau}\right)-v\left(N_{t+\tau}\right)-\frac{\psi_{\text{nr}}}{2}\left(\frac{W_{k,t+\tau}}{W_{k,t+\tau-1}}-1\right)^2-\frac{\zeta_{BG}}{2}\frac{(\varepsilon-1)Nu'(C)}{\left(\frac{W}{P}\right)}\left(\frac{W_{k,t+\tau}}{P_{t+\tau}}-\frac{W}{P}\right)^2\right).$$

Here ψ_{nr} parameterizes the degree of nominal rigidity, while ζ_{BG} captures the real-wage motive. The unions combine individual labor into tasks, which face demand

$$N_{kt}=\left(\frac{W_{kt}}{W_t}\right)^{-\varepsilon}N_t,$$

where $W_t=\left(\int W_{kt}^{1-\varepsilon}dk\right)^{\frac{1}{1-\varepsilon}}$ is the price index for aggregate employment services.

Each union is infinitesimal and therefore only takes into account its marginal effect on every household's consumption and labor supply. Household real earnings are

$$Z_t=\frac{1}{P_t}\int_0^1 W_{kt}\left(\frac{W_{kt}}{W_t}\right)^{-\varepsilon}N_t dk.$$

By the envelope theorem, we can evaluate indirect utility by assuming all income from the union wage change is consumed immediately. Then $\frac{\partial C_t}{\partial W_{kt}}=\frac{\partial Z_t}{\partial W_{kt}}$, where

$$\frac{\partial Z_t}{\partial W_{kt}}=\frac{1}{P_t}N_{kt}\left(1-\varepsilon\right).$$

On the other hand, total hours worked by household i are

$$N_{it} \equiv \int_0^1 \left(\frac{W_{kt}}{W_t} \right)^{-\varepsilon} N_t dk,$$

which falls when W_{kt} rises according to

$$\frac{\partial N_{it}}{\partial W_{kt}} = -\varepsilon \frac{N_{kt}}{W_{kt}}.$$

Therefore, the union's first-order condition gives

$$\left(\frac{W_{k,t}}{W_{k,t-1}} - 1 \right) \frac{W_{k,t}}{W_{k,t-1}} = \frac{\varepsilon}{\psi_{\text{nr}}} \left[N_{kt} v'(N_t) - \frac{\varepsilon - 1}{\varepsilon} \frac{N_{k,t} W_{k,t}}{P_t} u'(C_t) \right.$$
$$\left. - \zeta_{BG} \frac{N}{\frac{\varepsilon}{\varepsilon - 1} \left(\frac{W}{P} \right)} u'(C) \left(\frac{W_{k,t}}{P_t} - \frac{W}{P} \right) \frac{W_{k,t}}{P_t} \right]$$
$$+ \beta \left(\frac{W_{k,t+1}}{W_{k,t}} - 1 \right) \frac{W_{k,t+1}}{W_{k,t}}.$$

In equilibrium, all unions set the same wage: $W_{kt} = W_t$ and so $N_{kt} = N_t$. Define wage inflation as $\pi^w \equiv \frac{W_t}{W_{t-1}} - 1$. Then

$$\pi_t^w \left(1 + \pi_t^w \right) = \frac{\varepsilon}{\psi_{\text{nr}}} \left[N_t v'(N_t) - \frac{1}{\mu_w} Z_t u'(C_t) \right. \tag{A.1}$$
$$\left. - \frac{\zeta_{BG}}{\mu_w} u'(C) \frac{N}{N_t} \left(\frac{W_t}{P_t} - \frac{W}{P} \right) \frac{W_t}{P_t} / \frac{W}{P} \right] + \beta \pi_{t+1}^w \left(1 + \pi_{t+1}^w \right)$$

with $\mu_w = \frac{\varepsilon}{\varepsilon - 1}$. In the zero wage-inflation steady state

$$v'(N) = \frac{1}{\mu_w} u'(C) \frac{W}{P}.$$

Linearizing (A.1) around this steady state,

$$d\pi_t^w = \frac{\varepsilon}{\psi_{\text{nr}}} N\left[dv'(N_t) - \frac{1}{\mu_w} du'(C_t)\frac{W}{P} - (1+\zeta_{BG})\frac{1}{\mu_w} u'(C) d\left[\frac{W_t}{P_t}\right]\right] + \beta d\pi_{t+1}^w.$$

This also gives the first-order dynamics (and the steady state) of (20) above, with $\kappa_w = \frac{\varepsilon N v'(N)}{\psi_{\text{nr}}}$.

A.2 Comparison of the Real-Wage Targeting Motive to Blanchard and Galí (2007b)

In Blanchard and Galí (2007b), the (log) real wage evolves according to

$$w_t = \gamma w_{t-1} + (1-\gamma)\text{mrs}_t.$$

Consider instead a modification of this equation, where the lagged real wage is replaced by the steady-state value. Then, using hats to denote log deviations from steady state,

$$\hat{w}_t = (1-\gamma)\hat{\text{mrs}}_t.$$

Taking our wage equation (20) as $\theta_w \to 0$, gives

$$\mu_w \frac{v'(N_t)}{u'(C_t)} = (W_t / P_t)^{1+\zeta_{BG}}.$$

Taking logs, and with $\text{MRS}_t \equiv \frac{v'(N_t)}{u'(C_t)}$,

$$\hat{w}_t = \frac{1}{1+\zeta_{BG}} \hat{\text{mrs}}_t.$$

Blanchard and Galí (2007b) use values $\gamma = 0.6$ and $\gamma = 0.9$. So to match this, we would set

$$\zeta_{BG} = \frac{\iota}{1-\gamma} \in \{1.5, 9\}.$$

Our value lies in between those two.

Appendix B. Proofs

B.1 Proof of Proposition 2

In this section, we derive the "international Keynesian cross" shown in (32). To derive (32), we start from the general goods market clearing condition (27)

$$\bar{Y}_t = (1-\alpha)\left(\frac{P_{Ht}}{P_{HFt}}\right)^{-\eta}\left(\frac{P_{HFt}}{P_t}\right)^{-\eta_E} C_t + \alpha^*\left(\frac{P_{Ht}}{\mathcal{E}_t}\right)^{-\gamma} C^*, \tag{B.1}$$

where we, at this point, still allow for energy in production, $\xi_E > 0$. Consumption here can be written as an intertemporal consumption function[1]

$$C_t = \mathcal{C}_t\left(\{r_0, r_s^{\text{ante}}, Z_s\}\right), \tag{B.2}$$

where $Z_s = \frac{W_s}{P_s} N_s$ denotes aggregate labor income (2). This follows directly from (1).

In (B.2), we have made explicit the fact that aggregate demand for consumption C_t depends only on the initial ex-post return r_0, reflecting valuation effects, the time path of ex-ante real interest rates r_s^{ante} for $s \geq 0$ set by monetary policy (since $r_{t+1} = r_s^{\text{ante}}$ for all $t \geq 1$), and the path of real labor income Z_s for $s \geq 0$. We denote this general consumption function by $\mathcal{C}_t$.

We consider here the case of a constant real interest rate path, $r_s^{\text{ante}} = const = r_{ss}$, and will henceforth drop it from the consumption function (B.2). By the real UIP condition, (17) this also implies that

$$Q_t = Q_{ss}$$

and $d\log P_t = d\log \mathcal{E}_t$.

Next, we linearize (B.1), beginning with expressions for all relevant relative prices; then we linearize the left-hand side, followed by the right-hand side.

1. See Auclert and others (2023).

Relative prices. From (4), obtain

$$d\log P_{HFt} = \alpha_F d\log P_F + (1-\alpha_F) d\log P_H$$
$$d\log P_t = \alpha_E d\log \mathcal{E}_t + \alpha_E d\log P^*_{Et} + (1-\alpha_E) d\log P_{HF}.$$

Rearranging, we find

$$d\log \frac{P_{Ht}}{\mathcal{E}_t} = -\frac{\alpha_E}{1-\alpha} d\log P^*_{Et} \tag{B.3}$$

$$d\log \frac{P_{Ht}}{P_{HFt}} = -\frac{\alpha_E \alpha_F}{1-\alpha} d\log P^*_{Et} C^*_{Ht} = \alpha^* \left(\frac{P^*_{Ht}}{P^*} \right)^{-\gamma} C^* \tag{B.4}$$

$$d\log \frac{P_{HFt}}{P_t} = -\frac{\alpha_E}{1-\alpha_E} d\log P^*_{Et}. \tag{B.5}$$

Moreover, log-linearizing (13), we obtain

$$d\log P_{Ht} = (1-\xi_E) d\log W_t + \xi_E d\log P^*_{Et} + \xi_E d\log \mathcal{E}_t,$$

which lets us derive

$$d\log W_t - d\log P_{Et} = -\frac{1}{1-\xi_E} \frac{\alpha_E + 1 - \alpha}{1-\alpha} d\log P^*_{Et} \tag{B.6}$$

and

$$d\log \frac{W_t}{P_t} = -\frac{\alpha_E + \xi_E (1-\alpha)}{(1-\xi_E)(1-\alpha)} d\log P^*_{Et}. \tag{B.7}$$

Left-hand side of (A.2). We log-linearize the right-hand side as follows,

$$d\log \bar{Y}_t = (1-\xi_E) d\log Y_t + \xi_E d\log E_t.$$

Energy demand by domestic firms is given by

$$d\log E_t = d\log Y_t + v(\mathrm{dlog} W_t - \mathrm{dlog} P_{Et}),$$

so that we can write

$$d\log \bar{Y}_t = d\log Y_t + \xi_E v\left(\text{dlog}W_t - \text{dlog}P_{Et}\right).$$

Substituting in (B.6) and the steady-state expression $\bar{Y}_{ss} = \frac{1}{1-\tilde{\zeta}_E}$, we obtain for the left-hand side of (A.2),

$$\left(1-\xi_E\right) d\bar{Y}_t = dY_t - \frac{\xi_E}{1-\xi_E}\frac{\alpha_E + 1 - \alpha}{1-\alpha} v d\log P^*_{Et}. \tag{B.8}$$

Relative prices on the right-hand side of (B.1). For the right-hand side, we find

$$\left(1-\alpha\right)\left(\frac{P_{Ht}}{P_{HFt}}\right)^{-\eta}\left(\frac{P_{HFt}}{P_t}\right)^{-\eta_E} C_t + \alpha^*\left(\frac{P_{Ht}}{\mathcal{E}_t}\right)^{-\gamma} C^*$$

$$d\bar{Y}_t = -\left(1-\alpha\right)\eta d\log\frac{P_{Ht}}{P_{HFt}} - \left(1-\alpha\right)\eta_E d\log\frac{P_{HFt}}{P_t} + \left(1-\alpha\right) dC_t - \alpha^*\gamma d\log\frac{P_{Ht}}{\mathcal{E}_t}.$$

Substituting in (B.4), (B.5), (B.3), we arrive at

$$d\bar{Y}_t = \alpha_E\left(\alpha_F\eta + \left(1-\alpha_F\right)\eta_E\right) d\log P^*_{Et} + \alpha^*\gamma\frac{\alpha_E}{1-\alpha} d\log P^*_{Et} + \left(1-\alpha\right) dC_t. \tag{B.9}$$

Consumption response on the right-hand side of (A.2). In order to express dC_t in terms of primitives, observe that the valuation equation for assets, combined with (B.10), implies that share prices are

$$p_t = \frac{D_{t+1} + p_{t+1}}{1+r_t} = \text{PDV}\left(\left\{\left(\mu - 1\right) Z_s\right\}\right), \tag{B.10}$$

so that the initial revaluation r_0^p also only depends on the path of labor income Z_s. Following Auclert and others (2021a), we therefore can write the consumption function (B.2) simply as a function of Z_s,

$$\mathcal{C}_t = \mathcal{C}_t\left(\left\{Z_s\right\}\right),$$

whose (sequence-space) Jacobian we denote by

$$M_{t,s} \equiv \frac{\partial \mathcal{C}_t}{\partial Z_s}.$$

We stack the matrix as $\mathbf{M} \equiv \left(M_{t,s}\right)$. The exact shape of $\mathbf{M}$ is discussed in more detail in Auclert and others (2021a). With this notation, we can write, in vector notation,

$$d\mathbf{C} = \mathbf{M} \cdot d\log\mathbf{Z}, \tag{B.11}$$

where, using (B.7),

$$d\log Z_t = dY_t + d\log\frac{W_t}{P_t} = dY_t - \frac{\alpha_E + \xi_E(1-\alpha)}{(1-\xi_E)(1-\alpha)} d\log P_{Et}^*.$$

Thus,

$$d\mathbf{C} = -\frac{\alpha_E + \xi_E(1-\alpha)}{(1-\xi_E)(1-\alpha)}\mathbf{M} \cdot d\mathbf{P}_E^* + \mathbf{M}d\mathbf{Y}. \tag{B.12}$$

Equation (B.12) collapses to (31) in the special case of no energy usage in production, $\xi_E = 0$.

Combining left- and right-hand sides. Putting together (B.8), (B.9), (B.11), and the definition of χ in (30) we obtain the following equation,

$$d\mathbf{Y} = \left[(1-\xi_E)\frac{\alpha_E}{1-\alpha}\chi + \frac{\xi_E}{1-\xi_E}\left(1+\frac{\alpha_E}{1-\alpha}\right)\upsilon + \xi_E\frac{\alpha_E}{1-\alpha}\gamma\right] \tag{B.13}$$
$$d\log\mathbf{P}_E^* - \left(\alpha_E + \xi_E(1-\alpha)\right)\mathbf{M} \cdot d\mathbf{P}_E^* + (1-\xi_E)(1-\alpha)\mathbf{M}d\mathbf{Y}.$$

Setting $\xi_E = 0$, and hence $\alpha^* = \alpha$, we find that this collapses to (32).

B.2 Proof of Proposition 1

In the (complete-market) representative-agent model, the Backus-Smith condition (24) holds. Since the real exchange rate Q_t is constant, consumption is too. In other words, $d\mathbf{C} = 0$. Essentially, $\mathbf{M} = 0$ for the (complete-market) representative agent. This proves (28). (29) follows from (B.13) when we set $\mathbf{M} = 0$ and $\xi_E = 0$.

B.3 Proof of Proposition 3

Analogously to proposition 3 in Auclert and others (2021a) we solve the fixed point (32) for $d\mathbf{Y}$ to find

$$d\mathbf{Y} = \underbrace{\left(\sum_{k\geq 0}(1-\alpha)^k\mathbf{M}^k\right)}_{=(\mathbf{I}-(1-\alpha)\mathbf{M})^{-1}}\left(\frac{\alpha_E}{1-\alpha}\chi d\mathbf{P}_E^* - \alpha_E\mathbf{M}\cdot d\mathbf{P}_E^*\right).$$

We can rearrange this to (33). The results that $d\mathbf{Y} \lessgtr d\mathbf{Y}^{RA}$ and $d\mathbf{C} \lessgtr 0$ are equivalent to $\chi \lessgtr 1$ follow directly from $\mathbf{M} \geq 0$ and the assumption of a non-negative shock, $d\mathbf{P}_E^* \geq 0$.

B.4 Proof of Proposition 4

For (4), we set $\alpha_E = 0$ in (B.13). To get at the mapping between the "energy in production" and "energy in consumption" models, we denote by $\tilde{\alpha}_F$ the share of consumption going towards good F in the "energy in production" model. We then have the following consumption shares in the two models, across the three goods, where we unpack the H good into labor and (if $\xi_E > 0$) energy:

Table B.1 Consumption Shares in the Two Models

Consumption share by good	*"energy in production" model*	*"energy in consumption" model*
Domestic labor N	$(1-\alpha_E)(1-\alpha_F)$	$(1-\xi_E)(1-\tilde{\alpha}_F)$
F goods	$(1-\alpha_E)\alpha_F$	$\tilde{\alpha}_F$
E goods	α_E	$\xi_E(1-\tilde{\alpha}_F)$

Source: Authors' calculations.

To equalize the shares, we define in the "energy in production" model,

$$\tilde{\alpha}_F \equiv (1-\alpha_E)\alpha_F$$
$$\xi_E \equiv \frac{\alpha_E}{1-\tilde{\alpha}_F} = \frac{\alpha_E}{1-(1-\alpha_E)\alpha_F}.$$

It is straightforward to check that the domestic labor consumption share is equalized too. Notice that, with these definitions, we have that

$$\frac{\xi_E}{1-\xi_E} = \frac{\alpha_E}{1-\alpha}.$$

Thus, if $\nu = \chi$, the Keynesian cross equation (35) with energy in production is equivalent to that with energy in consumption (32).

Appendix C. Comparison with a TANK Model

For the two-agent complete-market model ("TA model"), we assume the household side of the model consists of a share $1 - \lambda$ of agents with unconstrained access to financial markets, denoted by superscript u, and a share λ with no access to financial markets, denoted by superscript c. The unconstrained agents behave just like the representative agent in section 2.1. So, we can characterize their consumption with the Backus-Smith condition,

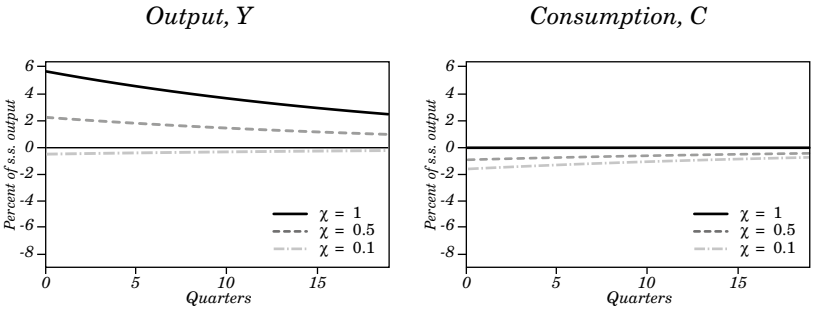
$$\left(c_t^u\right)^{-\sigma} = \frac{\left(c_{SS}^u\right)^{-\sigma}}{Q_t}.$$

The constrained agents consume their entire income each period, $c_t^c = Z_t$.

We suppose unions continue to split hours of work evenly between households. Aggregate consumption is the weighted average of these consumption responses,

$$C_t = (1-\lambda)c_{ss}^u + \lambda c_t^c.$$

Figure C1. Response to the Energy Price Shock in TA Model



Source: Authors' calculations.
Impulse responses in a two-agent model to the energy price shock $P^*_{E,t}$ displayed in figure 2. χ is the average substitution elasticity between energy and domestically produced goods. It is defined in (30).

And we set steady-state aggregate asset holdings, $A_{ss} = (1-\lambda)A^u_{ss}$, equal to those in the HA model. This gives rise to a household block characterized by the matrix of intertemporal MPCs,

$$\mathbf{M} = \lambda \mathbf{I}.$$

From Proposition 2, the impulse response of consumption is then

$$d\mathbf{C} = \underbrace{-\frac{\alpha_E}{1-\alpha}\lambda \cdot d\mathbf{P}^*_E}_{\text{Real-income channel}} + \underbrace{\lambda \cdot d\mathbf{Y}}_{\text{Multiplier}}$$

$$d\mathbf{Y} = \underbrace{\frac{\alpha_E}{1-\alpha}\chi \cdot d\mathbf{P}^*_E}_{\text{Exp. switching channel}} \underbrace{-\alpha_E \lambda \cdot d\mathbf{P}^*_E}_{\text{Real-income channel}} + \underbrace{(1-\alpha)\lambda \cdot d\mathbf{Y}}_{\text{Multiplier}}.$$

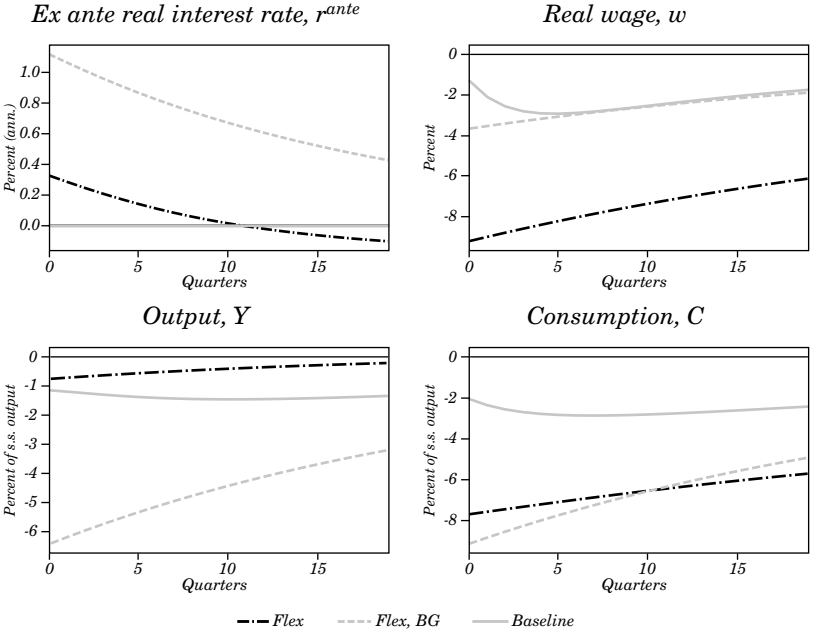
This has the solution

$$d\mathbf{Y} = \frac{\alpha_E}{1-\alpha}\frac{\chi - (1-\alpha)\lambda}{1-(1-\alpha)\lambda} \cdot d\mathbf{P}^*_E$$

$$d\mathbf{C} = \frac{\alpha_E}{1-\alpha}\frac{\lambda(\chi - 1)}{1-(1-\alpha)\lambda} \cdot d\mathbf{P}^*_E.$$

In figure C.1, we set $\lambda = 0.25$ and plot the response to the energy price shock without importer frictions, as in section 2. We see that the potential for declines in output and consumption is much more limited in this model.

Figure C2. Flexible Price Response to the Energy Price Shock



Source: Authors' calculations.
Note: This figure shows the impulse responses to the energy price shock P^*_{Et} displayed in figure 2 for the baseline model, the flexible price model with the real-wage friction (Flex, BG), and in the flexible price model without the real-wage friction (Flex).

Appendix D. Additional Model Outcomes

D.1 Flexible Price Allocation

In the section, we compare the response to the energy price shock in three cases: (1) the baseline case above, (2) the case with flexible prices but the real-wage stabilization motive, and (3) the case with flexible prices and no real-wage stabilization motive. The results are shown in figure C.2.

D.2 Real-Wage Stabilization with Taylor Rule vs. Real Rate Rule

In the main text, we show the inflation response under a real rate rule, where

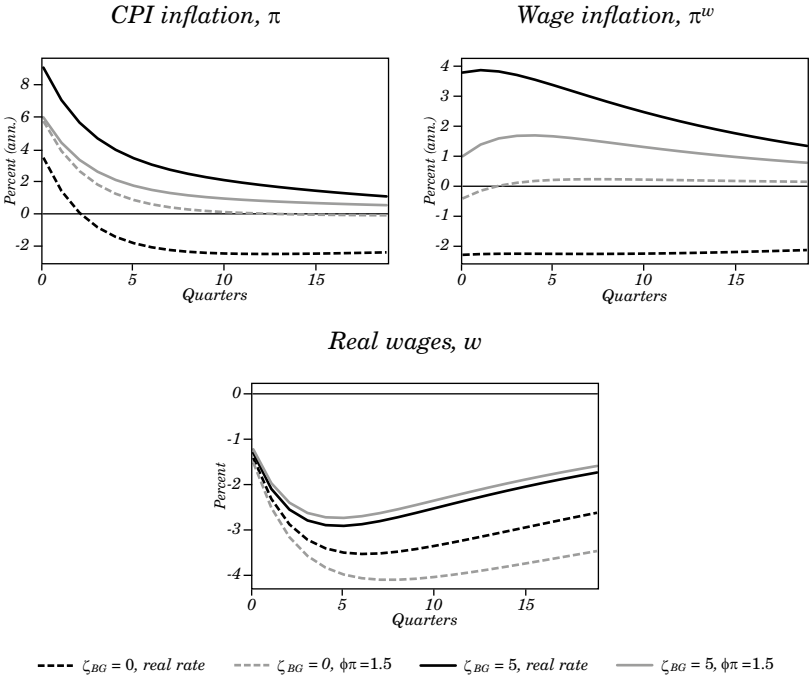
$$1 + i_t = \left(1 + r^*\right)\left(1 + \pi_{t+1}\right).$$

In figure D.1, we compare this to the response under the Taylor rule

$$1 + i_t = \left(1 + r^*\right)\left(1 + \phi_\pi \pi_t\right).$$

We see that the real-wage stabilization motive is more effective at raising real wages under the Taylor rule. Under the real rate rule, the effect is smaller, and in the absence of energy importer frictions, it would be zero.

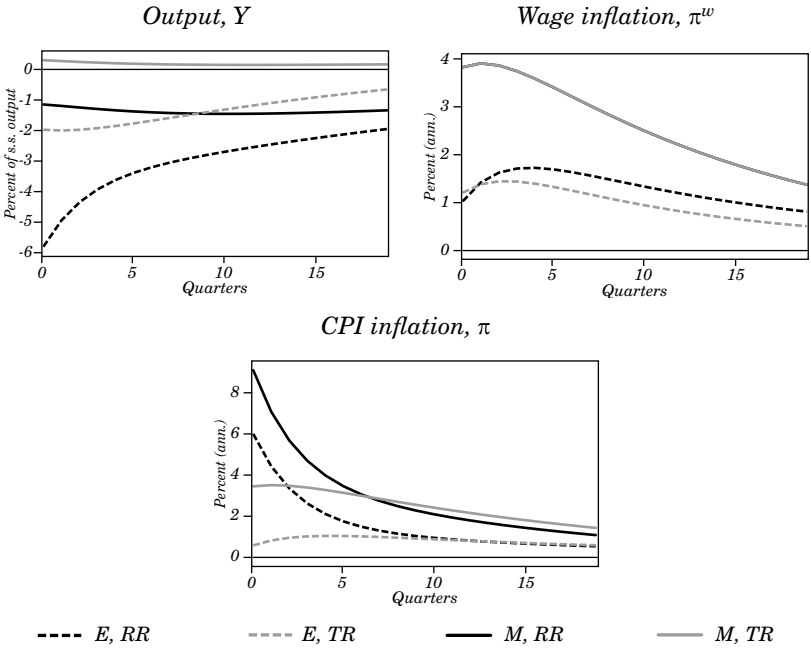
Figure D1. Real-Wage Stabilization with a Taylor Rule vs. a Real Rate Rule



Source: Authors' calculations.

Note: This figure shows the responses of prices and wages to an energy price shock, with and without the real-wage stabilization motive. It compares the response when the central bank follows a real rate rule against that when it follows a Taylor rule, with coefficient on current inflation ϕ_π.

Figure D2. Responses to an Energy Price Shock and a Markup Shock under Different Monetary Policy Rules



Source: Authors' calculations.
Note: This figure contrasts the response to the original energy price shock (E) with that to a markup shock (M) that leads to equivalent wage inflation (under our baseline real rate rule). It plots the responses to each shock under a real rate rule (RR) and a Taylor rule (TR) for monetary policy.

D.3 Markup shocks versus energy shocks

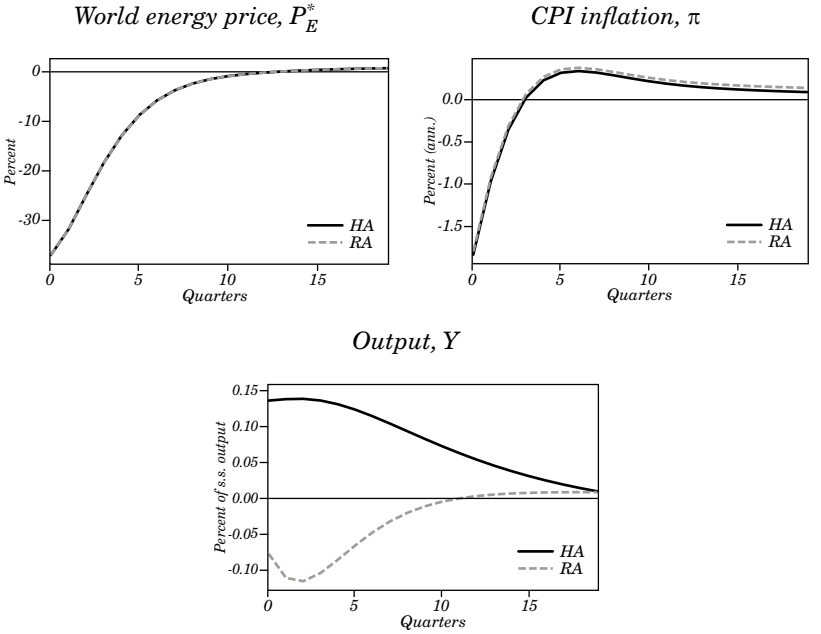
We now ask whether the interpretation of an energy price shock as a markup shock retains the results of our model. We suppose a union markup shock that induces the same path for wage inflation as under our energy price shock. We then compare the results in figure D.2. Under a real rate rule, both shocks are inflationary, but only the energy price shock leads output to contract. While switching to a Taylor rule does generate a decline in output in both models, it is significantly worse under the energy price shock.

D.4 Monetary Spillover in Different Models

In this section, we consider the impact on home of all other energy-importing countries tightening monetary policy and thereby lowering

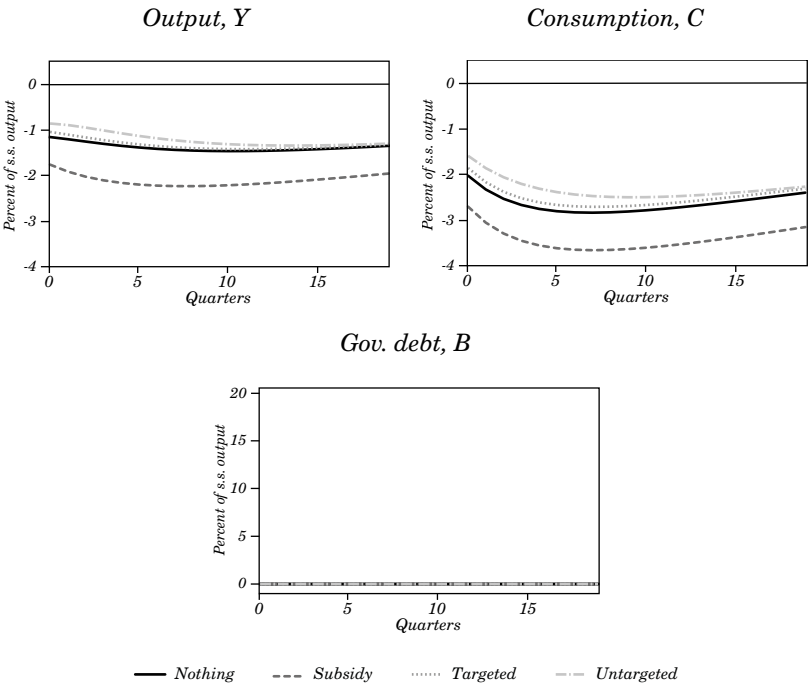
the world energy price. That is, we isolate the spillover channel. In the HA model, as discussed above, this shock leads to lower inflation and a boost in output, driven by the real-income channel. In the RA model, this same shock leads output to decline due to the expenditure switching channel. The results are shown in figure D.3.

Figure D3. Spillover Channel in the RA and HA Models



Source: Authors' calculations.
Note: This figure shows the impact of all other energy-importing countries tightening monetary as detailed in figure 11. It compares the response in the HA and RA models, for inflation and output.

Figure D4. Fiscal Policy with a Balanced Budget



Source: Authors' calculations.
Note: This figure compares the output and consumption responses to an energy price shock under no fiscal policy with the three fiscal policy programs explained in section 4.1, assuming a balanced budget throughout.

D.5 Balanced Budget Fiscal Policy

Here, we repeat the analysis in section 4, only now imposing a balanced budget at all dates: $B_t = B_{ss} = 0$ for all t. As we see in figure D.4, the three fiscal policies are now less effective at cushioning the fall in output and consumption. However, it remains the case that the untargeted transfer is most effective, on this measure, and the subsidy the least.

Measuring the Redistributive Effects of Monetary Policy: An Application to the Chilean Economy

Emiliano Luttini
The World Bank

Ernesto Pastén
Central Bank of Chile

Elisa Rubbo
University of Chicago

Active recent literature shows that monetary policy has heterogeneous effects on workers in different demographic groups.[1] In this paper, we build on the framework in Rubbo (2023) to estimate how monetary policy affects real incomes across demographic groups in the Chilean economy.

Specifically, Rubbo (2023) considers a multisector economy where different households are subject to different degrees of wage rigidity, are employed by industries with different price stickiness and capital intensity, and consume different bundles of goods. In this context, it provides an analytical expression for the effect of monetary policy on real incomes across households, depending on which industries they work in and which goods they consume. Rubbo (2023) shows that expansionary monetary policy increases employment relatively more for households that (i) have more flexible wages or work in

We thank the participants to the XXV Annual Conference of the Central Bank of Chile for their comments, and Ludwig Straub for an insightful discussion. We also thank Matías Pizarro and Rachel Coroseo for their excellent research assistance. Opinions and conclusions expressed in this paper are those of the authors' alone and do not necessarily represent the views of the Central Bank of Chile. All results have been reviewed to ensure no confidential data are disclosed.

1. See Coglianese and others (2023), Andersen and others (2022), Minton and Wheaton (2023).

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

sectors with more flexible prices; (ii) have more elastic labor supply or are complementary with other elastically supplied primary factors; and (iii) consume sticky-price goods. Moreover, controlling for the workers' own wage rigidity and labor-supply elasticity, monetary policy increases the relative wages of workers employed by sectors with more flexible prices or who are complementary with elastically supplied primary factors. Overall, nominal incomes unambiguously increase for these workers. Through its effect on consumption prices, expansionary monetary policy also increases the real income of households that purchase relatively more sticky-price goods or goods that rely on elastically supplied factors.

Rubbo (2023) summarizes these redistributive effects of monetary policy by using two cross-sectional multipliers, describing the relative response of employment and income across households. These multipliers account for all the channels described above and for their interactions through the input-output network and through general equilibrium effects.

We compute these two cross-sectional multipliers by using data for the Chilean economy to obtain the cumulative impulse responses of employment and income to a monetary shock across 50 demographic groups. Demographic groups are distinct by gender, age, and initial income quintiles. We construct measures of the exposure of each worker group to each industry through the employment and consumption channels by using detailed microdata available at the Central Bank of Chile. We first construct an input-output network of expenditure shares of each industry i on products from each industry j. We then combine the input-output network with measures of the expenditure shares of (i) each industry i on workers of each group g (to compute the exposure of workers to industries through the employment channel) and (ii) each worker group g on goods from each industry i (to measure the workers' exposure to each industry through the consumption channel). Finally, we merge these data with measures of price and wage adjustment probabilities, also obtained from microdata available at the Central Bank of Chile.

Our baseline results show significant heterogeneity in the degree of nominal rigidities that households face in the labor market, through both heterogeneous wage stickiness and heterogeneous price stickiness of the industries where they are employed. Older and higher-income workers tend to sort into industries with stickier prices. By contrast, we find little heterogeneity in the price stickiness of consumption goods across demographic groups. These differences result in a range

of cumulative employment responses that spans from 0.7percent for middle-aged, middle-income men to 5.3 percent for high-income men over 54, while income responses span from 0.8 percent to 5.4 percent.

We also account for heterogeneous capital intensity across industries. While capital intensity varies significantly across industries (with capital shares ranging from 0.1 to 0.8), the average employer's capital share is similar across demographic groups. This suggests that interacting demographic characteristics with the industry of employment would uncover even larger heterogeneity. Nonetheless, even with our currently available data, accounting for the presence of semifixed capital assets amplifies the cross-sectional dispersion of employment responses.

The paper is organized as follows: Borrowing from Rubbo (2023), section 1 describes the environment, and section 2 presents cross-sectional employment and income multipliers. Section 3 then illustrates the Chilean data, and section 4 presents the calibration results. Section 5 concludes.

Related literature. This paper provides an empirical implementation of the framework in Rubbo (2023) to the Chilean economy. Rubbo (2023) considers a New Keynesian model with heterogeneous agents to study how monetary policy affects different households depending on their exposure to different sectors of the economy. The HANK literature[2] also considers New Keynesian frameworks with heterogeneous agents. In HANK models, however, heterogeneity is driven by the agents' saving decisions as they face different discount factors and borrowing constraints, while in Rubbo (2023) and in this paper, monetary policy has heterogeneous effects through the households' real incomes.

The framework in Rubbo (2023) is similar to Baqaee and Farhi (2018), who study the propagation and aggregation of exogenous productivity and markup shocks in economies with multiple agents and a general input-output network. Different from Baqaee and Farhi (2018), in both Rubbo (2023) and our framework, markups change endogenously due to sticky prices.

A large literature studies the implications of input-output networks for aggregate monetary non-neutrality in single-factor models,[3] while

2. See Kaplan and others (2018), Auclert (2019), Auclert and others (2021).

3. See Basu (1995), Carvalho (2006), Nakamura and Steinsson (2010); Pastén and others (2019), La'O and Tahbaz-Salehi (2019); Rubbo (2020).

the currency union literature[4] introduces heterogeneous agents in simple Armington-style production structures. By using the framework in Rubbo (2023), this paper jointly accounts for heterogeneous agents and a quantitatively realistic input-output network and proposes a novel characterization of cross-sectional monetary non-neutrality which uses Phillips curve slopes as sufficient statistics.

Based on detailed micro-level data from Denmark and Sweden, Andersen and others (2022) and Coglianese and others (2021) empirically document heterogeneous effects of monetary policy depending on the individuals' employment and consumption. Minton and Wheaton (2023) also study the differential response of employment to monetary policy across U.S. States with different mandatory minimum wages. Their results are in line with the theoretical predictions of Rubbo (2023) and with the calibration results in this paper.

1. Environment

This section sets up a New Keynesian model, with monopolistic competition and sticky prices, featuring multiple heterogeneous industries, workers, and capital assets. It lays out the assumptions about preferences, production, and policy instruments, derives optimality conditions for consumers and producers, and defines the general equilibrium.

1.1 Final Users

1.1.1 Consumption

Preferences. There are N_w household groups, indexed by $h \in \{1,\ldots,N_w\}$ We denote the set of households by $\mathcal{N}_w$. Each group has a representative consumer-worker who supplies a distinct labor type and whose per-period preferences are described by the utility function

$$U_{ht} = \frac{C_{ht}^{1-\gamma_h}}{1-\gamma_h} - \frac{L_{ht}^{1+\varphi_h}}{1+\varphi_h}. \tag{1}$$

4. See Aoki (2001); Benigno (2004); Devereux and Engel (2003); Huang and Liu (2007); Gali and Monacelli (2008).

All households enjoy consumption (C) and dislike labor (L), with heterogeneous income effects on labor γ_h and Frisch elasticities of labor supply $\frac{1}{\varphi_h}$. Consumption aggregators $C_{ht} \equiv \mathcal{C}_h\left(c_{1ht},\ldots,c_{Nht}\right)$ are homothetic over the N goods produced in the economy, and can differ across groups.

Budget constraint. Each household h owns shares of all industries, as described by the ownership matrix Ξ, whose elements Ξ_{ih} denote the share of profits from sector i accruing to type- h agents. Hence the matrix Ξ, governs the allocation of the firms' profits Π_{it}, net of lump-sum taxes T_{it} paid by firms to the government.

Agents also own shares in the capital assets, as described by the matrix $\mathcal{Z}$, whose elements $\mathcal{Z}_{fh}$ denote the share of asset f owned by worker type h.

Agents maximize the present discounted value of per-period utility flows, with the same discount factor ρ, subject to group-specific budget constraints

$$P_{ht}^{C} C_{ht} \leq W_{ht} L_{ht} + \sum_{f} \mathcal{Z}_{fh} \frac{\varphi_f}{1+\varphi_f} R_f K_f + \sum_{i} \Xi_{ih}\left(\Pi_{it} - T_{it}\right) + T_{ht}, \tag{2}$$

where P_{ht}^{C} is the price index implied by the consumption aggregator $\mathcal{C}_h$, W_{ht} is the nominal wage earned by labor type h, $\frac{\varphi_f}{1+\varphi_f} R_f K_f$ is income from capital asset f (as explained below), and $\{T_h\}_{h=1}^{N_w}$ are zero-sum income transfers between agents. In the analysis below we will assume financial autarchy and consider the transfers $\{T_h\}_{h=1}^{N_w}$ as exogenous. If we allowed for borrowing and lending across agents, the transfers would be determined endogenously. Endogenous and exogenous transfers enter our expressions for cross-sectional employment in the same way, and their role is negligible quantitatively.

Consumption-leisure tradeoff. The optimal consumption-leisure tradeoff satisfies the first-order condition

$$\frac{W_{ht}}{P_{ht}^{C}} = C_{ht}^{\gamma_h} L_{ht}^{\varphi_h}, \tag{3}$$

where W_{ht} is the flexible wage. Our modeling of wage rigidities is detailed in section 1.2 below.

1.1.2 Investment

There are N_f capital assets in the economy, indexed by $f \in \{1,\ldots, N_f\}$. We denote the set of capital assets by N_f. Each capital asset is produced by combining a fixed endowment ($\overline{K}_f$) with an investment good I_f, according to the production function

$$K_f = \left[\left(1+\varphi_f\right) I_f\right]^{\frac{1}{1+\varphi_f}} \overline{K}_f. \tag{4}$$

For convenience, we assume that the investment component I_f fully depreciates from one period to the next, while the endowment component $\overline{K}_f$ never depreciates.

In turn, investment is produced with constant return to scale, using as inputs a combination of labor (L_{fh}), capital (K_{fg}), and intermediate goods (X_{fi}), according to the production function

$$I_f = G_f\left(\left\{L_{fh}\right\}, \left\{K_{fg}\right\}, \left\{X_{fi}\right\}\right). \tag{5}$$

There are N_f investment producing sectors, one for each asset type f, that sell the investment good at marginal cost P_f^I to capital retailers. Retailers purchase capital endowments from the agents, combine them with the investment good, and sell capital services to the firms for a rental rate R_f in a perfectly competitive market. Capital retailers are owned by the agents in proportion to their ownership shares $\mathcal{Z}$ in the capital endowments and rebate their profits accordingly.

Profit maximization yields the capital supply curves

$$U_f^{\varphi_f} = \frac{R_f}{P_f} \overline{K}_f, \tag{6}$$

where

$$U_f \equiv \left[\left(1+\varphi_f\right) I_f\right]^{\frac{1}{1+\varphi_f}} \tag{7}$$

can be interpreted as a measure of capital utilization. Profits are given by

$$\frac{\varphi_f}{1+\varphi_f} R_f \overline{K}_f, \tag{8}$$

while investment expenditures are equal to

$$\frac{1}{1+\varphi_f} R_f \bar{K}_f . \tag{9}$$

1.2 Production

There are N good-producing industries in the economy (indexed by $i, j \in \{1,\ldots, N\}$ Within each industry, there is a continuum of firms producing differentiated varieties.

All firms z in industry i have the same constant returns to scale production function (omitting time subscripts for legibility)

$$Y_{iz} = G_i\left(\{L_{ihz}\}, \{K_{ifz}\}, \{X_{ijz}\}\right), \tag{10}$$

where L_{ihz} is type- h labor hired by firm z in industry i, K_{ifz} is type- f capital used by firm z, and X_{ijz} is intermediate input j used by the firm.

Customers (consumers and other producers) buy a constant-elasticity-of-substitution (CES) bundle of sectoral varieties, with elasticity of substitution ϵ_i. The industry output is given by

$$Y_i = \left(\int Y_{if}^{\frac{\epsilon_i - 1}{\epsilon_i}} df \right)^{\frac{\epsilon_i}{\epsilon_i - 1}}, \tag{11}$$

and the implied sectoral price index is

$$P_i = \left(\int P_{if}^{1-\epsilon_i} df \right)^{\frac{1}{1-\epsilon_i}} . \tag{12}$$

We follow a standard practice in the literature and assume proportional input subsidies are in place to eliminate the markup distortions arising from monopolistic competition. This assumption eliminates the incentive to use expansionary monetary policy to reduce markups. Sectoral subsidies τ_i^* are given by

$$1 - \tau_i^* = \frac{\epsilon_i - 1}{\epsilon_i}. \tag{13}$$

All producers minimize costs given input prices. With constant returns to scale, marginal costs are the same for all firms within a sector i, and they all use inputs in the same proportions. The marginal

cost of sector i, denoted by MC_i, is the solution to the cost minimization problem (again omitting time subscripts for legibility)

$$MC_i = \min_{\{X_{ij}\},\{L_{ih}\}} \sum_h W_h L_{ih} + \sum_f R_f K_{if} + \sum_j P_j X_{ij} \text{ s.t. } G_i\left(\{L_{ih}\},\{K_{if}\},\{X_{ij}\}\right) = 1. \quad (14)$$

Price rigidities are modeled à la Calvo: in every sector i, a randomly selected fraction δ_i of firms can update their price at each given period. They set it to maximize the present discounted value of profits at each future period $t + s$, in the event that they are unable to update their price until $t + s$:

$$P_{it}^* = \frac{\epsilon_i\left(1-\tau_i^*\right)}{\epsilon_i - 1} \frac{\mathbb{E}_t \sum_s \left[SDF_{t+s} \left(1-\delta_i\right)^s Y_{ift+s}\left(P_{it}^*\right) MC_{it+s} \right]}{\mathbb{E}_t \sum_s \left[SDF_{t+s} \left(1-\delta_i\right)^s Y_{ift+s}\left(P_{it}^*\right) \right]}, \quad (15)$$

where $SDF_{t+s} = \rho^s \frac{U_{ct+s}}{U_{ct}} \frac{P_t^c}{P_{t+s}^c}$ is the households' stochastic discount factor, demand functions are given by $Y_{ift+s}\left(P_{ft}\right) = Y_{it}\left(\frac{P_{ft}}{P_{it}}\right)^{-\epsilon}$, and $\frac{\epsilon_i\left(1-\tau_i^*\right)}{\epsilon_i - 1} = 1$. The firms f that cannot adjust their price have $P_{ift} = P_{ift-1}$ and their markup $\mathcal{M}_{ift}$ must absorb any cost changes.

Factor marketplaces. To model sticky factor prices, we assume that primary factors (workers and capital assets) are first purchased by marketplaces, which then sell their services to producers in all the different sectors. Each marketplace deals with only one primary factor. Marketplaces are treated like any other industry. In particular there is a continuum of marketplaces for each type, with fixed unit mass, facing Calvo-style price rigidities.

1.2.1 Aggregation

Definition 1 introduces our notion of nominal GDP, while Definitions 2 and 3 allow us to compute infinitesimal changes in real GDP and the GDP deflator around an initial equilibrium (denoted by starred variables).

Definition 1. Nominal GDP is the sum of consumption and investment expenditures

$$GDP = \sum_h P_h^C C_h + \sum_f P_f^I I_f. \quad (16)$$

Definition 2. Infinitesimal changes in real GDP $d\log Y_t$ around an initial equilibrium (denoted with starred variables) are given by

$$d\log Y_t = \sum_{h\in\mathcal{N}_w} \frac{P_h^{*C} C_h^*}{GDP^*} d\log C_{ht} + \sum_{f\in\mathcal{N}_f} \frac{P_f^* I_f^*}{GDP^*} d\log I_{ft}. \tag{17}$$

Definition 3. Changes in the GDP deflator $d\log P_t^Y$ are defined as

$$d\log P_t^Y = \sum_h \frac{P_h^{*C} C_h^*}{GDP^*} d\log P_{ht}^C + \sum_f \frac{P_f^{*C} I_f^*}{GDP^*} d\log P_{ft}^C. \tag{18}$$

Remark 1. As a consequence of constant returns to scale, changes in real GDP equal the income weighted sum of changes in primary factor quantities:

$$d\log Y = \frac{\sum_{h\in\mathcal{N}_w} W_h^* L_h^* d\log L_h + \sum_{f\in\mathcal{N}_f} R_f^* K_f^* d\log K_f}{\sum_{h\in\mathcal{N}_w} W_h^* L_h^* + \sum_{f\in\mathcal{N}_f} R_f^* K_f^*}. \tag{19}$$

1.3 Monetary Policy

At each period, the economy is subject to an aggregate cash-in-advance constraint whereby nominal GDP cannot exceed the money supply M_t, chosen by the central bank:

$$P_t Y_t \leq M_t. \tag{20}$$

Seignorage revenues are distributed in proportion to the agents' consumption shares, so that—to a first order—seignorage rebates are exactly equal to the amount of new money that the agents need to purchase in order to finance consumption, and the two cancel out from the budget constraint.

Below, we sometimes use a static version of the cash-in-advance model, with $\rho = 0$, to provide intuition. In this model, firms enter each period t with pre-set prices, and only a fraction δ_i of producers in each sector can update them after the new money supply is announced. We usually assume that the economy enters period 0 in steady state, so that pre-set prices from $t = -1$ are equal across producers, as they expect real money balances to remain constant.

1.4 Equilibrium

The equilibrium concept adapts the definition in Baqaee and Farhi (2020) to account for the endogenous determination of markups given pricing frictions and shocks. Given sectoral markups, all markets must clear. In turn, the evolution of markups must be consistent with Calvo pricing and the realization of monetary policy.

Definition 4. At each period t, for given sectoral probabilities of price adjustment δ_i and money supply M_t, the general equilibrium is given by a vector of firm-level markups $\mathcal{M}_{ift}$, a vector of sectoral prices P_{it}, a vector of agent-specific nominal wages W_{ht}, a vector of labor supplies L_{ht}, a vector of capital supplies K_{ft}, a vector of sectoral outputs Y_{it}, a matrix of intermediate input quantities X_{ijt}, a matrix of final consumption C_{iht}, and a matrix of investment demand U_{ift} such that: (i) a fraction δ_i of firms in each sector i charges the profit-maximizing price given by (15); (ii) the markup charged by adjusting firms is given by the ratio of the profit-maximizing price and marginal costs, while the markups of nonadjusting firms are such that their price remains constant; (iii) households maximize utility subject to their budget constraint; (iv) investment producers maximize profits given input prices; (v) producers in each sector i minimize costs and charge the relevant markup; and (vi) markets for all goods and all primary factors clear.

Remark 2. This equilibrium concept nests the standard one with flexible prices, which is obtained as a special case when $\delta_i = 1$ for every sector i.

1.5 Log-Linearized Model

We log-linearize the model around an efficient equilibrium with flexible prices, unit nominal GDP $\left(\sum_h P_h^* C_h^* + \sum_f P_f^* I_f^* = 1\right)$ and zero transfers $\left(T_h^* = 0 \forall h\right)$. Monetary policy is the only source of shocks. Tables 1, 2, 3, and 4 introduce the variables and parameters that govern the dynamics of our economy.

Input-output definitions. To a first order, the production structure is fully characterized by equilibrium input and final use shares and by the relevant Allen elasticities of substitution. Our representation follows the same structure as the National Accounts and is summarized in table 1.

Table 1. Input-Output Definitions

Consumption shares	$\beta_C \in \mathbb{R}^{N\times H}, \beta_{ih} = \frac{P_i C_{ih}}{P_h^C C_h}$
Investment expenditure	$\beta_I \in \mathbb{R}^{N\times F}, \beta_{if} = \frac{P_i X_{if}}{P_f^I I_f}$
Labor shares	$\alpha_L \in \mathbb{R}^{N\times H}, \alpha_{ih} = \frac{W_h L_{ih}}{MC_i Y_i}$
Capital shares	$\alpha_K \in \mathbb{R}^{N\times F}, \alpha_{if} = \frac{R_f X_{if}}{MC_i Y_i}$
Input-output matrix	$\Omega \in \mathbb{R}^N \times \mathbb{R}^N, \Omega_{in} = \frac{P_n X_{in}}{MC_i Y_i}$
Substitution elasticities	$\{\theta_{jn}^i\}$ production ϵ_i between varieties

Source: Authors' calculations.

Table 2. Input-Output Definitions: Parameters

Leontief inverse	$(I-\Omega)^{-1}$
Domar weights	$\lambda \in \mathbb{R}^{N\times(H+F)}, \lambda^T = \beta^T (I-\Omega)^{-1}$
Final expenditure shares	$s_h = \frac{P_h^C C_h}{GDP}, s_f = \frac{P_f^I I_f}{GDP}$
Factor income shares	$\Lambda = \alpha^T \lambda s, \mathcal{L} \equiv \mathrm{diag}(\Lambda)$

Source: Authors' calculations.

Table 3. Other Parameters

Sector ownership	$\Xi \in \mathbb{R}^{N,H+F}, \Xi_{ih} \equiv \frac{\Pi_{ih} - T_{ih}}{\Pi_i - T_i}, \Xi_{if} \equiv 0$
Capital ownership	$\mathcal{Z} \in \mathbb{R}^{F,H}, \mathcal{Z}_{fh} \equiv \frac{R_{fh} K_{fh}}{R_f K_f}$
Factor supply elasticities	$\Phi \equiv \mathrm{diag}(\gamma_1,\ldots,\gamma_H,0,\ldots,0)\mathcal{W}\mathcal{L}$ $+\mathrm{diag}(\varphi_1^L,\ldots,\varphi_H^L,\varphi_1^K,\ldots,\varphi_F^K)$

Source: Authors' calculations.

The elasticity of substitution of sector i between inputs j and k is defined as

$$\theta^i_{jk} \equiv \frac{\mathrm{dlog}\dfrac{X_{ij}}{X_{ik}}}{\mathrm{dlog}\dfrac{P_j}{P_k}}. \tag{21}$$

With a slight abuse of notation, here X_{ij} and X_{ik} indifferently denote either primary factors or material inputs. Table 2 introduces useful derived parameters.

Factor supply and ownership shares. Table 3 summarizes the parameters which govern the supply of labor and capital, and the allocation of income from profits and capital assets.

In table 3 we denote the $N_w + N_f \times N_w + N_f$ matrix

$$\mathcal{W} \equiv \begin{pmatrix} I & \mathcal{Z}^T \left(I + \Phi_K\right)^{-1} \\ \mathbb{O} & \left(I + \Phi_K\right)^{-1} \end{pmatrix}, \tag{22}$$

whose elements $\mathcal{W}_{hn}$ can be interpreted as the fraction of factor n's income which accrues to factor h. Specifically, income from labor is entirely earned by the workers (corresponding to the identity matrix on the top left), while income from capital assets is divided between the profits of investment producers, which are rebated to households $\left(\mathcal{Z}^T\left(I + \Phi_K\right)^{-1}\right)$ and investment expenditures $\left(\left(I + \Phi_K\right)^{-1}\right)$.

Price rigidity parameters. As explained below, sectoral inflation rates depend on an increasing and convex function $\hat{\delta}_i$ of the Calvo parameters δ_i, given by

$$\hat{\delta}_i\left(\rho, \delta_i\right) \equiv \frac{\delta_i\left(1 - \rho\left(1 - \delta_i\right)\right)}{1 - \rho\delta_i\left(1 - \delta_i\right)}. \tag{23}$$

We denote by Δ the diagonal matrix collecting sectoral parameters $\left\{\hat{\delta}_i\right\}_{i=1}^N$.

Table 4. Model Variables

Employment gaps	$\ell_t = \begin{pmatrix} l_{1t} & \dots & l_{Ht}, & u_{1t} & \dots & u_{Ft} \end{pmatrix}^T$
Factor prices	$\mathbf{w} = \begin{pmatrix} w_1 & \dots & w_H, & r_1 & \dots & r_F \end{pmatrix}^T$
Good price inflation	$\pi_t = \begin{pmatrix} \pi_{1t} & \dots & \pi_{Nt} \end{pmatrix}^T, \pi_{it} \equiv p_{it} - p_{i,t-1}$

Source: Authors' calculations.

Notation. We include factor marketplaces and retailers among the production sectors. With this modeling choice, the matrices α, β and Ω take the specific form illustrated below. We ordered the good-producing sectors first, then the factor marketplaces, and lastly the final consumption and investment retailers.

$$\Omega = \begin{bmatrix} \Omega_P & \Omega_M & \mathbb{O}_{N,N_w+N_f} \\ \mathbb{O}_{N,N_w+N_f} & \mathbb{O}_{N_w+N_f} & \mathbb{O}_{N_w+N_f} \\ \Omega_C & \mathbb{O}_{N_w+N_f} & \mathbb{O}_{N_w+N_f} \end{bmatrix}, \alpha = \begin{bmatrix} \mathbb{O}_{N,N_w+N_f} \\ I \\ \mathbb{O}_{N_w+N_f} \end{bmatrix}, \beta = \begin{bmatrix} \mathbb{O}_{N,N_w+N_f} \\ \mathbb{O}_{N_w+N_f} \\ I_{N_w+N_f} \end{bmatrix} \quad (24)$$

Variables. Table 4 introduces the variables. Lower-case letters denote log deviations from a steady state with no monetary shocks.

Remark 3. Table 4 defines employment and utilization gaps for primary factors, rather than for each sector. These are sufficient to characterize the evolution of prices as well.

Definition 5. The aggregate output gap is given by

$$\bar{y} = \sum_{h \in \mathcal{N}_h} \Lambda_h l_h + \sum_{f \in \mathcal{N}_f} \Lambda_f u_f . \quad (25)$$

2. Cross-Sectional Multipliers

Following Rubbo (2023), the employment of primary factors changes according to the cross-sectional employment multiplier

$$\tilde{\ell} = \left(I - \mathcal{X}_\ell\right)^{-1} \left[1\bar{y} + \mathcal{D}_p \left(I - \mathcal{V}\right)\left(\boldsymbol{p}_{t-1} + \rho \mathbb{E}\boldsymbol{\pi}_{t+1} \right) \right]. \quad (26)$$

Equation (26) isolates a direct effect of increasing the aggregate output gap on cross-sectional employment (given by vector 1) and a

multiplier $\left(I-\mathcal{X}_{\ell}\right)^{-1}$ that captures the propagation of the shock in general equilibrium. Intuitively, through the direct effect, final users expand their demand for all goods proportionately. However even proportional changes in demand can have an effect on relative prices because the wage of inelastically supplied factors increases and the relative price of more flexible sectors also increases. In turn, this increases the demand for goods and factors that have become relatively cheaper.

This feedback—from demand to prices and back into demand—is captured by the multiplier

$$\mathcal{X}_{\ell} \equiv \mathrm{Cov}_{\mathbf{s}}\left(\mathcal{L}^{-1}\alpha^{T}\lambda, S^{-1}\mathcal{W}\mathcal{L}\right) + \mathcal{D}_{p}\,\kappa\Phi. \tag{27}$$

The first term in equation (27) says that employment increases for the primary factors that are used more intensively in the consumption basket of households whose income increased, as measured by the correlation between factor contents in consumption ($\alpha^{T}\lambda$) and the households' income from primary factors $\left(S^{-1}\mathcal{W}\mathcal{L}\right)$.

In the second term, the $\bar{N}\times\left(N_{w}+N_{f}\right)$ matrix $\kappa\Phi$ is the slope of sector-by-factor Phillips curves, which maps changes in the employment of primary factors into changes in prices. See the appendix for a definition of $\kappa\Phi$ in terms of primitives. The $\left(\bar{N}_{w}+N_{f}\right)\times N$ matrix $\mathcal{D}_{p}$ instead describes how changes in relative prices feed back into relative demand through income reallocation or substitution. See the appendix for a definition of $\mathcal{D}_{p}$ in terms of primitives. Intuitively, employment increases for the primary factors that are demanded by final users whose relative income increased or for factors whose relative price decreased or for whom the price of complementary factors decreased, and so on.

Employment also responds to past price changes and future expected inflation because they induce changes in current relative prices. The response is mediated by the $\bar{N}\times\bar{N}$ matrix $\mathcal{V}$ (defined in the appendix).

The income of each primary factor changes according to the cross-sectional income multiplier

$$\mathbf{w}+\ell = \left(I-\mathcal{X}_{I}\right)^{-1}\left[1\left(\bar{w}+\bar{\ell}\right) - \hat{\mathcal{D}}_{p}\left(I-\hat{\mathcal{V}}\right)\left(\boldsymbol{\chi}+\rho\mathbb{E}\boldsymbol{\pi}_{t+1}\right)\right]. \tag{28}$$

Again, monetary policy affects incomes through a direct effect and a general equilibrium multiplier

$$\mathcal{X}_{I} \equiv \mathrm{Cov}_{s}\left(\mathcal{L}^{-1}\alpha^{T}\lambda, S^{-1}\mathcal{W}\mathcal{L}\right) + \hat{\mathcal{D}}_{p}\,\hat{\kappa}\Phi. \tag{29}$$

There are two key differences between the employment and income multipliers. First, the elasticity of factor and good prices with respect to income is smaller than with respect to employment. This is intuitive: both changes in employment and income shift the labor-supply curve through a wealth effect, but changes in employment also move workers along their labor supply-curve. Second, unlike employment, income shares are not always decreasing in price. This is captured by the matrix $\hat{\mathcal{D}}_p$, which is different from $\mathcal{D}_p$ in (26). While higher prices reduce factor demand, they also increase income. These two effects exactly offset when demand is Cobb-Douglas. The direct effect on income prevails—and income shares are increasing in relative prices—when primary factors are complementary (in an aggregate sense). Vice versa, income shares are decreasing in relative prices when primary factors are substitutes.

Section 4 computes the employment and income multipliers in (26) and (28) for 50 demographic groups in the Chilean population, highlighting significant heterogeneity in the incidence of monetary policy.

3. Calibration

3.1 Data

We use confidential administrative data from the Chilean Internal Revenue Service (SII)) as well as detailed, disaggregated publicly available National Accounts data for Chile to calibrate several aspects of the model.

Cluster-level employment shares. We use firms' monthly payments to the unemployment insurance administrator, which reports the total labor compensation for each worker employed. A crucial characteristic of this dataset is that it contains demographic information at the workers level, such as age and gender. By using the universe of workers in this dataset in 2018, we compute the wages for quintiles of the distribution of labor compensation.

This way, we classify workers in each of the 111 industries into 50 clusters according to their combination of gender, labor compensation quintiles, and age categories: 18–24, 25–34, 45–54, and over 54 years old. We drop all workers under 18 and those over retirement age—60 for women and 65 for men. Then, we compute the whole-sample average of the share of workers in each cluster from the total number of individuals employed.

Cluster-level labor shares. By using the same classification as for employment shares described above, we compute the industry-level average of the share of labor compensation of each cluster over the total labor compensation for firms in each industry. We normalize these shares such that the total industry-level labor shares coincide with the same object computed from the 2017 National Accounts Input-Output Matrix data for the same 111 industry classification.[5] As the model is a closed economy, we adjust totals by excluding exports and imports.

Industry-level Input-Output linkages. From the 2017 Input-Output Matrix data, we compute the share of each industry in the total of intermediate inputs purchased from each of the 111 industries. These shares are normalized such that their sum for a given buying industry coincides with the share in costs of intermediate inputs reported in the National Accounts once exports and imports are excluded.

Industry-level capital shares. These shares are computed as the difference between one and the sum of the industry-level labor and intermediate inputs shares described above.

Industry-level consumption shares. Also the 2017 Input-Output Matrix data report the total use of industrial production as final consumption. We simply transform them to shares once exports and stock variation are excluded.

Industry-level nominal price rigidity. We use confidential V.A.T. electronic invoices at the transactional level from 2015 to 2022—the results are robust to exclude the COVID episode, the 2020–2022 period. This dataset reports quantities and prices of all products involved in firm-to-firm transactions. We interpret the degree of nominal price rigidity at the industry level as the industry average of the frequency of price adjustment. To compute it, we first follow the algorithm proposed by Acevedo and others (2022), which identifies individual products by matching the text field included in VAT invoices for the description of all products. Then we collapse this information at the daily frequency using the intra-day mode to obtain a daily time series of prices for each identified product for each selling firm. We compute the frequency of price changes each month, as the number of varieties with price changes divided by the total number of varieties in the firms' sales of each product. Then we classify firms into 111 industries using the official CAE-111 classification for National Accounts in Chile. Using this classification, we compute the average

5. https://www.bcentral.cl/contenido/-/detalle/cuentas-nacionales-chile-2013-2019-2019

frequency of price changes at the industry level by weighting each product sold by each selling firm by their share of nominal sales for the whole sample.[6]

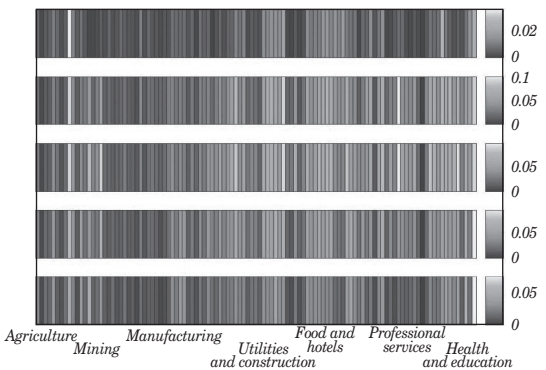
3.2 Summary Facts

We now discuss heterogeneity across Chilean population groups.

Employment shares. First, we show that different groups are likely to be employed by different industries. Figures 1 through 4 plot heatmaps of the probabilities that each group is hired by each sector. Different subplots correspond to different groups (either by income or by education), and different columns correspond to different industries

Price rigidity across industries. Figure 5 displays a bar chart of price adjustment probabilities across sectors, which shows substantial heterogeneity. Below, we show that different workers have different exposure to sticky-price vs. flex-price sectors through the employment and consumption channels.

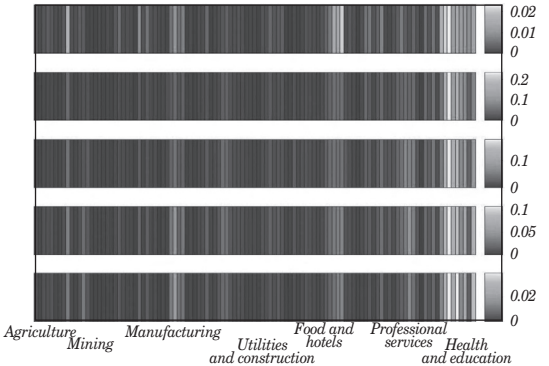
Figure 1. Employment by Age – Men



Source: Authors' calculations.
Note: Rows correspond to age quintiles, columns correspond to industries. Lighter colors correspond to higher employment shares.

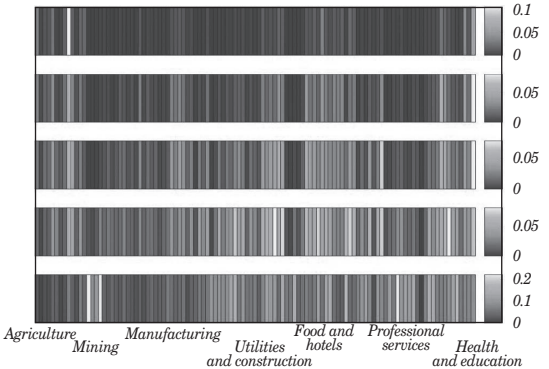
6. For those sectors for which electronic invoices do not have good information on prices, mostly services, we use data from the National Statistics Institute (INE) from January 2019 to March 2022 to calculate the frequency of price changes.

Figure 2. Employment by Age – Women



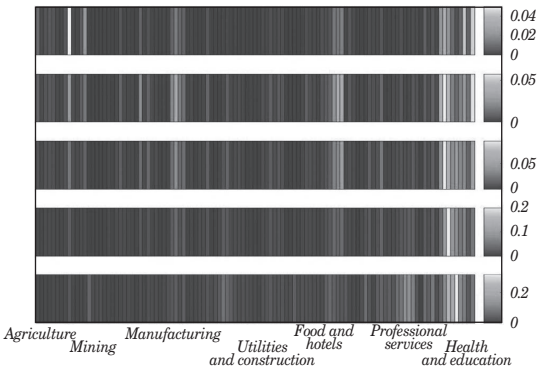
Source: Authors' calculations.
Note: Rows correspond to age quintiles, columns correspond to industries. Lighter colors correspond to higher employment shares.

Figure 3. Employment by Income – Men



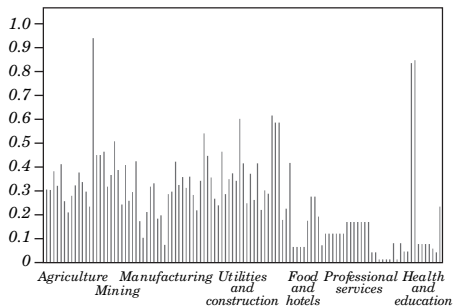
Source: Authors' calculations.
Note: Rows correspond to income quintiles, columns correspond to industries. Lighter colors correspond to higher employment shares.

Figure 4. Employment by Income – Women



Source: Authors' calculations.
Note: Rows correspond to age quintiles, columns correspond to industries. Lighter colors correspond to higher employment shares.

Figure 5. Price Adjustment Probabilities by Industry

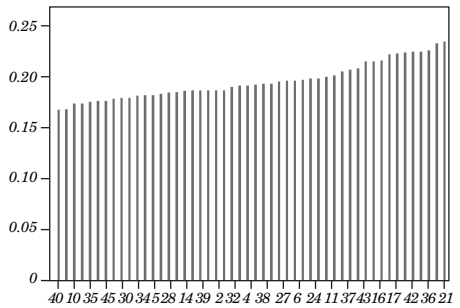


Source: Authors' calculations.

Price rigidity and workers. Figures 6 and 7 show which clusters of workers are more exposed to price rigidity through final consumption and through employment. As we discussed in section 2, the effect of monetary policy across different households depends crucially on these two objects.

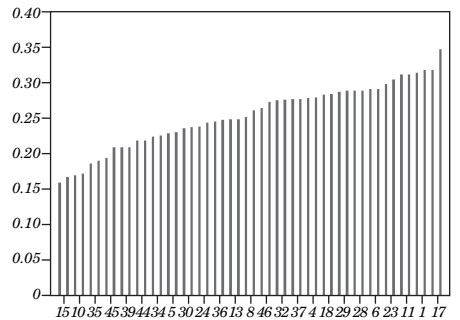
In figure 6 we display price rigidity of the consumption basket, weighting price rigidity of each economic sector by the consumption share of that sector in the consumption of goods of a given cluster. We find limited heterogeneity in this dimension.

Figure 6. Price Rigidity of Consumption Basket



Source: Authors' calculations.
Note: Cluster numbers on the x-axis correspond to men (1-25) or women (25-50), in ascending order by age groups, and by income quintile within age groups.

Figure 7. Price Rigidity of Employer Industries

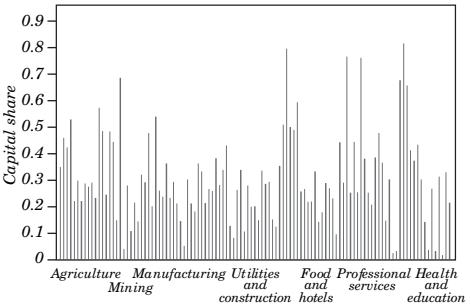


Source: Authors' calculations.
Note: Cluster numbers on the x-axis correspond to men (1-25) or women (25-50), in ascending order by age groups, and by income quintile within age groups.

Figure 7 displays price rigidity of the hiring sector, weighting price rigidity of each economic sector by the probability that a given cluster is hired by that sector. We find important heterogeneity as employer's price stickiness varies from 15 percent to 35 percent across clusters.

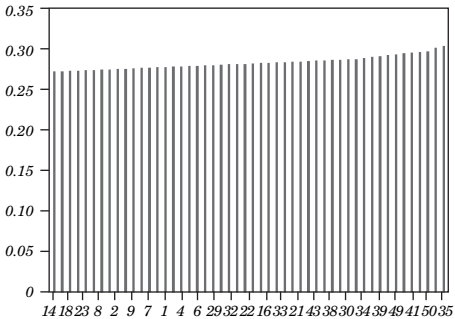
Capital shares across industries and workers. Finally, figure 8 displays a bar chart of capital shares across industries, showing substantial heterogeneity. Nonetheless, figure 9 shows that this generates only limited variation in the capital intensity of industries that hire different worker groups. This result suggests that interacting industry-by-demographic group would imply even more heterogeneous responses to monetary policy than we find in section 4.

Figure 8. Capital Shares across Industries, Sorted in Ascending Order



Source: Authors' calculations.

Figure 9. Capital Share of Employer Industries



Source: Authors' calculations.
Note: Cluster numbers on the x-axis correspond to men (1-25) or women (25-50), in ascending order by age groups, and by income quintile within age groups.

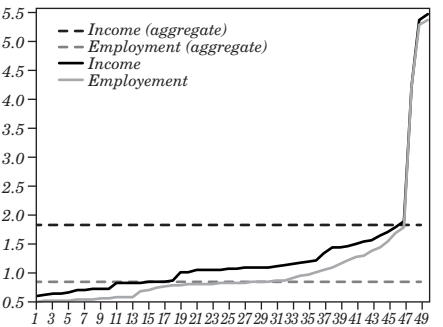
4. Quantitative Results

Figure 10 shows cumulative employment and income responses to a real money supply shock. The responses are normalized by the cumulative size of the shock and sorted in increasing order across demographic groups.

The employment responses display vast heterogeneity, ranging from 0.5 percent to more than 5 percent. The solid lines show employment and income of households only, while the dashed lines correspond to aggregate employment and income including capital assets.

Figure 11 compares cross-sectional and aggregate employment responses in the baseline calibration versus an otherwise identical model with no input-output linkages. As it is well known, the presence of input-output linkages increases monetary non-neutrality (i.e., it amplifies the response of employment to monetary shocks). In the Chilean economy, however, the effect of input-output linkages is very small due to a much smaller input-output multiplier than, for example, the United States.[7]

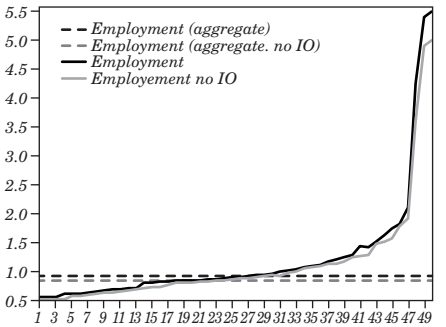
Figure 10. Cumulative Impulse Responses of Employment and Income to Monetary Shock



Source: Authors' calculations.

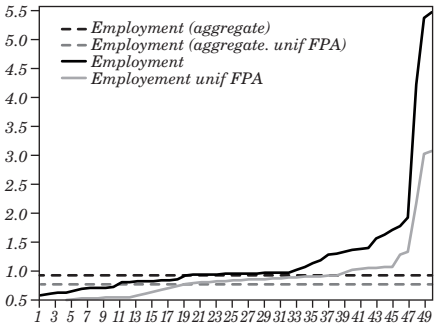
7. See Pastén and others (2019).

Figure 11. Employment Responses with and without Input-Output Linkages



Source: Authors' calculations.

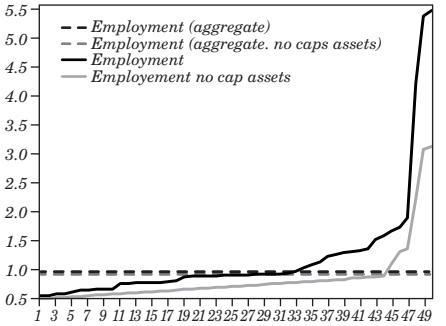
Figure 12. Employment Responses in the Baseline Model and in a Counterfactual Model with Uniform Price Adjustment Probabilities



Source: Authors' calculations.

Instead, figure 12 shows that heterogeneity in price adjustment frequencies across sectors is a major driver of cross-sectional non-neutrality. As it is well known from input-output New Keynesian models with a representative household, heterogeneous price adjustment frequencies increase the aggregate non-neutrality. Figure 12 shows that it also amplifies the cross-sectional dispersion of employment responses, as employment increases by more for demographic groups that are employed in sticky-price industries.

Figure 13. Employment Responses with and without Capital Assets



Source: Authors' calculations.

Finally, figure 13 compares cross-sectional and aggregate employment responses in the baseline calibration versus a model with no capital assets. The figure shows that ignoring the presence of semifixed capital assets would lead to underestimating the cross-sectional range of employment responses.

5. Conclusion

By using rich microdata from the Central Bank of Chile, we showed that Chilean households in different demographic groups are subject to different degrees of wage rigidity and are employed in industries with different price rigidity and capital shares. We then used the framework in Rubbo (2023) to compute the implications of heterogeneity for the response of employment and income to monetary policy across households. The model predicts significant heterogeneity in the response of employment and income across households, with larger employment effects for households that are employed by sticky-price sectors. These households have higher incomes and are in the higher age quintiles.

References

Andersen, A., E.T. Hansen, K. Huber, N. Johannesen, and L. Straub. 2022. "Disaggregated Economic Accounts." Technical Report W30630, National Bureau of Economic Research.

Aoki, K. 2001. "Optimal Monetary Policy Responses to Relative-Price Changes." *Journal of Monetary Economics* 48(1): 55–80.

Auclert, A. 2019. "Monetary Policy and the Redistribution Channel." *American Economic Review* 109(6): 2333–67.

Auclert, A., B. Bardoczy, M. Rognlie, and L. Straub (2021). "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models." *Econometrica* 89(5): 2375–408.

Baqaee, D.R. and E. Farhi. 2018. "Macroeconomics with Heterogeneous Agents and Input-Output Networks." NBER Working Paper No. 24684.

Baqaee, D.R. and E. Farhi. 2020. "Productivity and Misallocation in General Equilibrium." *Quarterly Journal of Economics* 135(1): 105–63.

Basu, S. 1995. "Intermediate Goods and Business Cycles: Implications for Productivity and Welfare." *American Economic Review* 85(3): 512–31.

Benigno, P. 2004. "Optimal Monetary Policy in a Currency Area." *Journal of International Economics* 63(2): 293–320.

Carvalho, C. 2006. "Heterogeneity in Price Stickiness and the Real Effects of Monetary Shocks." *B.E. Journal of Macroeconomics* 6(3): 1–58.

Coglianese, J., M. Olsson, and C. Patterson. 2023. "Monetary Policy and the Labor Market: A Quasi-Experiment in Sweden. University of Chicago, Becker Friedman Institute for Economics Working Paper No. 2023–123.

Devereux, M.B. and C. Engel. 2003. "Monetary Policy in the Open Economy Revisited: Price Setting and Exchange-Rate Flexibility." *Review of Economic Studies* 70(4): 765–83.

Gali, J. and T. Monacelli. 2008. "Optimal Monetary and Fiscal Policy in a Currency Union. *Journal of International Economics* 76(1): 116–32.

Huang, K.X.D. and Z. Liu. 2007. "Business Cycles with Staggered Prices and International Trade in Intermediate Inputs." *Journal of Monetary Economics* 54(4): 1271–89.

Kaplan, G., B. Moll, and G.L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

La'O, J. and A. Tahbaz-Salehi. 2019. "Optimal Monetary Policy in Production Networks." Working Paper No. 3488415, SSRN Electronic Journal.

Minton, R. and B. Wheaton. 2023. "Minimum Wages and the Rigid-Wage Channel of Monetary Policy." American Economic Association. American Economic Association.

Nakamura, E. and J. Steinsson. 2010. "Monetary Non-Neutrality in a Multisector Menu Cost Model." *Quarterly Journal of Economics* 125(3): 961–1013.

Pastén, E., R. Schoenle, and M. Weber. 2019. "The Propagation of Monetary Policy Shocks in a Heterogeneous Production Economy." *Journal of Monetary Economics* 116: 1–22.

Rubbo, E. 2020. "Networks, Phillips Curves, and Monetary Policy." *Econometrica* 91(4): 1417–55.

Rubbo, E. 2023. "Monetary Non-Neutrality in the Cross-Section." Working paper.

The Bank Lending Channel Across Time and Space

Dean Corbae
University of Wisconsin-Madison
National Bureau of Economic Research

Pablo D'Erasmo
Federal Reserve Bank of Philadelphia

We build a model of banking industry dynamics with imperfect competition to address the following question: How does monetary policy affect lending outcomes across time, given the spatial expansion of the banking industry?

Geographic expansion of the banking industry followed from the elimination of cross-state branching restrictions begun in the McFadden Act of 1927, which permitted national banks to branch only to the same extent as state banks, thus giving the states ultimate authority. While some states permitted such cross-state branching prior to 1994, the Riegle-Neal Act removed several obstacles to banks opening branches in other states and provided a uniform set of rules regarding banking in each state.

As we document in our data in section 1, following the signing of the Riegle-Neal Act there was rapid expansion of banks crossing state lines. What is interesting is how it has translated into banking concentration. Specifically, we document that by the mid-2000s the cross-section (top 4, top 5 - 35, top 36 - 2 percent) of commercial bank cross-state deposit expansion diverged significantly. This geographic expansion coincides with the rise of U.S. bank concentration and Herfindahl indices at the national and state levels.

The authors wish to thank Jean-Francois Houde and David Moreno for helpful comments and seminar participants at the Bank of Chile. The views expressed in this paper do not necessarily reflect those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

In section 1 we also document that, coincident with this geographic expansion, there have been important changes in the variance of deposit inflows, loan returns, and interest margins across bank sizes. An economy subject to shocks that are not perfectly correlated across space can explain these facts through geographic diversification. We also document that average costs decrease with bank size (suggestive of increasing returns to scale) which have been falling over time. These facts are consistent with a model of banking along the lines of the Diamond (1984) delegated monitoring model.[1]

Finally, section 1 examines how the bank lending channel along the lines of Kashyap and Stein (1995) and Kashyap and Stein (2000) has changed in the pre- versus post-reform data samples. We find that smaller banks exhibit greater sensitivity of their loan supply to increases in fed funds rates, consistent with a corporate finance view that smaller firms are subject to higher external finance costs. Across time we see that bigger banks have become less sensitive, while smaller banks have become more sensitive to policy hikes.

In sections 2 and 3, we build a model consistent with many of these facts. At its heart is a model of banking industry dynamics with imperfect competition as in some of our earlier work Corbae and D'Erasmo (2020) and Corbae and D'Erasmo (2021).[2] Along the lines of the 2020 paper, banks endogenously climb a size ladder consistent with higher costs to grow across space. The idea is that all banks start as state banks and Riegle-Neal lowered the cost of branching out on a regional and national basis. The equilibrium distribution of banks on the Besanko and Doraszelski (2004) ladder is solved by using the approximation techniques of Farias and others (2012).

After parameterizing the model in Section 4, we then use it to assess how geographic expansion over time affects the bank lending channel of monetary policy in Section 5. Geographic expansion has led to a skewed bank size distribution where big national banks account

1. Specifically, Diamond provides a framework where large banks arise to economize on the fixed costs of monitoring individual borrowers more efficiently than a large number of small depositors. Economies of scale in monitoring (decreasing average costs) induce size. The problem of monitoring the monitor is also solved by size; large, diversified banks can offer noncontingent (and hence incentive-compatible) deposit contracts. There are numerous empirical papers documenting the existence of scale economies in banking such as Berger and Mester (1997) or Berger and Hannan (1998). A large pool of depositors is also consistent with geographic diversification as described in Liang and Rhoades (1988).

2. A closely related paper by Aguirregabiria and others (2020) studies how geographic dispersion may prevent funding from flowing to high loan-demand areas. Also closely related is Gelman and others (2022) as well as Morelli and others (2023).

for a large share of the loan market. Monetary policy which raises the cost of external funding, e.g., a rise in fed funds, can have differential effects on the lending behavior of banks of different sizes along the lines of Kashyap and Stein (1995) distributed across different regions. Changes in monetary policy are a blunt instrument because it affects all regions instead of the affected region. As in Bellifemine and others (2022) and Wang and others (2022), we study the transmission of monetary policy in a model with bank heterogeneity and imperfect competition. We incorporate spatial differences and assess the bank lending channel across time. In particular, we conduct a counterfactual in Section 5 where we raise the cost of external finance and examine how it affects banks of different sizes across time.

1. Stylized Facts

In this section we present some data facts for a cross-section of the top 2 percent of banks across time. The data come from both the Fed's Consolidated Reports of Condition and Income for Commercial Banks, regularly called "Call Reports", which begin in 1984, and the Summary of Deposits of the Federal Deposit Insurance Corporation (FDIC), which begins in 1994.

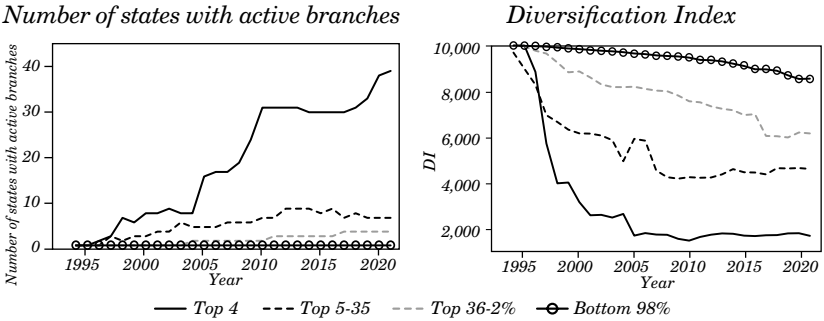
We examine geographic expansion following Riegle-Neal starting in 1994 in figure 1. The top panel graphs the number of states where a bank in each size category has an active branch. The bottom panel plots a diversification index. Let $\ell_{i,m,t}$ denote the amount of loans originated by lender i in market m in period t. Here we take m to be a state. The share of loans of lender i in state m in period t is $S_{i,m,t} = \frac{\ell_{i,m,t}}{\Sigma m \in M_{i,t} \ell_{i,m,t}}$ x 100 where $L_{i,t} = \Sigma m \in M_{i,t} \ell_{i,m,t}$ is the total amount of loans originated by lender i in period t and $M_{i,t}$ denotes the states in which lender i operates. We define a diversification index as follows:[3]

$$DI_{i,t} = \sum_{m \in M_{i,t}} s_{im,t}^2. \tag{1}$$

This index ranges between 0 and 10,000, and a smaller value indicates a more diversified lender. The bottom panel in figure 1 shows the (deposit-weighted) average of this diversification index within size categories. It is clear that there is a positive relationship between size and geographic diversification.

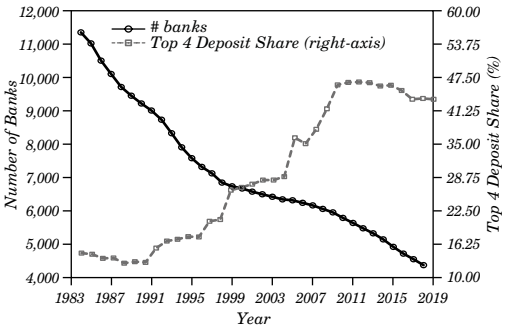
3. This measure is applied at the county level in Shin (2022).

Figure 1. Deposit Space and Size over time



Source: Source: Summary of Deposits.
Note: Banks are ranked according to deposits. Source: Summary of Deposits.

Figure 2. U.S. Banking Concentration

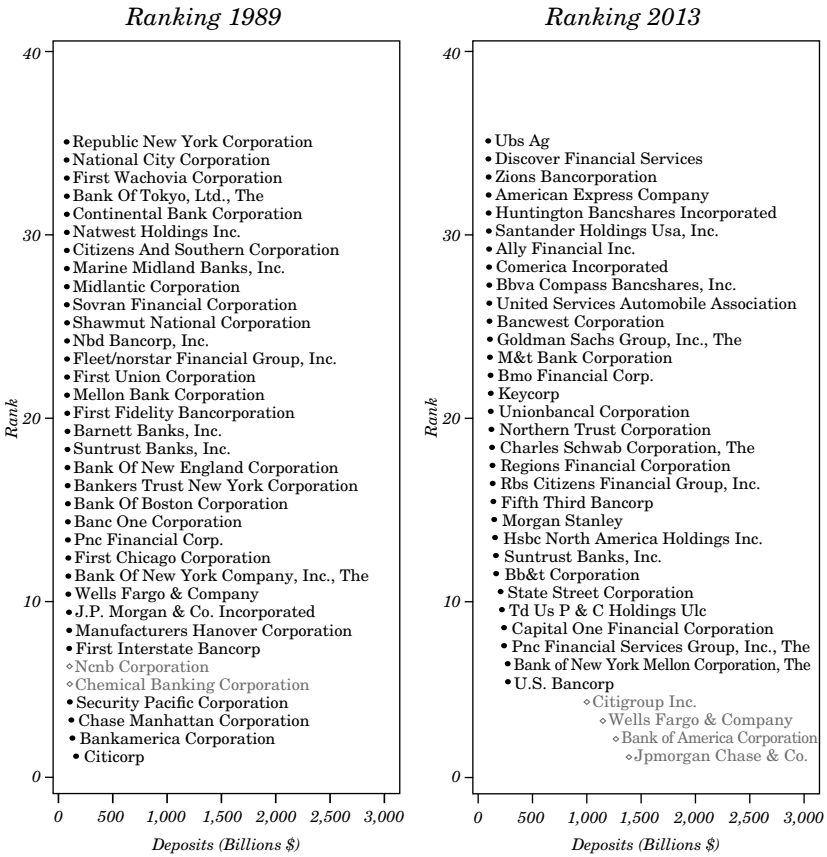


Source: Reports of Condition and Income (Call Reports).
Note: Banks are ranked according to deposits.

Figure 2 graphs the deposit market share of the top 4 banks across time. It shows that, prior to Riegle-Neal, market shares were relatively constant; then it shows a rapid transition following Riegle-Neal until the 2008 Financial Crisis, followed by another relatively constant share. Importantly, the fact that deposit shares in figure 2 are relatively constant during both the 'pre-reform' 1984–1992 and 'post-reform' 2011–2019 periods motivates our modeling choice of calibrating parameters consistent with long-run equilibria of those periods.

Figure 3 graphs how the top 35 individual banks ranked according to deposit size grew over time. Notable is the divergence of the top four from even the remaining top 5 - 35. This motivates our decision to categorize banks into three size bins (top 4, top 5 - 35, and top 36 - 2 percent).

Figure 3. Deposit Distribution Pre- and Post-Reform



Source: Reports of Condition and Income (Call Reports).
Note: Banks are ranked according to deposits. Deposits are in reported in real terms. Red bubbles identify banks that end up in the Top 4 of the distribution post-reform (2013).

As banks expand over space, they diversify shocks over their region. Here we explore how the variance of deposit inflows, loan returns, and interest margins vary across bank size and time (pre-reform 1984–1992 and the stationary portion post reform 2011–2019). Using our panel of commercial banks in the U.S., we estimate how that process of deposits evolves for bank holding companies of different sizes. We keep the same grouping convention that we described above. After controlling for firm and year fixed effects as well as a time trend, we estimate the following autoregressive process for log-deposits for bank i of type $\Theta \in \{s,r,n\}$ in period t:

$$\log\left(d^i_{\theta,t}\right) = \left(1-\rho^d_\theta\right)\overline{d}_\theta + \rho^d_\theta \log\left(d^i_{\theta,t-1}\right) + u^i_{\theta,t}, \tag{2}$$

where $d^i_{\theta,t}$ is the sum of deposits and other borrowings in period t for bank i, and $u^i_{\theta,t}$ is iid and distributed $N(0,\sigma^2_{\theta,u})$. Assuming the process is stationary, the variance of the deposit process is given by $\sigma_\theta = \frac{\sigma_{\theta,u}}{\left(1-\left(\rho^d_\theta\right)^2\right)^{1/2}}$. Since this is a dynamic model, we use the method proposed by Arellano and Bond (1991). Consistent with the evidence presented in figure 2, we estimate this process for the pre-reform period (1984–1992) and for the latest period in our sample (2009–2018).

Table 1 presents the results. The top panel provides estimates for the period prior to the passage of Riegle-Neal, which constrained bank branching to the state level (i.e. $\theta = s$), while the bottom panel provides estimates for a period of market share stability following Riegle-Neal, where some banks crossed their state borders to grew from $\theta = s$ to regional ($\theta = r$) and national ($\theta = n$) levels. Consistent with the diversification story in Diamond (1984) and empirical work such as Liang and Rhoades (1988), we find that the variance of deposit inflows σ_θ decreases as banks grow in size.

Corbae and D'Erasmo (2023) conduct the same analysis for loan returns, interest margins, and charge-off rates. Again it is apparent that the variance σ_θ of these variables decreases as banks grow in size. One interesting fact from that analysis is how close the average loan returns and margins are, which motivates our use of a Cournot-equilibrium concept with capacity constraints.

Table 1. Deposit Process Parameters

Size Group		*Pre-Reform (1984-1992)*			
Data	Model	d_θ	ρ_θ	$\sigma_{u,\theta}$	σ_θ
Top 2%	s	0.140	0.863	0.1773	0.3506

Size Group		*Post-Reform (2011-2019)*			
Data	Model	d_θ	ρ_θ	$\sigma_{u,\theta}$	σ_θ
Top 4	n	10.563	0.699	0.0306	0.0428
Top 5 - 35	r	1.000	0.764	0.0861	0.1333
Top 36 - 2%	s	0.138	0.761	0.1034	0.1595

Source: Call Reports.

Note: We study banks in the top 2 percent of the asset distribution. We group all banks ($\theta = s$) for the pre-reform period since regulation prevented them from expanding across state borders. For the post-reform period, we split this group and consider top 4 ($\theta = n$), top 5 - 35 ($\theta = r$), and top 36-2% ($\theta = s$). Average deposits are normalized to 1 for the top 5 - 35 group in the post-reform period. Average deposits d_θ is reported relative to this group.

With regard to the bank lending channel, we follow Kashyap and Stein (1995) and run the following specification:

$$\Delta y_{i,t} = \sum_{h=1}^{8} \beta_h \left(f_{t-h} - f_{t-h-1} \right) + \sum_{h=1}^{4} \alpha_h \Delta y_{i,t-h} + \gamma_j X_t + \phi x_{i,t} + a_i + \tau_t + Q_t + \in_{i,t}, \quad (3)$$

where $\Delta y_{i,t}$ denotes the growth rate of $y_{i,t}$ (loans) between quarter t and quarter $t-1$, f_{t-h} corresponds to the fed funds rate in period $t-h$, X_t captures aggregate variables (such as the inflation rate or changes in nominal GDP), $x_{i,t}$ are bank level controls that include the ratio of deposits to assets, the ratio of equity to assets, the ratio of cash and securities to assets. a_i is a bank fixed effect, τ_t is a year fixed effect, and Q_t is a quarter fixed effect.[4] Table 2 reports the value of $\Sigma_{h=1}^{8} \beta_h$ together with the p-value of the corresponding test of significance for the sum.

Table 2. Bank Lending Channel: Pre & Post by Bank Size

	Dep Var: Growth Loans	
	Pre-Reform	*Post-Reform*
Top 4	-0.2919	0.6480
Top 5 - 35	(0.22)	(0.76)
p-value	-0.2219	-1.2149
Top 36 - 2%	(0.03)	(0.07)
p-value	-0.1008	-1.3545
All (Top 2%)	(0.01)	(0.01)
p-value	-0.1212	-1.4272
ΔCPI	(0.00)	(0.00)
DeltaNGDP	yes	yes
Other Bank Controls	yes	yes
Bank FE	yes	yes
Period	84-92	11-19

Source: Reports of Condition and Income (Call Reports) and FRED, Federal Reserve Economic Data.
Note: *p*-values (for sum of coefficients) in parenthesis. Other bank controls include ratio of deposits to assets, the ratio of equity to assets, the ratio of cash and securities to assets.

4. See Appendix A.2 for a description of the data used to perform this analysis.

Importantly, table 2 documents that bigger banks are less sensitive to a rise in the fed funds rate than smaller banks across both pre-reform (1984–1992) and post-reform (2011–2019) periods. Specifically, the coefficients for top 4 banks are insignificant across both time periods and top 5 - 35 become insignificant at the five-percent level post reform. Interestingly, top 36 – 2 percent banks become more sensitive to rises in the fed funds rate. These results are consistent with the logic that bigger, more diversified banks have access to other sources of external funding and hence less sensitive to external funding via fed funds, as in Kashyap and Stein. Except for the smallest banks, the argument that more diversification through time makes banks less sensitive to fed funds shocks is also consistent with this idea.

In summary, expanded data analysis in Corbae and D'Erasmo (2023) documents:

1. Diversification across space grows with bank size over time as in figure 1.

2. The growth of concentration of the top 4 banks across time as in figures 2 and 3.

3. Relative stability of concentration prior to Riegle Neal in 1994 (i.e., 1984–1992) and after the Global Financial Crisis (i.e., 2011–2019), as is evident in figure 2.

4. Deviations from Zipf's law arising from growth of the right tail.

5. Rising deposit Herfindahl Indices at the national and state levels. The current average of state-level Herfindahl indices falls into the "moderately concentrated" designation by the Justice Department's Antitrust Division. The Herfindahl has grown by 80 percent from 1994. It also documents relative stability of Herfindahl indices prior to Riegle Neal in 1994 and after the Global Financial Crisis.

6. The growth in insured deposit funding and drop in variance of inflows across time by bank size, as is evident in table 1. The drop in variance associated with geographic expansion suggests geographic diversification.

7. The drop in variance of loan returns, interest margins, and charge-off rates respectively across time by bank size, again suggestive of geographic diversification.

8. The drop in average costs across time by bank size. The fact that average costs drop by bank size is consistent with increasing returns to scale at least over certain size ranges.

9. The cyclical nature of bank exit.

These data facts motivate our modeling choices:

1. Diversification and increasing returns are consistent with the delegated monitoring model of banks in Diamond (1984).

2. High levels of concentration motivate us to model the banking industry as imperfectly competitive.

3. We model bank growth with imperfect competition via a ladder along the lines of Besanko and Doraszelski (2004).

(a) We assume that pre-Riegle-Neal regulation made the cost of geographic expansion infinite while post-Riegle-Neal costs fall so that there is growth from state to regional to national consistent with figure 1.

(b) As banks grow, they expand their capacity and lower their variance of low-cost deposit inflows. We model this by bank-size-dependent Markov processes for exogenous deposit inflows consistent with the data analysis in table 1.

(c) As banks grow, they also bear lower costs of nondeposit external funding along the lines of standard models of corporate finance.

4. Banks Cournot-compete in the loan market subject to deposit capacity constraints. They also must compete with the nonbank funding sector.

5. There is endogenous bank exit across the business cycle (modeled here by the aggregate shock process). More variable interest margins and charge-off rates make smaller banks more susceptible to failure, especially in downturns.

(a) Endogenous bank exit also allows us to examine how monetary and regulatory policy can affect the bank size distribution and financial stability.

2. Model Environment

Time is discrete and there is an infinite horizon. There are two regions $j \in \{\mathcal{E}, \mathcal{W}\}$, for instance, 'east'" and 'west'. Each period, a mass B of ex-ante identical entrepreneurs who have a profitable project that needs to be funded (the potential borrowers) are born in each region. There is also a mass $H > B$ of identical households (the potential depositors) in each region that deposit their funds in the banking sector and finance banks and nonbanks $k \in \{\mathcal{B}, \mathcal{N}\}$ via equity injections, where $\mathcal{B}$ denotes traditional banks and $\mathcal{N}$ nonbanks. Financial intermediaries (banks and nonbanks) intermediate between potential borrowers and depositors.

To keep notation manageable, we let any beginning-of-period variable be denoted x and any end-of-period variable be denoted x'. Further, except where critical to understand the problem, we will not index by region j. For example, any decision rule taken in region j should be understood to depend on j.

2.1 Entrepreneurs

Ex-ante identical borrowers in region j demand loans in order to fund a risky project. The project requires one unit of investment (i.e., a loan either from a bank $k = \mathcal{B}$ or nonbank $k = \mathcal{N}$) at the beginning of period t. The entrepreneur chooses the scale R_k of the risky project in which they are investing those funds, which can be indexed on the lender type.[5] The project returns R_k at the end of the period according to:

$$\begin{cases} 1 + z_j' R_k & \text{with prob } p_j\left(R_k, z_j',\right) \\ 1 - \lambda & \text{with prob}\left[1 - p_j\left(R_k, z_j'\right)\right] \end{cases} \tag{4}$$

in the successful and unsuccessful states, respectively. That is, borrower gross returns are by $1 + z_j' R_k$ in the successful state and by $1 - \lambda'$ in the unsuccessful state, where z_j' is a regional-specific shock and λ is the fraction lost in default. The regional shocks z_j' are assumed to be independent over time and drawn from a bivariate normal distribution $F_z(\mu_z, \sigma_z, \rho_z)$ where μ_z denotes the mean, σ_z the standard deviation, and ρ_z covariance between regions. The success of a borrower's project, which occurs with probability $p_j(R_k, z_j')$, is independent across borrowers and time conditional on the borrower's choice of technology $R_k \geq 0$ and regional shock z_j'.

As for the likelihood of success or failure, a borrower who chooses to run a project with a higher return R_k has more risk of failure. Specifically $p_j(R_k, z_j')$ is assumed to be decreasing in R_k and increasing in z_j'. Thus, the technology exhibits a risk-return trade-off. Further, since R_k is a choice variable, project returns and failure rates are endogenously determined. While borrowers are ex ante identical, they are ex-post heterogeneous owing to the realizations of the shocks to

5. Note this is the first occurrence of the notation simplification we alluded to above; in general, since risky scale is a choice variable of the entrepreneur in region j, we would denote it $R_{k,j}$, but we neglect the j subject to keep notation manageable.

the return on their project. We envision borrowers either as firms choosing a technology that might not succeed or households choosing a house that might appreciate or depreciate.

The entrepreneur makes a discrete choice over which type of financial institution to borrow from $k \in \{\mathcal{B}, \mathcal{N}\}$. Bank and nonbank interest rates on their loans to the entrepreneur can differ. Taking the vector of interest rates $r_j = \{r_{\mathcal{B},j}, r_{\mathcal{N},j}\}$ on loans as given, entrepreneurs decide whether they want to fund a project given their outside option and then make a discrete choice over whether to borrow from a bank or nonbank in their region.

Once with a lender type k offering a loan at interest rate $r_{k,j}$, the entrepreneur chooses the risk-return tradeoff of their project $R_{k,j}$. This explains why we allow project choice to depend on k; since the borrower potentially faces different rates from different lenders, they may make different risk-return project choices. Following Buchak and others (2018), we assume that the value associated with financing the project with each type of lender in region j is subject to an unobservable idiosyncratic shock $\in = \{\in_{\mathcal{B}}, \in_{\mathcal{N}}\}$ affecting the value of taking a loan from each type of lender additively. We assume that $\in_k$ are iid shocks drawn from a type-one extreme-value distribution $F_{\in}(\in;\alpha)$ with scale parameter $1/\alpha$.

Borrowers have an outside option. At the beginning of period t, they receive a realization of their reservation utility of consumption $\omega \in [0, \overline{\omega}]$ if they decide not to run the project. These draw from distribution function $\Omega(\omega)$ are i.i.d. over time and across regions. This outside option leads to a downward-sloping aggregate demand for loans, while the extreme-value shocks determine loan demand across financial institution types.

There is limited liability on the part of the borrower at the project level so that the project return net of interest payments is bounded below at zero. If $r_{k,j}$ is the interest rate on a loan that the borrower faces, the borrower receives $\max\{z_j' R_k - r_{k,j}, 0\}$ in the successful state and 0 in the failure state. Specifically, in the unsuccessful state they receive $1 - \lambda$, which must be relinquished to the lender. Table 3 summarizes the risk-return tradeoff that the borrower faces. Since the choice of R_k is endogenous, changes in borrowing costs $r_{k,j}$ can affect the default frequencies on loans through a risk-shifting motive.

Table 3. Borrower's Problem (Conditional on Investing)

Borrower Chooses R_j	*Receive*	*Pay*	*Probability*		
				−	+
Success	$(1+z_j')R_k)$	$(1+r_{k,j})$	p	$(R_k,$	$z_j')$
Failure	$(1-\lambda)$	$min\{(1-\lambda),(1+r_{k,j})\}$	$1-p$	$(R_k$	$z_j')$

Both R_k and ω are private information to the entrepreneur. As in Bernanke and Gertler (1989), success or failure is also private information to the entrepreneur unless the loan is monitored by the lender.[6] With one-period loans, since reporting failure (and hence repayment of $1-\lambda < 1 + r_{k,j}$) is a dominant strategy in the absence of monitoring, loans must be monitored. Monitoring is costly as in Diamond (1984).

2.2 Households

In each region j, infinitely lived, risk-neutral households with discount factor β are endowed with one unit of the good each period. We assume households are sufficiently patient such that they choose to exercise their savings opportunities. In particular, households have access to an exogenous risk-free storage technology yielding $1 + \bar{r}$ between any two periods with $\bar{r} \geq 0$ and $\beta(1 + \bar{r}) = 1$. They can also choose to supply their endowment to a bank, a nonbank, or an individual entrepreneur. We assume that, after observing the deposit interest rate $r_{D,j}$, households who choose to deposit their earnings are randomly matched with a bank in their region at the beginning of any period t. Given deposit insurance, even if the bank fails, they receive their deposit with interest at the end of the period. Households can hold a portfolio of bank stocks yielding dividends (claims to bank cash flows) and can inject equity to banks. They can also invest in shares of the representative nonbank, which gives a claim to nonbank cash flows. They pay lump-sum taxes/transfers τ' at the end of any

6. While one interpretation of our entrepreneurs is that they are effectively one-period lived (born at the beginning of the period and dead at the end as in the OG model of Bernanke and Gertler (1989), we could have effectively modeled entrepreneurs as long-lived and added enough interperiod anonymity so that financial contracts are one-period lived as in Carlstrom and Fuerst (1997) and entrepreneurs are sufficiently impatient not to want to augment net worth.

period t which include a lump-sum tax τ'_D used to cover deposit insurance for failing banks. Finally, if a household wants to match directly with an entrepreneur (i.e., directly fund an entrepreneur's project), it must compete with bank loans. Hence, the household could not expect to receive more than the bank lending rate $r_{k,j}$ in successful states and must pay a monitoring cost. Since households can purchase claims to bank cash flows, and banks can more efficiently minimize costly monitoring along the lines of Diamond (1984), there is no benefit to matching directly with entrepreneurs.

2.3 Banks

We build a model along the lines of Ericson and Pakes (1995) where, within a region, banks Cournot-compete in a single-good market (loans) and there is endogenous entry and exit. As in Diamond (1984), banks exist in our environment to pool risk and economize on monitoring costs. We assume there are three types of banks: $\theta \in \Theta = \{s, r, n\}$ with size ranking $s < r < n$. We identify banks of type s with small state banks and banks of type r with bigger regional banks, both of which are constrained to operate in only one region. We associate banks of type n with large national banks operating across regions. There can be multiple banks of each type operating and banks of all types have some degree of market power.

To save on notation, each incumbent bank with the same state variables will be treated identically. We denote loans made by such a bank of type θ in region j in period t by ℓ_θ.[7] As in Corbae and D'Erasmo (2021), bank type θ determines the mean and variance of a bank's deposits $d_\theta \in D_\theta$. In particular, banks in the model face the deposit process we estimated in equation (2). To make our definition of type consistent with the data presented in table 1, the mean of the deposit process satisfies $\bar{d}_n > \bar{d}_r > \bar{d}_s$ so that higher types have a bigger funding base. Furthermore, also consistent with the data presented in table 1, the variance of deposits satisfies $\sigma_n \leq \sigma_r \leq \sigma_s$ so that bigger banks have lower variance consistent with diversification. We discretize the continuous deposit process d_θ in equation (2) into a finite support and denote its transition matrix by $G_\theta(d'_\theta, d_\theta)$. Unlike Corbae and D'Erasmo (2021), here deposits are the only source of funding besides seasoned

7. Again, since this is a choice variable, it should be understood that ℓ_θ also depends on j and there will be places where we make that explicit.

equity. Deposits are collected at the regional level, but we assume that national banks (n) can move deposits freely across regions. Since we do not take a stand on what a 'region' is both in the model and in the estimation in equation (1), to simplify on notation, we abstract from denoting the regional origin of d_θ.

Along the lines of Besanko and Doraszelski (2004), a given bank of type θ can invest $I_\theta \in \mathbb{R}_+$ to become a larger-type bank (i.e., a small local bank can invest to become a regional bank and a medium-sized regional bank can invest to become a large national bank). One can interpret this investment technology as a reduced-form way of capturing geographic expansion in ways that can also include mergers and acquisitions. We have assumed that prior to Riegle-Neal, all banks were restricted to operate only in their home state (i.e., of type $\theta = s$). After Riegle-Neal lowers the cost of geographic expansion, any $\theta = s$ bank can then invest I_s to transit to a bigger regional type $\theta' = r$ according to the following transition function:

$$(\theta' \mid \theta = s, I_s) = \begin{cases} \dfrac{(\alpha \cdot I_s \cdot (\Delta \bar{d}_{r,s})^{-\xi}}{1+\alpha I_s \cdot (\Delta \bar{d}_{r,s})^{-\xi}} & \text{if } \theta' = r \\ \dfrac{1}{1+\alpha \cdot I_s \cdot (\Delta \bar{d}_{r,s})^{-\xi}} & \text{if } \theta' = s \end{cases}, \tag{5}$$

where the parameters $\alpha > 0$ and $\xi > 0$ measure the effectiveness of investment (at $\iota_\theta = 0$ to be precise) and $\Delta \bar{d}_{r,s} = (\bar{d}_r - \bar{d}_s) > 0$. Since banks of type s are already heterogeneous via the deposit shock process, the bigger ones may have an incentive to bear the cost of growing to r while the smaller ones may remain of type s.

After state-level banks branch out to become regional, the following transition function for a type r bank governs whether it grows to become national (n), shrinks to become state (s), or remains regional:

$$T(\theta' \mid \theta = r, I_r) = \begin{cases} \dfrac{(1-\delta)\alpha \cdot I_r \cdot (\Delta \bar{d}_{n,r})^{-\xi}}{1+\alpha I_r \cdot (\Delta \bar{d}_{n,r})^{-\xi}} & \text{if } \theta' = n \\ \dfrac{1-\delta+\delta\alpha \cdot I_r \cdot (\Delta \bar{d}_{n,r})^{-\xi}}{1+\alpha I_r \cdot (\Delta \bar{d}_{n,r})^{-\xi}} & \text{if } \theta' = r \\ \dfrac{\delta}{1+\alpha \cdot I_r \cdot (\Delta \bar{d}_{n,r})^{-\xi}} & \text{if } \theta' = s \end{cases}, \tag{6}$$

where $\Delta \bar{d}_{n,r} = (\bar{d}_n - \bar{d}_r) > 0$.[8,9]

After the realization of θ, a given incumbent bank is randomly matched with a set of potential household depositors d_θ who receive deposit interest rate $r_{D,j}$ and then decide how many loans to extend. Regional and state banks $\theta \in \{r,s\}$ can fund an amount of loans larger than its deposits by using external borrowing $a_\theta < 0$ at rate $r^a(a_\theta) > \bar{r}$. If a bank chooses an amount of loans lower than its capacity constraint, the leftover deposits a_θ can be invested in the same risk-free technology that the households have access to with return equal to $\bar{r}$. The flow constraint for such regional and small banks is $\ell_{\theta,j} + a_\theta = d_\theta$. National banks are geographically diversified in the sense that they extend loans and receive deposits in both regions ($\Sigma_j \ell_{n,j} + a_n = d_n$). Note that, since the outside option for a household matched with a bank is to store at rate $\bar{r}$, we know that $r_{D,j} \geq \bar{r}$.

End-of-period static profits, associated with beginning-of-period deposits d_θ and lending $\ell_{\theta,j}$ in region j for an incumbent bank of type $\theta \in \{s,r\}$ in industry state μ (to be described below) depends on its end-of-period state $s_j = (\mu, z'_j)$ given by

8. This specification nests Besanko and Doraszelski (2004) when $\xi_{1\theta} = 0$.

9. Finally, since a national bank cannot grow higher, its transition function is given by

$$T(\theta' \mid \theta = n, I_n) = \begin{cases} \dfrac{1-\delta+\alpha \cdot I_n \cdot (\Delta \bar{d}_{n,r})^{-\xi}}{1+\alpha I_n \cdot (\Delta \bar{d}_{n,r})^{-\xi}} & \text{if } \theta' = n \\ \dfrac{\delta}{1+\alpha \cdot I_n \cdot (\Delta \bar{d}_{n,r})^{-\xi}} & \text{if } \theta' = r \end{cases}. \tag{7}$$

We assume that a national bank that reduces its size is randomly assigned to a region $j \in \{e,w\}$ with probabilty .

$$\pi_\theta\left(d_\theta, s_j\right) = \left[p_j\left(R_{\mathcal{B}}, z_j'\right) r_{\mathcal{B},j}\left(\boldsymbol{\mu}\right) - \left(1 - p_j\left(R_{\mathcal{B}}, z_j'\right)\right)\lambda\right]\ell_{\theta,j} - c_\theta\left(\ell_{\theta,j}\right) + \left(\bar{r}\,1_{\{a_\theta \geq 0\}} + r^a 1_{\{a_\theta < 0\}}\right) a_\theta - r_{D,j} d_\theta - c_{F,\theta}, \tag{8}$$

where $p_j(\cdot)$ denotes the fraction of bank loans that are repaid at the end of the period in region j (an endogenous object that is consistent with the borrower's problem), $r_{\mathcal{B},j}(\cdot)$ is the Cournot-equilibrium interest rate on bank loans in region j, $c_\theta(\ell_{\theta,j})$ is the marginal cost of extending $\ell_{\theta,j}$ loans, and $C_{F,\theta}$ is the fixed operating costs. Profits for an incumbent bank of type $\theta = \boldsymbol{n}$ in state $\mathbf{s} = (\boldsymbol{\mu}, z_j', z_{-j}')$ is given by

$$\pi_{\boldsymbol{n}}\left(d_{\boldsymbol{n}}, \mathbf{s}\right) = \sum_j \left\{\left[p_j\left(R_{\mathcal{B}}, z_j'\right) r_{\mathcal{B},j}\left(\boldsymbol{\mu}\right) - \left(1 - p_j\left(R_{\mathcal{B}}, z_j'\right)\right)\lambda\right]\ell_{\boldsymbol{n},j} - c_{\boldsymbol{n}}\left(\ell_{\boldsymbol{n},j}\right)\right\} + \left(\bar{r}\,1_{\{a_{\boldsymbol{n}} \geq 0\}} + r^a 1_{\{a_{\boldsymbol{n}} < 0\}}\right) a_{\boldsymbol{n}} - r_{D,j} d_{\boldsymbol{n}} - c_{F,\boldsymbol{n}}, \tag{9}$$

Given profits $\pi_\theta(d_\theta, s_j)$ for $\theta \in \{\mathcal{S}, \mathcal{N}\}$ and $\pi_{\boldsymbol{n}}(d_{\boldsymbol{n}}, s)$, banks can choose to exit. To keep the computation of the model tractable, we also incorporate a type-specific exogenous probability of exit ρ_θ^x. If a bank decides to continue, it then decides how much to invest in order to improve its capacity to collect deposits. Banks can finance investment with internal funds (π_θ) or by issuing equity e_θ whenever $I_\theta > \pi_\theta$. That is, $e_\theta = \max\{I_\theta - \pi_\theta, 0\}$. Issuing equity is costly with cost function given by $\varsigma_\theta(e_\theta)$. For tractability, unlike Corbae and D'Erasmo (2021), here we assume that banks cannot retain earnings.[10] Dividends net of equity injections for a bank of type for $\theta \in \{\mathcal{S}, \mathcal{N}\}$ in state (d_θ, s_j) are given by

$$\mathcal{D}_\theta\left(d_\theta, s_j\right) = \pi_\theta\left(d_\theta, s_j\right) - I_\theta - 1_{\{e_\theta > 0\}}\varsigma_\theta\left(e_\theta\right). \tag{10}$$

A similar equation can be written for national bank dividends $\mathcal{D}_{\boldsymbol{n}}(d_{\boldsymbol{n}}, s)$.

The objective function of the bank is to maximize the expected present discounted value of future dividends net of equity injections with discount factor β. It is important to note that, while deposits conditional on bank size (d_θ) are exogenous, external finance is endogenous, since bank size (θ) via investment (I_θ) and seasoned equity (e_θ) are endogenous.

10. Since one of the objectives of Corbae and D'Erasmo (2021) was to understand the role of capital buffers, we allowed for the endogenous retention of earnings which augmented a bank's capital. Here we endogenize bank size by allowing banks to invest I_θ to change θ; e.g., become bigger.

We assume there is limited liability and that incumbent banks have the option to exit after extending loans. The value of exiting for a bank of type $\theta \in \{\mathfrak{s},\mathfrak{v}\}$ is given by

$$\max\Big\{0,\zeta_\theta\Big[1+p_j\left(R_{\mathcal{B}},z_j'\right)r_{\mathcal{B},j}\left(\boldsymbol{\mu}\right)-\left(1-p_j\left(R_{\mathcal{B}},z_j'\right)\lambda\right]\ell_{\theta,j} \tag{11}$$
$$-c_\theta\left(\ell_{\theta,j}\right)-c_{F,\theta}-\left(1+r_{D,j}\right)d_\theta+\zeta_\theta 1_{\{a_n\geq 0\}}\left(1+\bar{r}\right)a_\theta+1_{\{a_n<0\}}\left(1+r^a\right)a_\theta\Big\},$$

where ζ_θ captures the recovery rate of a bank's assets at the exit stage. This induces an exit decision rule $x_\theta(d_\theta,s_j)$ for $\theta \in \{\mathfrak{s},\mathfrak{v}\}$. A similar equation can be written for the national bank.

We consider an entry process similar to Farias and others (2012). At time period t, there are a finite but large number of potential entrants. Potential entrants make entry decisions simultaneously, are short-lived, and do not consider the option of delaying entry. Entrants bear a positive entry cost κ funded by an initial equity injection by households and base their entry decision on the net present value of entering today. Entrants do not earn profits in the period they decide to enter. They appear in the following period in state $(\theta' = \mathfrak{s},d_{\mathfrak{s}}')$ (i.e., we assume that all entrants start as a small bank), where $d_{\mathfrak{s}}'$ is drawn from $\overline{G}_{\mathfrak{s}}(d_{\mathfrak{s}})$the invariant distribution Gbar associated with $G(d_{\mathfrak{s}}',d_{\mathfrak{s}})$. We denote the number of entrants $N_{e,j}$, which is determined endogenously in equilibrium.

In summary, the simple balance sheet of a bank in our environment is given by book assets equal loans (ℓ_θ), storage (a_θ), and fixed capital (κ_θ) while book liabilities equal deposits (d_θ) and equity injections (e_θ).

If all banks in a given state $(\theta,d_\theta) \in \{\Theta \times D_\theta\}$ are treated symmetrically, then the cross-sectional distribution $\boldsymbol{\mu}$ specifies the number of banks across state and region. More specifically,

$$\boldsymbol{\mu}=\left\{\left\{\mu_{n,*}\left(d_n\right)\right\}_{d_n\in D_n},\left\{\mu_{\theta,j}\left(d_\theta\right)\right\}_{\theta\in\{\mathfrak{v},\mathfrak{s}\},j\in\{e,w\},d_\theta\in D_\theta}\right\}. \tag{12}$$

We let N denote the number of incumbent banks at time period t, that is,

$$N=\sum_{d_n\in D_n}\mu_{n,*}\left(d_n\right)+\sum_{\theta\in\{\mathfrak{v},\mathfrak{s}\},j\in\{e,w\},d_\theta\in D_\theta}\mu_{\theta,j}\left(d_\theta\right). \tag{13}$$

Further, the law of motion for the industry state is denoted

$$\boldsymbol{\mu}' = \mathcal{H}(\boldsymbol{\mu}, N_e), \tag{14}$$

where N_e denotes the number of entrants, and the transition function $\mathcal{H}$ is defined explicitly below in equation (33).

2.4 Nonbank Lenders

A representative national nonbank that discounts the future at rate β specializes in extending loans to entrepreneurs (in both regions) in a perfectly competitive market. To keep the analysis simple, the nonbank is financed with equity $e_{\mathcal{N}}$ raised from the household sector and is not subject to limited liability. When lending to entrepreneurs, nonbanks face a marginal monitoring cost $c_{\mathcal{N}}$. Like banks, the representative nonbank can diversify entrepreneurs' idiosyncratic risk, but it is subject to regional fluctuations.

Let $\pi_{\mathcal{N}}(\mathbf{s})$ denote the end-of-period profits of the nonbank after the realization of regional shocks associated with its current lending $\ell_{\mathcal{N},j}$ given by

$$\pi_{\mathcal{N}}(\mathbf{s}) = \sum_j \left\{ \left[p_j\left(R_{\mathcal{N}}, z_j'\right) r_{\mathcal{N},j}(\boldsymbol{\mu}) - \left(1 - p_j\left(R_{\mathcal{N}}, z_j'\right)\right)\lambda \right] - c_{\mathcal{N}} \right\} \ell_{\mathcal{N},j} \tag{15}$$

subject to flow constraint $S_{\mathcal{N}} e_{\mathcal{N}} = \Sigma_j \ell_{\mathcal{N},j}$, where $(\boldsymbol{\mu}, z'_j, z'_{-j})$ and $S_{\mathcal{N}}$ are household shareholdings of the nonbank. Since the nonbank operates in a perfectly competitive market it takes the regional interest rate $r_{\mathcal{N},j}$ as given. The nonbank issues dividends according to $D_{\mathcal{N}} = \pi_{\mathcal{N}}(\mathbf{s})$.

The objective function of the nonbank is to maximize the expected present discounted value of future cash flows to households with discount factor β. We assume that there is free entry into the nonbank sector and, to simplify the analysis, we set the entry cost to zero.

2.5 Government Budget Constraint

The government collects lump-sum taxes to cover the cost of deposit insurance. Post-liquidation net transfers are given by

$$\begin{aligned}\Delta'\left(d_\theta, s_j\right) = {} & \left(1 + r_{D,\theta}\right) d_\theta \\ & - \zeta_\theta \left[1 + p \cdot r_{\mathcal{B}}\left(Z, \mu\right) - \left(1 - p\right)\lambda'\right] \ell_{\theta,j} - \zeta_\theta \left(1 + \bar{r}\right)\left(d_\theta - \ell_{\theta,j}\right),\end{aligned}$$

where $\zeta_\theta \leq 1$ is the post-liquidation value of the bank's asset portfolio. Aggregate taxes are given by

$$\tau_D'(\boldsymbol{s}) \cdot H = \sum_{\theta, d_\theta, j} \left[\int_{\lambda'} x\left(d_\theta, s_j\right) \max\left\{0, \Delta'\left(d_\theta, s_j\right)\right\} \mu_\theta\left(d_\theta\right) df\left(\lambda'\right) \right].$$

2.6 Information

There is asymmetric information on the part of borrowers and lenders (banks, nonbanks, and households). Only borrowers know the riskiness of the project they choose (R_k) and their outside option (ω). Success or failure of their project is only observable after bearing the monitoring cost. To maintain consistency with payoffs between project choice and outside option, they receive a perfect unobservable signal about their outside option at the beginning of the period. Other information is observable.

2.7 Timing

In any period t, the timing of events is as follows:

1. At the beginning of the period

(a) Bank type θ and the mass of depositors that the bank is matched with d_θ are realized given household asset decisions. That determines the industry state (i.e., cross-sectional distribution $\boldsymbol{\mu}$).

(b) After observing ω, borrowers choose whether to invest in the risky technology or to choose their outside option $\iota \in \{0,1\}$ and, if so, they draw $\in$.

(c) Those borrowers who choose to undertake a project choose the type of lender $k \in \{\mathcal{B}, \mathcal{N}\}$ and the level of technology R_k.

(d) Banks and the representative nonbank choose how many loans to extend. In addition, banks choose how many deposits to accept and how many securities to acquire, and nonbanks receive their equity injections from households.

(e) The loan market is cleared determining $r_j = \{r_{\mathcal{B},j}, r_{\mathcal{N},j}\}$

2. At the end of the period, z_j' is realized:

(a) Project returns for entrepreneurs are determined.

(b) The portfolio of performing and nonperforming loans is determined via project returns and p_j resulting in a realization of $\pi_\theta(d_\theta, s_j)$, $\pi_n(d_n, \mathbf{s})$, and $\pi_{\mathcal{N}}(\mathbf{s})$.

(c) Bank exit x_θ and entry e choices are made.

(d) Bank investment I_θ is chosen together with dividend payments and equity injections. Dividends net of equity injections for the representative nonbank are also determined.

(e) Households pay taxes τ' to fund deposit insurance and consume.

3. Equilibrium

3.1 Entrepreneur Problem

Every period, given $\boldsymbol{r}_j = \{r_{\mathcal{B},j}, r_{\mathcal{N},j}\}$ and ω, entrepreneurs located in region j choose whether ($\iota = 1$) or not ($\iota = 0$) to operate their technology. Conditional on choosing $\iota = 1$, entrepreneurs observe $\epsilon = \{\epsilon_{\mathcal{B}}, \epsilon_{\mathcal{N}}\}$ and then choose which type of lender $K \in \{\mathcal{B},\mathcal{N}\}$to borrow from and the scale of the technology to operate R_k to solve

$$\max_{\{\iota\}} (1-\iota)\cdot\omega + \iota \cdot E_\epsilon\left[\Pi_E\left(\boldsymbol{r}_j, \epsilon\right)\right], \tag{16}$$

where the value of investing (conditional on ϵ) is

$$\begin{aligned}\Pi_E\left(\boldsymbol{r}_j, \epsilon\right) = \max_{\{K, R_K\}}\Big\{ & 1_{\{K=\mathcal{B}\}} E_{z_j'}\left[\pi_E\left(r_{\mathcal{B},j}, R_{\mathcal{B}}, z_j'\right) + \epsilon_{\mathcal{B}}\right] \\ & + 1_{\{K=\mathcal{N}\}} E_{z_j'}\left[\pi_E\left(r_{\mathcal{N},j}, R_{\mathcal{N}}, z_j'\right) + \epsilon_{\mathcal{N}}\right]\Big\},\end{aligned} \tag{17}$$

where $1_{\{\cdot\}}$ is an indicator function that takes the value one if the argument $\{\cdot\}$ is true and zero otherwise, and

$$\pi_E\left(r_{K,j}, R_K, z_j'\right) = \begin{cases} \max\left\{0, z_j' R_K - r_{K,j}\right\} & \text{with prob } p_j\left(R_K, z_j'\right) \\ \max\left\{0, -\left(\lambda + r_{K,j}\right)\right\} & \text{with prob } 1 - p_j\left(R_K, z_j'\right) \end{cases}$$

The solution to (17) implies that the share of borrowers choosing a loan from a lender of type K in region j is

$$s_{K,j}\left(\mathbf{r}_j\right) = \frac{\exp\left(\alpha E_{z_j'}\left[\pi_E\left(r_{K,j}, R_K, z_j'\right)\right]\right)}{\sum_{\hat{K}\in\{\mathcal{B},\mathcal{N}\}} \exp\left(\alpha E_{z_j'}\left[\pi_E\left(r_{\hat{K},j}, R_{\hat{K}}, z_j'\right)\right]\right)}. \tag{18}$$

The expected value of taking out a loan in region j is[11]

$$V_{E,j}\left(\mathbf{r}_j\right) = \int \Pi_E\left(r_j, \epsilon\right) dF_\epsilon\left(\epsilon;\alpha\right). \tag{19}$$

If the entrepreneur undertakes the project financed by lender type K, then an application of the envelope theorem implies

$$\frac{\partial E_{z_j'}\left[\pi_E\left(r_{K,j}, R_K, z_j'\right)\right]}{\partial r_{K,j}} = -E_{z_j'}\left[p_j\left(R_K, z_j'\right)\right] < 0. \tag{20}$$

Thus, participating borrowers (i.e., those who choose to run a project rather than take the outside option) are worse off the higher the interest rate on loans is.

This has implications for the aggregate demand for loans determined by the participation decision (i.e., $\omega \le V_{E,j}(\mathbf{r}_j)$). In particular, the total demand for loans in region j is given by

$$L_j^d\left(\mathbf{r}_j\right) = \int_0^{\overline{\omega}} 1_{\left\{\omega \le V_{E,j}\left(\mathbf{r}_j\right)\right\}} d\Omega(\omega). \tag{21}$$

Then loan demand for commercial banks in region j is given by

$$L_{B,j}^d\left(\mathbf{r}_j\right) = s_{B,j}\left(\mathbf{r}_j\right) L_j^d\left(\mathbf{r}_j\right). \tag{22}$$

In that case, everything else equal, (20) implies $\frac{\partial L_{B,j}^d\left(\mathbf{r}_j\right)}{\partial r_{B,j}} < 0$. That is, the bank loan-demand curve is downward sloping. Furthermore, bank market shares are decreasing in bank lending rates (i.e., $\frac{\partial s_{B,j}\left(\mathbf{r}_j\right)}{\partial r_{B,j}} < 0$) and aggregate loan demand decreases with an increase in bank lending rates (i.e., $\frac{\partial L_j^d\left(\mathbf{r}_j\right)}{\partial r_{B,j}} \le 0$).

3.2 Incumbent Bank Problem

As in Ericson and Pakes (1995), we consider symmetric equilibrium in the sense that all banks in the same region and individual state d_θ are treated identically. Since a bank's individual state lies in a finite set, the industry state μ is a counting measure.

11. The expected value of taking out a loan has a convenient closed form: $\frac{\gamma_E}{\alpha} + \frac{1}{\alpha}\ln\left(\Sigma_k \exp\left(\alpha E_{z_j'}\left[\pi_E\left(r_{K,j}, R_K, z_j'\right)\right]\right)\right)$ where γ_E is Euler's constant.

After being exogenously matched with d_θ potential depositors and offering them a take-it-or-leave-it deposit-rate offer $r_{D,j}$, an incumbent bank of type θ chooses loans $\ell_{\theta,j}$ in order to maximize profits. Given the outside storage option for a household is $\bar{r}$, the bank deposit rate $r_{D,j} = \bar{r}$ in all regions. In this way, we are abstracting from important deposit-side competition in order to focus on the bank lending channel. After profits are realized, banks can choose to exit setting $x_\theta = 1$ or choose to remain $x_\theta = 0$. When choosing its loan supply a small or regional bank in region j solves

$$\ell_{\theta,j}(d_\theta, \boldsymbol{\mu}) = \arg\max_{\ell_{\theta,j}+a_\theta=d_\theta} E_{z_j'}\left[\pi_\theta(d_\theta, s_j)\right]. \tag{23}$$

Similarly, when choosing its loan supply across regions, a national bank solves

$$\left\{\ell_{n,j}(d_n, \boldsymbol{\mu})\right\}_{j\in\{e,w\}} = \arg\max_{\{\ell_{n,j}\}_{j\in\{e,w\}}} E_{z_j', z_{-j}'}\left[\pi_n(d_n, \boldsymbol{s})\right] \tag{24}$$

subject to

$$\sum_i \ell_{n,j} + a_n = d_n. \tag{25}$$

Given that all banks have some degree of market power, a bank takes into account that its loan supply affects the loan interest rate in its region and that other banks will best respond to its loan supply. The first-order condition for a small or regional bank in problem (23) with respect to ℓ is

$$E_{z_j'}\Bigg[\underbrace{\left(p_j r_{B,j}(\boldsymbol{\mu}) - (1-p_j)\lambda - \frac{dc_\theta}{d\ell_{\theta,j}}\right)}_{(+)\text{ or }(-)} + \ell_{\theta,j}\Bigg(\underbrace{p_j}_{(+)} + \underbrace{\frac{\partial p_j}{\partial R_B}\frac{\partial R_B}{\partial r_{B,j}(\boldsymbol{\mu})}}_{(-)}(r_{B,j}(\boldsymbol{\mu}) + \lambda)\Bigg)$$
$$-\mathbf{1}_{\{(d_\theta - \ell_{\theta,j})\geq 0\}} r - \mathbf{1}_{\{(d_\theta - \ell_{\theta,j})\geq 0\}} r^a (d_\theta - \ell_{\theta,j}) = 0,$$
$$\underbrace{\frac{dr_{B,j}(\boldsymbol{\mu})}{d\ell_{\theta,j}}}_{(-)}\Bigg] \tag{26}$$

where $p_j \equiv p_j(R_B, z_j')$. The first bracket represents the marginal change in profits from extending an extra unit of loans. The second bracket

corresponds to the marginal change in profits due to a bank's influence on the interest rate it faces. This term depends on the bank's market power. A change in interest rates also endogenously affects the fraction of delinquent loans faced by banks (i.e., the term $\frac{\partial p_j}{\partial R_{\mathcal{B}}} \frac{dR_{\mathcal{B}}}{dr_{\mathcal{B},j}} < 0$). Given limited liability, entrepreneurs take on more risk when their financing costs rise. The last two terms represent the marginal cost of uninsured external borrowing for the bank. When the bank accumulates securities ($1_{\{a_\theta = d_\theta - \ell_{\theta,j} \geq 0\}}$), the marginal cost is given by the opportunity cost of the loan (what the bank could receive from storage). When the bank uses external borrowing to extend loans beyond its deposit base ($1_{\{a_\theta = d_\theta - \ell_{\theta,j} < 0\}}$), the marginal cost is given by the cost of external funds. A similar condition holds for the national bank.

Changes in the loan interest rate (i.e., $\frac{dr_{\mathcal{B},j}}{d\ell_{\theta,j}}$) in (26) are derived from the market clearing condition $L^d_{\mathcal{B},j}(\mathbf{r}_j) = L^s_{\mathcal{B},j}(\mu)$, where $L^s_{\mathcal{B},j}(\mathbf{r}_j)$ is given above in (22) and $L^s_{\mathcal{B},j}(\mu)$ denotes the total supply of loans given by

$$L^s_{\mathcal{B},j}(\boldsymbol{\mu}) = \sum_{\theta, d_\theta} \ell_{\theta,j}(d_\theta, \boldsymbol{\mu}) \mu_{\theta,j}(d_\theta) \tag{27}$$

For a given bank distribution $\boldsymbol{\mu}$, changes in the loan supply $\ell_{\theta,j}$ of a given bank have a direct effect on the aggregate loan supply but also an indirect effect via changes in the response of its competitors.

After loans have been extended, the value of an incumbent state or regional bank $\theta \in \{\mathcal{S}, \mathcal{R}\}$ in region j at the exit stage 2c is

$$V_\theta(d_\theta, s_j) = \max_{x \in \{0,1\}} \left\{ V^{x=0}(d_\theta, s_j), V^{x=1}(d_\theta, s_j) \right\} \tag{28}$$

where $s_j = (\boldsymbol{\mu}, z'_j)$, $V^{x=1}(d_\theta, s_j)$ is defined in equation (11) and

$$\begin{aligned} V^{x=0}(d_\theta, s_j) \;=\; & \max \Big\{ \pi_\theta(d_\theta, s_j) - I - 1_{\{\ell_\theta > 0\}} \cdot \varsigma_\theta(\ell_\theta) \\ & + \beta \rho^x_\theta E_{\theta', d'_\theta, s'_j | d_\theta, s_j} \left[V(d'_\theta, s'_j) \right] + \beta(1 - \rho^x) E_{\theta', d'_\theta, s'_j | d_\theta, s_j} \left[V^{x=1}(d'_\theta, s'_j) \right] \Big\} \end{aligned} \tag{29}$$

subject to

$$\ell_\theta = \max \{ I - \pi_\theta, 0 \} \tag{30}$$

and the transition functions $T(\theta' \mid \theta, I)$ and $\mu' = \mathcal{H}(\mu, N_e)$. A similar problem holds for the national bank with $\pi_n(d_n, s)$ substituting for $\pi_\theta(d_\theta, s_j)$ in (29).

3.3 Bank Entry

The value of an entrant in region j net of entry costs in the industry state μ is

$$V_{e,j}(\mu) = -\kappa + \beta E\left[V_s\left(d_s', s_j'\right)\right]. \tag{31}$$

Recall that entrants do not operate in the period they enter and, consistent with the data, we assume they all start small (i.e., with $\theta = s$). Potential entrants will decide to enter if $V_{e,j}(\mu) \geq 0$. The number of entrants $N_{e,j}$ is determined endogenously in equilibrium. Free entry implies that

$$V_{e,j}(\mu) \times N_{e,j} = 0. \tag{32}$$

That is, in equilibrium, either the value of entry is zero, the number of entrants is zero, or both. The total value of entrants is given by $N_e = \sum_j N_{e,j}$.

3.4 Evolution of the Cross-Sectional Bank Size Distribution

The distribution of banks evolves according to $\mu' = \mathcal{H}(\mu, N_e)$, where each component is given by:

$$\begin{aligned}\mu'_\theta(d'_\theta) = &\sum_{\theta \in \{n,s\}, j \in \{e,w\}, d_\theta \in D_\theta} \left(1 - x(d_\theta, s_j)\right)\left(1 - \rho^x_\theta\right) T\left(\theta' \mid \theta, I(d_\theta, s_j)\right) G_\theta(d'_\theta, d_\theta)\, \mu_\theta(d_\theta) \\ &+ \sum_{d_n \in D_n} \left(1 - x(d_n, s)\right)\left(1 - \rho^x_n\right) T\left(\theta' \mid n, I(d_n, s)\right) G_n(d'_\theta, d_n)\, \mu_n(d_n) \\ &+ N_{e,j} \sum_{j, d_s} G_s(d_s)\end{aligned} \tag{33}$$

where $\overline{G}_s(d_s)$ is the distribution from which deposits for entrants are drawn. Equation (33) makes clear how the law of motion for the distribution of banks is affected by entry (N_e) and exit (X) decisions as well as the bank size investment decision (I).

3.5 Nonbank Problem

The representative nonbank operates in a competitive industry, so when making lending decisions, it takes the loan interest rate $r_{\mathcal{N},j}$ as given. Taking into account that $\beta(1+\bar{r}) = 1$, the first-order condition of the nonbank with respect to $\ell_{\mathcal{N},j}$ is given by

$$\bar{r} = E_{z'_j}\left[p\left(R_{\mathcal{N}}\left(r_{\mathcal{N},j}\right), z'_j\right) - \left(1 - p\left(R_{\mathcal{N}}\left(r_{\mathcal{N},j}\right), z'_j\right)\right)\lambda\right] - c_{\mathcal{N}}, \tag{34}$$

where $R_{\mathcal{N},j}(r_{\mathcal{N}})$ is the optimal choice of technology by the entrepreneur in region j when taking a loan from a nonbank facing interest rate $r_{\mathcal{N},j}$. Equation (34) is one equation in one unknown which pins down the interest rate $r_{\mathcal{N},j}$ of the nonbank sector.[12] Evaluating the nonbank loan demand at this price we can determine the level of lending of the nonbank. Equation (34) also makes clear that the expected net return between a bank deposit and nonbank investment is equalized, with the spread depending on $c_{\mathcal{N}}$. However, while bank deposits guarantee a risk-free return (since there is deposit insurance), equity injections in a nonbank are subject to regional risk.

3.6 Definition of Equilibrium

A pure strategy Markov perfect industry equilibrium (MPIE) is:[13]

1. $\{\iota_j, K_j, R_{k,j}\}$ are consistent with entrepreneur optimization inducing an aggregate loan-demand function $L_j^d(\mathbf{r}_j)$.

2. $\{\ell_{\theta,j}, I_{\theta,j}, x_{\theta,j}, e_{\theta,j}, V_\theta\}$ are consistent with bank optimization inducing an aggregate loan-supply function $L_{\mathcal{B},j}^s$.

3. Free entry is satisfied.

4. The law of motion for the industry state $\boldsymbol{\mu}' = \mathcal{H}(\boldsymbol{\mu}, \{N_j^e\}_j)$ induces a sequence of cross-sectional distributions that are consistent with entry, exit, and investment decision rules.

5. The vector of interest rate $r_j(\boldsymbol{\mu})$ is such that the loan market clears.

6. Stock prices are consistent with bank valuation V_θ.

7. Taxes $\tau'_D(s)$ cover the cost of deposit insurance.

12. The fact that $r_{\mathcal{N},j}$ is independent of the entire distribution of banks is a form of block recursivity as in Menzio and Shi (2010).

13. See Corbae and D'Erasmo (2023) for the statement of the household problem.

4. Parameterization

When solving for the model moments, note that despite shocks d_θ which are i.i.d. across banks, the fact that banks are not of measure zero induces aggregate uncertainty.

Thus, we use the computational methods in Ifrach and Weintraub (2017).[14] In particular, the approximation methods allow for there to be strategically important (dominant) banks. We think of rising concentration as occurring between two long-run stochastic equilibria (one coinciding with pre-Riegle-Neal and one coinciding with the period following the Great Recession) that lead to different dynamics for the bank distribution following a decline in branching costs.

A model period is one year. Our main source for bank level variables (and aggregates derived from them) are the Call Reports.[15] We aggregate commercial bank level information to the bank holding company level. As discussed above, moments from the Call Report data are computed beginning in 1984, due to an overhaul of the data in that year.

Given that prior to the passing of the Riegle-Neal Interstate Banking and Branching Efficiency Act in 1994 there was only one type of bank in operation (s small banks), we calibrate the model to the post-reform period 2009–2018, where restrictions on opening bank branches across state lines were not in place. In the latter period, banks of three types operate $\theta \in \{s, r, n\}$, which allows us to obtain estimates for the investment transition matrix. We focus on the top 2 percent of banks when sorted by assets and identify those in the top 4 with those with $\theta = n$, those in the top 5 - 35 with $\theta = r$, and those in the top 36 to 2 percent with $\theta = s$.

We parameterize the stochastic process for the borrower's project as follows. For each borrower, let $y^e = a - bR + \varepsilon_e$, where ε_e is iid (across agents and time) and drawn from $N(z'_j, \sigma^2_\varepsilon)$. We define success to be

14. Appendix A.1 describes the solution algorithm we use to approximate a Markov-Perfect Equilibrium.

15. Source: FDIC, Call and Thrift Financial Reports, balance sheet, and income statement items. See Appendix A.2 for a description of the data.

the event that $y^e > 0$, so in states with higher ε_e success is more likely. Then

$$\begin{aligned} p\left(R, z_j'\right) &= 1 - \Pr\left(y^e \leq 0 \# R, z_j'\right) \\ &= 1 - \Pr\left(\varepsilon_e \leq -a + bR\right) \\ &= \Phi_{z_j'}\left(a - bR\right), \end{aligned}$$

where $\Phi_{z_j'}(x)$ is a normal cumulative distribution function with mean z_j' and variance σ_ε^2. The stochastic process for the borrower outside option, $\Omega(\omega)$, is simply taken to be the uniform distribution $[0,\overline{\omega}]$.

We assume that the regional shock is distributed iid normal with zero mean and variance σ_z^2. We set the cross-correlation between regions ρ_z to 0.01. We discretize the regional shock and let z_j take two values $z_{j,t} \in \mathcal{Z}_j = \{z_L, z_H\}$ with $z_H > 0$ and $z_L = -z_H$.

To reduce the number of parameters to calibrate, and as we do not have enough information on the liquidation value of the assets of large banks since we do not observe liquidations in the largest category, we set $\zeta_\theta = \zeta$ and calibrate ζ using data from the FDIC. We parameterize the equity-issuance cost function $\varsigma_\theta(e_\theta) = (\varsigma_\theta^0 + \varsigma_\theta^1 e_\theta)$ where $e_\theta = \max\{0, -(\pi_\theta - \iota_\theta)\}$ and the cost of extending loans $c_\theta(\ell_\theta) = c_\theta^0 \ell_\theta + c_\theta^1 \ell_\theta^2$. We assume that the total cost of external borrowing is $r^a(a_\theta) = r_0^a a_\theta + r_1^a a_\theta^2$ when $1_{\{a_\theta < 0\}}$.

As part of the calibration exercise, post reform, we estimate transition probabilities between banks of different sizes. In particular, we estimate transition matrices by counting the number of banks in each bin-year and dividing by the total number of banks of each type in a given year. We then take the time-series average of the corresponding bin for each period. For example, to compute the fraction of banks that remain in state s, we first count how many s banks in period t are still of type s in period $t+1$. Let this number be $N_t^{s,s}$. Then, evaluated at the equilibrium level of investment $I_\theta(d_\theta, s_j)$, the value in $T_t(\theta' \mid \theta, I_\theta(d_\theta, s_j))$ equals $\frac{N_t^{s,s}}{N_{s,t}}$ where $N_{s,t}$ corresponds to all banks of type s in period t. The reported value in table 6 corresponds to the time average of $T_t(\theta' \mid \theta, I_\theta(d_\theta, s_j))$. The failure state incorporates the transition to a bank outside the top 2 percent .

We calibrate the entry cost κ by choosing the average number of entrants in each region $N_{e,j}$ that results in an average number of banks equal to 103 (the average number of banks in the top 2 percent

post reform). We assume that the number of potential entrants is the same across regions. In the experiments that follow, κ is kept constant to this value and $N_{e,j}$ adjust to satisfy the equilibrium conditions. In addition, we set the origination cost for nonbanks to match the fraction of bank lending to total credit.

Table 4 presents the parameters of the model and the targets that were used. We use several moments from our panel of banks in the U.S. and the estimates of the deposit process presented in table 1. Entries above the line correspond to parameters chosen outside the model, while entries below the line correspond to parameters chosen within the model by simulated method of moments. Table 5 presents a set of data moments together with their model-generated counterparts for the post-reform period (i.e., the period used in the calibration). Moments above the line correspond to those used in the calibration procedure and those below the line are untargeted moments. In all we have 26 parameters and 26 targeted moments. Given that there is symmetry in the underlying stochastic processes and parameter values in the model, the two regions yield similar long-run averages in the tables.

Table 4. Parameters and Targets

Parameter		*Value*	*Target*
Deposit Interest Rate (%)	r^D	0.005	Avg Interest Expense Deposits
Mean Charge-off Rate	μ_λ	0.314	Avg Charge-off Rate
Exit Value Recovery	ζ	0.804	Recovery Value Bank Failures (FDIC)
Bank Discount Factor	β	0.995	$1/(1+\bar{r})$
Correlation Regional Shocks	$\rho_{\mathcal{Z}}$	0.01	regional correlation of default frequency
Measure Borrowers	B	320.0	Bank Loans to Output Ratio
Borrower Success Prob. Function	a	4.291	Avg. Borrower Return
Borrower Success Prob. Function	b	28.94	Avg. Default Frequency
Borrower Success Prob. Function	σ_e	0.107	Avg. Loan Interest Rate
Outside Option	$\bar{\omega}$	0.462	Elasticity of Loan Demand
Std. Dev Reg Shocks	σ_z	0.020	Std Dev Loan Returns
Linear Cost Loans $\mathfrak{s}$	$c^0_{\mathfrak{s}}$	0.001	Avg Net Mg Expense $\mathfrak{s}$
Quadratic Cost Loans $\mathfrak{s}$	$c^1_{\mathfrak{s}}$	0.025	Elasticity Mg Expense $\mathfrak{s}$
Fixed Operating Cost $\mathfrak{s}$	$C_{F,\mathfrak{s}}$	0.001	Fixed Cost / Loans $\mathfrak{s}$
Linear Cost Loans μ	$c^0_{\mathfrak{y}}$	0.001	Avg Net Mg Expense $\mathfrak{y}$
Quadratic Cost Loans μ	$c^1_{\mathfrak{y}}$	0.003	Elasticity Mg Expense $\mathfrak{y}$
Fixed Operating Cost $\mathfrak{y}$	$C_{F,\mathfrak{y}}$	0.005	Fixed Cost / Loans $\mathfrak{y}$
Linear Cost Loans $\mathfrak{n}$	$c^0_{\mathfrak{n}}$	0.003	Avg Net Mg Expense $\mathfrak{n}$
Quadratic Cost Loans $\mathfrak{n}$	$c^1_{\mathfrak{n}}$	0.001	Elasticity Mg Expense $\mathfrak{n}$
Fixed Operating Cost $\mathfrak{n}$	$C_{F,\mathfrak{n}}$	0.010	Fixed Cost / Loans $\mathfrak{n}$
Proportional Cost Loans $\mathcal{N}$	$C_{\mathcal{N}}$	0.034	Share Bank Loans / Total Loans
Transition Probability Function	α	100.00	Loan Market Share $\mathfrak{s}$
Transition Probability Function	δ	0.600	Fraction of Banks $\mathfrak{s}$
Transition Probability Function	ξ	0.850	Transition $\mathfrak{s}$ to $\mathfrak{y}$
Fixed Equity-Issuance Costs $\mathfrak{s}$	$\zeta^0_{\mathfrak{s}}$	0.001	Avg Equity Issuance $\mathfrak{s}$
Proportional Equity-Issuance Costs $\mathfrak{s}$	$\zeta^1_{\mathfrak{s}}$	0.050	Fract $\mathfrak{s}$ Banks Issue Equity
Fixed Equity-Issuance Costs $\mathfrak{s}$	$\zeta^0_{\mathfrak{y}}$	0.005	Avg Equity-Issuance *mathcal* $\mathfrak{y}$
Proportional Equity-Issuance Costs μ	$\zeta^1_{\mathfrak{y}}$	0.025	Fract $\mathfrak{y}$ Banks Issue Equity
Fixed Equity-Issuance Costs $\mathfrak{n}$	$\zeta^0_{\mathfrak{n}}$	0.020	Avg Equity Issuance $\mathfrak{n}$
Proportional Equity-Issuance Costs n	$\zeta^1_{\mathfrak{n}}$	0.010	Fract $\mathfrak{n}$ Banks Issue Equity
Entry Cost	k	0.083	Total Number of Banks
Borrowing Cost Function	r^a_0	0.012	Spread Fed Funds to Deposit Cost
Borrowing Cost Function	r^a_1	0.008	Fed Funds / Assets

Source: Author's calculations.

Note: The entry cost is set as part of the equilibrium selection. In particular, in the baseline case, the entry cost is the one that satisfies the zero-entry condition for the value of entrants $N_{e,j} = 0.75$ (on average) that provides the best fit of the model. This entry cost is kept constant when running the experiments presented in section 5.

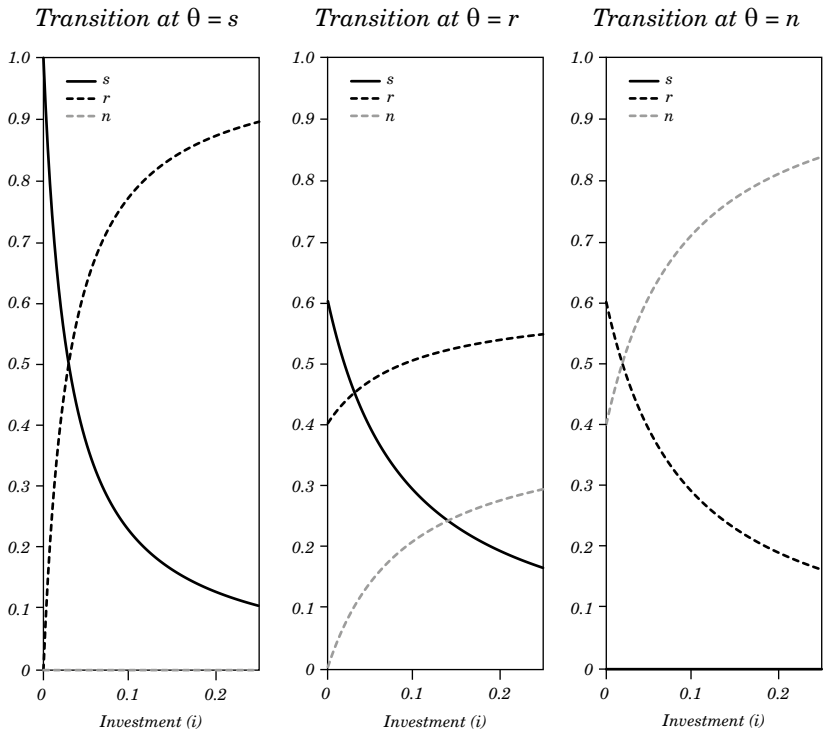
Table 5. Data & Model Moments Post Reform

Moments (%)	*Data*	*Model*
Charge-off Rate	0.685	0.39
Std Dev Charge-off Rate	0.20	0.20
Avg. Borrower Return	12.94	13.81
Avg. Default Frequency	2.09	1.25
Loan Interest Rate	2.971	2.59
Elasticity of Loan Demand	-1.1	-1.17
Avg. Net Mg Expense s	1.402	1.06
Elasticity Mg Expense s	0.875	1.83
Fixed Cost / Loans s	0.444	0.52
Avg. Net Mg Expense r	0.904	0.66
Elasticity Mg Expense r	0.940	1.69
Fixed Cost / Loans r	0.583	0.46
Avg. Net Mg Expense n	0.228	0.36
Elasticity Mg Expense n	1.05	1.08
Fixed Cost / Loans n	0.585	0.10
Loan Market Share s	16.04	20.58
Fraction of Banks s	69.35	70.80
Avg. Equity Issuance s	0.044	0.01
Fract s Banks Issue Equity	6.86	32.99
Avg. Equity Issuance r	0.04	0.00
Fract r Banks Issue Equity	4.23	0.00
Avg. Equity Issuance n	0.004	0.00
Fract n Banks Issue Equity	2.17	0.00
Bank Loans to Output Ratio	60.34	78.96
Share Bank Loans / Total Loans	50.00	78.64
Transition s to r	2.10	20.91
Spread Fed Funds to Deposit Cost	0.65	0.65
Fed Funds / Assets	2.16	15.29
Total Number of Banks	103	194.66
Exit (Failure) Rate	3.93	0.77
Deposit to Output Ratio	57.78	67.34
Markup	74.33	106.61
Avg. Net Interest Margin	4.18	2.06
Avg. Cost s	1.85	1.58
Avg. Cost r	1.487	1.12
Avg. Cost n	0.813	0.45
Fraction of Banks r	27.15	26.92
Fraction of Banks n	3.5	2.28
Loan Market Share r	50.22	44.25
Loan Market Share n	33.74	35.18
Number of Banks s	68	137.81
Number of Banks r	31	52.41
Number of Banks n	4	4.44
Deposit Market Share s	12.94	20.87
Deposit Market Share r	36.84	39.29
Deposit Market Share n_{25}	50.22	39.83

Note: Moments above the line correspond to calibration targets.

Figure 4 presents the estimated transition probability function for a given level of investment. The figure illustrates that the probability of growing (shrinking) is increasing (decreasing) in bank investment. Table 6 provides the resulting transition matrix across types evaluated at the equilibrium level of investment $I_{\theta}(d_{\theta},s_j)$.[16] The table illustrates that failure rates in the model are decreasing in size as in the data. Further, it illustrates that size is generally persistent.

Figure 4. Calibrated Transition Probabilities (Post Reform)



16. The initial year of the post-reform period used to estimate this matrix differs from that presented in table 5. The small number of bank types we consider prevents us from using the 2009–2018 period and obtain a meaningful transition matrix.

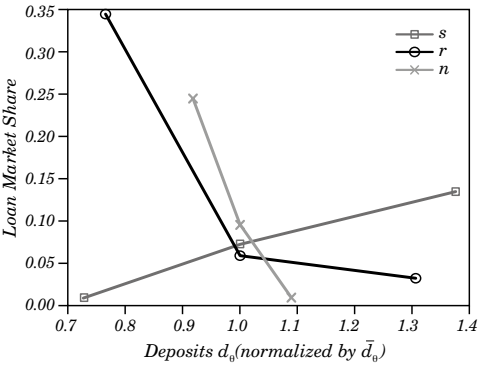
Table 6. Bank-Type Transition Matrix $T(\theta' \mid \theta, I_\theta(d_\theta, s_j))$

	Data: Post - Reform Period (1994-2019)				*Model: Post - Reform Period (1994-2019)*			
	$\theta' = s$	$\theta' = r$	$\theta' = n$	Failure	$\theta' = s$	$\theta' = r$	$\theta' = n$	Failure
Entrant	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
$\theta = s$	0.92	0.02	0.00	0.05	0.78	0.21	0.00	0.01
$\theta = r$	0.03	0.97	0.01	0.00	0.55	0.42	0.03	0.00
$\theta = n$	0.00	0.02	0.98	0.00	0.00	0.40	0.60	0.00

Note: We study banks in the top two percent of the asset distribution. We consider the top 4 (θ_n), the top 5 - 35 (θ_r), and the rest.

Figure 5 presents the distribution of banks for the equilibrium post reforms. The ranking of the variance of deposit inflows reflecting geographic diversification of funding shocks (i.e., highest variance for state banks and lowest for national banks) from table 1 is evident in the support of the distribution in figure 5. The market shares by bank size reflect the number of banks and the loan decisions (conditional on size). In an equilibrium where all banks extend loans equal to the amount of deposits they take, by construction, the shape of the loan distribution would derive from the shape of the deposit distribution. As deposit inflows are normally distributed, loan market shares would be distributed normal as well. Transitions from s to r and from r to n increase the number of banks with the lowest deposit value conditional on being of type r or n, as it is more likely to start with the lowest value of deposits when transitioning upwards. In addition, it is the case that in this equilibrium, most banks of type s and r extend less loans than the deposits they take (which is particularly important for the highest level of deposits), thus reducing the market share of this class of banks even further.

Figure 5. Distribution Banks Post Reform



Source: Author's calculations.

5. The Bank Lending Channel

The bank lending channel of monetary policy suggests that banks play a special role in the transmission of monetary policy. The channel works through how monetary policy effects on the cost of external funding. The corporate finance approach to the bank lending channel, as elucidated in Kashyap and Stein (1995) and Kashyap and Stein (2000), posits that larger banks are less sensitive to increases in fed funds rates since they have easier access to external funding. Thus, bigger banks lower their loan supply less than smaller banks in response to a rise in external funding costs like fed funds.

5.1 Model Mechanism

Here we describe how the bank lending channel works in the context of our model. There are three sources of external funding in our model: insured deposits d_θ at rate $\bar{r}$, fed borrowing $a_\theta = d_\theta - \ell_\theta < 0$ at rate $r^a > \bar{r}$, and equity e_θ at cost $\varsigma_\theta(e_\theta)$. Bank-type heterogeneity affects the sources of external funding in several ways.

First, the type $\theta \in \{s,r,n\}$ dependent deposit process $G_\theta(d'_\theta, d_\theta)$ provides a bank with a cheap source of funds. In particular, while we assume that banks of different types face the same insured borrowing cost $\bar{r}$, its mean deposit base is increasing in size and variance is decreasing in size as in the data in table 1. This means that bigger banks have a larger source of FDIC-subsidized external funding. In our model, banks always use this cheap source of external finance first

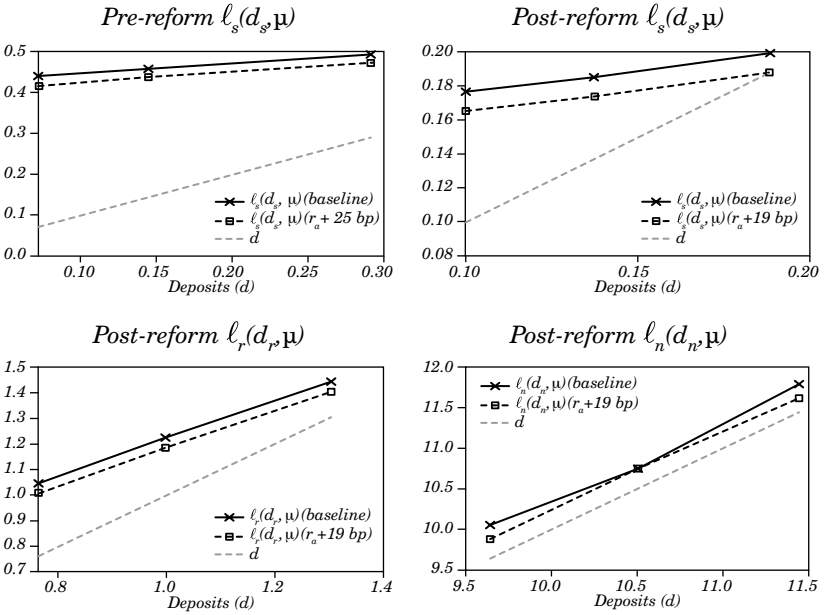
(as in a pecking order theory) when making loans, i.e., when solving (23) and (24).

Second, bank investment in type growth $I_\theta(d_\theta,s_j)$ is a function of profitability $\pi_\theta(d_\theta,s_j)$ (which is increasing in deposit base) and its type-dependent equity-issuance costs $\varsigma_\theta(e_\theta)$ in (29). There is a pecking order here as well—banks first use internal funds π_θ and then issue equity when $I > \pi_\theta$. Type $\mathfrak{s}$ banks with a high realization of deposits in a given state s_j (and hence high profits $\pi_{\mathfrak{s}}(d_{\mathfrak{s}}, s_j)$) are more likely to grow to regional types $\mathfrak{r}$. Since bigger banks face lower proportional equity-issuance costs in $\varsigma_{\mathfrak{s}}(e) > \varsigma_{\mathfrak{r}}(e) > \varsigma_{\mathfrak{n}}(e)$ as in table 4, bigger banks are more likely to stay big, enjoying low-cost deposit funding.

These two model elements provide the mechanism by which a rise in the fed funds rate may translate into more sensitivity by small banks than large banks as in Kashyap and Stein. In particular, fed fund borrowing in our model $a_\theta = d_\theta - \ell_\theta < 0$ is decreasing in d_θ for a given amount of loans ℓ_θ extended. Thus, banks with a small amount of low-cost $\bar{r}$ deposits are more exposed to borrowing at the higher fed funds rate $r^a > \bar{r}$. Thus, the cost of supplying a given amount of loans, which solves the first-order condition (26), is decreasing in deposits d_θ; since bigger banks enjoy a bigger deposit base, they are less sensitive to a rise in external borrowing via fed funds.

We illustrate this by graphing loan decision rules $\ell_\theta(d_\theta, \mu)$ for a 25 basis point rise in fed funds in the pre-reform equilibrium (top panel) and for a 19 basis point (same proportional) rise in the post-reform equilibrium (bottom panels) in figure 6. There are several things to note: (i) the difference between the decision rule ℓ and the 45-degree line illustrates fed fund borrowing, (ii) higher deposit bases lower external funding via fed funds, (iii) a rise in fed funds rates lowers loan supply more for small banks than bigger ones.

Figure 6. Lending Channel: Pre & Post-Reform Loan Decision Rules



Source: Author's calculations.

5.2 Simulation Results

We implement this policy experiment by raising the external funding cost r^a by 25 basis points in the pre-Riegle-Neal equilibrium of our model (from $r^a = 0.0125$ to $r^a = 0.015$) as well as in the post-reform, where the policy rate goes from $r^a = 0.0115$ to $r^a = 0.0134$ (the same proportional change). We evaluate the effect of this policy in the short run ($T = 1$, the response on impact, and $T = 2$, the average over two years). The starting point for the simulation is the long-run average distribution.[17]

Referring back to the first-order condition for loan choice in (26), conditional on a bank borrowing (i.e., $\mathbf{1}_{\{a_\theta<0\}}$), a rise in external funding

17. Since we focus on the short-run effects of this policy, we assume that banks do not update their beliefs about the industry state and expect to compete, in the short run, against the long-run industry from before the policy change. In Corbae and D'Erasmo (2021) we allow beliefs to be updated stochastically from the initial set of beliefs to the new long-run set of beliefs consistent with the policy change.

costs through a rise in the fed fund rate raises the marginal cost of making loans and hence lowers the 'intensive' margin of bank loans $\ell_{\theta,j}(d_\theta, \mu)$. As evident in table 7, the rise in r^a leads to lower average loans both pre and post reform. Importantly, we see larger banks are less sensitive (i.e., decrease loan supply less in response) to the rise in funding costs than smaller banks, as in Kashyap and Stein. While average loans (the intensive margin) drop, there can be changes in the distribution (the 'extensive' margin) that can induce interesting general equilibrium effects. Specifically, while aggregate loans decrease in the short run ($T = 1$) both pre and post reform, given the large post-reform interest margin at $T = 1$, we see large growth in the number of regional banks out of small. Such extensive margin changes can change the dynamics of aggregate loans and interest rates.

Table 7. Bank Lending Channel Pre & Post Reform

	Pre - Reform			*Post - Reform*		
		r_0^a + 0.25%			r_0^a + 0.19%	
	Baseline	T=1Δ%	T=2Δ%	Baseline	T=1Δ%	T=2Δ%
Avg. Def Freq.	1.58	-13.24	0.47	1.25	-12.84	-1.67
Loan Int. Rate	4.73	2.26	2.29	2.59	8.67	-3.44
Interest Margin	3.41	3.39	3.12	2.06	11.00	-4.25
Bank Loan Supply	82.43	-4.83	-4.90	129.06	-6.15	2.11
Bank Loans to Output	59.46	-4.42	-4.33	78.96	-4.89	1.37
Bank Loans / Total Loans	59.16	-4.27	-4.33	78.64	-4.75	1.38
Avg. Loans s	0.46	-4.59	-4.59	0.19	-5.10	-5.07
Avg. Loans r	-	-	-	1.09	-3.26	-3.95
Avg. Loans n	-	-	-	10.26	-1.09	-1.13
Loan Mkt Share s	100.00	0.00	0.00	20.52	2.21	-13.79
Loan Mkt Share r	-	-	-	44.19	-1.89	9.97
Loan Mkt Share n	-	-	-	35.28	1.09	-4.46
Total Number of Banks	178.09	-0.26	-0.26	194.66	-0.63	-0.62
N s Banks	178.09	-0.26	-0.26	137.81	1.08	-7.27
N r Banks	-	-	-	52.41	-4.82	16.91
N n Banks	-	-	-	4.44	-4.10	-1.15

Source: Author's calculations.
Note: Pre-reform r_0^a = 1.50% and Post-reform r_0^a = 1.15%. The increase in the policy rate corresponds to an increase of 25 bp of in the pre-reform (a 33% increase). T denotes the number of periods used to compute reported averages.

Table 8. Bank Lending Channel Post Reform (Baseline vs No Investment Update)

	Post - Reform		
	Baseline	r_0^a + 0.19%	
		Equilibrium $T = 2\Delta\%$	Partial Equilibrium $T = 2\Delta\%$
Avg. Def Freq.	1.25	-1.67	0.88
Loan Int. Rate	2.59	-3.44	6.78
Interest Margin	2.06	-4.25	8.42
Bank Loan Supply	129.06	2.11	-4.77
Bank Loans to Output	78.96	1.37	-3.66
Bank Loans / Total Loans	78.64	1.38	-3.66
Avg. Loans s	0.19	-5.07	-5.13
Avg. Loans r	1.09	-3.95	-3.38
Avg. Loans n	10.26	-1.13	-1.12
Loan Mkt Share s	20.52	-13.79	-0.51
Loan Mkt Share r	44.19	9.97	-0.31
Loan Mkt Share n	35.28	-4.46	0.69
Total Number of Banks	194.66	-0.62	-0.63
N s Banks	137.81	-7.27	-0.13
N r Banks	52.41	16.91	-1.75
N n Banks	4.44	-1.15	-3.03

Source: Author's calculations.
Note: Pre-reform r_0^a = 1.50% and Post-reform r_0^a = 1.15%. The increase in the policy rate corresponds to an increase of 25 bp in the pre-reform (a 33% increase). T denotes the number of periods used to compute reported averages.

To attempt to decompose intensive versus extensive margin effects associated with monetary policy, table 8 shows how changes in fed funds rates affect the incentive to grow. There, we perform a counterfactual in the last column, where we compute changes two years after the increase in r^a, where we use investment decision rules from our baseline in the simulation to compare to equilibrium changes in the center column, i.e., just those from the last column of table 7.[18] Importantly, when controlling for the extensive growth margin, bank aggregate loan supply falls by nearly five percent in the counterfactual at $T = 2$, while aggregate bank loan supply increases by over two

18. Since investment comes at the end of any period, investment outcomes for are chosen prior to the shock so nothing is changed in for this counterfactual.

percent when using the equilibrium investment rules. This arises from growth in the number of regional banks despite the fact that average loans for each type fall. Thus, monetary policy has an impact on aggregate lending not only through the intensive margin but also via its effects on the composition of the banking industry.

6. Conclusion

In this paper we examine the consequences of geographic expansion and rising bank concentration for the bank lending channel. The model is consistent with smaller banks being more sensitive to monetary policy contractions, compatible with a corporate finance approach to banking as elucidated in Kashyap and Stein (2000). The model makes clear that monetary policy can also affect growth dynamics in the banking industry with implications for competition. Corbae and D'Erasmo (2023) provide a more general analysis of the data and implications of the model for financial stability and monetary policy.

References

Aguirregabiria, V., R. Clark, and H. Wang. 2020. "The Geographic Flow of Bank Funding and Access to Credit: Branch Networks, Local Synergies, and Competition." CEPR Discussion Paper No. DP13741. CEPR Press, Paris & London.

Arellano, M. and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58(2): 277–97.

Bellifemine, M., R.Jamilov, and T. Monacelli. 2022. "Hbank: Monetary Policy with Heterogeneous Banks." Source? CEPR Discussion Paper No. 17129. CEPR Press, Paris & London.

Berger, A.N. and T.H. Hannan. 1998. "The Efficiency Cost of Market Power in the Banking Industry: A Test of the 'Quiet Life' and Related Hypotheses." *Review of Economics and Statistics* 80(3): 454–65.

Berger, A.N., L.F. Klapper, and R. Turk-Ariss. 2009. "Bank Competition and Financial Stability." *Journal of Financial Services Research* 35(2): 99–118.

Berger, A.N. and L.J. Mester. 1997. "Inside the Black Box: What Explains Differences in the Efficiencies of Financial Institutions?" *Journal of Banking and Finance* 21(7): 895–947.

Bernanke, B. and M. Gertler. 1989. "Agency Costs, Net Worth, and Business Fluctuations." *American Economic Review* 79(1): 14–31.

Besanko, D. and U. Doraszelski. 2004. "Capacity Dynamics and Endogenous Asymmetries in Firm Size." *RAND Journal of Economics*: 23–49.

Buchak, G., G. Matvos, T. Piskorski, and A. Seru. 2018. "Fintech, Regulatory Arbitrage, and the Rise of Shadow Banks." *Journal of Financial Economics* 130: 453–83.

Carlstrom, C. and T. Fuerst. 1997. "Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis." *American Economic Review* 87(5): 893–910.

Corbae, D. and P. D'Erasmo. 2020. "Rising Bank Concentration." *Journal of Economic Dynamics and Control*. Volume 115.

Corbae, Dean and Pablo D'Erasmo. 2021. "Capital Buffers in a Quantitative Model of Banking Industry Dynamics." Econometrica 89:2975-3023.

Corbae, D. and P. D'Erasmo. 2023. "A Quantitative Model of Banking Industry Dynamics across Time and Space." Manuscript.

Den Haan, W.J., S.W. Sumner, and G.M. Yamashiro. 2007. "Bank Loan Portfolios and the Monetary Transmission Mechanism." *Journal of Monetary Economics* 54(3): 904–924.

Diamond, D.W. 1984. "Financial Intermediation and Delegated Monitoring." *Review of Economic Studies* 51(3): 393–414.

Ericson, R. and A. Pakes. 1995. "Markov Perfect Industry Dynamics: A Framework for Empirical Work." *Review of Economic Studies* 62(1): 53–82.

Farias, V., D. Saure, and G.Y. Weintraub. 2012. "An Approximate Dynamic Programming Approach to Solving Dynamic Oligopoly Models." *The RAND Journal of Economics* 43(2): 253–82.

Gelman, M., I. Goldstein, and A. MacKinlay. 2022. "Bank Diversification and Lending Resiliency." Available at SSRN 4147790.

Ifrach, B. and G.Y. Weintraub. 2017. "A Framework for Dynamic Oligopoly in Concentrated Industries." *Review of Economic Studies* 84(3): 1106–50.

Kashyap, A.K. and J.C. Stein. 1995. "The Impact of Monetary Policy on Bank Balance Sheets." In *Carnegie-Rochester Conference Series on Public Policy* 42: 151–95.

Kashyap, A.K. and J.C. Stein. 2000. "What Do a Million Observations on Banks Say about the Transmission of Monetary Policy?" *American Economic Review* 90(3): 407–28.

Liang, N. and S.A. Rhoades. 1988. "Geographic Diversification and Risk in Banking." *Journal of Economics and Business* 40(4): 271–84.

Menzio, G. and S. Shi. 2010. "Block Recursive Equilibria for Stochastic Models of Search on the Job." *Journal of Economic Theory* 145(4): 1453–94.

Morelli, J., M. Moretti, and V. Venkateswaran. 2023. "Geographical Expansion in U.S. Banking: A Structural Evaluation." Manuscript.

Shin, Y. 2022. "Deposit Market Competition and Equity Capital Issuance." Manuscript.

Wang, Y., T.M. Whited, Y. Wu, and K. Xiao. 2022. "Bank Market Power and Monetary Policy Transmission: Evidence from a Structural Estimation." *Journal of Finance* 77(4): 2093–141.

Weintraub, G.Y., C.L. Benkard, and B. Van Roy. 2008. "Markov Perfect Industry Dynamics with Many Firms." *Econometrica* 76(6): 1375–411.

Appendix A

A.1 Solution Algorithm

The analysis of Markov-Perfect Equilibrium with imperfect competition is generally limited to industries with just a few firms, less realistic than the number of banks we consider in this paper. The main restriction is that, since firms have market power, their decision rules are a function of the decision rules of all their competitors. Even if one were to restrict to symmetric strategies in which decision rules become a function of the industry state as we do, the number of industry states to be considered quickly becomes very large.

For this reason, we solve the model by adapting the approach in Farias and others (2012) to an environment with aggregate and regional shocks. The algorithm approximates a Markov-Perfect Equilibrium by assuming that firms, at each time, make decisions based on their own state and the average industry state (conditional on a set of finite moments) that prevail in equilibrium. This reduces the computational cost considerably since firms' decision rules are not explicitly a function of the sequence of industry states, but rather a function of the long-run average distribution. The results in Weintraub and others (2008) and Farias and others (2012) establish conditions under which this approximation works well asymptotically.

In our application, we approximate the industry state by assuming that it equals the average cross-sectional distribution. That is, when maximizing profits, banks choose the optimal level of loans, deposits, and securities competing against the long-run average distribution of banks. We denote that distribution by $\bar{\boldsymbol{\mu}}$. We quantitatively show that in our setup, where banks do not accumulate assets and regional shocks are i.i.d., this assumption approximates the observed distribution very well. Note that, given that we know the investment and exit decision rules for each bank type, we do not need to approximate the transition from $\bar{\boldsymbol{\mu}}$ to $\bar{\boldsymbol{\mu}}'$, i.e., unlike the Krusell-Smith method. Instead, we simply apply the transition operator $\boldsymbol{\mu}' = \mathcal{H}(\bar{\boldsymbol{\mu}}, N_e)$ in equation (33) from the text.

To find an equilibrium we perform the following steps:

1. Solve the problem of the entrepreneur (16)-(17) and derive the total loan-demand function (21). Given that the extreme-value distribution implies bank and nonbank market shares given in (18), we can calculate bank loan demand as in (22).

2. Solve the problem of the nonbank (34) to obtain the residual loan demand for bank loans.

3. Set tolerances $\in^{\ell}$,$\in^{I}$,$\in^{x}$,$\in^{e}$, and $\in^{\mu}$ to small values. Start with a number of entrants $N^{e,g}$ where iteration $g = 0$ is an initial guess.

4. Guess an investment decision rule $I^{h}(\cdot)$ and an exit decision rule $X^{h}(\cdot)$, where iteration $h = 0$ is an initial guess.

5. Using $N^{e,g}$, $I^{h}(\cdot)$, and $X^{h}(\cdot)$ and a large sequence of shocks $\{z_{j,t}, z_{-j,t}\}_{t=1}^{T}$, simulate the distribution of banks $\{\mu_t\}_{t=1}^{T}$. Discard the initial 250 periods and compute the average industry state $\bar{\mu}^{h}$ by taking the average of the observed distribution.[19]

6. Obtain an equilibrium in the loan market:

a. Guess a loan decision rule $\ell^{k}(\cdot)$ where iteration $k = 0$ is an initial guess.

b. For each $\{\theta,d\}$, given that the industry state $\bar{\mu}^{h}$ and $\ell^{k}(\cdot)$ determines the loan-supply function of a bank's competitors, obtain the best response $\ell^{k+1}(\cdot)$ by maximizing profits in equation (23).

c. Compute $\Delta^{\ell} = \|\ell^{k+1}(\cdot) - \ell^{k}(\cdot)\|$.

d. If $\Delta^{\ell} < \in^{\ell}$, an equilibrium in the loan market has been found, so continue to the next step. If not, return to step with the updated loan decision rule $\ell^{k+1}(\cdot)$.

7. Solve the bank problem to obtain investment and exit rules:

a. For each $\{\theta,d,s_j\}$, solve the bank problem in (29) to obtain $I^{h+1}(\cdot)$ and in (28) with $V^{x=1}(\cdot)$ given in (11) to obtain $X^{h+1}(\cdot)$.

b. Using $I^{h+1}(\cdot)$ and $X^{h+1}(\cdot)$, compute a new long-run industry state $\bar{\mu}^{h+1}$ using the transition operator in equation (33).

c. Compute $\Delta^{I} = \|I^{h+1}(\cdot) - I^{h}(\cdot)\|$, $\Delta^{x} = \|x^{h+1}(\cdot) - x^{h}(\cdot)\|$, and $\Delta^{\mu} = \|\bar{\mu}^{h+1} - \bar{\mu}^{h}\|$.

d. If $\Delta^{I} < \in^{I}$, $\Delta^{x} < \in^{x}$, and $\Delta^{\mu} < \in^{\mu}$ continue to the next step. If not, return to step with the updated industry state $\bar{\mu}^{h+1}$.

8. Obtain the value of an entrant (net of entry costs) $V^{e}(\bar{\mu}^{h+1})$ in equation (31). If $\|V^{e}(\bar{\mu}^{h+1})\| < \in^{e}$, an equilibrium has been found. If not, update the number of entrants $N^{e,g+1}$ and return to step 5 with the updated number of entrants. The update of $N^{e,g}$ is done taking into account the value of $V^{e}(\bar{\mu}^{h+1})$. If $V^{e}(\bar{\mu}^{h+1}) > 0$, set $N^{e,g+1} > N^{e,g}$. If $V^{e}(\bar{\mu}^{h+1}) < 0$, set $N^{e,g+1} < N^{e,g}$.

19. Note that, to simulate the distribution, you need an initial distribution. We assume that the distribution in period 1 equals one with a number $N^{e,g}$ of θ_3 banks. As we discard the initial periods to compute the average distribution, the selection of this initial distribution is not quantitatively relevant.

9. A final check on the equilibrium is how well the 'average' industry (conditional on Z) approximates the observed distribution along the equilibrium path. We compute the average distance between the observed distribution $\{\mu_t\}(t=1)^T$ and the average distribution μ and the values are small.

While the algorithm just described has been proven to converge, we also experimented with a slightly modified version, where we evaluate the value of the entrant for many possible values of the number of entrants and define an equilibrium as one where the condition in point 8 is satisfied. This modified version of the algorithm is more costly computationally but robust.

A.2 Data Description

As in Corbae and D'Erasmo (2021), we compile a large panel of banks from 1984 to 2019 using data for the last quarter of each year.[20] The source for the data is the Call Reports that banks submit to the Federal Reserve each quarter.[21] Call Report data are available for all banks regulated by the Federal Reserve System, the FDIC, and the Comptroller of the Currency. All financial data are on an individual-bank basis.

We consolidate individual commercial banks to the bank holding company level and retain those bank holding companies and commercial banks (if there is not top holder) for which the share of assets allocated to commercial banking (including depository trust companies, credit card companies with commercial bank charters, private banks, development banks, limited charter banks, and foreign banks) is higher than 25 percent. We follow Kashyap and Stein (2000) and Den Haan and others (2007) in constructing consistent time series for our variables of interest. Finally, we only include banks located within the fifty states and the District of Columbia. In addition to information from the Call Reports, we identify bank failures using public data from the FDIC.[22] We also identify mergers and acquisitions

20. There was a major overhaul to the Call Report format in 1984. Since 1984, banks are, in general, required to provide more detailed data concerning assets and liabilities. Due to changes in definitions and the creation of new variables after 1984, some of the variables are only available after this date.

21. Balance sheet and income statement items can be found at the FFIEC website.

22. Data is available at the FDIC website.

using the Transformation Table on the FFIEC website. We identify 'events' where the acquired and acquiring firms are commonly owned in some form before the acquisition (i.e., the listed merger is only a corporate reorganization) and discard these events from the merger sample.

To deflate balance-sheet and income-statement variables, we use the CPI index. When we report weighted aggregate time series, we use the asset market share as the weight. To control for the effect of a small number of outliers, when constructing the loan returns, cost of funds, charge-off rates, and related series, we eliminate observations in the top and bottom 1 percent of the distribution of each variable. We also control for the effects of bank entry, exit, and mergers by not considering the initial period, the final period, or the merger period (if relevant) of any given bank.

To analyze the bank lending channel, we follow Kashyap and Stein (1995) and extend our annual data to quarterly frequency. As before, we work with data from 1984 to 2019. We discard observations for banks that are involved in a merger, de novo banks in the period they enter, as well as the final period of banks that fail. Variables are defined as in Corbae and D'Erasmo (2021). We follow Kashyap and Stein (1995) and, for each of our size categories, we run the following specification (in panel fixed-effects form)

$$\Delta y_{i,t} = \sum_{h=1}^{8} \beta_h (f_{t-h} - f_{t-h-1}) + \sum_{h=1}^{4} \alpha_h \Delta y_{i,t-h} + \gamma_j X_t + \phi x_{i,t} + a_i + \tau_t + Q_t + \in_{i,t}, \tag{A.2.1}$$

where $\Delta y_{i,t}$ denotes the growth rate of $y_{i,t}$ (total loans net of C&I loans) between quarter t and quarter $t-1$, f_{t-h} corresponds to the fed funds rate in period $t-h$, X_t captures aggregate variables (such as the inflation rate or changes in nominal GDP), $x_{i,t}$ bank level controls that include the ratio of deposits to assets and the ratio of cash and securities to assets. a_i is a bank fixed effect, τ_t is a year fixed effect, and Q_t is a quarter fixed effect. The fed funds come from FRED and correspond to the end-of-quarter value.

A.3 Cost Estimation

We estimate the marginal cost of producing a loan $c_\theta(\ell_{\theta,t})$ and the fixed cost $C_{F,\theta}$ following the empirical literature on banking.[23,24] The marginal cost is derived from an estimate of marginal net expenses that is defined to be marginal noninterest expenses net of marginal noninterest income. Marginal noninterest expenses for bank j are derived from the following trans-log cost function:

$$\begin{aligned}\log NIE_t^j = {} & g_1 \log(w_t^j) + h_1 \log(\ell_t^j) + g_2 \log(q_t^j) + g_3 \log(w_t^j)^2 \\ & + h_2 [\log(\ell_t^j)]^2 + g_4 \log(q_t^j)^2 + h_3 \log(\ell_t^j) \log(q_t^j) + h_4 \log(\ell_t^j) \log(W_t^j) \\ & + g_5 \log(q_t^j) \log(W_t^j) + g_6^1 t + g_6^2 t^2 + g_{8,t} + g_9^j + \in_t^j,\end{aligned} \tag{A.3.2}$$

where $NIE_{\theta,t}^j$ is noninterest expenses (calculated as total expenses minus the interest expense on deposits, the interest expense on fed funds purchased, and expenses on premises and fixed assets), g_9^j is a bank fixed effect, W_t^j corresponds to input prices (labor expenses), ℓ_t^j corresponds to real loans (one of the two bank j's outputs), q_t^j represents safe securities (the second bank output), the t regressor refers to a time trend, and $k_{8,t}$ refers to time fixed effects. We estimate this equation by panel fixed effects with robust standard errors clustered by bank.[25] Noninterest marginal expenses are then computed as:

$$\begin{aligned}\text{Mg Non-Int Exp.} \equiv \frac{\partial NIE_t^j}{\partial \ell_t^j} = {} & \frac{NIE_t^j}{\ell_t^j} \left[h_1 + 2h_2 \log\left(\ell_t^j\right) \right. \\ & \left. + h_3 \log\left(q_{jt}\right) + h_4 \log\left(w_t^j\right) \right].\end{aligned} \tag{A.3.3}$$

Marginal noninterest income (Mg NonInt Inc.) is estimated by using an equation similar to equation (A.3.2) (without input prices), where the left-hand side corresponds to total noninterest income. Net marginal expenses (Net Exp.) are computed as the difference between marginal noninterest expenses and marginal noninterest income. The

23. See, for example, Berger and others (2009) and our previous paper Corbae and D'Erasmo (2021).

24. The marginal cost estimated is also used to compute our measure of markups and the Lerner Index.

25. We eliminate bank-year observations in which the bank organization is involved in a merger or the bank is flagged as being an entrant or a failing bank. We only use banks with three or more observations in the sample.

fixed cost $C_{F,0}$ is estimated as the total cost on expenses of premises and fixed assets. Table 5 presents the estimated average net expense, the fixed cost, as well as the average cost by bank size.

Estimating HANK for Central Banks

Sushant Acharya
Bank of Canada / Centre for Economic Policy Research

Marco Del Negro
Federal Reserve Bank of New York Centre for Economic Policy Research

Aidan Gleich
Federal Reserve Bank of New York

Ethan Matlin
Harvard University

Reca Sarfati
Massachusetts Institute of Technology

William Chen
Massachusetts Institute of Technology

Keshav Dogra
Federal Reserve Bank of New York

Shlok Goyal
Harvard University

Donggyu Lee
Federal Reserve Bank of New York

Sikata Sengupta
University of Pennsylvania

Central banks are very interested in investigating questions surrounding inequality and its relationship with monetary policy. This is arguably for very good reasons. First of all, inequality has become a central issue in many countries. It is therefore important to ask how central-bank policies affect inequality. Second, even if central bankers

We thank participants at the XXV Annual Conference of the Central Bank of Chile "Heterogeneity in Macroeconomics: Implications for Monetary Policy," for which this paper was written, and especially our discussant, Markus Kirchner, for helpful comments. We also thank Mikkel Plagborg-Møller for useful feedback and Pranay Gundam, Ramya Nallamotu, and Brian Pacula for research assistance. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Bank of Canada, the Federal Reserve Bank of New York, or the Federal Reserve System.

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

were not concerned with the answer to the above question, they ought to be concerned with the fact that inequality changes the transmission mechanism of monetary policy, as forcefully argued in Kaplan and others (2018) and Ahn and others (2018). Several central banks have indeed shown interest in these topics (in fact, the title of the conference on which this volume is based is "Heterogeneity in Macroeconomics: Implications for Monetary Policy") and a few have begun to develop models that speak to the interaction of monetary policy and inequality such as heterogeneous agent New Keynesian (HANK) models following the seminal work by Kaplan and others (2018).

Models serve many purposes, and for some of these purposes a model's ability to fit the data—that is, to adequately describe the data from a quantitative point of view—is important, especially for central banks. After all, the popularity of representative-agent DSGE models, such as Smets and Wouters (2007), henceforth, SW, since the beginning of the century is largely due to these models' ability to forecast with an accuracy that is at least comparable to that of other models previously used in central banks such as vector autoregressions. Even if forecasting is not the main purpose of a model—and arguably it is not the main purpose of DSGE models—, it is a way to test its reliability in providing answers to quantitative questions: forecasting accuracy lends quantitative credibility.

These considerations prompt us to ask: What is the forecasting accuracy of HANK models? To the extent that these models have a more realistic transmission mechanism than representative-agent models, one would hope that this translates into a better forecasting performance. This is particularly true for aggregate consumption, since the main difference between SW-type DSGEs and HANK models is the replacement of the representative-agent Euler equation, which determines consumption in standard DSGEs, with the aggregation of individual households' consumption policy functions. These consumption policy functions depend, among other things, on the wealth distribution in the economy, that is, on inequality. This is the first paper to our knowledge that provides an assessment of the out-of-sample forecasting accuracy of HANK models.

From a computational point of view, the task of performing an out-of-sample forecasting accuracy exercise is not trivial, as it involves estimating a HANK model over and over for each of the several vintages of data for which we want to compute forecasts. Concretely, our forecasting exercise begins in the first quarter of 2000 and ends in the last quarter of 2019, for a total of 80 periods. For each period we

estimate the model by using Bayesian methods—the same approach used by SW and much of the HANK literature. Each estimation is very costly in computational terms for HANK if one calculates the likelihood by using the Kalman filter since these models have a very large state-space which includes the distribution of wealth (both liquid and illiquid, in a two-asset HANK model) across households.[1]

All of the growing literature estimating HANK models using Bayesian methods[2] use the standard Markov chain Monte Carlo approach followed in the representative DSGE model literature to obtain draws from the posterior distribution,[3] and featured in popular packages such as Dynare. This approach has two drawbacks. First, it cannot be naturally parallelized, being a Markov chain-based algorithm. Second, one has to start every new estimation from scratch. For example, if one just estimated the model up to 2000.Q1 and then adds only one more quarter of data, with Markov chain methods, one has to start the Markov chain anew even though one suspects that the posterior distribution may not be all that different.

This paper deviates from this trend and uses a Monte Carlo method that can be readily parallelized—Sequential Monte Carlo. This parallelization makes it feasible to estimate models even when each likelihood computation takes a substantial amount of time. This method has another crucial advantage, namely, that models can be estimated 'online'. What online estimation means is that the swarm of particles describing the posterior distribution computed for the estimation up to 2000.Q1 can be used to jump-start the estimation with one or more quarters of data, thereby making it considerably faster. This online feature is what makes repeated estimation, and therefore our forecasting accuracy exercise, possible.[4] While these methods are not new,[5] one contribution of this paper is to explain how and why they work to an audience with little or no background in Monte Carlo methods, so that this paper may serve as a blueprint for central-bank researchers planning to estimate HANK models and

1. The advantage of the so-called sequence-space Jacobian approach to solving and estimating HANK models championed by Auclert and others (2021) is that it circumvents the issue of the large state-space associated with carrying around a set of distributions.

2. See Winberry (2018), Auclert and others (2021), Bayer and others (2022), and Lee (2021), among others.

3. For example, see An and Schorfheide (2007).

4. The methods described in this paper can be used in the context of limited information approaches, such as those used by Hagedorn and others (2018), who estimate a HANK model using impulse-response matching as in Christiano and others (2005).

5. See Cai and others (2021).

use them in routine policy analysis and forecasting exercises. We also plan to share the code used in our forecasting exercise at the GitHub.com page.

As anticipated above, the other contribution of this paper is to provide a forecasting accuracy assessment of a HANK model. While several HANK models have been developed, in this paper, we use that of Bayer and others (2022), henceforth, BBL. We do this because, in their frontier contribution, the authors put particular care in the empirical fit of their model, making sure that they include all the shocks and frictions that make SW-type models empirically successful. In other words, the BBL model is the closest thing to a HANK version of SW. We then ask: Does this model forecast macro time series better than the original SW? Unfortunately, the answer from our preliminary investigation is no. For some series such as inflation, the forecasting performance is similar. For other series, notably for consumption growth, the accuracy for the HANK model is much worse than for the representative-agent model, which is particularly disappointing for the reasons discussed above.

What are the reasons for, and the implications of, the relatively worse forecasting performance of this HANK model compared to SW? We suspect that one key reason is that many parameters in HANK—namely those affecting the model's steady state—are still calibrated. This is not necessarily for a philosophical choice on the part of the HANK modelers, but because recomputing the steady state is extremely costly. If this suspicion is correct, these findings pose a computational challenge to HANK researchers interested in estimation. For sure the findings should be interpreted as a motivation to do more research on HANK models, as opposed to sticking to representative-agent models. Inequality is one of the critical issues of our times—no matter the forecasting performance of HANK models. The fact that the latter can be improved is a stimulus for further efforts, especially from central-bank researchers who want to use these models for quantitative purposes.

In the remainder of the paper, section 1 presents BBL's model and solution approach to make the paper self-contained, section 2 describes the Sequential Monte Carlo algorithm and the online estimation approach used to perform the forecasting exercise, section 3 discusses the results, and section 4 concludes.

1. Model

This paper employs the HANK model developed by BBL, which augments standard New Keynesian DSGE models, such as those presented in SW or Christiano and others (2005), with heterogeneous agents and incomplete markets. The model incorporates standard shocks and frictions utilized in DSGE models. Moreover, it is also capable of reproducing notable characteristics of household heterogeneity that are deemed important in the literature such as heterogeneous wealth and income composition and the presence of wealthy hand-to-mouth households. BBL show that, when the model is estimated on aggregate data, it can reproduce the business-cycle dynamics of aggregate data as well as of observed U.S. inequality. As the model is entirely taken from BBL, we will provide only a brief description of the model environment below in order to make the paper self-contained.[6]

1.1 Households

There exists a unit mass of infinitely-lived households, indexed by i, that maximize their lifetime utility,

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u\left(c_{it}, n_{it} \mid h_{it}\right), \tag{1}$$

where β denotes the discount factor, c_{it} denotes consumption, and n_{it} denotes hours worked. The instantaneous utility function $u(\cdot)$ follows Greenwood and others (1988),

$$u\left(c_{it}, n_{it} \# h_{it}\right) = \frac{x_{it}^{1-\xi} - 1}{1-\xi}, x_{it} = c_{it} - h_{it}^{1-\tau^P} \frac{n_{it}^{1+\gamma}}{1+\gamma}, \tag{2}$$

6. We use the version of model available at the GitHub.com page as of June 2022, when we began this project. The latest version of the model, as described in Bayer and others (2022), has two minor differences compared to the model adopted in this paper. First, the latest version of the model has a different formulation for the liquid asset return. Specifically, BBL assume that entrepreneurs sell claims to a fraction of profits as liquid shares, and the liquid asset return is the weighted average of the interest on government bonds and the return on these shares, which consists of profit payouts and the realized capital gain. In addition, BBL assume that time-varying income risks respond to output growth, which makes income risks either procyclical or countercyclical. These modifications allow BBL to better explain inequality series and their income risk estimates with their model.

where $\bar{\tau}^P$ is the steady-state level of tax progressivity, ζ is the coefficient of relative risk aversion, γ is the inverse of the Frisch elasticity, and h_{it} is idiosyncratic labor productivity. There are two types of households: workers ($h_{it} \neq 0$) and entrepreneurs ($h_{it} = 0$). Idiosyncratic productivity $h_{it} = \frac{\tilde{h}_{it}}{\int \tilde{h}_{it} di}$ evolves as follows:

$$\tilde{h}_{it} = \begin{cases} \exp\left(\rho_h \log \tilde{h}_{it-1} + \epsilon_{it}^h \right) & \text{with probability } 1-\zeta \text{ if } h_{it-1} \neq 0, \\ 1 & \text{with probability } \iota \text{ if } h_{it-1} = 0, \\ 0 & \text{else.} \end{cases} \tag{3}$$

The above equation implies that workers become entrepreneurs with the probability ζ or continue to be workers with the probability $1 - \zeta$. While being workers, labor productivity h_{it} evolves according to an AR (1) process in logs with an autocorrelation coefficient ρ_h. The shocks ϵ_{it}^h are normally distributed with variance $\sigma_{h,t}^2$. This variance changes over time according to the following process:

$$\sigma_{h,t}^2 = \bar{\sigma}_h^2 \exp(\hat{s}_t) \tag{4}$$

$$\hat{s}_{t+1} = \rho_s \hat{s}_t + \epsilon_t^\sigma, \tag{5}$$

where the shocks ϵ_t^σ follow a normal distribution with zero mean and the standard deviation σ_s.[7]

Workers earn wage income $w_t h_{it} n_{it}$, where w_t is the real wage paid to households by labor unions. In addition, rents from unions Π_t^U are equally distributed among workers. Entrepreneurs become workers with the probability ι or maintain their entrepreneur status with the probability $1 - \iota$. When entrepreneurs become workers, their productivity becomes one. Entrepreneurs do not supply labor and, instead, receive profits Π_t^F generated in the firm sector, except for rents of unions.

7. In the latest version of BBL, they assume that the level of income risks is affected by the output growth, i.e., $\hat{s}_{t+1} = \rho_s \hat{s}_t + \Sigma_Y \frac{Y_{t+1}}{Y_t} + \epsilon_t^\sigma$. Depending on the sign of the coefficient Σ_Y, idiosyncratic income risks are either pro- or counter-cyclical in the model. This setup allows BBL to better capture the dynamics of income risks with their model.

Markets are incomplete, and households self-insure against income risks by saving in two types of assets: illiquid capital and liquid bonds. Capital as an asset is illiquid in the sense that only a fraction λ of households are allowed to adjust their capital holdings in each period. In contrast, households can freely adjust their liquid bond holdings.

The household's budget constraint can be written as

$$c_{it} + b_{it+1} + q_t k_{it+1} = b_{it} \frac{R_{it}}{\pi_t} + \left(q_t + r_t\right) k_{it} + \left(1 - \tau_t^L\right)\left(w_t h_{it} n_{it}\right)^{1-\bar{\tau}_t^P} + \mathbb{1}_{h_{it} \neq 0}\left(1 - \tau_t\right) \Pi_t^U + \mathbb{1}_{h_{it}=0}\left(1 - \tau_t^L\right)\left(\Pi_t^E\right)^{1-\bar{\tau}^P}, \; k_{it+1} \geq 0, \; b_{it+1} \geq \underline{b}, \tag{6}$$

where b_{it} is real liquid bonds, k_{it} is capital stock, q_t is the price of capital, r_t is dividend on capital holdings, $\pi_t = \frac{P_t}{P_{t-1}}$ is the gross inflation rate, and $\underline{b}<0$ is an exogenous borrowing limit. Workers' labor income and entrepreneurs' profit income are taxed progressively. The two tax rates, τ_t^L and τ_t^P, determine the degree of tax progressivity. The union profit is taxed uniformly at the average tax rate τ_t. Finally, the return on the liquid assets R_{it} depends on whether households are borrowers:

$$R_{it} = \begin{cases} A_t R_t^b & \text{if } b_{it} \geq 0 \\ A_t R_t^b + \overline{R} & \text{if } b_{it} < 0. \end{cases} \tag{7}$$

The coefficient A_t is the so-called "risk-premium shock" (see SW), which reflects intermediation efficiency, and $\overline{R}$ is the borrowing premium. R_t^b is the nominal interest rate on government bonds, which is determined by the monetary authority.[8]

Since households may or may not be able to adjust their illiquid asset holdings, the household's problem is characterized by three functions. The value function V_t^a when households are allowed to adjust their capital holdings, the function V_t^n when households are not allowed to adjust, and the expected value in the next period $\mathbb{W}_{t+1}$

$$V_t^a\left(b,k,h\right) = \max_{b_a', k'} u\left[\left(x\left(b, b_a', k, k', h\right)\right)\right] + \beta \mathbb{E}_t \mathbb{W}_{t+1}\left(b_a', k', h'\right), \tag{8}$$

8. In the model described in Bayer and others (2022), they assume that entrepreneurs sell claims to a fraction ω_Π of profits at the price of q_t^Π as liquid shares, and these shares become a part of the household's liquid asset portfolio as well. Thus, the liquid asset return is the weighted average of the return on government bonds and the return on profit shares. Consequently, dynamics of profit shares also affect the liquid asset return in the model.

$$V_t^n\left(b,k,h\right) = \max_{b_n} u\left[\left(x\left(b,b_n',k,k,h\right)\right)\right] + \beta \mathbb{E}_t \mathbb{W}_{t+1}\left(b_n',k,h'\right), \tag{9}$$

$$\mathbb{W}_{t+1}\left(b',k',h'\right) = \lambda V_{t+1}^a\left(b',k',h'\right) + \left(1-\lambda\right)V_{t+1}^n\left(b',k',h'\right), \tag{10}$$

where $x(b,b',k,k',h) = c\ (b,b',k,k',h) - h^{1-\bar{\tau}^P}\ \frac{n(w)^{1+\gamma}}{1+\gamma}$ is the household's composite demand for goods and leisure.[9] Maximization is subject to the budget constraint described above.

1.2 Firms

The firm sector comprises four types of firms; 1) final goods producers, 2) intermediate goods producers, 3) capital producers, and 4) labor packers. Final goods producers transform intermediate goods into final consumption goods. Intermediate goods producers produce differentiated goods using capital and labor service as inputs. Capital producers transform final goods into new capital stock, subject to adjustment frictions, and rent out capital to intermediate goods producers. Labor packers combine differentiated labor supplied by unions and rent out homogeneous labor services to intermediate goods producers. Intermediate goods producers and unions operate under a monopolistically competitive environment and set prices subject to nominal rigidity à la Calvo (1983).

2.2.1 Final Goods Producers

Final goods producers combine differentiated intermediate goods and make final consumption goods according to a CES aggregation technology:

$$Y_t = \left(\int y_{jt}^{\frac{\eta_t - 1}{\eta_t}} dj\right)^{\frac{\eta_t}{\eta_t - 1}}, \tag{11}$$

where y_{jt} is intermediate good j and η_t is the time-varying elasticity of substitution. Profit maximization yields the following individual good demand and the aggregate price index

9. Because of the specific form of Greenwood-Hercowitz-Huffman (GHH) preference used in the model, all workers supply the same amount of labor, depending on the level of real wage only.

$$y_{jt} = \left(\frac{p_{jt}}{P_t} \right)^{-\eta_t} Y_t \tag{12}$$

$$P_t = \int p_{jt}^{1-\eta_t} dj, \tag{13}$$

where p_{jt} is individual good j's price.

1.3 Intermediate Goods Producers

There is a continuum of intermediate goods firms, indexed by j, that produce differentiated goods, using capital and labor services, according to a constant return-to-scale production functions,

$$y_{j,t} = Z_t N_{jt}^{\alpha} \left(u_{jt} K_{jt} \right)^{1-\alpha}, \tag{14}$$

where α is the labor share in production, Z_t is total factor productivity that follows an AR(1) process in logs, N_{jt} is labor input, and $u_{jt}K_{jt}$ is capital input with the utilization rate u_{jt}. Capital depreciation rate depends on the degree of utilization according to $\delta\left(u_{jt}\right) = \delta_0 + \delta_1\left(u_{jt} - 1\right) + \frac{\delta_2}{2}\left(u_{jt} - 1\right)^2$. First-order conditions associated with the cost minimization are as follows:

$$w_t^F = \alpha \mathrm{mc}_{jt} Z_t \left(\frac{u_{jt} K_{jt}}{N_{jt}} \right)^{1-\alpha} \tag{15}$$

$$r_t + q_t \delta\left(u_{jt}\right) = u_{jt}\left(1-\alpha\right)\mathrm{mc}_{jt} Z_t \left(\frac{N_{jt}}{u_{jt} K_{jt}} \right)^{\alpha} \tag{16}$$

$$q_t \left[\delta_1 + \delta_2 \left(u_{jt} - 1 \right) \right] = (1-\alpha)\mathrm{mc}_{jt} Z_t \left(\frac{N_{jt}}{u_{jt} K_{jt}} \right)^{\alpha}, \tag{17}$$

where mc_{jt} is the marginal cost of production of firm j. Since the production function exhibits constant return-to-scale, the above optimality conditions imply that marginal costs are identical across producers, i.e., $\mathrm{mc}_{jt} = \mathrm{mc}_t$.

Firms operate under monopolistically competitive environments and set prices for their goods subject to price adjustment frictions à la Calvo (1983): only a fraction $1 - \lambda_y$ of firms can adjust their prices,

while the rest index their prices to the steady-state inflation rate $\bar{\pi}$. Firms maximize the present value of real profits,

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \lambda_Y^t \left(1-\tau_t^L\right) Y_t^{1-\tau_t^P} \left\{ \left(\frac{p_{jt}\,\bar{\pi}^t}{P_t} - \mathrm{mc}_t \right) \left(\frac{p_{jt}\,\bar{\pi}^t}{P_t} \right)^{-\eta_t} \right\}^{1-\tau_t^P}. \tag{18}$$

The corresponding optimality condition, with a first-order approximation, implies the following Phillips curve:

$$\log\left(\frac{\pi_t}{\bar{\pi}}\right) = \beta \mathbb{E}_t \log\left(\frac{\pi_{t+1}}{\bar{\pi}}\right) + \kappa_Y \left(\mathrm{mc}_t - \frac{1}{\mu_t^Y}\right), \tag{19}$$

where $\kappa_Y = \dfrac{(1-\lambda_Y)(1-\lambda_Y\beta)}{\lambda_Y}$ is the slope of Phillips curve, and $\mu_t^Y = \dfrac{\eta_t}{\eta_t - 1}$ is the target markup. The target markup follows an AR (1) process with shock $\in_t^{\mu^Y}$.

2.3.1 Capital Producers

Capital producers transform final goods into new capital stock, subject to adjustment frictions, while taking the price of capital q_t as given. They maximize

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t I_t \left\{ q_t \Psi_t \left[1 - \frac{\phi}{2} \left(\log \frac{I_t}{I_{t-1}} \right)^2 \right] - 1 \right\}, \tag{20}$$

where ϕ governs the degree of investment adjustment frictions, and ψ_t represents marginal efficiency of investment à la Justiniano and others (2011), which follows an AR(1) process in logs with innovation $\in_t^{\psi}$. Up to first order, the optimality condition for the maximization problem is

$$q_t \Psi_t \left[1 - \phi \log \frac{I_t}{I_{t-1}} \right] = 1 - \beta \mathbb{E}_t \left[q_{t+1} \Psi_{t+1} \phi \log \left(\frac{I_{t+1}}{I_t} \right) \right]. \tag{21}$$

Finally, the law of motion for aggregate capital is given by

$$K_t - \left(1 - \delta\left(u_t\right)\right) K_{t-1} = \Psi_t \left[1 - \frac{\phi}{2} \left(\log \frac{I_t}{I_{t-1}} \right)^2 \right]. \tag{22}$$

2.3.2 Unions and Labor Packers

There exists a unit mass of unions, indexed by l, which purchase labor services from workers and sell a different variety of labor to labor packers. Labor packers combine a different variety of labor into homogeneous labor input according to the following CES aggregation technology:

$$N_t = \left(\int \hat{n}_{lt}^{\frac{\zeta_t - 1}{\zeta_t}} \, dj \right)^{\frac{\zeta_t}{\zeta_t - 1}}, \tag{23}$$

where $\hat{n}_{lt}$ is a variety l labor service and ζ_t is the elasticity of substitution. Labor packers' cost minimization implies the following demand for each variety l of labor services:

$$\hat{n}_{lt} = \left(\frac{W_{lt}}{W_t^F} \right)^{-\zeta_t} N_t, \tag{24}$$

where W_{lt} is the nominal wage set by union l and W_t^F is the nominal wage at which labor packers sell labor input to intermediate goods producers.

Unions have monopolistic power and maximize their stream of profits by setting prices w_{lt} for their labor variety, subject to nominal rigidity à la Calvo (1983). Specifically, only $1 - \lambda_w$ fraction of unions can adjust wages, while the rest of unions index wages to the steady-state wage inflation rate. Thus, they maximize

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \lambda_w^t \frac{W_t^F}{P_t} N_t \left\{ \left(\frac{W_{lt}^{\bar{\pi}_W^t}}{W_t^F} - \frac{W_t}{W_t^F} \right) \left(\frac{W_{lt}^{\bar{\pi}_W^t}}{W_t^F} \right)^{-\zeta_t} \right\}. \tag{25}$$

From the optimality condition for the maximization problem, we obtain the wage Phillips curve under a first-order approximation:

$$\log\left(\frac{\pi_t^W}{\bar{\pi}_W} \right) = \beta \mathbb{E}_t \log\left(\frac{\pi_{t+1}^W}{\bar{\pi}_W} \right) + \kappa_\omega \left(\mathrm{mc}_t^\omega - \frac{1}{\mu_t^\omega} \right), \tag{26}$$

where $\kappa_\omega = \frac{(1-\lambda_w)(1-\lambda_w\beta)}{\lambda_W}$ is the slope of Phillips curve and $\pi_t^W \equiv \frac{W_t^F}{W_{t-1}^F} = \frac{w_t^F}{w_{t-1}^F}\pi_t^Y$ is the gross wage inflation rate with w_t and w_t^F being the real wages for households and firms, respectively. $\mathrm{mc}_t^\omega = \frac{w_t}{w_t^F}$ is the actual and $\frac{1}{\mu_t^\omega} = \frac{\zeta_t - 1}{\zeta_t}$ is the target markdown of wages that unions pay to households relative to wages they charge to intermediate goods producers. The target markdown follows an AR(1) process in logs that is subject to the wage markup shock $\in_t^w$.

1.4 Governments

The government sector consists of a fiscal and a monetary authority. The fiscal authority issues government bonds, levies taxes, and makes government purchases. The issuance of government bonds is governed by the following rule:

$$\frac{B_{t+1}}{B_t} = \left(\frac{B_t}{\bar{B}}\right)^{-\gamma_B}\left(\frac{\pi_t}{\bar{\pi}}\right)^{\gamma_\pi}\left(\frac{Y_t}{Y_{t-1}}\right)^{\gamma_Y} D_t,\ D_t = D_{t-1}^{\rho_D}\exp \in_t^D, \tag{27}$$

where D_t represents the government structural deficit, which evolves exogenously as an AR(1) process subject to the shock $\in_t^D$. The parameters γ_B,γ_π, and γ_Y represent how sensitively the deficit responds to the existing debt, the evolution of the inflation rate, and the output growth, respectively. The government also sets the average tax rate according to the rule:

$$\frac{\tau_t}{\bar{\tau}} = \left(\frac{\tau_{t-1}}{\bar{\tau}}\right)^{\rho_\tau}\left(\frac{B_t}{B_{t-1}}\right)^{(1-\rho_\tau)\gamma_B^\tau}\left(\frac{Y_t}{Y_{t-1}}\right)^{(1-\rho_\tau)\gamma_Y^\tau}. \tag{28}$$

The fiscal authority ensures that the average tax rate equals the target tax rate τ_t by adjusting the level parameter τ_L,

$$\tau_t = \frac{\mathbb{E}_t\left(w_t n_{it} h_{it} + \mathbb{1}_{h_{it}=0}\Pi_t^E\right) - \tau_L \mathbb{E}_t\left(w_t n_{it} h_{it} + \mathbb{1}_{h_{it}=0}\Pi_t^E\right)^{\tau^P}}{\mathbb{E}_t\left(w_t n_{it} h_{it} + 1_{h_{it}=0}\Pi_t^E\right)}. \tag{29}$$

The total tax revenue is $T_t = \tau_t\left(w_t n_{it} h_{it} + \mathbb{1}_{h_{it}\neq 0}\Pi_t^U + \mathbb{1}_{h_{it}=0}\Pi_t^E\right)$ and government purchases are determined by the balanced budget constraint, i.e., $G_t = B_{t+1} + T_t - R_t^b/\pi_t B_t$.

The monetary authority determines the nominal interest rate on government bonds according to the following Taylor rule with interest-rate smoothing:

$$\frac{R_{t+1}^b}{R_t^b} = \left(\frac{R_t^b}{\overline{R}^b}\right)^{\rho_R} \left(\frac{\pi_t}{\overline{\pi}}\right)^{(1-\rho_R)\theta_\pi} \left(\frac{Y_t}{Y_{t-1}}\right)^{(1-\rho_R)\theta_Y} \epsilon_t^R, \tag{30}$$

where $\overline{R}^b$ is the steady-state nominal interest rate. The coefficients ϕ_π and ϕ_Y represents the sensitivity of the policy rate to the evolution of price and output gap, respectively. The parameres ρ_R represents the degree of interest-rate smoothing.

1.5 Market Clearing Conditions

The model has four markets; goods, labor, liquid and illiquid asset markets. The only liquid asset in the model is government bonds.[10] Thus, the liquid asset market clearing condition is given by

$$B_{t+1} = B_{t+1}^d \equiv \int \left\{\lambda b_a^*(b,k,h) + (1-\lambda) b_n^*(b,k,h)\right\} d\phi_t(b,k,h), \tag{31}$$

where b_a^* and b_n^* are the optimal liquid asset choice of adjusting and nonadjusting households with liquid asset holding b, illiquid asset holding k, and productivity level h, respectively, and ϕ_t is the distribution of households over the idiosyncratic state space. The left-hand and right-hand sides of the above equation represent the aggregate liquid asset supply and demand, respectively. Similarly, the illiquid asset, i.e., capital, market clearing condition is given by

$$K_{t+1} = K_{t+1}^d \equiv \int \left\{\lambda k_a^*(b,k,h) + (1-\lambda) k_n^*(b,k,h)\right\} d\phi_t(b,k,h), \tag{32}$$

with k_a^* and k_n^* being optimal capital holding of adjusting and nonadjusting households with liquid asset holding b, illiquid asset holding k, and productivity level h.

The labor market clears when the following equation holds:

$$\int \hat{n}_{lt}\, dl = D_t^w N_t = \int h_{it} n_{it} d\phi_t(b,k,h), \tag{33}$$

10. In contrast, the most recent version of BBL has two kinds of liquid assets, government bonds and profits shares.

where $D_t^w = \int \left(\frac{W_{lt}}{W_t^F} \right)^{-\zeta_t} dl$ is the dispersion of wages set by unions and N_t is the aggregate labor input. The first two items of the above equation represent the demand of intermediate goods producers for a variety of labor, while the last item is the aggregate labor variety supplied by households. Once assets and labor markets clear, the goods market also clears because of Walras's law.

1.6 Numerical Method

Following Reiter (2009), BBL solve the model using a linearized solution technique. the first step is to write the equilibrium as a system of nonlinear difference equations as follows:

$$\mathbb{E}_t F\left(X_t^*, X_{t+1}^*\right) = 0, \tag{34}$$

where X_t^* is a vector of state and control variables in period t. Then, BBL linearize the above system around the nonstochastic steady-state and applies a standard perturbation method such as the one proposed by Klein (2000). However, without any further treatments, applying a standard perturbation method is infeasible since the size of the above system is very large due to many idiosyncratic state variables such as asset holdings, productivity levels, and working statuses. Thus, BBL follow Bayer and Luetticke (2020) and reduce the size of the two biggest components of the system, i.e., value functions and household distributions.

For the value functions, BBL use a discrete cosine transform (DCT), as proposed by Bayer and Luetticke (2020). All the value functions are written as linear interpolants based on a set of nodal values, and these nodal values are represented by DCT coefficients of Chebyshev polynomials as follows:

$$\hat{\mathbb{W}}_{b/k,t}\left(b_i, k_j, h_l\right) = \sum_{p,q,r} \theta_{\mathbb{W}_{b/k},t}^{p,r} T_p(i) T_q(j) T_r(l), \tag{35}$$

where $\hat{W}_{b/k}$ is the partial derivative of the continuation value $\mathbb{W}$ with respect to bond b, (capital k) holdings, $T_{p/q/r}(\cdot)$ are Chebyshev polynomials, and $\theta_{\mathbb{W}}^{p,q,r}$ are the corresponding DCT coefficients. In the above expression, BBL force very small coefficients to be zero in order to achieve size reduction. That is, they only keep enough of these

coefficients so as to approximate the original value functions with a certain threshold level of precision. In perturbing the system, they perturb these coefficients instead of the function values themselves.

BBL reduce the size of the distribution in a similar way. For the distribution, they only keep marginal distributions F_t^b, F_t^k, and F_t^h in the system and use a copula $C_t(\cdot)$, a functional relationship between marginals and the joint distribution, to recover the joint distribution from marginals. Then, the copula $C_t(\cdot)$ at time t is approximated by using Chebyshev polynomials,

$$\hat{C}_t\left(F_i^b, F_j^k, F_l^h\right) = \sum_{p,q,r} \theta_{C,t}^{p,q,r} T_p(i) T_q(j) T_r(l), \tag{36}$$

where $\hat{C}(\cdot)$ is the deviation of the copula at time t from its steady-state counterpart. Again, BBL reduce the size of the system by keeping only a small number of DCT coefficients $\theta_C^{p,q,r}$.

After the state space reduction, the dimension of the system decreases substantially and one can find a linearized solution rather quickly. However, for estimation, further acceleration of the solution method is required since one needs to efficiently evaluate the model's likelihood. To this end, BBL follow Bayer and Luetticke (2020) and only estimate a subset of parameters that do not affect the households' problem. BBL first partition X^* into the part related to household choices f and the aggregate part X,

$$\mathbb{E}_t F\left(X_t^*, X_{t+1}^*\right) = \mathbb{E}_t F\left(f_t, X_t, f_{t+1}, X_{t+1}\right), \tag{37}$$

Then, they obtain the following linearized system:

$$\begin{bmatrix} A_{ff} & A_{fX} \\ A_{Xf} & A_{XX} \end{bmatrix} \begin{bmatrix} f_t \\ X_t \end{bmatrix} = -\mathbb{E}_t \begin{bmatrix} B_{ff} & B_{fX} \\ B_{Xf} & B_{XX} \end{bmatrix} \begin{bmatrix} f_{t+1} \\ X_{t+1} \end{bmatrix}. \tag{38}$$

If only the parameters that do not affect the household problem are estimated, one only needs to update A_{XX} and B_{XX} during the estimation. Since the size of aggregate blocks A_{XX} and B_{XX} is relatively small, one can update the Jacobian rather quickly.

Finally, BBL perform a further model reduction, which relies on a factor representation of the idiosyncratic model part, i.e., the part related to household choices. Once they define objects in a way such

that $B_{fX} = B_{Xf} = 0$, they reduce the size of the system by applying a singular value decomposition (SVD) on the idiosyncratic model part. Specifically, they rewrite the linearized system as

$$\begin{bmatrix} B_{ff}^{-1}A_{ff} & B_{ff}^{-1}A_{fX} \\ A_{Xf} & A_{XX} \end{bmatrix} \begin{bmatrix} f_t \\ X_t \end{bmatrix} = \begin{bmatrix} \tilde{A}_{ff} & \tilde{A}_{fX} \\ \tilde{A}_{Xf} & \tilde{A}_{XX} \end{bmatrix} \begin{bmatrix} f_t \\ X_t \end{bmatrix} = -\mathbb{E}_t \begin{bmatrix} f_{t+1} \\ B_{XX}X_{t+1} \end{bmatrix}. \tag{39}$$

Then, by applying an SVD on $\tilde{\mathrm{A}}_{ff}$, i.e., $\tilde{\mathrm{A}}_{ff} = U\Sigma V$, and the Eckart-Young-Mirsky theorem, they obtain

$$\begin{bmatrix} V_1'U\pounds_1 & V_1'\tilde{A}_{fX} \\ \tilde{A}_{Xf}V_1 & \tilde{A}_{XX} \end{bmatrix} \begin{bmatrix} Y_t \\ X_t \end{bmatrix} \approx -\mathbb{E}_t \begin{bmatrix} Y_{t+1} \\ B_{XX}X_{t+1} \end{bmatrix}, \tag{40}$$

where V_1 refers to the rows in V that correspond to the largest singular values and $Y_t = V_{ft}'$. Since $\tilde{\mathrm{A}}_{ff}$ is independent of the estimated parameters, the SVD needs to be performed only infrequently. With this second-stage model reduction, the size of the model decreases drastically once again, and the QZ-decomposition needed to solve the system can take place within a relatively short amount of time, which makes the estimation feasible.

2. Online Estimation of HANK Models

The goal of this section is to describe a Monte Carlo approach that makes what we call 'online' estimation of HANK models possible. By online estimation, we mean estimation that can be conducted without starting from scratch as the dataset changes because, say, a new quarter of data is available. If estimating a model from scratch is nowadays a relatively trivial computational task for (linear) medium-scale DSGE models of the size of Smets and Wouters (2007), it becomes much more time-consuming and computer-intensive when the size of the state space becomes very large, as is the case for HANK models.

Online estimation can be useful to central-bank researchers who would like to use HANK models for forecasting. It can also be useful for academics who intend to run pseudo out-of-sample forecasting comparisons to assess the forecasting ability of HANK models, as we do in this paper, as these comparisons involve re-estimating the

model(s) for each vintage of data.[11] Finally, online estimation can also be used to quickly re-estimate a model after small changes such as, for instance, modifications of the prior, or any other relatively minor (or perhaps even major) alterations of the model.

The first part of this section describes the estimation problem and why the way it is currently handled by popular DSGE estimation packages such as Dynare may not be ideal for HANK models. The following subsection provides an intuitive description of an alternative estimation method—Sequential Monte Carlo (henceforth, SMC)—and explains why this approach is suitable for online estimation. While this section borrows much of the material from Cai and others (2021), it strives to be accessible to an audience with little or no prior knowledge of Monte Carlo methods.[12]

2.1 Bayesian Estimation of HANK Models by Using State-space Methods

The solution of the log-linearized version of the model described in section 1 produces the following transition equation:

$$s_t = T(\theta)s_{t-1} + R(\theta)\varepsilon_t, t = 1,..,T, \tag{41}$$

where s_t is the vector of states, θ is the parameter vector, and the shocks ε_t are independently and identically distributed according to $\varepsilon_t \sim N(0, Q(\theta))$. The measurement equation

$$y_t = Z(\theta)s_t + D(\theta) + u_t, t = 1,..,T \tag{42}$$

connects the latent states s_t to the vector of observables Y_t, where the measurement error shocks are independently and identically distributed according to $u_t \sim N(0, H(\theta))$. The likelihood of this linear, Gaussian state-space model $p(y_{1:T}|\theta)$ can be readily computed via the Kalman filter, where we use the notation $Y_{1:T}$ to denote the sequence of

11. Edge and Gürkaynak (2010), and Del Negro and Schorfheide (2013), are examples of forecasting comparisons using medium-scale DSGEs.

12. A terrific introduction to such methods is provided in textbooks such as Gelman and others (1995), Geweke (2005), and Herbst and Schorfheide (2015), where the latter focuses specifically on DSGE model estimation. We refer the reader to these textbooks for a more formal treatment of the ideas described below.

observations $\{y1, \ldots, y_T\}$. Using Bayes' law, the posterior distribution of the parameters $p(\theta|y_{1:T})$ is obtained from

$$p(\theta \mid y_{1:T}) = \frac{p(y_{1:T} \mid \theta)\, p(\theta)}{\int p(y_{1:T} \mid \theta)\, p(\theta)\, d\theta}, \tag{43}$$

where $p(\theta)$ represents our prior for the parameters.[13] The discussion so far applies to any log-linearized DSGE model and follows closely An and Schorfheide (2007), Del Negro and Schorfheide (2010), and Herbst and Schorfheide (2015). The peculiarity of HANK models is that the state-space vector s_t is extremely large, thus making the Kalman filter and hence the evaluation of the likelihood $p(y_{1:T}|\theta)$ very costly.[14]

Since the posterior $p(\theta|y_{1:T})$ does not follow any known distribution, we need Monte Carlo methods in order to obtain draws from it and describe the results of our inference on θ—that is, tabulate the posterior mean, the 90 percent posterior coverage intervals, et cetera. The most standard Monte Carlo algorithm used for this purpose when estimating DSGE models, and used in Dynare, is the Random-Walk Metropolis-Hastings (RWMH) algorithm, which is the so-called Markov chain algorithm in that it produces a chain of draws from the posterior distribution $\{\theta^{(1)}, \ldots, \theta^{(j)}, \ldots, \theta^{(J)}\}$. Loosely speaking, the algorithm works as follows: in order to obtain the draw $\theta^{(j)}$, you take the previous draw $\theta^{(j-1)}$, add some randomness to generate a proposal θ^* and then either accept (that is set $\theta^{(j)} = \theta^*$) or reject (that is set $\theta^{(j)} = \theta^{(j-1)}$) this proposal according to a formula that guarantees convergence of the chain to the desired ergodic distribution, that is, $p(\theta|y_{1:T})$.[15]

This is the algorithm used by almost all papers doing Bayesian estimation of DSGE models, including BBL, and, for medium-sized models, this algorithm has shown to work reasonably well. It has a few downsides however: 1) it is well known that RWMH may get stuck and fail to explore the entirety of the parameter space, especially in

13. Del Negro and Schorfheide (2009) discuss the choice of priors for DSGE models, and Müller (2012) provides an easy way to assess their influence on the results.

14. Herbst (2015) shows how the so-called "Chandrasekhar Recursions" formulas can substantially reduce the computational burden of evaluating the likelihood. One issue with these formulas is that they are far less generous than standard formulas in terms of accommodating missing data, which is why we do not use them in this paper.

15. Again, see An and Schorfheide (2007), Del Negro and Schorfheide (2010), or Herbst and Schorfheide (2015).

the presence of multi-modality;[16] 2) it cannot be parallelized, since it is a Markov chain; and 3) one has to start from scratch for any new estimation, even if the changes in the estimation settings are relatively minor so that one would not expect a major change in the posterior distribution (e.g., adding one more quarter of data). These issues are particularly serious for HANK models because their posterior distribution is harder to evaluate. For instance, one approach to dealing with problem (1) amounts to running very long chains, which increases the chances of visiting the entirety of the parameter space. Of course this approach is less appealing when computing $p(\theta^{(j)}|y_{1:T})$ is very costly. Similarly, the fact that the algorithm cannot be parallelized limits the extent to which one can take advantage of computer power to speed up the algorithm. While recent developments in Monte Carlo methods, such as Hamiltonian Monte Carlo,[17] have made Markov chain methods more efficient and to some extent amenable to parallelization, problem (3)—the fact that one has to start each estimation from scratch—makes SMC methods appealing. We describe these methods in the next section.

2.2 The Sequential Monte Carlo Algorithm

In order to appreciate how and why Sequential Monte Carlo works, it may be useful to take a brief detour into the early history of Monte Carlo methods and discuss an approach called Importance Sampling.[18] Let's say you do not know how to draw from the posterior $p(\theta|y_{1:T})$, but you can draw very efficiently from a proposal distribution $q(\theta)$. For example, $q(\theta)$ could be a Gaussian with mean $\hat{\theta} = \text{argmax}_{\theta} p(\theta|y_{1:T})$, the peak of the posterior, and with variance proportional to the negative of the inverse of the numerical second derivative of the posterior evaluated at $\hat{\theta}$. Then you can obtain $\{\theta^{(1)}, \ldots, \theta^{(j)}, \ldots, \theta^{(J)}\}$ independent draws from $q(\theta)$ and assign to each of these draws a weight $W^{(j)} = w^{(j)} / \left(\frac{1}{J} \sum_{j=1}^{J} w^{(j)} \right)$, where

$$w^{(j)} = p\left(y_{1:T} \mid \theta^{(j)}\right) p\left(\theta^{(j)}\right) / q\left(\theta^{(j)}\right) \propto p\left(\theta^{(j)} | y_{1:T}|\right) / q\left(\theta^{(j)}\right).$$

16. See, for instance, Herbst and Schorfheide (2014).

17. See Duane and others (1987); Neal and others (2011); Stan Development Team (2015), henceforth, HMC.

18. See Hammersley and Morton (1954) for an early example and the textbooks mentioned in footnote 13 for a more modern treatment.

In other words, the idea behind Importance Sampling is to draw from $q(\theta)$ and then do a change of measure from $q(\cdot)$ to the so-called target distribution (the actual posterior) by reweighing these draws. Note that the denominator in (43) is irrelevant in the computation of $w^{(j)}$ since it does not depend on θ, and that the $W^{(j)}$ are in any case re-normalized to sum up to J (the choice of J as the normalization constant, as opposed to the more conventional 1, is driven by numerical reasons). Given the swarm of particles $\left\{\theta^{(j)}, W^{(j)}\right\}_{j=1}^{J}$ produced by this approach, one can then approximate any object of interest $h(\theta)$ using the Monte Carlo average

$$\bar{h}_J = \frac{1}{J}\sum_{j=1}^{J} W_n^{(j)} h\left(\theta^{(j)}\right),$$

where for instance $h(\theta) = \theta$ if one wants to compute the mean.

This may sound like a very reasonable approach except that the accuracy of this approximation does not just depend on J, which can be easily increased, but also on the effective particle sample size

$$\widehat{ESS} = J / \left(\frac{1}{J}\sum_{j=1}^{J}\left(W^{(j)}\right)^2\right).$$

In other words, if $q(\theta)$ is a good proposal—in the example above, if the posterior is approximately Gaussian—, then for J reasonably large Importance Sampling delivers a good approximation of the object of interest: all the weights $W^{(j)}$ will be similar in magnitude and the effective sample size $\widehat{ESS}$ will not be much lower than J. If it is not a good approximation, then most of the weights will be close to zero, and $\widehat{ESS}$ will be much lower than J. In this situation, Importance Sampling fails. When the posterior is irregular, as is the case for many DSGEs, coming up with a good (global) approximation is nearly impossible, and this may partly explain why in DSGE estimation these methods have been abandoned in favor of Markov chain approaches such as RWMH.[19]

SMC brings Importance Sampling and the use a swarms of particles back into play for DSGE estimation thanks to two ideas. The

19. Importance Sampling-inspired approaches have remained very popular for filtering problems however, such as the particle filter—see Fernández-Villaverde and Rubio-Ramírez, 2007.

first idea is that if you can pick the posterior you want to approximate, then the problem of choosing a suitable proposal becomes much easier. For instance, if the posterior is

$$p_n\left(\theta \mid y_{1:T}\right) = \frac{p\left(y_{1:T} \mid \theta\right)^{\phi_n} p\left(\theta\right)}{\int p\left(y_{1:T} \mid \theta\right)^{\phi_n} p\left(\theta\right) d\theta}, \tag{44}$$

with ϕ_n being a very small number, then the prior $p(\theta)$ is likely to work pretty well as a proposal: by construction, the target is almost the same as the proposal. Of course, $p_n(\theta|y_{1:T})$ constructed with ϕ_n close to zero is not what we want to approximate in the end. So we can increase ϕ_n progressively toward 1, and use the $n-1$ swarm as a proposal for the stage n target, making sure that at each stage n the target and the proposal remain reasonably close.

If the swarm of particles is still that generated by the prior, all this slicing into intermediate steps would amount to nothing: the prior is a poor proposal for the eventual posterior, and the effective sample size will likely still be very low. But the second idea, which borrows from Markov chain methods, comes to the rescue: from one stage to the other, particles can travel. Just like a single particle in RWMH travels around the posterior, and naturally tends to visit regions of the parameter space where the posterior places non-negligible mass, so can each of the particles in the swarm $\left\{\theta_n^{(j)}, W_n^{(j)}\right\}_{j=1}^{J}$. In other words, the particles can adapt as ϕ_n increases toward 1, so that in the end we have a good approximation of the posterior distribution.[20]

Formally, the SMC algorithm goes as follows:

Algorithm 1 (SMC Algorithm).

1. Initialization. ($\phi_0 = 0$). Draw the initial particles from the prior: $\theta_1^{(j)} \overset{iid}{\sim} p\left(\theta\right)$ and $W_1^{(j)} = 1, j = 1, ..., J$.

20. A little bit of history: In the statistics literature, Chopin (2002) showed how to adapt particle filtering techniques to conduct posterior inference for a static parameter vector. John Geweke played an important role popularizing these techniques in economics—e.g., Durham and Geweke (2014)—, and Creal (2007) was the first paper that applied SMC techniques to posterior inference in a DSGE model. Herbst and Schorfheide (2014) was quite impactful, as it showed that a properly tailored SMC algorithm delivers more reliable posterior inference for the Smets and Wouters (2007) DSGE model with loose priors and a multimodal posterior distribution than the widely used RWMH algorithm. They also provide some convergence results for an adaptive version of the algorithm building on Chopin (2004).

2. Recursion. For $n = 1, ..., N_\phi$,

(a) Correction. Re-weight the particles from stage n – 1 by defining the incremental weights

$$\tilde{w}_n^{(j)} = p\left(y_{1:T} \mid \theta_{n-1}^{(j)}\right)^{\phi_n - \phi_{n-1}} \tag{45}$$

and the normalized weights

$$\tilde{W}_n^{(j)} = \frac{\tilde{w}_n^{(j)} W_{n-1}^{(j)}}{\frac{1}{J}\sum_{j=1}^{J} \tilde{w}_n^{(j)} W_{n-1}^{(j)}}, j = 1, \ldots, J. \tag{46}$$

(b) Selection (Optional). Resample the swarm of particles $\left\{\theta_{n-1}^{(j)}, \tilde{W}_n^{(j)}\right\}_{j=1}^{J}$ and denote resampled particles by $\left\{\hat{\theta}^{(j)}, W_n^{(j)}\right\}_{j=1}^{J}$, where $W_n^{(j)} = 1$ for all j.

(c) Mutation. Propagate the particles $\left\{\hat{\theta}_i, W_n^{(j)}\right\}$ via N_{MH} steps of an MH algorithm with transition density $\theta_n^{(j)} \sim K_n\left(\theta_n \mid \hat{\theta}_n^{(j)}; \zeta_n\right)$ and stationary distribution $p_n(\theta|y_{1:T})$.[21]

3. For $n = N_\phi (\phi_{N\phi} = 1)$, the final Importance Sampling approximation of $\mathbb{E}_\pi\left[h(\theta)\right]$ is given by:

$$\bar{h}_{N_\phi, N} = \sum_{j=1}^{J} h\left(\theta_{N_\phi}^{(j)}\right) W_{N_\phi}^{(j)}. \tag{48}$$

Step 2a is the same as in Importance Sampling, where the proposal is the previous stage's posterior $p_{n-1}(\theta|y_{1:T})$ and the target is $p_n(\theta|y_{1:T})$. Step 2c is the Metropolis-Hastings step, where each particle is given a chance to adapt to the new posterior. Step 2b needs some discussion.

21. The transition kernel $K_n(\theta_n \mid \hat{\theta}_n; \zeta_n)$ needs to have the following invariance property:

$$p_n(\theta_n \mid y_{1:T}) = \int K_n(\theta_n \mid \hat{\theta}_n; \zeta_n)\, p_n(\hat{\theta}_n \mid y_{1:T}) d\hat{\theta}_n. \tag{47}$$

Thus, if $\hat{\theta}_n^{(j)}$ is a draw from the stage n posterior $p_n(\theta_n \mid y_{1:T})$, then so is $\theta_n^{(j)}$. The MH accept-reject probabilities ensure that such property is satisfied. In our application we follow Herbst and Schorfheide (2014) and Cai and others (2021) in our choice of $K_n(\theta_n \mid \hat{\theta}_n; \zeta_n)$, but developments in MC algorithms, such as HMC, can be used to make this step, and hence the whole SMC algorithm, more efficient. Farkas and Tatar (2020) is an example of a paper that combines SMC with HMC.

Its purpose is to make sure that, if the weights of the particles in the swarm become very uneven, and effective particle sample size

$$\widehat{ESS}_n = J/\left(\frac{1}{J}\sum_{j=1}^{J}\left(\tilde{W}_n^{(j)}\right)^2\right),$$

falls below a threshold $\underline{J}$, a new swarm of particles is generated from the old swarm so that all the particles have the same weight.[22]

One aspect of the algorithm we have not yet discussed is the number of stages N_ϕ as well as the schedule $\{\phi_1, \ldots, \phi_{N_\phi}\}$. In estimating the Smets and Wouters (2007) model, Herbst and Schorfheide (2014) find that $N_\phi = 500$ and a schedule given by the function $\phi_n=(n/N_\phi)^{2.1}$ works well. The convexity of the schedule implies that ϕ_n increases very slowly at the beginning and faster at the end. Of course, it is far from obvious that whatever schedule works well for the Smets and Wouters (2007) model also works well for a HANK or any other DSGE model. In this respect, Cai and others (2021) improve upon Herbst and Schorfheide (2014) by making the schedule $\{\phi_1, \ldots, \phi_{N_\phi}\}$ adaptive—that is, endogenous to the difficulty of the problem. Recall that the ESS measures, loosely speaking, the deterioration of the quality of the swarm $\left\{\theta_n^{(j)}, W_n^{(j)}\right\}_{j=1}^{J}$: if ESS is low, the swarm has essentially 'lost' most of its particles as the weights have become very uneven. Adaptation is then achieved by setting at each stage $\phi_n=\phi$ where ϕ solves

$$\widehat{ESS}(\phi) - (1-\alpha)\widehat{ESS}_{n-1} = 0,$$

where

$$\tilde{w}^{(j)}(\phi) = \left[p\left(y_{1:T} \mid \theta_{n-1}^{(j)}\right)\right]^{\phi-\phi_{n-1}}, \tilde{W}^{(j)}(\phi) = \frac{\tilde{w}^{(j)}(\phi) W_{n-1}^{(j)}}{\frac{1}{J}\sum_{j=1}^{J}\tilde{w}^{(j)}(\phi) W_{n-1}^{(j)}}, \widehat{ESS}(\phi)$$

$$= N/\left(\frac{1}{J}\sum_{j=1}^{J}\left(\tilde{W}_n^{(j)}(\phi)\right)^2\right).$$

22. Loosely speaking, particles with relatively large weight $\tilde{W}_n^{(j)}$ —that is, that are in high posterior regions of the parameter space—are given an opportunity to 'procreate' (generate a number of children that is in expected values proportional $\tilde{W}_n^{(j)}$), while particles with relatively small weight ($\tilde{W}_n^{(j)}$ <1) —that is, that are in regions of the parameter space with very little mass—are 'killed' with probability 1–$\tilde{W}_n^{(j)}$. There are many resampling schemes—see Liu (2001) or Cappé and others (2005). We use systematic resampling in the applications below.

The above formulas can be understood as follows: Pick a desired deterioration α of the effective sample size between stages n–1 and n, and set ϕ_n so as to achieve exactly such deterioration.[23] The parameter α expresses the degree of 'carefulness' of the researchers, bearing in mind that lower α's imply a longer estimation time.[24] Once α is chosen, the schedule becomes endogenous to the difficulty of the problem as measured by the deterioration of the ESS.

This section concludes with a description of the some of virtues of SMC. First, for a suitably large choice of the size of the swarm J, it is robust to irregular shapes of the posterior such as multi-modality, as shown in Herbst and Schorfheide (2014) and Cai and others (2021) among others. This is because the initial swarm $\{\theta_0^{(j)}, W_0^{(j)}\}_{j=1}^J$ is drawn from the prior and hence covers for large enough J any region of the parameter space where the prior places non-negligible mass. Hence if the posterior has many modes, there ought to be some initial particles in the neighborhood of such modes. Second, most of the SMC steps, such as the computation of the incremental weights in Step 2a and, most importantly, the mutation step in Step 2c, can be parallelized. Third, the algorithm produces an approximation of the marginal likelihood as a by-product. In fact, using the definitions of $\tilde{w}_n^{(j)}$ and $\tilde{W}_{n-1}^{(j)}$ one can see that

$$\frac{1}{J}\sum_{i=1}^{N}\tilde{w}_n^{(j)}\,\tilde{W}_{n-1}^{(j)} \approx \int \frac{p\left(y_{1:T} \mid \theta\right)^{\phi_n}}{p\left(y_{1:T} \mid \theta\right)^{\phi_{n-1}}}\left[\frac{p\left(y_{1:T} \mid \theta\right)^{\phi_{n-1}}}{\int p\left(y_{1:T} \mid \theta\right)^{\phi_{n-1}}\, p(\theta)d\theta}\right]d\theta \tag{49}$$

$$= \frac{\int p\left(y_{1:T} \mid \theta\right)^{\phi_n}\, p(\theta)d\theta}{\int p\left(y_{1:T} \mid \theta\right)^{\phi_{n-1}}\, p(\theta)d\theta}.$$

This implies that the product $\prod_{n=1}^{N_\phi}\left(\frac{1}{J}\sum_{j=1}^{J}\tilde{w}_n^{(j)}\, W_{n-1}^{(j)}\right)$ approximates the marginal likelihood as long as the prior is proper ($\int p(\theta)d\theta=1$) since all the terms cancel out except for $\int p(y_{1:T}|\theta)p(\theta)d\theta / \int p(\theta)d\theta$. Fourth, and perhaps most important for this application, the final swarm of particles $\{\theta_{N\phi}^{(j)}, W_{N\phi}^{(j)}\}_{j=1}^J$ can be reused, making recursive estimation of the model very convenient. We are going to turn to this feature next.

23. See Cai and others (2021) for a more detailed description.

24. In the light of the results in Cai and others (2021), we choose α=5 percent in this application.

2.3 Online Estimation

Imagine you have run the SMC algorithm 1 and have a swarm of particles $\{\theta^{(j)}, W^{(j)}\}_{j=1}^{J}$ that approximates well the posterior $p(\theta \mid y_{1:T})$. Expression (45) in Step 2a of the algorithm can be generalized as

$$\tilde{w}_n^{(j)} = \frac{p_n\left(Y \mid \theta_{n-1}^{(j)}\right)}{p_{n-1}\left(Y \mid \theta_{n-1}^{(j)}\right)}, \tag{50}$$

where we now use the more generic notation Y for $y_{1:T}$ for reasons that will soon become apparent. Note that in (45) we considered the special case where the stage- n likelihood $p_n(\mathrm{Y} \mid \theta) = p(\mathrm{Y} \mid \theta)^{\phi_n}$.

Imagine that you now want to obtain the posterior for a different model $\tilde{p}(\cdot \mid \theta)$ (but with the same parameter vector θ) estimated on a different dataset $\tilde{\mathrm{Y}}$:

$$\tilde{p}(\theta \mid \tilde{Y}) = \frac{\tilde{p}(\tilde{Y} \mid \theta)\, p(\theta)}{\int \tilde{p}(\tilde{Y} \mid \theta)\, p(\theta) d\theta}. \tag{51}$$

The simplest possible case is the one where the model is the same ($\tilde{p}(\cdot \mid \theta) = p(\cdot \mid \theta)$), and the dataset has one more time series observation ($\tilde{\mathrm{Y}} = y_{1:T+1}$), but the algorithm can accommodate situations where the data has been revised, or the model changed. Draws for the posterior $\tilde{p}(\theta \mid \tilde{\mathrm{Y}})$ can be readily obtained from algorithm 1 after replacing expression (45) with expression (50) and using the stage- n likelihood function[25]

$$\tilde{p}_n(\tilde{Y} \mid \theta) = \tilde{p}(\tilde{Y} \mid \theta)^{\phi_n}\, p(Y \mid \theta)^{1-\phi_n}. \tag{52}$$

In other words, we use the posterior distribution $p(\theta \mid \mathrm{Y})$ as a 'bridge' to obtain the new posterior $\tilde{p}(\theta \mid \tilde{\mathrm{Y}})$, as opposed to starting from the prior distribution. To the extent that the differences between $\tilde{p}(\tilde{\mathrm{Y}} \mid \theta)$ and $p(\mathrm{Y} \mid \theta)$ are not large, the swarm from $p(\theta \mid \mathrm{Y})$ should offer a fairly good starting point for the SMC algorithm.[26] This is the approach we use to estimate the BBL HANK model recursively. In particular, we start from the end-of-sample estimation $p(\theta \mid y_{1:T})$ and then proceed

25. See Cai and others (2021).

26. The initialization step in algorithm 1 needs to be modified so that the swarm $\{\theta^{(j)}, W^{(j)}\}_{j=1}^{J}$, possibly after a selection step 2b so that all the $W^{(j)}$'s equal 1, replaces the swarm drawn from the prior.

backward using formula (52) with $\tilde{Y}=y_{1:T-\tau}$ and $Y=y_{1:T-\tau+1}$, for $\tau=1,\ldots,\bar{\tau}$. We should stress that doing the online recursive estimation backward or forward—that is, starting from $p(\theta \mid y_{1:T-\bar{\tau}})$ and using this as a bridge to obtain $p(\theta \mid y_{1:T-\bar{\tau}+1})$, and so on—should make no difference, as both procedures recover $p(\theta \mid y_{1:T-\tau})$[27]

We conclude this section by highlighting some of the potentials of this approach besides the online estimation of HANK models. Mlikota and Schorfheide (2022) introduce the notion of "model tempering". If a model is very costly to estimate from scratch, one can save a lot of time by first estimating a coarser version of the model that is much cheaper to estimate (e.g., the linearized version of a nonlinear model), and then using that as a bridge to estimate the full model. Mlikota and Schorfheide (2022) use this approach to estimate a nonlinear model.

3. Results

This section describes the forecasting results and is divided into three parts. The first part describes the setup of the exercise, including the data. The second part discusses the estimates of the parameters, focusing on the differences between the original BBL results and those obtained using the SMC algorithm. The last part covers the forecasting horse race between BBL and SW.

3.1 Setup

For our exercise, we use the dataset made available by BBL online at the GitHub.com page as of June 2022, when we began this project. This dataset comprises the seven variables used by SW in the estimation of their model, namely the growth rate of per-capita real output, consumption, investment, and wages, the logarithm of hours worked per capita, GDP deflator inflation, and the federal funds rate.[28] In their database, these variables are available at the quarterly frequency from 1954.Q3 to 2019.Q4. In addition, BBL estimate their model by adding four variables that reflect various aspects of inequality and are not used in standard representative-agent DSGE estimation. These are the wealth and income shares of the top 10 percent, estimates of tax progressivity constructed following Ferriere

27. In particular there is no sense in which the backward procedure introduces any hindsight bias: by the time that ϕ_n in (52) reaches 1, the posterior draws no longer condition on $Y=y_{1:T-\tau+1}$.

28. During the zero-lower-bound period, the authors use the shadow rate measure created by Wu and Xia (2016).

and Navarro (2018), and estimates of idiosyncratic income risk from Bayer and others (2019). The top 10 percent shares are available annually from 1954 to 2019, the tax progressivity measure is available annually from 1954 to 2017, and the idiosyncratic income risk measure is available from 1983.Q1 to 2013.Q1.[29] The likelihood computation of the state space model easily accommodates missing data.

BBL demean all the time series prior to estimation. While this is not standard practice in the DSGE estimation and forecasting literature or in central banks' practice,[30] we follow BBL because adding a constant would imply introducing steady-state growth and inflation and therefore altering their model. We chose not to do this in order to remain as close as possible to BBL's specification. This choice has two implications. First, we have to use their dataset also for the forecasting exercise—that is, the demeaned data is what the model's forecasts are evaluated upon. Second, we estimate the competitor in the horse race—the SW model—also on demeaned data, which implies that we drop the constant from SW's measurement equations.[31]

The out-of-sample forecasting exercise begins in the first quarter of 2000 (in the notation of section 2, $T-\bar{\tau} = 2000Q1$) and ends in the last quarter of 2019 (T=2019Q4), for a total of 80 periods. In order to avoid hindsight bias, for each period we re-estimate the model using only data available up to that period.[32] For each model M_m under consideration (BBL, SW), we then generate horizon-h mean forecasts $[y_{T-\tau+h}|y_{1:T-\tau}, M_m]$ for the variables of interest using the state space model consisting of equations (41) and (42), and compare these forecasts with actual outcomes $y_{T-\tau+h}$.[33] As discussed above in section 2, we estimate the model in period T$-\tau$ using the posterior distribution for T$-\tau$+1 as a bridge in the SMC algorithm. For the sake of robustness, we start this process from two different posterior

29. See Bayer and others (2022) for a more detailed description of the dataset.

30. For example, see Del Negro and Schorfheide (2013), Cai and others (2019).

31. We should note that BBL have a representative-agent version of their HANK model, which are worse fits for the seven macro variables than the heterogeneous agents version in terms of marginal likelihood. Given that the the actual SW model is available, this is what we use in the horse race as the alternative to the BBL HANK model. From the perspective of a central bank choosing whether to use a representative-agent or a HANK model for predictions, arguably the choice is between SW and BBL.

32. In the forecasting literature it is customary to perform so-called pseudo real-time forecasting, where the data vintage available the time $T-\tau$ is used for estimation, as opposed to the revised data (here, the T vintage). The demeaning of the data and the fact that there are no vintages for the inequality series makes this pseudo real-time exercise not possible.

33. In order to compute the expectation $\mathbb{E}[y_{T-\tau+h}|y_{1:T-\tau}, M_m]$ using (41) and (42), only the filtered states $s_{T-\tau|T-\tau} = \mathbb{E}[s_{T-\tau}|y_{1:T-\tau}, M_m]$ are needed, which are obtained from the Kalman filter.

distributions $p(\theta \mid y_{1:T})$. One distribution consists of the draws made available by BBL at the GitHub.com page (which is the reason we do the online estimation backward since these draws are only available for T=2019Q4), and the other is based on an SMC estimation starting from the prior. For all SMC estimations, we use a swarm of J=10k particles.[34] The next section discusses these posterior estimates in some detail.

3.2 Estimation

This section presents the results from our estimation. Specifically, we discuss the prior distribution, which is the same as BBL's, and posteriors from the full-sample SMC estimation using the eleven variables described in the previous section. Also, we show the posteriors from SMC estimation when using only seven aggregate variables and excluding data on inequality, tax progressivity estimates, and income risk estimates. For comparison, we present BBL's estimation results obtained from BBL's GitHub.com page as of June 2022. [35] As mentioned above, in order to make the estimation feasible BBL calibrate, as opposed to estimate, several of the model's parameters. Table 1 shows the values of these calibrated parameters.

3.2.1 Priors

For parameters related to monetary policy, BBL impose normal distribution with a mean of 1.7 and standard deviation of 0.3 for θ_π, while imposing normal distribution with a mean of 0.13 and standard deviation of 0.05 for θ_Y. For the interest-rate smoothing parameter ρ_R, they assume a beta distribution with parameters (0.5,0.2).

Regarding fiscal policy, the debt-feedback parameter γ_B in the bond issuance rule is assumed to follow a gamma distribution with a mean of 0.10 and standard deviation of 0.08, which implies that the prior for the autocorrelation of government debt is 0.9. For the responsiveness of government debt to inflation and output growth, γ_π and γ_Y, they impose standard normal distributions. Similarly, they assume beta distributions with a mean of 0.5 and standard deviation of 0.2 for the autoregressive parameters in the tax rules, ρ_P, and ρ_τ. The feedback

34. We obtain nearly identical results when using 1k particles, at least in terms of RMSEs.

35. As mentioned, BBL made a few changes to their model and calibrated parameters since June 2022. Hence these MH estimates do not replicate the results presented in the most recent version of the paper.

parameters for average tax rates, γ_Y^{τ}, and γ_B^{τ}, are assumed to follow standard normal distributions.

For the structural shocks, BBL assume beta distributions with a mean of 0.5 and a standard deviation of 0.2 for the autocorrelation parameters and inverse-gamma distributions with a mean of 0.001 and a standard deviation of 0.02 for standard deviations of shocks. Finally, for idiosyncratic income risks, BBL impose a beta distribution with a mean of 0.7 and a standard deviation of 0.2 for autocorrelation parameters and a gamma distribution with a mean of 0.65 and a standard deviation of 0.03.

Table 1. Calibration

Par.	*Value*	*Description*
Households: Income process		
ρ_h	0.980	Persistence labor productivity
σ_h	0.120	Std. dev. labor productivity
ι	0.063	Trans. prob. from entrepreneurs to workers
ζ	1/3750	Trans. prob. from workers to entrepreneurs
Households: Financial frictions		
λ	0.095	Portfolio adj. prob.
$\overline{R}$	0.017	Borrowing premium
Households: Preferences		
β	0.984	Discount factor
ξ	4.000	Relative risk aversion
γ	2.000	Inverse of Frisch elasticity
Firms		
α	0.682	Share of labor
δ_0	0.022	Depreciation rate
$\overline{\eta}$	11.000	Elasticity of substitution
$\overline{\zeta}$	11.000	Elasticity of substitution
Government		
$\overline{\tau}^L$	0.175	Tax rate level
$\overline{\tau}^P$	0.12	Tax progressivity
$\overline{R}^b$	1	Gross nominal rate
$\overline{\pi}$	1	Steady-state inflation rate

Source: Authors' calculations.

Regarding variable capital utilization, BBL assume a gamma distribution with a mean of 5.0 and standard deviation of 2.0 for $\delta s = \delta_2/\delta_1$. Similarly, they impose a gamma distribution with a mean of 4.0 and standard deviation of 2.0 for ϕ, the parameter that governs investment adjustment costs. For the slopes of price and wage Phillips curves, K_Y and K_w, they adopt Gamma priors with a mean of 0.10 and standard deviation of 0.03. The prior mean for these parameters implies that the average duration of price and wage is four quarters.

3.2.2 Posteriors

The posterior distributions using the full sample (T=2019Q4) are displayed in tables 2 and 3. Column 5 of each table shows BBL's original posteriors that they obtained using the RWMH algorithm, referred to as the MH estimation hereafter. Columns 6 and 7 show the posterior distributions we obtained via the SMC approach using eleven and seven variables, respectively, referred to as the 11 and 7 var SMC estimations hereafter. In the 7 var SMC estimation, we follow BBL and shut down income risk and tax progressivity shocks as we do not use the related data in the estimation.

The posteriors from the 11 and 7 var SMC estimations exhibit only small differences, which is consistent with BBL's findings. Adding data on inequality to the estimation does not significantly affect the results for the parameters that govern the aggregate dynamics of the model. The investment adjustment cost is estimated to be higher in the 7 var SMC estimation, but otherwise the posterior distributions are close to each other.

Posteriors from the MH and SMC estimations are also broadly similar for many parameters, but exhibit differences for some parameters, which we discuss in the remainder of this section. Starting with the parameters of the monetary policy rule, posteriors from the SMC estimations imply a slightly higher interest-rate inertia and lower sensitivities of the interest rate with respect to the inflation rate and output growth relative to those from the SMC estimation. The interest-rate smoothing parameter is 0.82 and 0.84 at the mean in the SMC estimations, while the posterior mean is 0.79 in the MH estimation. The Taylor rule coefficient on inflation is relatively low in the SMC estimations, with the 10 to 90 percentile range being from 1.53 to 1.80 in the 11 var estimation and 1.36 to 1.77 in the 7 var

estimation. In contrast, the corresponding range is from 2.04 to 2.42 in the MH estimation. The coefficient on output growth is around 0.2 at the mean in the SMC estimations, while it is a bit higher at 0.29 in the MH estimation.

In the case of fiscal policy parameters, differences between MH and SMC estimations are more pronounced. Regarding the bond issuance rule, the SMC estimations imply much less persistence of the structural deficit with an autoregressive coefficient of around 0.81 at the posterior mean, while the mean is 0.97 in the MH estimation. Also, posteriors from the SMC estimations imply a much stronger countercyclical response of government debt to the inflation rate and output growth. The elasticities of the bond issuance with respect to inflation and output growth are -2.74 and -0.74 at the posterior mean in the 11 var SMC estimation and -3.97 and -1.22 in the 7 var SMC estimation, as opposed to -1.60 and -0.35 in the MH estimation. Posteriors for parameters governing tax rules show even larger differences. While posteriors from the SMC estimations imply countercyclical tax rate responses with respect to the growth rate of government debt, the posterior from the MH estimation implies procyclical responses. Also, tax rates are estimated to be more persistent in the SMC estimations than in the MH estimation.

Table 2. Prior and Posterior Distributions: Policies and Frictions

Par	Dist	Prior		Posterior		
		Mean	*Std. Dev*	*BBL (MH)*	*BBL (SMC)*	*BBL (7 Var)*
Monetary policy						
ρ_R	Beta	0.50	0.20	0.785 (0.754,0.814)	0.818 (0.793,0.842)	0.841 (0.819,0.864)
σ_R	Inv-Gamma	0.10	2.00	0.243 (0.224,0.269)	0.206 (0.191,0.220))	0.210 (0.195,0.226)
θ_π	Normal	1.70	0.30	2.237 (2.044,2.424)	1.670 (1.532,1.804)	1.570 (1.357,1.773)
θ_Y	Normal	0.13	0.05	0.287 (0.223,0.361)	0.212 (0.158,0.261)	0.187 (0.117,0.258)
Fiscal policy: deficit						
ρ_D	Beta	0.50	0.20	0.965 (0.950,0.980)	0.790 (0.760,0.816)	0.775 (0.738,0.811)
σ_D	Inv-Gamma	0.10	2.00	0.310 (0.277,0.342)	0.632 (0.548,0.715)	0.994 (0.842,1.159)
γ_B	Gamma	0.10	0.08	0.031 (0.008,0.047)	0.039 (0.021,0.056)	0.025 (0.005,0.043)
γ_π	Normal	0.00	1.00	-1.601 (-1.778,-1.452)	-2.739 (-3.091,-2.379)	-3.969 (-4.529,-3.397)
γ_Y	Normal	0.00	1.00	-0.350 (-0.418,-0.309)	-0.736 (-0.826,-0.631)	-1.222 (-1.405,-1.029)
Fiscal policy: taxes						
ρ_τ	Beta	0.50	0.20	0.653 (0.440,0.961)	0.809 (0.731,0.885)	0.702 (0.508,0.886)
γ_B^τ	Normal	0.00	1.00	0.166 (0.110,0.217)	-1.765 (-2.079,-1.387)	-1.222 (-1.483,-0.937)
γ_Y^τ	Normal	0.00	1.00	-0.148 (-0.410,0.038)	-0.329 (-1.736,1.228)	1.152 (-0.203,2.507)
Income risk						
ρ_S	Beta	0.50	0.20	0.663 (0.606,0.727)	0.917 (0.627,0.995)	-
σ_S	Gamma	65.00	30.00	64.08 (55.91,71.06)	57.67 (51.87,64.59)	-
Frictions						
δ_S	Gamma	5.00	2.00	0.456 (0.278,0.631)	1.800 (1.419,2.227)	2.345 (1.885,2.796)
ϕ	Gamma	4.00	2.00	0.787 (0.373,1.244)	3.532 (2.899,4.164)	6.928 (5.799,8.066)
K_Y	Gamma	0.10	0.03	0.111 (0.094,0.125)	0.116 (0.103,0.128)	0.114 (0.099,0.128)
K_W	Gamma	0.10	0.03	0.112 (0.095,0.128)	0.126 (0.111,0.142)	0.113 (0.096,0.128)

Source: Authors' calculations.
Notes: The table shows prior and posterior distributions of the estimated parameters. 10 and 90 percentile of the distributions are in parenthesis.

Table 3. Prior and Posterior Distributions: Structural Shocks

Par	*Dist*	*Prior*		*Posterior*		
		Mean	*Std. Dev*	*BBL (MH)*	*BBL (SMC)*	*BBL (7 Var)*
Structural shocks						
ρ_A	Beta	0.50	0.20	0.954 (0.925,0.976)	0.976 (0.970,0.982)	0.984 (0.976,0.993)
σ_A	Inv-Gamma	0.10	2.00	0.162 (0.133,0.194)	0.078 (0.059,0.094)	0.062 (0.041,0.081)
ρ_Z	Beta	0.50	0.20	0.998 (0.996,0.999)	0.967 (0.958,0.978)	0.973 (0.963,0.983)
σ_Z	Inv-Gamma	0.10	2.00	0.569 (0.526,0.624)	0.616 (0.573,0.662)	0.612 (0.569,0.659)
ρ_Ψ	Beta	0.50	0.20	0.848 (0.790,0.904)	0.486 (0.422,0.568)	0.416 (0.338,0.495)
σ_Ψ	Inv-Gamma	0.10	2.00	3.814 (2.820,4.982)	12.25 (9.905,14.77)	24.29 (20.58,28.13)
ρ_{μ_p}	Beta	0.50	0.20	0.862 (0.824,0.907)	0.968 (0.955,0.983)	0.967 (0.953,0.980)
σ_{μ_p}	Inv-Gamma	0.10	2.00	1.563 (1.404,1.714)	1.410 (1.296,1.520)	1.452 (1.316,1.597)
ρ_{μ_w}	Beta	0.50	0.20	0.862 (0.826,0.907)	0.898 (0.867,0.931)	0.878 (0.845,0.911)
σ_{μ_w}	Inv-Gamma	0.10	2.00	6.142 (5.385,6.916)	4.732 (4.043,5.292)	5.079 (4.399,5.753)
ρ_P	Beta	0.50	0.20	0.961 (0.943,0.981)	0.975 (0.958,0.990)	-
σ_P	Inv-Gamma	0.10	2.00	3.534 (2.938,4.192)	3.647 (3.011,4.202)	-

Source: Authors' calculations.
Notes: The table shows prior and posterior distributions of the estimated parameters. 10 and 90 percentile of the distributions are in parenthesis. Standard deviations are multiplied by 100 for readability.

Table 4. Prior and Posterior Distributions: Policies and Frictions (2000.Q1 estimation)

Par	Dist	Prior		Posterior		
		Mean	*Std. Dev*	*BBL (MH)*	*Backward from MH*	*Backward from SMC*
Monetary policy						
ρ_R	Beta	0.50	0.20	0.785 (0.754,0.814)	0.763 (0.729,0.797)	0.796 (0.762,0.831)
σ_R	Inv-Gamma	0.10	2.00	0.243 (0.224,0.269)	0.272 (0.244,0.302)	0.243 (0.220,0.264)
θ_π	Normal	1.70	0.30	2.237 (2.044,2.424)	1.941 (1.692,2.169)	1.566 (1.365,1.746)
θ_Y	Normal	0.13	0.05	0.287 (0.223,0.361)	0.198 (0.129,0.270)	0.178 (0.105,0.245)
Fiscal policy: deficit						
ρ_D	Beta	0.50	0.20	0.965 (0.950,0.980)	0.954 (0.918,0.992)	0.824 (0.783,0.860)
σ_D	Inv-Gamma	0.10	2.00	0.310 (0.277,0.342)	0.424 (0.332,0.505)	0.681 (0.546,0.822)
γ_B	Gamma	0.10	0.08	0.031 (0.008,0.047)	0.035 (0.008,0.061)	0.028 (0.006,0.048)
γ_π	Normal	0.00	1.00	-1.601 (-1.778,-1.452)	-2.0330 (-2.358,-1.707)	-2.898 (-3.402,-2.417)
γ_Y	Normal	0.00	1.00	-0.350 (-0.418,-0.309)	-0.435 (-0.562,-0.297)	-0.756 (-0.891,-0.615)
Fiscal policy: taxes						
ρ_τ	Beta	0.50	0.20	0.653 (0.440,0.961)	0.544 (0.346,0.743)	0.718 (0.584,0.853)
γ_B^τ	Normal	0.00	1.00	0.166 (0.110,0.217)	0.0774 (-0.083,0.212)	-1.386 (-1.738,-1.018)
γ_Y^τ	Normal	0.00	1.00	-0.148 (-0.410,0.038)	-2.170 (-3.555,-0.913)	-0.050 (-1.482,1.336)
Income risk						
ρ_S	Beta	0.50	0.20	0.663 (0.606,0.727)	0.633 (0.520,0.746)	0.615 (0.498,0.729)
σ_S	Gamma	65.00	30.00	64.08 (55.91,71.06)	52.93 (43.64,61.43)	49.69 (42.10,56.87)
Frictions						
δ_S	Gamma	5.00	2.00	0.456 (0.278,0.631)	0.474 (0.268,0.655)	1.916 (1.335,2.474)
ϕ	Gamma	4.00	2.00	0.787 (0.373,1.244)	2.554 (1.774,3.316)	3.820 (2.955,4.647)
K_Y	Gamma	0.10	0.03	0.111 (0.094,0.125)	0.121 (0.104,0.138)	0.128 (0.110,0.144)
K_W	Gamma	0.10	0.03	0.112 (0.095,0.128)	0.097 (0.083,0.113)	0.097 (0.083,0.113)

Source: Authors' calculations.

Notes: The table shows prior and posterior distributions of the estimated parameters. 10 and 90 percentile of the distributions are in parenthesis.

Table 5. Prior and Posterior Distributions: Structural Shocks (2000.Q1 estimation)

Par	*Dist*	*Prior*		*Posterior*		
		Mean	*Std. Dev*	*BBL (MH)*	*Backward from MH*	*Backward from SMC*
Structural shocks						
ρ_A	Beta	0.50	0.20	0.954 (0.925,0.976)	0.932 (0.886,0.984)	0.976 (0.964,0.987)
σ_A	Inv-Gamma	0.10	2.00	0.162 (0.133,0.194)	0.215 (0.146,0.279)	0.085 (0.056,0.117)
ρ_Z	Beta	0.50	0.20	0.998 (0.996,0.999)	0.993 (0.990,0.996)	0.962 (0.947,0.977)
σ_Z	Inv-Gamma	0.10	2.00	0.569 (0.526,0.624)	0.652 (0.580,0.715)	0.645 (0.589,0.701)
ρ_Ψ	Beta	0.50	0.20	0.848 (0.790,0.904)	0.631 (0.543,0.734)	0.438 (0.348,0.526)
σ_Ψ	Inv-Gamma	0.10	2.00	3.814 (2.820,4.982)	9.771 (6.856,12.49)	15.64 (12.08,18.91)
ρ_{μ_p}	Beta	0.50	0.20	0.862 (0.824,0.907)	0.862 (0.815,0.913)	0.849 (0.773,0.937)
σ_{μ_p}	Inv-Gamma	0.10	2.00	1.563 (1.404,1.714)	1.301 (1.121,1.480)	1.277 (1.077,1.461)
ρ_{μ_w}	Beta	0.50	0.20	0.862 (0.826,0.907)	0.895 (0.863,0.930)	0.917 (0.886,0.950)
σ_{μ_w}	Inv-Gamma	0.10	2.00	6.142 (5.385,6.916)	4.404 (3.840,5.033)	4.173 (3.628,4.717)
ρ_P	Beta	0.50	0.20	0.961 (0.943,0.981)	0.949 (0.928,0.971)	0.961 (0.940,0.982)
σ_P	Inv-Gamma	0.10	2.00	3.534 (2.938,4.192)	3.835 (3.093,4.545)	3.686 (3.015,4.341)

Source: Authors' calculations.
Notes: The table shows prior and posterior distributions of the estimated parameters. 10 and 90 percentile of the distributions are in parenthesis. Standard deviations are multiplied by 100 for readability.

Among the parameters governing the model's frictions, the posterior distributions of variable capital depreciation and investment adjustment cost parameters show significant differences. In the MH estimation, the capital depreciation parameter is 0.46 at the posterior mean. In contrast, the means for the same parameter are 1.80 and 2.35 in the 11 and 7 var SMC estimations, respectively. The posterior mean for the capital adjustment cost parameter is 3.53 and 6.93 in the two SMC estimations and only 0.79 in the MH estimation.

The posterior distributions for the rest of the parameters, including the parameters describing income risk, the slope of the price and wage Phillips curves, and the structural shocks, are broadly similar. The autocorrelation of the MEI shock is the only exception. In the SMC estimations, MEI shocks are estimated to be not very persistent with an autocorrelation of around 0.5 at the posterior mean, while in the MH estimation, the posterior mean for this parameter is 0.85.

The differences between the MH and the SMC estimation results obviously lead to the question as to which method is the most accurate. This is a nontrivial question to address since doing so would involve repeated (independent) estimations of the HANK model as done for instance in Cai and others (2021). This is computationally very costly. We therefore sidestep this issue entirely and use both the SMC and MH estimations in our forecasting comparison exercise. By this we mean that we obtain two different sets of backward bridge estimations using the approach described in section 2.3—one starting from the SMC and one starting from the MH draws. It turns out that the accuracy of the BBL model estimated using SMC is better than that using the MH draws. While this evidence is no proof that the SMC estimation is more reliable, it seems to point in that direction.

Finally, in tables 4 and 5, we present the posterior distributions from the estimations using the data up to 2000.Q1, which we obtain using the approach described in section 3. Columns 6 and 7 show the posterior distributions from the backward estimation starting from the 2019.Q4 MH and SMC draws, respectively. For comparison, we also show the posteriors from the original 2019.Q4 BBL's estimation in column 4. The 2000.Q1 posterior is close to the MH BBL posterior when the original MH draws were used as a starting point. Similarly, when using the full sample SMC estimation result as a starting point, the 2000.Q1 results are close to the estimates obtained for the 2019.Q4 SMC estimation.

3.3 Assessing HANK's Out-of-sample Forecasting Accuracy

Figure 1 shows the results of the horse race between BBL and SW focusing on four variables of interest: output, consumption, investment growth, and the GDP deflator inflation. For each of these variables, the figure displays the root mean square errors (RMSEs), expressed in percent, computed as

$$RMSE_{i,h,\mathcal{M}_m} = \frac{1}{\bar{\tau} - h + 1} \sum_{\tau=h}^{\bar{\tau}} \left(y_{i,T-\tau+h} - \mathbb{E}\left[y_{i,T-\tau+h} \mid y_{1:T-\tau}, \mathcal{M}_m \right] \right)^2, \tag{53}$$

where i indicates the variable being forecast, h is the forecast horizon, which ranges from 1 to 7 quarters ahead, and M_m is the model. The model set is M={BBL,SW}, with the BBL RMSEs shown by the solid grey line and the SW RMSEs by the solid black line. The BBL model is referred to as BBL (SMC), because it uses the posterior computed from the online estimation starting from the SMC draws, as opposed to the original MH draws from BBL.[36]

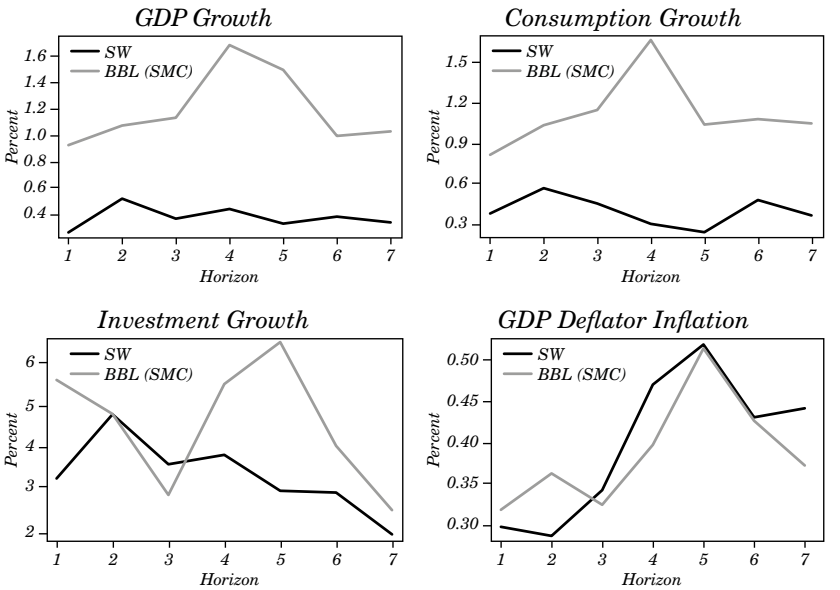
For the variables measuring real activity, in particular output and consumption growth, the results of the horse race are not very kind to the BBL model. This is especially true for consumption growth, where the RMSEs are roughly between about two (for both short and longer horizons) and six (h=4) times larger for BBL. The differences in forecasting performance for consumption growth largely translate into similar differences in RMSEs for GDP growth, given that consumption represents the largest component of GDP. For investment growth, the forecasting accuracy of the two models is similar for shorter horizons but is again worse for BBL for medium horizons. One piece of good news for the BBL model is that, for the GDP deflator inflation, its RMSEs are comparable to those of SW for all forecast horizons.

The much worse forecasting performance for BBL compared to SW for consumption growth is particularly disappointing. The key difference between HANK and SW-type models is the following: in HANK models the representative-agent Euler equation, which determines consumption in standard DSGEs, is replaced with the aggregation of individual households' consumption policy functions.

36. The SMC-based estimation performs better than the MH-based one, as shown later.

These consumption policy functions reflect inequality in both income and wealth: poor agents are hand-to-mouth, or close to, and have a high marginal propensity to consume out of income, while richer agents can substitute intertemporally and have low marginal propensities to consume. The BBL version we use to compute the RMSEs in figure 1 includes, among the observables used in the estimation (and forecasting), those reflecting inequality such as the top 10 percent shares in income and wealth. One would have hoped that this much more realistic view of the world translated into a better quantitative understanding of the behavior of aggregate consumption and hence a better forecasting performance. This does not seem to be the case, at least for the BBL model.

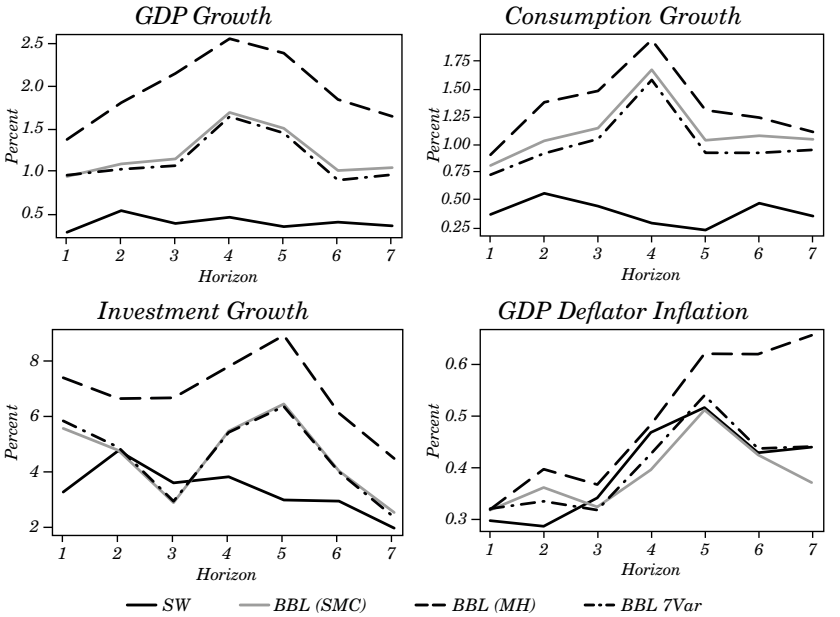
Figure 1. RMSEs: BBL vs SW



Source: Authors' calculations.
Note: The figure plots computed using expression (53) for the BBL (solid grey lines) and the SW (solid black lines) models. The BBL model is referred to as BBL(SMC) as it uses the posterior computed from the online estimation starting from the SMC draws.

Before discussing possible reasons for these findings, we show in figure 2 that the results are robust to using the results from i) the online estimation starting from the Metropolis-Hastings (MH) draws, which we refer to as BBL (MH) (dashed black lines), and ii) the model estimated using only the seven aggregate macro variables, and no measure of inequality, as observables (BBL 7Var, dash-and-dotted black lines). We find that the RMSEs obtained using the MH draws are uniformly worse than those obtained from the SMC draws. We also find that the RMSEs for the eleven- and the seven-variable BBL are almost indistinguishable from one another. This is somewhat disappointing from the perspective of the HANK literature, as it suggests that measures of inequality matter little for the dynamics of macroeconomic aggregates, at least for this model. The result is reminiscent of the findings in Chang and others (2021), who use functional vector autoregressions to argue that there is limited feedback between inequality and aggregate macro time series.

Figure 2. RMSEs: Robustness



Source: Authors' calculations.
Note: The figure plots computed using expression (53) for the seven-variable BBL model (BBL 7Var, dash-and-dotted black lines), the BBL model using the posterior computed from the online estimation starting from the MH draws and referred to as BBL(MH) (dashed black lines), the BBL model using the posterior computed from the online estimation starting from the SMC draws, and referred to as BBL(SMC) (solid grey lines), and the SW model (solid black lines).

What are the possible reasons for these somewhat negative results? First, while BBL is a priori an ideal candidate for this forecasting comparison given that it incorporates SW's shocks and frictions, perhaps other HANK models may perform better than BBL from a forecasting point of view. Seen from this perspective, the results in this paper are an invitation to HANK modelers to use the methodology (and the code) described in this paper to see how well their model fares in terms of forecasting accuracy.

Second, the good—at least relative to VARs—forecasting performance of representative-agent DSGEs à la SW was not achieved overnight but resulted from a decade of advancement in modeling, crystallized in Christiano and others (2005).[37] It may be that HANK models need to go through a similar process. There is also evidence[38] that some of the reasonable forecasting performance of representative-agent DSGEs is due to features like habit persistence that, according to some, i) may not have particularly strong micro-foundations, and ii) may be difficult to replicate in HANK models.

Finally, as mentioned in the introduction and discussed in section 3.2, the parameters in HANK affecting the model's steady state are calibrated, not estimated. This is for a computational reason: recomputing the steady state is extremely costly. However, the estimated DSGE literature has shown that not estimating parameters, perhaps not too surprisingly, hurts the fit of DSGE models and their forecasting performance. If this is the reason why BBL forecast worse than SW, these findings pose a computational challenge to HANK researchers interested in estimation: finding ways of computing the steady state more efficiently and/or using estimation algorithms that do not require recomputing the steady state too many times.

37. There is a perception among macroeconomists that the reasonable forecasting performance of DSGEs is the result of hindsight: model features are chosen ex post so that these models produce reasonably good RMSEs. More than ten years of actual (ex ante) forecasting with DSGE models at the NY Fed arguably shows that this perception is unfounded—Cai and others (2019).

38. For example, Del Negro and others (2007).

4. Conclusion

This paper had two objectives. One was to provide a toolkit for efficient repeated estimation of HANK models that can be used by researchers at central banks and in academia. We argued that online estimation using Sequential Monte Carlo provides such a toolkit and explained how it works. The second objective was to "kick the tires" of HANK models by comparing the out-of-sample forecasting accuracy of a prominent example of such models, Bayer and others (2022), to that of the Smets and Wouters (2007) model. HANK models did not fare too well: their forecasting performance for real activity variables, especially GDP and consumption growth, is notably inferior to that of SW. The results for consumption are particularly disappointing, given that the main difference between SW-type DSGEs and HANK models is the replacement of the representative-agent Euler equation with the aggregation of individual households' consumption policy functions, which reflects inequality.

These findings should be interpreted as a motivation to do more research on HANK models. First, no matter the forecasting performance of HANK models, inequality is one of the critical issues of our times and features prominently in the transmission of policies. There are questions, such as investigating the effect on growth and inflation of the government transfers during the COVID pandemic, that representative-agent models simply cannot adequately answer. Kaplan and others (2020) and Auclert and others (2023) are recent examples of quantitative research based on HANK models that focus on some of these salient policy issues. Second, since all models are misspecified, model diversity should play an important role for policymakers who use models to inform their decisions. Finally, the fact that the forecasting performance of HANK models can be improved is a just stimulus for further efforts, in terms of both modeling and making computations more efficient.

References

Ahn, S., G. Kaplan, B. Moll, T. Winberry, and C. Wolf. 2018. "When Inequality Matters for Macro and Macro Matters for Inequality." *NBER macroeconomics annual* 32(1): 1-75.

An, S. and F. Schorfheide. 2007. "Bayesian Analysis of DSGE Models." *Econometric Reviews* 26:(2–4); 113–172.

Auclert, A., B. Bardóczy, M. Rognlie, and L. Straub. 2021. "Using the Sequencespace Jacobian to Solve and Estimate Heterogeneous-agent Models." *Econometrica* 89(5): 2375–408.

Auclert, A., M. Rognlie, and L. Straub. 2023. "The Trickling Up of Excess Savings." Technical Report, National Bureau of Economic Research.

Bayer, C. and R. Luetticke. 2020. "Solving Heterogeneous Agent Models in Discrete Time with Many Idiosyncratic States by Perturbation Methods." *Quantitative Economics* 11: 1253–88.

Bayer, C., B. Born, and R. Luetticke. 2023. "Shocks, Frictions, and Inequality in U.S. Business Cycles."

Bayer, C., R. Luetticke, L. Pham-Dao, and V. Tjaden, 2019. "Precautionary Savings, Illiquid Assets, and the Aggregate Consequences of Shocks to Household Income Risk." *Econometrica* 87(1): 255–90.

Cai, M., M. Del Negro, E. Herbst, E. Matlin, R. Sarfati, and F. Schorfheide. 2021. "Online Estimation of DSGE Models." *The Econometrics Journal* 24 (1): C33–C58.

Cai, M., M. Del Negro, M.P. Giannoni, A. Gupta, P. Li, and E. Moszkowski. 2019. "DSGE Forecasts of the Lost Recovery." *International Journal of Forecasting* 35(4): 1770–89.

Calvo, G. 1983. "Staggered Prices in a Utility Maximizing Framework." *Journal of Monetary Economics* 12(3): 383–98.

Cappé, O., E. Moulines, and T. Ryden. 2005. *Inference in Hidden Markov Models*. New York, NY: Springer Verlag.

Chang, M., X. Chen, and F. Schorfheide. 2021. "Heterogeneity and Aggregate Fluctuations." Technical Report, National Bureau of Economic Research.

Chopin, N. 2002. "A Sequential Particle Filter for Static Models." *Biometrika* 89(3): 539–51.

Chopin, N. 2004. "Central Limit Theorem for Sequential Monte Carlo Methods and its Application to Bayesian Inference." Annals of Statistics 32(6): 2385–411.

Christiano, L.J., M. Eichenbaum, and C.L. Evans. 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy* 113: 1–45.

Creal, D. 2007. "Sequential Monte-Carlo Samplers for Bayesian DSGE Models." Manuscript, University Chicago Booth.

Del Negro, M. and F. Schorfheide. 2009. "Monetary Policy with Potentially Misspecified Models." *American Economic Review* 99(4): 1415–50.

Del Negro, M. and F. Schorfheide. 2010. "Bayesian Macroeconometrics." In *Handbook of Bayesian Econometrics*, edited by H.K. van Dijk, G. Koop, and J. Geweke. Oxford University Press, 2010.

Del Negro, M. and F. Schorfheide. 2013. "DSGE Model-Based Forecasting." In *Handbook of Economic Forecasting*, vol. 2, edited by G. Elliott and A. Timmermann. Elsevier.

Del Negro, M., F. Schorfheide, F. Smets, and R. Wouters. 2007. "On the Fit of New Keynesian Models." *Journal of Business and Economic Statistics* 25(2): 123–62.

Duane, S., A.D. Kennedy, B.J. Pendleton, and D. Roweth. 1987. "Hybrid Monte Carlo." *Physics letters B* 195(2): 216–22.

Durham, G. and J. Geweke. 2014. "Adaptive Sequential Posterior Simulators for Massively Parallel Computing Environments." In *Advances in Econometrics*, vol. 34, edited by I. Jeliazkov and D. Poirier. West Yorkshire, U.K.: Emerald Group Publishing Limited.

Edge, R. and R. Gürkaynak. 2010. "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?" Brookings Papers of Economic Activity forthcoming.

Farkas, M. and B. Tatar. 2020. "Bayesian Estimation of DSGE Models with Hamiltonian Monte Carlo." Technical Report, IMFS Working Paper Series.

Fernández-Villaverde, J. and J.F. Rubio-Ramírez. 2007. "Estimating Macroeconomic Models: A likelihood approach." *The Review of Economic Studies* 74(4): 1059–87.

Ferriere, A. and G. Navarro. 2018. "The Heterogeneous Effects of Government Spending: It's All About Taxes." FRB International Finance Discussion Paper 1237.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin, Bayesian data analysis, Chapman and Hall/CRC, 1995.

Geweke, J. 2005. *Contemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: John Wiley & Sons.

Greenwood, J., Z. Hercowitz, and G.W. Huffman. 1988. "Investment, Capacity Utilization, and the Real Business Cycle." *American Economic Review* 78(3): 402–17.

Hagedorn, M., I. Manovskii, and K. Mitman. 2018. "Monetary Policy in Incomplete Market Models: Theory and Evidence." Technical Report, University of Pennsylvania.

Hammersley, J.M. and K.W. Morton. 1954. "Poor Man's Monte Carlo." *Journal of the Royal Statistical Society: Series B (Methodological)* 16 (1): 23–38.

Herbst, E. 2015. "Using the 'Chandrasekhar Recursions' for Likelihood Evaluation of DSGE Models." *Computational Economics* 45(4): 693–705.

Herbst, E. and F. Schorfheide. 2014. "Sequential Monte Carlo Sampling for DSGE Models." *Journal of Applied Econometrics* 29(7): 1073–98.

Herbst, E. and F. Schorfheide. 2015. "Bayesian Estimation of DSGE Models," Princeton University Press, 2015.

Justiniano, A., G.E. Primiceri, and A.ndrea Tambalotti. 2011. "Investment Shocks and the Relative Price of Investment." *Review of Economic Dynamics* 14(1): 102–21.

Kaplan, G., B. Moll, and G.L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

Kaplan, G., B. Moll, and G.L. Violante. 2020. "The Great Lockdown and the Big Stimulus: Tracing the Pandemic Possibility Frontier for the U.S." Technical Report, National Bureau of Economic Research.

Klein, P. 2000. "Using the Generalized Schur Form to Solve a Multivariate Linear Rational Expectations Model." *Journal of Economic Dynamics and Control* 24(10): 1405–23.

Lee, D. 2021. "Quantitative Easing and Inequality." Technical Report, Manuscript.

Liu, J.S. 2001. *Monte Carlo Strategies in Scientific Computing*. New York, NY: Springer Verlag.

Mlikota, M. and F. Schorfheide. 2022. "Sequential Monte Carlo with Model Tempering." arXiv preprint arXiv:2202.07070.

Müller, U.K. 2012. "Measuring Prior Sensitivity and Prior Informativeness in Large Bayesian Models." *Journal of Monetary Economics* 59(6): 581–97.

Neal, R.M. 2011. "MCMC Using Hamiltonian Dynamics." Handbook of Markov Chain Monte Carlo 2(11): 2.

Reiter, Michael. 2009. "Solving Heterogeneous-Agent Models by Projection and Perturbation." *Journal of Economic Dynamics and Control* 33(3): 649–65.

Smets, F. and R. Wouters. 2007. "Shocks and Frictions in U.S. Business Cycles: A Bayesian DSGE Approach." *American Economic Review* 97(3): 586–606.

Stan Development Team. 2015. "Stan: A C++ Library for Probability and Sampling, version 2.8. 0."

Winberry, T. 2018. "A Method for Solving and Estimating Heterogeneous Agent Macro Models." *Quantitative Economics* 9(3): 1123–51.

Wu, J.C. and F. Dora Xia. 2016. "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound." *Journal of Money, Credit and Banking* 48(2-3): 253–91.

From Micro to Macro Hysteresis: Long-Run Effects of Monetary Policy

Felipe Alves
Bank of Canada

Giovanni L. Violante
Princeton University
Centre for Economic Policy Research
Institute for Fiscal Studies
National Bureau of Economic Research

The traditional view of macroeconomic dynamics is that aggregate time series can be decomposed into a long-run component (the trend, or the deterministic steady state) and an orthogonal short-run component (the business cycle) which fluctuates around the trend. Quantitative dynamic stochastic general equilibrium (DSGE) models used for research and policy analysis fit into this description and, consistently with this view, routinely assume that transitory shocks have no long-term effects on aggregates.[1] An alternative view of business cycles is that of macroeconomic hysteresis. According to this interpretation, the economy's long-run dynamics are not driven by an exogenous trend, but are instead a function of the entire history of shocks hitting the economy. Under this hypothesis, transitory shocks have permanent effects on the level of economic activity.[2,3]

The hysteresis view of aggregate fluctuations can be traced back to Okun (1973)—and later Tobin (1980)— who argued that recessions

We thank Jordi Galì for useful comments.

1. In models where exogenous total factor productivity (TFP) growth drives the trend and TFP is subject to permanent shocks, an innovation to productivity has an impact on the growth rate, but other shocks do not.

2. See Cerra and others, 2023, for a recent survey of macro hysteresis.

3. This idea is also linked to the notion of macroeconomic resilience to shocks (the opposite of hysteresis) articulated in Brunnermeier (2021).

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

could, through erosion of human capital among the labor force, leave potentially permanent scars on the economy. In a similar fashion, Blanchard and Summers (1986) used the hysteresis view to describe the experience of European labor markets during the 1970s and 1980s, when the unemployment rate seemed to have permanently settled at a higher level after a series of negative cyclical shocks.[4] After losing center stage for about two decades, the conjecture that transitory shocks can lead to permanent, or at least very persistent, effects on aggregates reemerged with the Great Recession, after which the U.S. and euro area economies suffered a slow and protracted recovery.[5]

Whereas this earlier work lacked well-identified empirical evidence on hysteresis and the channels through which they arise, there is now a sizable and growing body of work—at both the micro and macro level—backing up the idea that short-run fluctuations can lead to nearly permanent effects on aggregates. At the micro level, we have accumulated considerable evidence of negative hysteresis of recessions on individual labor-market outcomes. This work includes findings of long-lasting impacts on earnings and participation stemming from aggregate labor-market conditions upon college graduation,[6] local business-cycle fluctuations,[7] and job displacements.[8] There is also suggestive evidence that these scarring effects are felt disproportionately among disadvantaged groups of workers. For instance, Yagan (2019) reports highly uneven impacts of the Great Recession on employment and earnings, with low-wage workers suffering the most. Cajner and others (2021) find that labor-force participation of young (ages 16 to 24) and black workers exhibit a much larger and more persistent response to local business-cycle fluctuation compared to prime-age and white workers.

At the macro level, much of the available evidence of negative hysteresis focuses on monetary policy shocks, as they constitute a well-identified example of transitory demand shocks. Blanchard and others (2015) analyze more than 20 international recessions driven by large contractionary monetary policy shocks. They find that two thirds

4. Ljungqvist and Sargent (1998) developed a model to microfound such unemployment hysteresis where the interaction between a generous welfare state and skill decay during nonemployment plays a central role.

5. Reifschneider and others (2015) highlight that the estimates of long-run growth had been systematically revised downward following the Great Recession.

6. See Rothstein (2023).

7. See Yagan (2019) and Cajner and others (2021).

8. See Davis and Von Wachter (2011) and Guvenen and others (2017).

of these episodes are associated with a permanently lower output level and some of them even with permanently lower output growth. Applying local projection instrumental-variable techniques on long panel data for over a century for 17 countries, Jordà and others (2020) uncover very long-lasting effects of monetary policy shocks. Ma and Zimmermann (2023) show that monetary policy, through its impact on innovation activity, affects the productive capacity of the economy in the long term. Furlanetto and others (2021) use local projection methods to identify generic demand shocks as those innovations that lead output and inflation to comove positively in the short run. They conclude that a subset of these shocks has also an impact in the long run and are quantitatively important in the U.S., in particular when the Great Recession is included in the sample.

In this paper, we connect these two pieces of literature by developing a macroeconomic hysteresis model built on the micro evidence that job losses lead to persistently lower individual earnings through a combination of skill decay and abandonment of the labor force. We then use the model to investigate whether the long-term negative effects of recessions on individual job prospects carry over to the overall economy. In other words, we examine whether the labor-market micro-level sources of negative hysteresis we feed into the model give rise to macro hysteresis in response to transitory aggregate shocks. In line with much of the macroeconomic empirical literature discussed above, we focus on the aggregate economy response to a short-lived contractionary monetary policy shock.

Our model merges the standard heterogeneous-agent New Keynesian (HANK) incomplete-market framework with a three-state labor market featuring search frictions and endogenous labor supply at the extensive margin. Labor-market frictions prevent full employment—some employed individuals who want to keep working are forcefully separated, and others searching for a job do not find it. Crucially, separation and job-finding rates depend on individual skill levels—in our calibration, workers at the bottom of the distribution are both less likely to find a job when searching for one and more likely to lose it when employed, as in the data. Besides facing labor-market frictions that give rise to unemployment, workers also make labor-supply decisions by choosing whether or not to participate in the labor market at the prevailing equilibrium wage. Workers' productivity evolves stochastically according to a process that depends on their labor-market status—skills grow during employment through returns to experience but gradually depreciate when the worker is not employed.

Each of these three model ingredients—i.e., labor-market frictions, participation decisions, and skill dynamics—is disciplined by micro evidence. We estimate the dependence of job-finding and separation rates on workers' skill levels from the Basic Monthly Current Population Survey (CPS) merged with the Annual Social and Economic Supplement (ASEC), following the approach of Heathcote and others (2020). To get participation dynamics that resemble the data, we target what Hobijn and Şahin (2021) call the "attachment wedge", i.e., the difference between the unemployment-to-nonparticipation (UN) and the employment-to-nonparticipation (EN) flows in the data. The fact that workers are much more likely to drop out of participation during unemployment than employment spells (UN > EN) creates, mechanically, a downward pressure on the participation rate during downturns when the pool of unemployed workers rises sharply. Finally, we calibrate skill losses during nonemployment to match the large and persistent earnings losses upon displacement documented by Topel (1990), Jacobson and others (1993), and Davis and von Wachter (2011).

Our main experiment studies the long-run effects of a transitory unanticipated contractionary monetary policy shock that reduces total labor income by 1 percent in the first year following the shock. In the short run, the shock causes both an increase in job separation and a decline in job-finding rates. As workers flow into and remain stuck in unemployment, their skills depreciate, thus making job opportunities less likely to arrive and wages upon re-employment less attractive. Ten years after the shock, long after the surprise to the monetary policy rule has died out, participation and labor productivity are still depressed by 0.06 ppt and 00.11 percent, respectively. Together, these two components add up to a 0.20 percent reduction in total labor income. Thus a temporary shortfall of aggregate demand disrupts aggregate supply over the long run in our model.[9] Importantly, the effects at the bottom of the wage distribution are much stronger than their average ones—a decade after the monetary policy shock, the scarring of labor income for workers in the lowest skill quartile is almost ten times the average scarring effect. Hysteresis, we find, operates disproportionately through low-wage workers.

9. Eventually, all labor-market variables return to their steady state, so the shock does not have permanent effects in the very long run in our model. But since the adverse effects of the shock survive far after the shock itself has already died out, we treat these long-lasting negative effects of transitory shocks as evidence of macro hysteresis within the model.

Despite the long shadow cast on output, the shock generates only short-lived movements in inflation, which quickly goes back to the target following the monetary shock. The reason for this is the decline in labor productivity and labor-force participation, which generate inflationary pressures that offset the long-run deflationary pressures coming from the persistent decline in output.

Related literature. Our paper is related to several strands in the literature. First, our emphasis on the joint dynamics of unemployment and labor-force participation relates to the recent literature extending general equilibrium business-cycle models to incorporate frictional labor markets and an endogenous participation margin. Contributions in this literature include Galí and others (2012), Shimer (2013), Krusell and others (2017), Christiano and others (2021), and Cairó and others (2022).[10] None of these papers studies hysteresis at the macro level.

Our paper also relates to the small but growing literature where macro hysteresis originates from the labor market.[11] Chang and others (2002) develop a model where skill accumulation through past work experience (i.e., learning-by-doing) gives rise propagation mechanism through labor productivity that resembles our channel. Galí (2022) incorporates an insider-outsider model of the labor market within a New Keynesian framework and shows that the inefficiently high wage arising in equilibrium can be a source of macro hysteresis. Abbritti and others (2021) develop a similar logic in a model with downward wage rigidity and endogenous growth. Acharya and others (2022) analyze the impact of monetary policy in a search and matching model where skill depreciation of unemployed workers can lead to steady-state multiplicity.

With respect to this body of work, our contribution is twofold. First, we develop our insights within a state-of-the-art framework that combines elements of heterogeneous-agent models with elements of New Keynesian models. This class of HANK models is becoming a workhorse for quantitative fiscal and monetary policy analysis. Relative to the representative-agent framework, the heterogeneity makes the mapping between the model and the relevant cross-sectional

10. Like Krusell and others (2017) and Cairó and others (2022), we focus on matching worker flows and the implied employment, unemployment, and labor-market participation dynamics.

11. A parallel literature explores the impact of business cycles on long-term growth through innovation and technological change, e.g., Comin and Gertler (2006), Bianchi and others (2019), Fornaro and Wolf (2020), Gaillard and Wangner (2023).

evidence much easier to draw.[12] Second, we highlight the role of skill losses upon displacement and participation decisions as complementary sources of macro hysteresis.

Finally, our paper heavily builds upon the empirical literature documenting scarring effects of recessions on individual labor-market trajectories, especially among low-wage workers. Aaronson and others (2019) and Cajner and others (2017) show that low-wage workers are more exposed to aggregate fluctuations. Kahn (2010) and Rothstein (2023) uncover evidence of persistently depressed labor-market outcomes for individuals who enter the labor market in recessions. Davis and Von Wachter (2011) show that long-term earnings losses are worse when job displacement occurs in a recession. Guvenen and others (2017) and Athey and others (2023) find that such earnings losses are most severe at the bottom of the distribution because they lead to detachment from the labor force. Yagan (2019), Rinz (2022), and Hershbein and Stuart (2020) all present evidence of strong persistence in local labor-market outcomes in the aftermath of the Great Recession. They conclude that human capital decay is an important mechanism generating negative hysteresis on labor earnings, with stronger impacts for low-wage workers. Furlanetto and others (2021) document that negative hysteresis propagates almost exclusively through lower employment and labor-force participation and that these effects are especially strong at the bottom of the wage distribution. Finally, Lepetit (2023) provides evidence that, in response to demand shocks, the slope of the inflation-output relationship is much flatter at long horizons than at short ones, consistent with our model's prediction.

The rest of the paper is organized as follows. Section 1 outlines the model. Section 2 describes its parameterization. Section 3 discusses the results of our model's simulations. Section 4 concludes and examines the implications of our findings for the optimal design of monetary policy.

12. In this aspect, our framework shares many of the same ingredients with Krusell and others (2017), who also develop a heterogeneous-agent model with search frictions and endogenous participation. Their focus is on matching the behavior of gross worker flows. Relative to their analysis, our paper makes the following two improvements. First, we extend our analysis to a general equilibrium environment with nominal-wage rigidities, whereas theirs is done in partial equilibrium. Next, we rely on a more extensive set of micro evidence to calibrate the model's labor-market frictions and skill dynamics, which we show to matter for the quantitative aggregate implications of a monetary policy shock.

1. Model

The structure of the model follows closely the framework we have developed in previous work.[13]

1.1 Households

Time is continuous and indexed by t. The economy is populated by a continuum of infinitely lived households (or individuals) with measure one, who discount the future at rate $\rho > 0$.

Individuals can be in one of three mutually exclusive labor-market states s_t: employed and earning labor income, $(s_t = e)$, unemployed and searching for a job $(s_t = u)$, outside the labor force, $(s_t = n)$. Among the unemployed, we distinguish between those who are eligible $(u = u_1)$ and not eligible $(u = u_0)$ for unemployment insurance (UI) benefits. Workers gain eligibility only if they are laid off from work. They then lose it at some constant rate which reflects benefit duration. Among those out of the labor force, we distinguish between 'active' nonparticipants $(n = n_1)$ and 'passive' nonparticipants $(n = n_0)$. The former can, at a lower rate than the unemployed, find jobs and enter employment, while the latter cannot.[14]

Households derive utility from consumption c_t and incur disutility from the effort cost κ_s associated with being in labor-market status s (the extensive margin) and from the effort cost of working h_t hours (the intensive margin). We specify the following functional form for period utility

$$\mathrm{u}^s\left(c_t, h_t\right) = \log c_t - \psi \frac{h_t^{1+\frac{1}{\sigma}}}{1+\frac{1}{\sigma}} - \kappa^s, \tag{1}$$

where $\sigma > 0$ is the Frisch elasticity of labor supply. We assume that $\kappa^e > \kappa^u > \kappa^n \geq 0$.

13. See Alves and Violante (2024).

14. This differentiation captures the heterogeneity in the pool of nonparticipants (Hall and Kudlyak, 2019), where some individuals are able and willing to work, while others are unable to accept any job offer (e.g., because they are sick) or are discouraged from searching further.

Table 1. Transition Matrix Across the Five Employment States

	e	u_1	u_0	n_1	n_0
e	$\ddots$	λ_{zt}^{eu}	$\times$	$\trianglerighteq$	η^{en_0}
u_1	$\lambda_{zt}^{ue} \cdot \triangleright$	$\ddots$	$\eta^{u_1u_0}$	$\trianglerighteq$	η^{un_0}
u_0	$\lambda_{zt}^{ue} \cdot \triangleright$	$\times$	$\ddots$	$\trianglerighteq$	η^{un_0}
n_1	$\lambda_{zt}^{ne} \cdot \triangleright$	$\times$	$\trianglerighteq$	$\ddots$	$\eta^{n_1n_0}$
n_0	$\times$	$\times$	$\times$	$\eta^{n_0n_1}$	$\ddots$

Source: Authors' calculations.
Note: The × symbol means that transition cannot happen. The symbol $\trianglerighteq$ means that an endogenous participation decision moves the individual in that state. The $\triangleright$ symbol means that an endogenous job acceptance decision moves the individual into employment. $\lambda_{zt}^{ss'}$ and $\eta^{ss'}$ are exogenous Poisson rates. The diagonal dots stand for the negative of the sum of all the other entries on that line.

Each individual is endowed with efficiency units of labor (or skills) z evolving according to an Ornstein-Uhlenbeck diffusion process which depends on labor-market status s_t:

$$d\log z_t = \left\{-\rho_z \log z_t + \mathbb{I}_{\{s_t=e\}}\delta_z^+ - \mathbb{I}_{\{s_t\neq e\}}\delta_z^-\right\}dt + \sigma_z d\mathcal{W}_t. \tag{2}$$

When workers are employed ($s_t = e$), skills drift up at rate $\delta_z^+ > 0$ and, when they are not employed ($s_t = u,n$), they drift down at rate $\delta_z^- < 0$. The parameter $\rho_z > 0$ measures the degree of mean reversion in skill dynamics, the standard deviation σ_z determines uncertainty about future realizations, and $\mathcal{W}_t$ is a Wiener process.

Every period individuals can transition across states through a combination of exogenous Poisson rates and optimal mobility decisions. Table 1 describes all the possible transitions and their endogenous/exogenous nature.

At any date , employed and unemployed workers can decide to quit the labor force and enter active nonparticipation (rows 1, 2, 3 of table 1). Similarly, an active nonparticipant can choose to re-enter the labor force as an unemployed ineligible for UI (row 4). Employed workers who decide to remain attached can still be laid off, and thus move from e to u at an exogenous rate λ_{zt}^{eu} which depends on the worker's skill level z (row 1). Unemployed workers who choose to remain in the labor force draw an employment opportunity at an exogenous rate λ_{zt}^{ue} and decide

whether to accept it or not (rows 2 and 3).[15] UI benefits can expire at rate $\eta^{u_1 u_0}$, and an eligible unemployed becomes ineligible (row 2). Also active participants receive job opportunities at rate λ_{zt}^{ne} and decide whether to accept them or not (row 4). All workers can exogenously move into passive nonparticipation at rate η^{s,n_0} (rows 1,2,3,4). At rate $\eta^{n_0 n_1}$, passive nonparticipants become active again (row 5).

Employed individuals earn labor income $w_t h_t z_t$, where w_t is the real wage per effective hour. Eligible unemployed receive benefits $b(z_t)$. We let UI benefits be a function of current worker productivity z_t, as a proxy for actual replacement rates. Both types of income are taxed at a proportional rate t. Every household is entitled to a lump-sum transfer ϕ. Households can save through a financial asset a_t with rate of return r_t, but cannot borrow.

Household problem. The vector (s, a, z) fully summarizes the individual state variables. The dynamic problem solved by the household at time t is a mix of an optimal control problem, the choice of $c_t > 0$, and two optimal stopping problems: a continuous one, the participation decision $\mathfrak{p}_t^s \in \{0,1\}$, and one arising at random Poisson jump times, the job acceptance decision $\mathfrak{f}_t^s \in \{0,1\}$. The stochastic nature of the problem is due to both the Poisson arrival rates that determine transitions across labor-market states and the diffusion that describes the evolution of skills z_t. Conditional on these realizations, wealth evolves deterministically. Let $v_t^s(a,z)$ be the value at date t of an individual with employment state s, wealth a, and productivity z.

Consider the problem of the active nonparticipant (η_1):

$$
\begin{aligned}
v_0^{n_1}(a_0, z_0) = \max_{\{c_t\}_{t\geq 0}, \tau^*} \mathbb{E}_0 \Bigg[& \int_0^{\tau^{\min}} e^{-\rho t} \mathfrak{u}^n(c_t, h_t)\, dt + \mathbb{I}_{\{\tau^{\min} = \tau^e\}} e^{-\rho \tau^e} \\
& \max\left\{ v_{\tau^e}^e\left(a_{\tau^e}, z_{\tau^e}\right), v_{\tau^e}^{n_1}\left(a_{\tau^e}, z_{\tau^e}\right) \right\} \\
& + \mathbb{I}_{\{\tau^{\min} = \tau^*\}} e^{-\rho \tau^*} \left(v_{\tau^*}^{u_0}\left(a_{\tau^*}, z_{\tau^*}\right) - \vartheta \right) + \mathbb{I}_{\{\tau^{\min} = \tau^n\}} e^{-\rho \tau^{n_0}} v_{\tau^{n_0}}^{n_0}\left(a_{\tau^{n_0}}, z_{\tau^{n_0}}\right) \Bigg] \\
& \text{s.t.} \\
& c_t + \dot{a}_t = r_t a_t + \phi \\
& a_t \geq 0
\end{aligned}
\tag{3}
$$

15. The unemployed ineligible for UI always accept job offers because in equilibrium there is a unique wage per effective hours and, if they did not want to work, they would choose nonparticipation where the fixed cost κ^s is lower. Eligible unemployed instead may turn down job opportunities if UI benefits are generous enough.

Active nonparticipants receive job opportunities at rate λ_{zt}^{ne}, with τ^e being the first arrival time of this event. Conditional on receiving this job offer, they choose whether to accept it $\left(\mathfrak{f}_t^{n_1}=1\right)$ or not $\left(\mathfrak{f}_t^{n_1}=0\right)$. At every instant, the nonparticipant also chooses whether to remain unattached $\left(\mathfrak{p}_t^{n_1}=0\right)$ or re-enter the labor force $\left(\mathfrak{p}_t^{n_1}=1\right)$ in which case they become unemployed without UI benefits ($u = u_0$) We assume that re-entering the labor force involves a small fixed switching cost ϑ.[16] The optimal stopping time τ^* represents the first instant in which the choice $\mathfrak{p}_t^{n_1}$ switches from 0 to 1. Finally, at rate $\eta^{n_1 n_0}$ (with τ^{n_0} being the first arrival rate of this shock) active nonparticipants become passive nonparticipants. The conditional expectation reflects the uncertainty in transition rates and skill dynamics. In addition to the participation and job acceptance decisions, at every instant, the worker chooses its consumption flow c_t. The last two lines of this problem state the budget constraint (in real terms) and the borrowing limit.

Problems for passive nonparticipants, ineligible unemployed, eligible unemployed, and employed workers are analogous and are described in detail in appendix A.

1.2 Firms

Final-good producers. A competitive representative final-good producer aggregates a continuum of intermediate inputs indexed by $j \in [0,1]$ with technology

$$Y_t = \left(\int_0^1 y_{jt}^{\frac{v-1}{v}} dj \right)^{\frac{v}{v-1}}, \tag{4}$$

where $v > 0$ is the elasticity of substitution across inputs. This firm takes prices as given and solves

$$\max_{\{y_{jt}\}} P_t Y_t - \int_0^1 p_{jt} y_{jt} dj \tag{5}$$

16. The presence of a small switching cost is mostly a technical assumption to avoid 'chattering', i.e., infinitely fast switching between n_1 and u_0, in the optimal solution of the problem. For all other participation decisions, this problem does not arise because switching back can only occur upon the realization of Poisson shocks.

subject to (4). Cost minimization implies that demand for intermediate good j at price p_{jt} is

$$y_{jt}=\left(\frac{p_{jt}}{P_t}\right)^{-\upsilon} Y_t, \text{ where } P_t=\left(\int_0^1 p_{jt}^{1-\upsilon}dj\right)^{\frac{1}{1-\upsilon}} \tag{6}$$

is the price of the final good and the numeraire of the economy.

Intermediate-good producers. A continuum of measure one of monopolistically competitive firms produces the intermediate goods using labor. Production requires hiring labor on a continuum of tasks indexed by $k \in [0,1]$. Each firm j hires labor services (efficiency-weighted hours) ℓ_{jkt} on every task k, combines them into a final labor input ℓ_{jt} using a Dixit-Stiglitz aggregator with elasticity of substitution ε, and produces the intermediate good according to the linear technology $y_{jt} = \alpha\ell_{jt}$ Every period firms face a fixed operating cost χ expressed in terms of final good. At every date t, these firms take the task-specific wage as given, and maximize profits by solving

$$\max_{p_{jt},\{\ell_{jkt}\}_k}\left(\frac{p_{jt}}{P_t}\right)y_{jt}-\int_0^1 w_{kt}\ell_{jkt}dk-\chi \tag{7}$$

s.t.

$$y_{jt}=\alpha\ell_{jt}$$

$$\ell_{jt}=\left[\int_0^1 \ell_{jkt}^{\frac{\varepsilon-1}{\varepsilon}}dk\right]^{\frac{\varepsilon}{\varepsilon-1}},$$

$$y_{jt}=\left(\frac{p_{jt}}{P_t}\right)^{-\upsilon} Y_t$$

where w_{kt} is the real wage on task k. Cost minimization yields the relative demand of labor for task k

$$\ell_{jkt}=\left(\frac{w_{kt}}{w_t}\right)^{-\varepsilon}\ell_{jt}, \tag{8}$$

where w_t is the Dixit-Stiglitz real aggregate wage index $w_t=\left[\int_0^1 w_{kt}^{1-\varepsilon}dk\right]^{\frac{1}{1-\varepsilon}}$ that satisfies $\int_0^1 w_{kt}\ell_{jkt}dk = w_t\ell_{jt}$. The profit-maximizing price-setting

decision yields the standard expression whereby the relative price equals a markup over the marginal cost of production

$$\frac{p_{jt}}{P_t} = \frac{v}{v-1}\left(\frac{w_t}{\alpha}\right). \tag{9}$$

In a symmetric equilibrium with $p_{jt} = P_t$, all firms produce the same amount $y_{jt} = Y_t$ with labor $\ell_{jt} = \ell_t = \alpha^{-1}Y_t$.

From the assumption of constant returns to scale in production, imposing $p_{jt} = P_t$ in (9) implies that the equilibrium aggregate real wage per effective hour is constant over time. As a consequence, price inflation equals wage inflation and the real wage is constant.

Finally, the real aggregate profits of the production sector are

$$\Pi_t = Y_t - w_t\ell_t - \chi. \tag{10}$$

Every period, profits are paid as dividends to the mutual fund that owns all intermediate producers.

1.3 Wage Setting

This block of the model adapts the wage-setting mechanism of Erceg and others (2000)—i.e., the standard New Keynesian sticky wage model—to a heterogeneous-agent economy. We follow closely the approach of Auclert and others (2018, 2023) with the needed modifications due to our continuous time formulation and the presence of the extensive margin in labor supply.

Every worker i at date t supplies hours on each task k. The nominal wage ω_{kt} per effective hour worked on task k is set by a union that represents all workers on that particular task. By adhering to the union, each employed worker agrees to supply, at that wage, the same number of hours h_{kt} to producers. The problem of each union k is:

$$\max_{\{\omega_{kt}\}_{t\geq 0}} \int_0^\infty e^{-\rho t}\left[\int_{s_{it}=e} \mathfrak{u}^e\left(c_{it}, h_{it}\right)di - \frac{\Theta}{2}\left(\frac{\dot{\omega}_{kt}}{\omega_{kt}} - \pi^*\right)^2\right]dt \tag{11}$$

s.t.

$$h_{it} = \int_0^1 h_{kt}\,dk$$

$$c_{it} + \dot{a}_{it} = r_t a_{it} + (1-t)\frac{1}{P_t} z_{it}\int_0^1 \omega_{kt}h_{kt}\,dk + \phi$$

$$h_{kt}\int_{s_{it}=e} z_{it}di = \ell_{kt}\left(\frac{\omega_{kt}}{\omega_t}\right)^{-\varepsilon}\ell_t.$$

At every date t, the union sets the nominal wage ω_{kt} in order to maximize the welfare of its current members (all individuals employed at date t) subject to Rotemberg-style quadratic costs of adjusting the nominal wage, in utility terms, with scaling parameter Θ. Let inflation be denoted by $\pi_t = \dot{P}_t / P_t$. This cost is expressed in terms of deviations of nominal-wage growth from the central bank's inflation target, the deterministic steady-state trend inflation rate π^*. The first constraint faced by the union states that the total hours worked by an employed worker equal the sum of hours worked on each task. The second constraint is the budget constraint of employed workers. The third one states that contractual effective hours required by the union from its workers must equal the firm's demand for task k effective labor, ℓ_{kt}.[17] Because each task-specific union is 'small' (there is a continuum of tasks) the impact of a union's wage on individual income or firm's employment is negligible. As a result, the union takes as given all individual decisions and the firm's labor demand curves for their task.[18]

In a symmetric equilibrium where all unions charge the same nominal wage $\omega_{kt} = \omega_t$, the amount of labor demanded for all tasks is the same $\ell_{kt} = \ell_t$, and, since unions represent the same set of workers, the number of hours worked on each task is equalized $h_{kt} = h_t$. Combining this with the production function of intermediate-good producers, we arrive at an aggregate production function

$$Y_t = \alpha \left(\int_{s_{it}=e} z_{it} di \right) h_t = \alpha \, Z_t^e H_t, \tag{12}$$

where $Z_t^e = \left(\frac{1}{E_t} \int_{s_{it}=e} z_{it} di \right)$ denotes average labor productivity among the employed, E_t is aggregate employment, and $H_t = h_t E_t$ is aggregate hours worked.

17. Note that the right-hand side of this latter constraint equals (8).

18. Huo and Ríos-Rull (2020) criticize the representative-agent New Keynesian (RANK) model featuring nominal-wage rigidity because, in the equilibrium of that model, workers may end up being forced to supply hours against their will (thus violating the principle of voluntary exchange) and would be better off not working. They suggest a resolution based on a different equilibrium concept. We propose a different solution: in our model, unions offer all workers an employment contract that specifies a non-negotiable pair of wages and hours, but workers can always voluntarily choose not to participate in it and remain nonemployed.

The solution to the unions' wage-setting problem yields the wage Phillips curve

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \frac{\epsilon}{\Theta} H_t \left[\psi h_t^{\frac{1}{\sigma}} - \left(\frac{\epsilon - 1}{\epsilon} \right) (1 - \mathfrak{t}) w_t \left(\frac{1}{E_t} \int_{s_{it}=e} \frac{1}{c_{it}} z_{it} di \right) \right], \tag{13}$$

where π_t is aggregate (wage and price) inflation rate. See Alves and Violante (2024) for a detailed derivation.

The term in the square brackets of equation (13) captures unions' incentives to raise or decrease nominal wages. When the marginal disutility of an extra hour of work exceeds the productivity-weighted marginal utility generated by the (markup-augmented after-tax) income derived from this additional hour of work, unions will push up nominal wages to reduce labor demand and close the gap between these two margins. Another useful interpretation of the term in brackets relates to the notion of the labor wedge, as discussed in Dávila and Schaab (2023). Defining the aggregate labor wedge as

$$H_t \left[\left(\frac{\epsilon - 1}{\epsilon} \right) (1 - \mathfrak{t}) w_t \left(\frac{1}{E_t} \int_{s_{it}=e} \frac{1}{c_{it}} z_{it} di \right) - \psi h_t^{\frac{1}{\sigma}} \right],$$

we conclude that unions increase (decrease) their nominal wages whenever the aggregate labor wedge is negative (positive), that is, whenever the measured gains from asking its members to work an additional hour stands below (above) the marginal disutility of an extra hour of work.

1.4 Mutual Fund

A competitive risk-neutral mutual fund owns all intermediate-good firms and holds all debt issued by the government.[19] Let X_t^m denote the shares of the intermediate-good producers held by the mutual fund, q_t the unit share price, Π_t per-share dividends (or profits), B_t^m the amount of government bonds held by the fund, and r_t^b the real interest rate on government bonds. In appendix B, we show that the equilibrium must satisfy the following no-arbitrage condition

19. The setup in this section follows closely Alves and others (2020).

$$r_t = \frac{\Pi_t + \dot{q}_t}{q_t} = r_t^b, \tag{14}$$

which holds at every t, except when a shock hits the economy.[20] The value of the fund, denoted by A_t, is given by $A_t = q_t X_t^m + B_t^m$.

1.5 Government

Let G_t be the units of the final goods purchased by the government (fiscal authority) at time t, ϕ lump-sum transfers, b UI benefits, t the labor income tax, and $B_t^g > 0$ outstanding real government debt. The government faces the following intertemporal budget constraint:

$$G_t + \phi + (1-\mathrm{t}) \int_{s_{it}=u} b(z_{it})\, di + r_t^b B_t^g = \mathrm{t} w_t h_t \int_{s_{it}=e} z_{it} di + \dot{B}_t^g. \tag{15}$$

Outside of steady state, we assume that the government follows the passive fiscal policy rule:

$$G_t = G^* - \beta_B \left(B_t^g - B^* \right), \ \beta_B > 0, \tag{16}$$

where the superscript * denotes steady-state values. Thus, following an aggregate shock debt adjusts to satisfy the government budget constraint, and government expenditures respond to deviations of debt from its steady-state level to keep debt from growing too quickly.

1.6 Monetary Authority

The monetary authority sets the nominal interest rate t according to a rule that reacts to deviations of inflation from its targets with some inertia

$$\frac{dl_t}{dt} = -\beta_l \left(l_t - l^* - \beta_\pi \left(\pi_t - \pi^* \right) \right). \tag{17}$$

We let l^* denote the steady-state nominal rate and $\pi_t = \dot{P}_t / P_t$ the aggregate inflation rate at date t. The coefficients β_π capture the

20. In this case, the price q_t features a jump.

strength of the policy response to deviations of inflation from target π^*. The coefficient β_l captures the degree of interest rate smoothing. The nominal interest rate and the real interest rate on government bonds r_t^b are linked through the Fisher equation $r_t^b = i_t - \pi_t$.

1.7 Equilibrium

An equilibrium for this economy is defined as time paths for household consumption decisions $\left\{c_t^s(a,z)\right\}_{t\geq 0}$ for $s \in \left\{e, u_0, u_1, n_0, n_1\right\}$, participation and job offer acceptance decisions $\left\{\mathfrak{p}_t^s(a,z), \mathfrak{f}_t^s(a,z)\right\}_{t\geq 0}$ for all s, unions' nominal wage setting $\left\{\omega_{kt}\right\}_{t\geq 0}$ for all labor types k, intermediate producers' hiring decisions $\left\{\ell_{kt}\right\}_{t\geq 0}$ for all k, mutual fund allocations between equity and government bonds $\left\{X_t^m, B_t^m\right\}_{t\geq 0}$, real rates of return on the mutual fund and on government bonds $\left\{r_t, r_t^b\right\}_{t\geq 0}$, firms' share price $\{q_t\}_{t\geq 0}$, fiscal variables (taxes, transfers, UI benefits, expenditures, and debt) $\left\{\mathrm{t}, \phi, b(z), G_t, B_t^g\right\}_{t\geq 0}$, nominal interest rates $\{i_t\}_{t\geq 0}$, aggregate output, consumption, profits, contractual hours worked, and inflation $\left\{Y_t, C_t, \Pi_t, h_t, \pi_t\right\}_{t\geq 0}$, and measures of households $\left\{\mu_t^s(a,z)\right\}_{t\geq 0}$ for all s such that at every t: (i) households optimize; (ii) final-good and intermediate-good producers solve (5) and (7), respectively; (iii) unions solve (11) and inflation satisfies the Phillips curve in (13) (iv) the mutual fund maximizes profits; (v) the government budget constraint (15) holds; (vi) the fiscal and monetary authorities follow their policy rules (16) and (17); (vii) the sequence of distributions satisfies aggregate consistency conditions, and (viii) all good and asset markets clear.

Besides the continuum of intermediate-good and labor-varieties markets, there are five other markets in our economy: the intermediate firms' shares market, the government bond market, the mutual fund shares market, the final-good market, and the labor market. The first three markets clear when, respectively

$$\begin{aligned} &X_t^m = 1 \\ &B_t^m = B_t^g \\ &A_t := \sum_{s\in\{e,u,n\}} \int a_t d\mu_t^s = q_t + B_t^g \end{aligned} \quad ,$$

where, without loss of generality, we normalized the measure of firms' shares to 1. These market clearing conditions, together with the no-arbitrage condition (14) and the definition of firm profits (10), determine firm share prices, real interest rates, and aggregate profits.

The final-good market clears when

$$Y_t = C_t + G_t + \chi.$$

The labor market is frictional with workers in one of the three labor-market states: employment, unemployment, and nonparticipation.

A stationary equilibrium is a particular case of our definition where—absent aggregate shocks—all decisions, prices, aggregate variables, and distributions are time-invariant.

2. Parameterization

Preferences. The discount rate ρ is set to target a ratio of mean wealth to annual earnings of 0.56, corresponding to the amount of liquid wealth immediately available for consumption smoothing among U.S. households.[21] This choice allows the model to match a sizable quarterly aggregate marginal propensity to consume of 0.10 without adding illiquid assets or preference heterogeneity. We set $\gamma = 1$ (log-utility over consumption expenditures) and $\sigma = 1$ (quadratic disutility of hours worked).

Working entails a variable and a fixed cost. The variable disutility parameter ψ is set so that there is no inflationary pressure beyond trend inflation in steady state. The fixed disutility of work k^e is set to match the sensitivity of *en* flows as discussed below. The disutility cost of searching k^u is set to match the observation that jobseekers spend less than 30 minutes per day searching.[22] The flow utility of nonparticipation k^n is normalized to zero.[23]

Productivity dynamics. The mean reversion parameter ρ_z is set to -0.0017, corresponding to an annual autocorrelation of $\exp(-12 \times \rho_z) = 0.98$. The negative drift δ^- is set to match the evidence of earnings losses upon displacement from Davis and Von Wachter (2011). Specifically, we target the estimate that laid-off workers still earn on average 15 percent less than their control group 10 years after separation. As a normalization, we set the positive drift δ^+ so that the average skill level of the employed is 1.

21. See Kaplan and Violante (2022).

22. See Faberman and others (2017).

23. The switching ϑ is set to a very small number to make the optimal stopping problem well behaved.

Table 2. Transition Matrix Across the Five Employment States

Parameter		*Value*	*Target*
Preferences			
Discount rate	ρ	0.0060	Liquid wealth to annual earnings (0.56)
Risk aversion	γ	1.00	External
Labor-supply elasticity	σ	1.00	External
Utility weight on hours	ψ	0.8579	No wage inflationary pressures at SS
Disutility of working	k^e	0.9147	Sensitivity of en flows
Disutility of searching	k^u	0.0379	30 minutes per day searching
Disutility of nonparticipation	k^n	0	Normalization
Productivity process			
Skill mean reversion	ρ_z	-0.0017	External
Skill drift while employed	δ^+	0.0016	Normalization of average skill level to 1
Skill drift while nonemployed	δ^-	-0.0262	10-Year earnings losses from displacement (15%)
Skill diffusion	σ_z	0.0288	P90-P50 hourly wage ratio (3)
Labor-market frictions			
Job-separation rate out of E	$\{\lambda_i^{eu}\}_{i=1}^3$	{0.008,0.051,-2.490}	Average labor-market flows
Job-finding rate out of U	$\{\lambda_i^{ue}\}_{i=1}^3$	{0.375,-0.229,-6.123}	Average labor-market flows
Job-finding rate out of N	$\{\lambda_i^{ne}\}_{i=1}^3$	{0.214,-0.131,-6.123}	Average labor-market flows
Passive nonparticipation rate during E	η^{en_0}	0.007	Average labor-market flows
Passive nonparticipation rate during U/N	η^{un_0}, $\eta^{n_1n_0}$	0.099	Average labor-market flows

Table 2. Transition Matrix Across the Five Employment States (continued)

Parameter		*Value*	*Target*
Labor-market frictions			
Passive nonparticipation exit rate	$\eta^{n_0 n_1}$	0.339	Average labor-market flows
Elasticity of job-finding rates to hours	-	15.00	Sensitivity of *ue* flows to MP shock
Elasticity of job-separation rates to hours	-	5.00	Sensitivity of *ue* flows to MP shock
Taxes and transfers			
UI replacement rate	$\bar{b}$	0.50	External
UI expiration rate	$\eta^{u_1 u_0}$	0.167	Average duration of UI (six months)
Lump-sum transfer	ϕ	0.055	6% of annual average earnings
Labor tax rate	t	0.2	External
Technology and Price / Wage Setting			
Firm productivity	α	1.38	Normalization
Firm fixed cost	χ	0.12	Steady-state real rate of 3%
Price/Wage markups	ν, ε	10	External
Wage adjustment cost	Θ	6,667	Slope of wage Phillips curve (0.015 quarterly)
Fiscal and monetary policy			
Trend inflation	π^*	2%	Fed's inflation target
Taylor rule persistence	β_ι	0.07	Response of u to MP shock
Taylor rule reaction to inflation	β_π	2.25	External
Government expenditures response to debt	βB	0.10	External

Source: Authors' calculations.
Note: The corresponding targeted moments are discussed in the main text. The model period is one month.

Finally, we choose the standard deviation σ_z to match a 90-50 wage ratio of 3, the value for the 2019 CPS.[24,25]

Labor-market frictions. The estimation and calibration of the labor-market frictions are based on Alves and Violante (2024). We leave the detailed discussion to that paper but provide a short summary of our strategy here.

Going back to the transition matrix in table 1, the model features seven rates to calibrate. The separation rate λ^{eu}_{zt}, the job-finding rates for unemployed λ^{ue}_{zt}, and the job-finding rate for active nonparticipants λ^{ne}_{zt} vary with time t and are allowed to depend on the worker's skill level z. In the steady state, we model their dependence on worker's skill level z as

$$\lambda^{ss'}(z) = \lambda_0^{ss'} + \lambda_1^{ss'} \exp\left(\lambda_2^{ss'} z\right) \tag{18}$$

We choose the coefficients in (18) in two steps. In the first step, we use data on transition rates across the workers' wage distribution to get an estimate of $\lambda_0^{ss'}$, $\lambda_1^{ss'}$, $\lambda_2^{ss'}$ for *eu*, *ue*, and *ne*.[26] These estimated coefficients determine the 'shape' of transition rates along worker's skill level. In the second step, which takes place during the calibration, we rescale the first-stage $\lambda_0^{ss'}$, $\lambda_1^{ss'}$ coefficients to target average worker flows *eu*, *ue*, and *ne* measured from the CPS.

The exogenous $\eta^{ss'}$ rates to and from the passive nonparticipant state do not depend on time nor on worker's skill z. These are set as follows. We set the transition rate from employment to passive nonemployment η^{en_0} to match the average level and sensitivity of the

24. See Heathcote and others (2023).

25. We target the 90-50 ratio because earnings variation at the top of the distribution is more directly associated with productivity variation, which is what we aim to measure, compared to the rest of the distribution where the extensive margin of labor supply plays a bigger role.

26. We do not use ne transitions directly in our estimation because the job acceptance decisions from nonparticipants create a wedge between job-finding rates out of nonparticipation $\lambda^{ne}(z)$ (our object of interest) and the observed *ne* flows (our empirical measure). Instead, we impose that the job-finding rate out of nonparticipation shares the same shape as the job-finding rate out of unemployment.

en flows in response to a monetary policy shock.[27] For the transition rate between unemployment and passive nonemployment η^{un_0}, we set it to match the average *un* flow.[28] Finally, we choose the transition rate from passive to active nonparticipation $\eta^{n_0 n_1}$ to match the flows out of participation, since workers have to be active nonparticipants before becoming jobseekers.

Taxes and transfers. We assume that unemployment benefits are given by $b(z_{it}) = \bar{b} w_t h_t z_{it}$, and set the UI replacement rate $\bar{b}$ to 0.5 of individual earnings. We set the rate $\eta^{u_1 u_0}$ to 0.167 to reflect an average UI benefits duration of 6 months. The proportional tax rate t is set to 0.2 and the lump-sum transfer ϕ is set to match 6 percent of average earnings in steady state.[29] The amount of government debt is set to equal one fourth of total equity.[30] Government expenditures are set residually to satisfy the budget constraint in steady state.

Production and price setting. Firm productivity α is set so that the after-tax hourly wage per efficiency unit in steady state is normalized to 1. The fixed operating cost χ affects the value of equity and, therefore, the size of the aggregate supply of liquid wealth. We set χ so that, given the household demand curve, the annual real interest rate that clears the asset market is 2 percent.

Both elasticities of substitution across labor types (ε) and across intermediate goods (ν) are set to 10, which implies wage and price markups around 10 percent. The nominal-wage adjustment cost Θ is set to match a slope of the structural wage Phillips curve (the semi-elasticity of inflation to deviations of marginal rate of substitution from the real wage) of 0.015 quarterly as recently estimated by Del Negro and others (2020).

27. In the model, both k^e and η^{en_0} are potential sources of *en* flows. Increasing the utility cost of working k^e raises the likelihood that a worker decides to leave employment to nonparticipation after a negative skill shock. Similarly, increasing the transition rate to passive nonparticipation η^{en_0} mechanically induces a flow towards passive nonparticipation. However, these two sources of *en* transitions hold very distinct implications for the sensitivity of *en* to the monetary policy shock. If we rely solely on the disutility k^e to match the average flows, we find a counterfactually strong positive *en* response following a contractionary monetary policy shock, as workers adjust their labor supply to counteract the negative wealth effects of the shock. If we rely solely on forced transitions η^{en_0} instead, then all movements from employment to nonparticipation are exogenous and we find no en response following a monetary policy shock. To discipline the relative importance of endogenous and exogenous *en* flows, we thus use the estimated *en* response in Graves and others (2023).

28. We set $\eta^{n_1 n_0} = \eta^{un_0}$ which corresponds to the assumption that all nonemployed workers transition into inactive nonparticipation at the same rate.

29. See Alves and Violante (2024).

30. See the 2019 Flow of Funds, table B.101.h Balance Sheet of Households.

Monetary and fiscal policy. We set steady-state (trend) inflation rate π^* at 2 percent. In our inflation targeting (IT) rule (17), we set the reaction coefficient on deviations of inflation from its trend to $\beta_\pi = 2.25$. The interest rate smoothing parameter is set to $\beta_\iota = 0.07$ to match the empirical persistence of the deviations of the unemployment rate after a monetary policy shock, as estimated by Graves and others (2023). Namely, in the data, unemployment returns to its pre-shock value after 4–5 years.[31] Finally, in the fiscal rule (16), we set $\beta_B = 0.1$.

Cyclicality of frictions. We model out of steady-state fluctuations of labor-market frictions (separation and job-finding rates) in a mechanical way. Specifically, we make the entire job-finding and separation rates functions $\lambda_t^{eu'}(z)$, $\lambda_t^{ne'}(z)$ and $\lambda_t^{ue'}(z)$ fluctuate in proportion to changes in the average hours per worker. This approach allows us to capture the heterogeneous fluctuations in job-finding rates and separation rates over the business cycle across skill levels without complicating the model further.

3. Results

The results are organized as follows. In section 3.1, we compare our model against recent empirical estimates of the effect of monetary policy surprises on labor-market variables. As we show, our calibration captures well the estimated responses of workers' stocks and flows to a monetary policy shock, including the response of the flows along the participation margin. In section 3.2, we focus on the long-run impact (10 years after the shock) of a transitory monetary policy contractionary shock for the dynamics of earnings and inflation. In subsection 3.2.1, we compute the long-run impact of the shock on aggregate labor earnings and explore its drivers along the skill distribution. In subsection 3.2.2, we use the Phillips curve (13) to investigate the short and long-run dynamics of inflation. Overall, we find that the micro-level sources of scarring present in the labor market do spill over to the macro economy, with a transitory contractionary monetary policy shock leading to long-lasting negative effects on aggregate earnings but not on inflation.

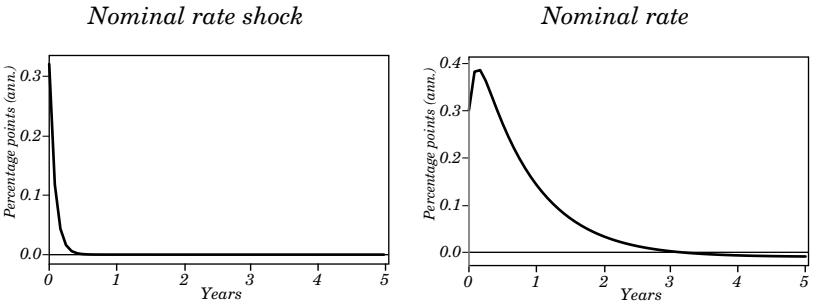
31. See figure C1.

3.1 Monetary Policy Transmission in a Frictional Labor Market

We study the impulse response to a (unanticipated) extremely transitory negative shock to the Taylor rule (17). In what follows, we compare the model's impulse responses to the ones estimated by Graves and others (2023) by using a high-frequency identification strategy, i.e., variation in rates in a narrow time window around announcements and Fed Chairs' speeches. Their results are reproduced in appendix C for reference.

To ease the comparison with Graves and others (2023), the size of the monetary shock in this section is chosen to match their estimated peak effect for the unemployment rate of 0.20 percentage points. The time paths for the shock and the nominal rate are illustrated in figure 1. This target implies a shock of 30 basis points that after a quarter is almost completely reabsorbed. The deviations of the nominal rate, plotted in the right-hand side panel, persist for longer due to the inertial reaction embedded in the rule (17).[32]

Figure 1. Monetary Policy Shock



Source: Authors' calculations.
Left panel: Monetary policy shock. Right panel: Path for the nominal interest rate implied by the Taylor rule (17).

32. Recall that the inertial parameter has been chosen to match the estimated persistence of unemployment rate deviations to the identified monetary policy shock. The internal propagation mechanism of the model generates persistence that goes well beyond the mechanical one due to the inertial Taylor rule.

Impulse-response functions to a monetary policy shock. Figure 2 plots the impulse-response functions (IRFs) for inflation, hours worked, unemployment rate, and output. As expected, the unexpected spike in nominal rates leads to a recession: unemployment rises, hours worked and output fall, and so does inflation. Note, however, that even though inflation, hours worked, and unemployment revert quite quickly to their pre-shock values, output is much slower to recover and remains depressed five years after the monetary shock. [33]

Figure 3 displays the IRFs of a number of labor-market stock and flow variables to this surprise increase in the policy rate. We begin with the stocks. The response of the unemployment rate, participation rate, and employment to population ratio are all consistent with the estimated VAR responses in Graves and others (2023).[34] In line with their results, the unemployment rate reacts sooner and displays the strongest response among all labor-market stocks. In contrast, the response in labor-force participation is weaker and takes more time to materialize: its trough is around one fifth of the peak in unemployment and occurs roughly a year later. As in the data, the dynamics of participation are very persistent: participation is still depressed five years after the shock, when the unemployment rate has already converged back to its steady state.

Turning to labor-market flows, all six flows move in the same direction as the estimates of Graves and others (2023).[35] Upon a monetary contraction, unemployment inflows (eu and nu) rise, unemployment outflows (ue and un) fall, and flows between employment and nonparticipation (en and ne) also decline. The response of ue and eu flows are mechanical, given the way we model fluctuations in labor-market frictions.[36] Other flows are mediated by individual labor-supply reactions. Importantly, the model reproduces the negative response of participation exit flows (en and un flows)

33. We discuss the long-run scars on output and earnings, as well as their sources, in the next section.

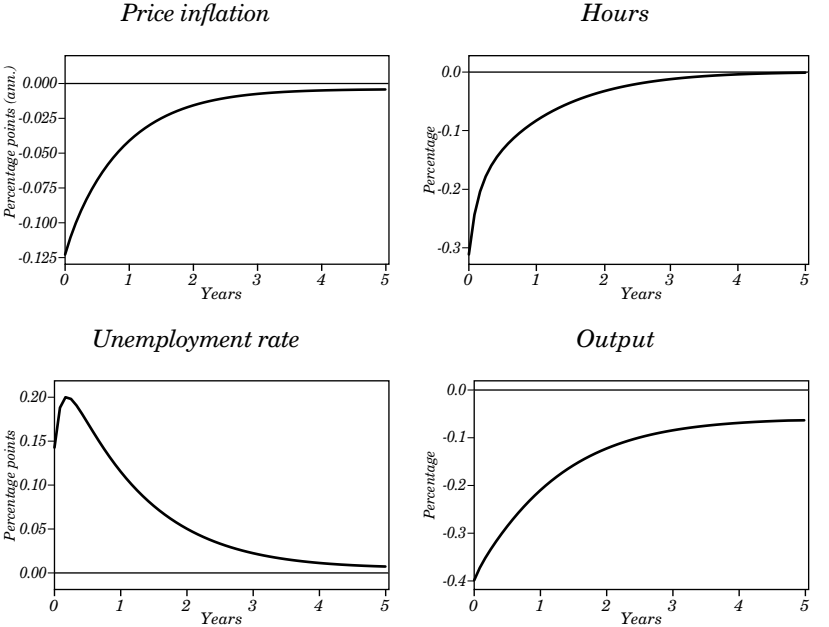
34. See their figure 5 reproduced in figure C1.

35. See their figure 6 reproduced in figure C2.

36. Recall that job-finding and separation rates are proportional to hours worked. Since hours are procyclical with respect to monetary shocks, the chosen elasticities of frictions to hours guarantee that ue and eu flows respond negatively and positively to the shock.

which is responsible for the initial increase in labor-force attachment at the start of the recessions.[37]

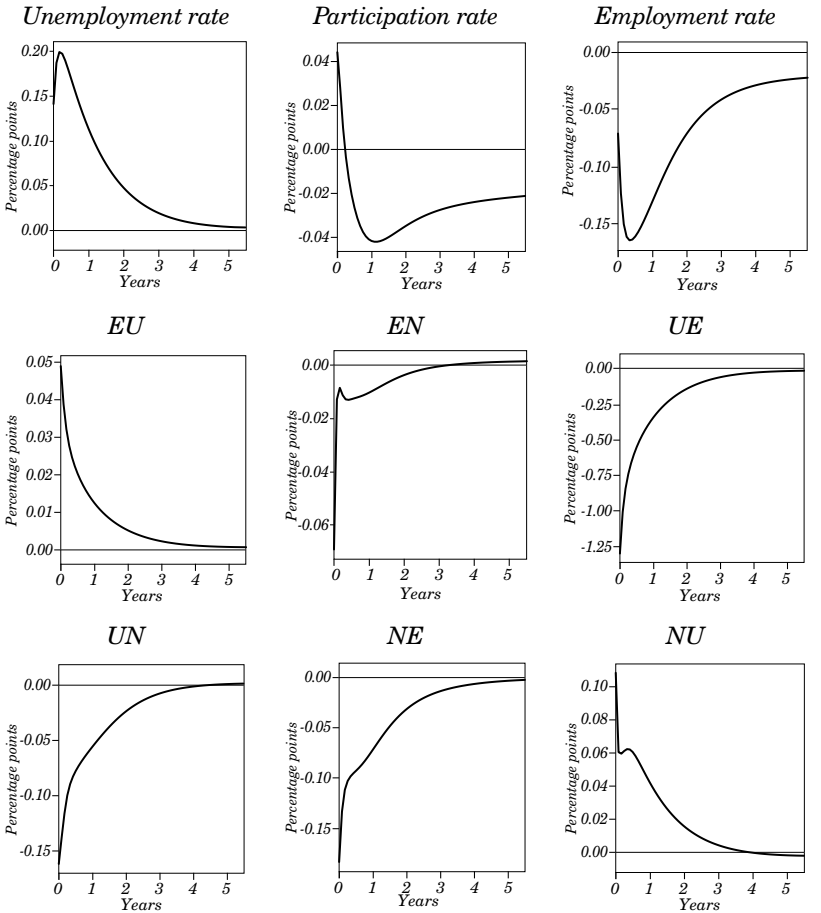
Figure 2. Model's IRFs to a Contractionary Monetary Policy Shock



Source: Authors' calculations.

37. Apart from an initial spike on *en* flows, the model also does a good job quantitatively at matching the magnitude of the flow response out of a monetary contraction compared to the estimated VAR responses in Graves and others (2023). (See their figure 6 reproduced in figure C2).

Figure 3. Model's Labor-Market Stocks and Flows IRFs to a Contractionary Monetary Policy Shock

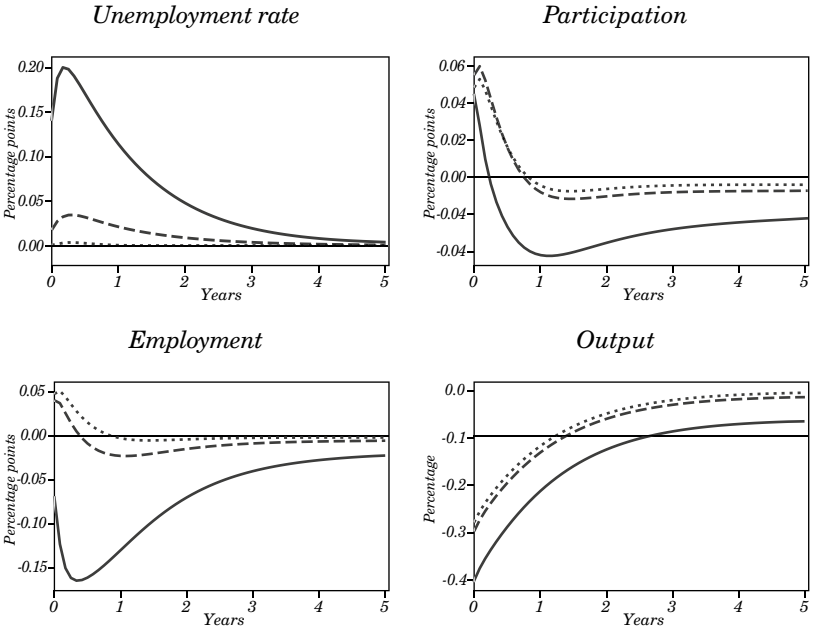


Source: Authors' calculations.
Note: *E* denotes employment, *U* unemployment, and *N* nonparticipation.

Our model is thus not only consistent with the empirical response of employment, unemployment, and nonparticipation but also with the underlying flows between the three labor-market states. Matching the dynamics of labor-market flows is important as these provide additional information about the relative role of labor-market frictions versus workers' labor-supply decisions in overall fluctuations of labor-market stocks.[38] For instance, the labor-market flows' reaction is crucial to understanding the weak negative response of labor-force participation to a contractionary monetary policy shock. As discussed in Elsby and others (2015) and Graves and others (2023), the procyclical movement in labor force participation exit rates (*en* and *un*) tends to push participation up in recessions. Working against this force, fluctuations in the *eu* and *ue* rates, which are determined mostly by fluctuations in labor-market frictions, exert a strong negative pressure on participation during downturns. Even though these flows do not affect participation directly, they induce a sharp countercyclical response of the unemployment rate. As unemployed workers are more likely than employed ones to drop out of the labor force—i.e., the *un* flow is, on average, much larger than the *en* flow—a large increase in the unemployment pool exerts, over time, downward pressure on participation. The effect of these two countervailing forces shows up on the model implied dynamics of the participation rate, as depicted in figure 3. At impact, the upward pressure from the *un* and *en* responses dominates, causing participation to increase initially. This effect dissipates quickly and, less than a year after the shock, the labor-force participation rate falls below trend, where it remains persistently depressed for the entire plotted horizon.

38. We are not the first to tackle the task of developing a labor-market model consistent with the joint dynamics of labor-market stocks and flows. In a standard representative-agent model, Cairó and others (2022) show that the opportunity cost of employment needs to be significantly more procyclical than the returns to work in order for the baseline model to match the procyclicality of participation flows *en* and *un*. Working with a heterogeneous-agent model similar to ours, Krusell and others (2017) highlight the importance of wealth heterogeneity and composition effects to explain the cyclicality of labor-market flows. Looking specifically at the worker flows dynamics conditional on a monetary policy shock, Graves and others (2023) also appeal to wealth effects in order to justify of the procyclical reaction of *en* and *un* flows.

Figure 4. Counterfactual Exploring the Importance of Fluctuations in Job-Finding and Separation Rates



Source: Authors' calculations.
Note: Solid line: baseline model IRFs. Dotted line: counterfactual IRFs with both job finding and separation rates fixed at their steady-state level. Dashed line: counterfactual IRFs with job-finding rate fixed at its steady-state level, but job-separation rates varying as in baseline.

Importance of job-finding and separation rates. Matching the cyclicality in job-finding and separation rates is thus crucial to the model's success in generating the right dynamics of labor-force participation and the other stocks. Next, we assess the relative importance of these two margins. We do so by computing an IRF to the same monetary policy shock while holding the job-finding and separation rates fixed at their steady-state levels, i.e., we set the elasticities of job-finding and separation rates (first jointly, then separately) to hours worked to zero. Figure 4 shows the outcome of this counterfactual exercise.

When we fixed both transition rates at their steady-state levels (dotted black line), unemployment, participation, and output display very different dynamics compared to the baseline case (solid line).

The unemployment rate shows almost no response to the shock, while participation features a strong counterfactual positive response without any significant decline below steady state thereafter. In addition, output falls less on impact than in the baseline and, importantly, it displays no long-lasting scarring effects. This result highlights that, besides being important for the response of labor-market stocks, fluctuations in the job-finding and separation rates are also the driving force behind the macro hysteresis.

Between job-finding and separation rates, which one contributes the most to the response of labor-market variables in our baseline calibration? The dashed line in figure 4 computes the IRF when the job-finding rate is kept constant, and the separation rate is allowed to move with hours worked. Thus, the difference between the dotted line (where both transition rates are kept fixed) and the dashed line (where job-finding rate only is held fixed) measures the role of fluctuations in the job separation rates. For all the variables in the figure, the dashed line is very close to the black line, indicating that it is fluctuations in the job-finding rate, through their impact on the unemployment pool, which are the main driving force of the response of the real economy to a contractionary monetary policy shock.

3.2 Long-Run Effects of Monetary Policy

In this section, we explore how the micro-level sources of hysteresis in the labor market lead to the hysteresis at the aggregate level following a transient monetary policy shock. We start by looking at the behavior of labor earnings to the monetary policy contraction. To better gauge the magnitude of the long-run effects, the simulations in this section are computed under a monetary policy shock that reduces total labor income by 1 percent over the first year.

3.2.1 Long-Run Labor Earnings

Aggregate labor income in our model $W_t = w\ell_t$ can be written as

$$W_t = wZ_t^e h_t (1 - u_t) LFPR_t, \tag{19}$$

where w is the constant real wage, Z_t^e is average labor productivity, h_t is average hours of employed workers, u_t is the unemployment

rate, and $LFPR_t$ is the labor-force participation rate.[39] In what follows, we separate terms in (19) into three, reflecting the different channels through which the monetary policy shock affect workers' earnings. The first is labor productivity Z_t^e, which is driven both by the individual skill dynamics and the selection of workers into employment. The second term in the combined effect of hours worked and the employment rate $h_t\,(1 - u_t)$. Since the unemployment rate is essentially determined by the job-finding and separation rates, which, in turn, are a function of hours worked, this term can be thought of as capturing changes in earnings driven by short-run fluctuations in firms' labor demand. The third term is simply the labor force participation $LFPR_t$.

Figure 5 shows the response of total labor income (and its decomposition into the three channels) one year and ten years after the monetary policy shock. To analyze the differential effects of the shock on low and high-wage workers, we also plot the earnings responses separately for the top and bottom quartiles of the workers' skill distribution.[40] Comparing the responses in the first year following the shock, we find that total labor income falls roughly twice more at the bottom quartile than it does at the top. Through the labor income decomposition, we see that the reaction of hours and the unemployment rate (the dark gray portion of the bars) is main force pushing down income in the aggregate and across the skill distribution.[41]

Ten years after the shock, aggregate labor income is still depressed by 0.2% (one fifth of the first-year decline). The income scarring is particularly acute at the bottom, with total labor income for the lowest quartile barely recovering from the first-year decline—for this group, ten-year ahead labor earnings are still 1.5% below steady-state, a reduction in earnings fifteen times larger than at the top quartile. The drivers of these long-run losses are also very different from those operating in the short-run. While short-run earnings losses are driven mostly by hours and unemployment dynamics, long-run

39. To arrive at this decomposition, start by noting that total labor income $w\ell_t = w\int_{s_{it}=e} z_{it} h_t di$, which can be expressed as $wZ_t^e h_t E_t$. Next use that total employment $E_t = (1 - u_t)\, LFPR_t$ which delivers expression (19).

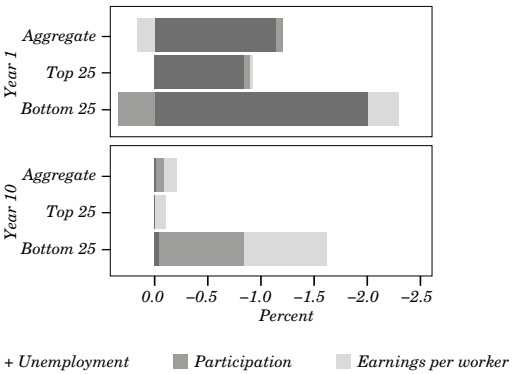
40. Clearly, equation (19) holds also at the group level, e.g., across all workers in the bottom quartile of the skill distribution at every t.

41. Participation shows a different dynamic at the bottom and the top, with a rise in participation at the bottom quartile moderating earnings losses for that group. However, these are small compared to the changes induced by movements in hours and unemployment.

earnings are depressed through a combination of weaker labor-force participation and productivity. Interestingly, participation falls only for the bottom of the skill distribution—as low-skilled workers go through unemployment they also become more likely to persistently exit the labor force.

Figure 6 offers another way to visualize what we have just discussed. The figure plots IRFs for total labor income and its three components over the first ten years after the shock. The solid line denotes outcomes for the whole population, while the dashed and dotted lines show outcomes for the bottom and top quartiles of the skill distribution respectively. The IRFs for the quartiles confirm our previous observation that the long-run impact of the shock is much stronger at the bottom of the skill distribution.

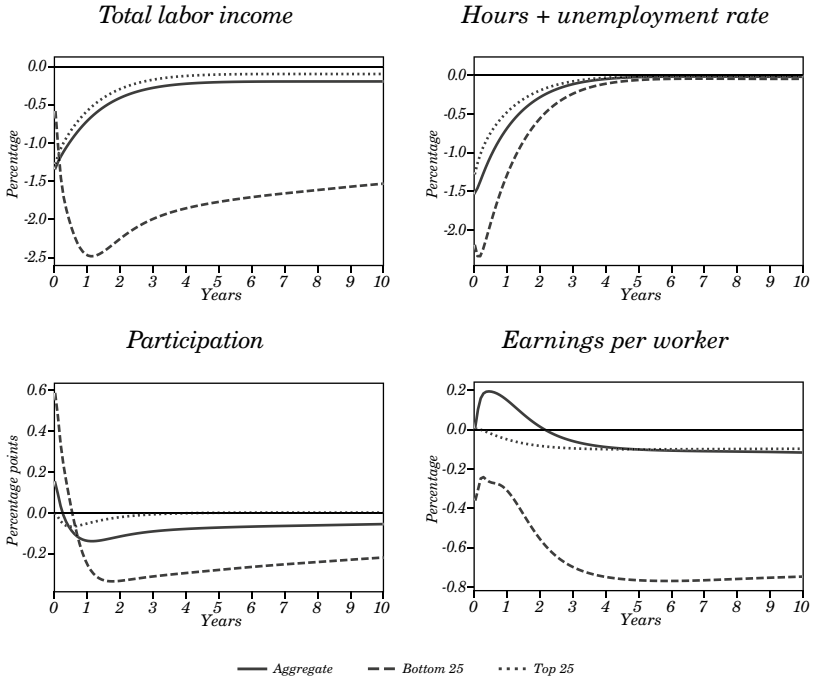
Figure 5. Decomposition of the Response of Labor Income



Source: Authors' calculations.

Note: Decomposition of the response of total labor income for the whole population, the population with skills in the top quartile (top 25), and the population with skills in the bottom quartile (bottom 25), at two different horizons (years 1 and 10 after the shock) See the discussion of equation (19) for a description of the three components. The size of the contractionary monetary policy shock is normalized so that total earnings drop by 1% over the first year following the shock.

Figure 6. IRFs to a Contractionary Monetary Policy Shock of Total Labor Income and its Three Components



Source: Authors' calculations.
Note: Solid line: aggregate. Dashed line: bottom quartile of the population by skill. Dotted line: top quartile of the population by skill. The size of the contractionary monetary policy shock is normalized so that total earnings drop by 1% over the first year following the shock.

3.2.2 Long-Run Inflation Dynamics

The Phillips curve we derived in equation (13) reveals that inflationary pressures are associated with an aggregate notion of the labor wedge. Log-linearizing the labor wedge around the steady state and substituting the result back into our wage Phillips curve, we obtain the following expression for the dynamics of inflation:

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \kappa\Big[\left(\xi d\log Y_t + d\log C_t\right) - \xi d\log LFPR_t - \left(\xi+1\right)d\log Z_t^e + \left(d\log \tilde{C}_t^e - d\log C_t\right)\Big]' \tag{20}$$

where Y_t is aggregate output, C_t is aggregate consumption, $LFPR_t$ is the labor-force participation rate, Z_t^e is the average labor productivity, and $\tilde{C}_t^e$ is the productivity-weighted consumption of employed workers. Log-deviations of X from steady state are denoted by $d\log X$. See appendix D for a detailed derivation.

Drivers of inflation dynamics. Expression (20) identifies four drivers of inflation dynamics in our model. The first term $(\xi d\log Y_t + d\log C_t)$ combines movements in aggregate output Y_t and household total consumption C_t, and is equivalent to the marginal rate of substitution between leisure and consumption of an 'as-if' representative-agent with log utility over consumption and inverse Frisch elasticity ξ.[42] This term is procyclical and leads to deflationary pressures upon a contractionary monetary policy shock.

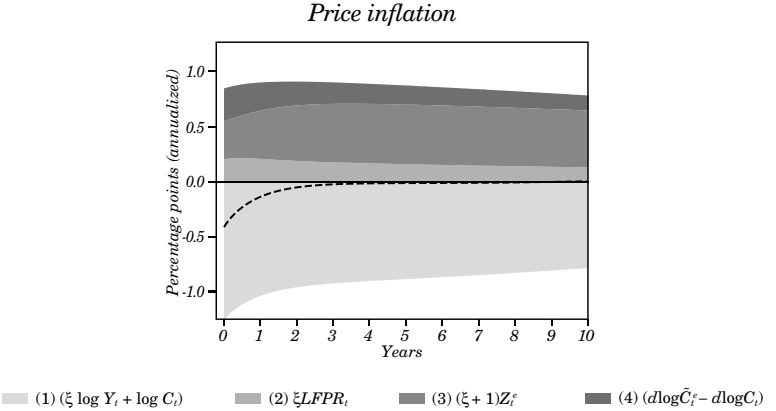
The other three components are germane to our heterogeneous-agent model with endogenous participation and state-dependent skill dynamics. In contrast to the first term, they all create inflationary pressures following a contractionary monetary policy shock. The second term is driven by movements in labor-force participation $d\log LFPR_t$. Intuitively, a rise in the participation rate lowers inflationary pressures by increasing the supply of potential workers. Average labor productivity $d\log Z_t^e$, which shows up in the third term, has a similar effect: a decrease in average labor productivity following a monetary contraction adds inflationary pressures as workers need to supply more hours to produce the same amount of the final good. The fourth term $(d\log \tilde{C}_t^e - d\log C_t)$ denotes the gap between productivity-weighted consumption of the employed and total consumption. This component captures the idea that the labor union, when setting nominal wages, cares only about the marginal utility of union 'insiders.' In a recession driven by a contractionary monetary policy (or a demand) shock, this term is positive, making unions less willing to lower nominal wages in response to a reduction

42. Imagine an economy with linear production technology on hours $Y_t = H_t$ and a representative agent with utility over consumption and hours worked given by $U(C,H) = \log(C) - \psi\frac{H^{1+\xi}}{1+\xi}$. In this case, the marginal rate of substitution between leisure and consumption, already using the production technology to substitute hours for output, is given by $MRS_t = \psi Y_t^{\xi} C_t$, or in log-linear deviations from steady state, $d\log MRS_t = (\xi d\log Y_t + d\log C_t)$.

in the demand for their labor task.[43] This extra degree of nominal wage rigidity induced by union's behavior is reminiscent of the classical insider-outsider model.[44]

Decomposition of inflation dynamics. Figure 7 shows the decomposition of inflation dynamics in the four terms in equation (20). As we previewed above, in response to a contractionary monetary policy shock, the first term in the decomposition is deflationary, while the remaining three terms are all inflationary. Quantitatively, the inflationary pressures coming from the last three terms are quite strong.[45] In particular, even though output and consumption in the long run remain depressed—which, through the first term, exerts a persistent deflationary pressure—inflation returns to target very quickly as the other terms keep pushing inflation up.

Figure 7. Decomposition of the Response of Inflation to a Contractionary Monetary Policy Shock at Different Horizons



Source: Authors' calculations.
Note: The dashed line is the model's aggregate price inflation. The four shaded areas correspond to the four terms in equation (20). The size of the contractionary monetary policy shock is normalized so that total earnings drop by 1% over the first year following the shock.

43. To see this, remember that high-wages workers are less likely to lose their jobs in recessions, making the employment pool during contractions skewed towards high-skilled workers. This composition effect explains why the consumption of employed workers $\tilde{C}_t^e$ falls less than total consumption C_t^e during downturns.

44. See Galí (2022).

45. The inflationary effect of productivity is smaller in the short run because of the positive selection effect in labor-force participation.

4. Conclusions

We have developed a model where the transmission mechanism that impresses such long memory to the macroeconomy operates through the labor market, according to Okun (1973) hypothesis. During economic downturns, many workers are displaced from their jobs. As they spend time unemployed, they are subject to large and persistent skill losses which lead some of them to transition into nonparticipation. Nonparticipation tends to be a long-lasting state that fuels further skill deterioration and crystallizes disattachment from the labor force. This vicious circle is the reason why even a very transitory demand shock, such as a contractionary monetary policy shock, leads to a long-run decline in labor productivity, labor-force participation, and output. These chronic effects do not extend to persistent deflation because lower productivity and deficient labor supply represent inflationary forces that counteract the protracted decline in aggregate demand.

Going forward, there are at least three interrelated issues that we have not tackled in this paper. First, because of the way we numerically solve for the equilibrium dynamics of the model, our environment features both scarring effects of negative temporary shocks and uplifting effects of positive transitory ones. An expansionary demand shock that pulls more people into employment allows them to gradually build their productivity and reinforces attachment to the labor force. Whether hysteresis is only negative or also positive is an empirical question that remains to be settled.[46]

Second, the model has implications for the optimal conduct of monetary policy. In light of our results that in the model short-lived negative demand shocks leave persistent scars on output, but they do not necessarily generate deflationary pressures, an inflation-focused central bank may do too little for the economy.[47] Putting excessive weight on inflation (or deflation) would lead a central bank to stop responding to the shock too quickly thus allowing it to persistently damage the real economy. Optimal monetary policy should be more aggressive early on to moderate the increase in unemployment, which is the source of the hysteresis. In addition, according to our model, a rule that responds to output (or participation) is more suitable than one

46. Bluedorn and Leigh (2019) find some empirical support for the positive hysteresis hypothesis based on long-run revisions of professional forecasters when positive news about the labor market is released.

47. See Galí (2022).

that responds to unemployment because it incorporates deviations of productivity and labor-force participation from their efficient (flexible price) level. In addition, even once the post-shock hysteresis has taken place and the damage to the real economy has occurred, our model implies that keeping monetary policy expansionary for a while can fully revert the underutilization of labor without a surge in inflation.

Finally, we have shown that hysteresis is much more severe at the bottom of the skill distribution. The reason is that it is low-wage workers who are marginally attached to the labor force and those for whom earnings losses upon displacement are the largest.[48] Since the new framework of the Fed[49] reinterprets its full employment mandate as broad-based and inclusive, policymakers should be especially aware of the long-run effects, both positive and negative, that untamed shocks can have on the more disadvantaged groups.

48. See Athey and others (2023), Cajner and others (2017), Guvenen and others (2017), and Yagan (2019).

49. See Powell (2020).

References

Aaronson, S.R., M.C. Daly, W.L. Wascher, and D.W. Wilcox. 2019. "Okun Revisited: Who Benefits Most from a Strong Economy?" *Brookings Papers on Economic Activity* 2019(1): 333–404.

Abbritti, M., A. Consolo, and S. Weber. 2021. "Endogenous Growth, Downward Wage Rigidity and Optimal Inflation." ECB Working Papers No. 2021/2635.

Acharya, S., J. Bengui, K. Dogra, and S.L. Wee. 2022. "Slow Recoveries and Unemployment Traps: Monetary Policy in a Time of Hysteresis." *Economic Journal* 132(646): 2007–47.

Alves, F., G. Kaplan, B. Moll, and G.L. Violante. 2020. "A Further Look at the Propagation of Monetary Policy Shocks in HANK." *Journal of Money, Credit and Banking* 52(S2): 521–59.

Alves, F. and G.L. Violante. 2024. "Some Like It Hot: Monetary Policy under Okun's Hypothesis." Princeton University.

Athey, S., L.K. Simon, O.N. Skans, J.V., and Y. Yakymovych. 2023. "The Heterogeneous Earnings Impact of Job Loss across Workers, Establishments, and Markets." Working Papers, Stanford University.

Auclert, A., B. Bardóczy, and M. Rognlie. 2023. "MPCs, MPEs and Multipliers: A Trilemma for New Keynesian Models." *Review of Economics and Statistics* 105(3): 700–12.

Auclert, A., M. Rognlie, and L. Straub. 2018. "The Intertemporal Keynesian Cross." NBER Working Papers No. 25020.

Bianchi, F., H. Kung, and G. Morales. 2019. "Growth, Slowdowns, and Recoveries." *Journal of Monetary Economics* 101(C): 47–63.

Blanchard, O.J., E. Cerutti, and L. Summers. 2015. "Inflation and Activity – Two Explorations and their Monetary Policy Implications." NBER Working Papers No. 21726.

Blanchard, O.J. and L.H. Summers. 1986. "Hysteresis and the European Unemployment Problem." *NBER Macroeconomics Annual* 1: 15–78.

Bluedorn, J.C. and D. Leigh. 2019. "Hysteresis in Labor Markets? Evidence from Professional Long-Term Forecasts." IMF Working Papers No. 2019/114.

Brunnermeier, M.K. 2021. *The Resilient Society.* Endeavor Literary Press.

Cairó, I., S. Fujita, and C. Morales-Jiménez. 2022. "The Cyclicality of Labor Force Participation Flows: The Role of Labor Supply Elasticities and Wage Rigidity." *Review of Economic Dynamics* 43: 197–216.

Cajner, T., J. Coglianese, and J. Montes. 2021. "The Long-Lived Cyclicality of the Labor Force Participation Rate." U.S. Federal Reserve Discussion Papers No. 2021/047.

Cajner, Tomaz, Tyler Radler, David Ratner, and Ivan Vidangos. 2017. "Racial gaps in labor market outcomes in the last four decades and over the business cycle." FED Working Paper.

Cerra, V., A. Fatás, and S.C. Saxena. 2023. "Hysteresis and Business Cycles." *Journal of Economic Literature* 61(1): 181–225.

Chang, Y., J.F. Gomes, and F. Schorfheide. 2002. "Learning-by-Doing as a Propagation Mechanism." *American Economic Review* 92(5): 1498–520.

Christiano, L.J., M. Trabandt, and K. Walentin. 2021. "Involuntary Unemployment and the Business Cycle." *Review of Economic Dynamics* 39: 26–54

Comin, D. and M. Gertler. 2006. "Medium-Term Business Cycles." *American Economic Review* 96(3): 523–51.

Davis, S.J. and T. von Wachter. 2011. "Recessions and the Costs of Job Loss." *Brookings Papers on Economic Activity* 42(2): 1–72.

Del Negro, M., M. Lenza, G.E. Primiceri, and A. Tambalotti. 2020. "What's Up with the Phillips Curve?" *Brookings Papers on Economic Activity* 51(1): 301–73.

Dávila, E. and A. Schaab. 2023. "Optimal Monetary Policy with Heterogeneous Agents: Discretion, Commitment, and Timeless Policy." NBER Working Papers No. 30961.

Elsby, M.W.L., B. Hobijn, and A. Şahin. 2015. "On the Importance of the Participation Margin for Labor Market Fluctuations." *Journal of Monetary Economics* 72: 64–82.

Erceg, C.J., D. W. Henderson, and A.T. Levin. 2000. "Optimal Monetary Policy with Staggered Wage and Price Contracts." *Journal of Monetary Economics* 46(2): 281–313.

Faberman, R.J., A.I. Mueller, A. Şahin, and G. Topa. 2017. "Job Search Behavior among the Employed and Nonemployed." NBER Working Papers No. 23731.

Fornaro, L. and M. Wolf. 2020. "The Scars of Supply Shocks." CEPR Discussion Paper No. DP15423.

Furlanetto, F., A. Lepetit, Ø. Robstad, J. Rubio-Ramírez, and P. Ulvedal. 2021. "Estimating Hysteresis Effects." CEPR Discussion Paper No. DP16558.

Gaillard, A. and P. Wangner. 2023. "Inequality, Business Cycles, and Growth: A Unified Approach to Stabilization Policies." Brown University.

Galí, J. 2022. "Insider-Outsider Labor Markets, Hysteresis, and Monetary Policy." *Journal of Money, Credit and Banking* 54(S1): 53–88.

Galí, J., F. Smets, and R. Wouters. 2012. "Unemployment in an Estimated New Keynesian Model." *NBER Macroeconomics Annual* 26(1): 329–60.

Graves, S., C. Huckfeldt, and E.T. Swanson. 2023. "The Labor Demand and Labor Supply Channels of Monetary Policy." NBER Working Papers No. 31770.

Guvenen, F., F. Karahan, S. Ozkan, and J. Song. 2017. "Heterogeneous Scarring Effects of Full-Year Nonemployment." *American Economic Review* 107(5): 369–73.

Hall, R.E. and M. Kudlyak. 2019. "Job-Finding and Job-Losing: A Comprehensive Model of Heterogeneous Individual Labor-Market Dynamics." NBER Working Papers No. 25625.

Heathcote, J., F. Perri, and G.L. Violante. 2020. "The Rise of U.S. Earnings Inequality: Does the Cycle Drive the Trend?" *Review of Economic Dynamics* 37: S181–S204.

Heathcote, J., F. Perri, G.L. Violante, and L, Zhang. 2023. "More Unequal We Stand? Inequality Dynamics in the United States, 1967-2021." *Review of Economic Dynamics* 50: 235–66.

Hershbein, B. and B.A. Stuart. 2020. "Recessions and Local Labor Market Hysteresis." Upjohn Institute Working Paper No. 20–325.

Hobijn, B. and A. Şahin. 2021. "Maximum Employment and the Participation Cycle." NBER Working Papers No. 29222.

Huo, Z. and J.V. Ríos-Rull. 2020. "Sticky Wage Models and Labor Supply Constraints." *American Economic Journal: Macroeconomics* 12(3): 284–318.

Jacobson, L.S., R.J. LaLonde, and D.G. Sullivan. 1993. "Earnings Losses of Displaced Workers." *American Economic Review* 83(4): 685–709.

Jordà, Ò., S.R. Singh, and A.M. Taylor. 2020. "The Long-Run Effects of Monetary Policy." NBER Working Papers No. 26666.

Kahn, L.B. 2010. "The Long-Term Labor Market Consequences of Graduating from College in a Bad Economy." *Labour Economics* 17(2): 303–16.

Kaplan, G. and G.L. Violante. 2022. "The Marginal Propensity to Consume in Heterogeneous Agent Models." *Annual Review of Economics* 14(1): 747–75.

Krusell, P., T. Mukoyama, R. Rogerson, and A. Şahin. 2017. "Gross Worker Flows over the Business Cycle." *American Economic Review* 107(11): 3447–76.

Lepetit, A. 2023. "Hysteresis, Inflation Dynamics, and the Changing Phillips Correlation." Federal Reserve Board Working Paper.

Ljungqvist, L., and T.J. Sargent. 1998. "The European Unemployment Dilemma." *Journal of Political Economy* 106(3): 514–50.

Ma, Y. and K. Zimmermann. 2023. "Monetary Policy and Innovation." NBER Working Papers No. 31698.

Okun, A.M. 1973. "Upward Mobility in a High-Pressure Economy." *Brookings Papers on Economic Activity* 1973(1): 207–61.

Powell, J.H. 2020. "New Economic Challenges and the Fed's Monetary Policy Review." Remarks at the Jackson Hole Symposium sponsored by the Federal Reserve Bank of Kansas City.

Reifschneider, D., W. Wascher, and D. Wilcox. 2015. "Aggregate Supply in the United States: Recent Developments and Implications for the Conduct of Monetary Policy." *IMF Economic Review* 63(1): 71–109.

Rinz, K. 2022. "Did Timing Matter? Life Cycle Differences in Effects of Exposure to the Great Recession." *Journal of Labor Economics* 40(3): 703–35.

Rothstein, J. 2023. "The Lost Generation? Labor Market Outcomes for Post-Great-Recession Entrants." Journal of Human Resources 58(5): 1452–79.

Shimer, R. 2013. "Job Search, Labor-Force Participation, and Wage Rigidities." In *Advances in Economics and Econometrics*, edited by D. Acemoglu, M. Arellano, and E. Dekel. Cambridge, MA: Cambridge University Press.

Tobin, J. 1980. "Stabilization Policy Ten Years After." *Brookings Papers on Economic Activity* 1980(1): 19–89.

Topel, R. 1990. "Specific Capital and Unemployment: Measuring the Costs and Consequences of Job Loss." *Carnegie-Rochester Conference Series on Public Policy* 33: 181–214, Elsevier.

Yagan, D. 2019. "Employment Hysteresis from the Great Recession." *Journal of Political Economy* 127(5): 2505–58.

APPENDICES

Appendix A. Household Optimization Problems

We presented the problem of the active nonparticipant in the main text. Here we describe all others.

Consider the problem of the passive nonparticipant (n_0):

$$v_0^{n_0}(a_0, z_0) = \max_{\{c_t\}_{t\geq 0}} \mathbb{E}_0 \int_0^{\tau^{n_1}} e^{-\rho t} u^n(c_t, h_t)\, dt + e^{-\rho \tau^{n_1}} v_{\tau^{n_1}}^{n_1}\left(a_{\tau^{n_1}}, z_{\tau^{n_1}}\right) \tag{A1}$$

$$c_t + \dot{a}_t = r_t a_t + \phi$$

$$a_t \geq 0.$$

Passive nonparticipants do not receive any job offers. At rate η_1, with τ^{n_1} being the first arrival rate of this event, they become active nonparticipants (state n_1). The conditional expectation reflects the uncertainty in transition rates and skill dynamics. In addition to the participation decision $\mathfrak{p}_t^{n_0}$, at every instant, the worker chooses its consumption flow c_t. The last two lines of this problem state the budget constraint (in real terms) and the borrowing limit.

The problem of an unemployed household who is not eligible for UI benefits is:

$$v_0^{u_0}(a_0, z_0) = \max_{\{c_t\}_{t\geq 0}, \tau^*} \mathbb{E}_0 \left[\int_0^{\tau^{\min}} e^{-\rho t} u^u(c_t, h_t)\, dt + \mathbb{I}_{\left\{\tau^{\min} = \tau^e\right\}} e^{-\rho \tau^e} v_{\tau^e}^e\left(a_{\tau^e}, z_{\tau^e}\right) \right. \tag{A2}$$

$$\left. + \mathbb{I}_{\left\{\tau^{\min} = \tau^*\right\}} e^{-\rho \tau} v_{\tau^*}^{n_1}\left(a_{\tau^*}, z_{\tau^*}\right) + \mathbb{I}_{\left\{\tau^{\min} = \tau^n\right\}} e^{-\rho \tau^{n_0}} v\tau_{\tau^{n_0}}^{n_0}\left(a_{\tau^{n_0}}, \tau_{\tau^{n_0}}\right) \right]$$

s.t.

$$c_t + \dot{a}_t = r_t a_t + \phi$$

$$a_t \geq 0.$$

Ineligible unemployed workers receive a job offer at rate λ_{zt}^{ue} (with τ^e being the first arrival time of this event) and always take it. At any time τ^* during the unemployment spell, the individual can exit the labor force ($\mathfrak{p}_t^u = 0$). Finally, at rate η_0 (with τ^{n_0} being the first arrival rate of this shock) they can become passive nonparticipants.

The problem of an unemployed household who is eligible for UI benefits is:

$$\nu_0^{u_1}(a_0,z_0)=\max_{\{c_t\}_{t\geq 0},\tau^*}\mathbb{E}_0\Bigg[\int_0^{\tau^{\min}}e^{-\rho t}u^u(c_t,h_t)\,dt+\mathbb{I}_{\{\tau^{\min}=\tau^e\}}e^{-\rho\tau^e} \tag{A3}$$

$$\max\left\{\nu_{\tau^e}^e\left(a_{\tau^e},z_{\tau^e}\right),\nu_{\tau^e}^{u_1}\left(a_{\tau^e},z_{\tau^e}\right)\right\}+\mathbb{I}_{\{\tau^{\min}=\tau^*\}}e^{-\rho\tau^*}\nu_{\tau^*}^{n_1}\left(a_{\tau^*},z_{\tau^*}\right)$$

$$+\mathbb{I}_{\{\tau^{\min}=\tau^u\}}e^{-\rho\tau^{u1}}\nu\tau_{\tau^u0}^{u_0}\left(a_{\tau^u0},z_{\tau^u0}\right)+\mathbb{I}_{\{\tau^{\min}=\tau^{n_0}\}}e^{-\rho\tau^{n_0}}\nu_{\tau^{n_0}}^{n_0}\left(a_{\tau^{n_0}},z_{\tau^{n_0}}\right)\Bigg]$$

s.t.

$$c_t+\dot{a}_t=r_ta_t+(1-t)b(z_t)+\phi$$

$$a_t\geq 0.$$

Besides receiving job opportunities and choosing whether to take them, choosing to drop out of the labor force and exogenously switching to passive nonparticipant status, the eligible unemployed could lose their entitlement to UI benefit at rate $\eta^{u_1u_0}$, with τ^{u_0} being the first arrival time of this event.

Finally, the problem of the employed household is:

$$\nu_0^e(a,z)=\max_{\{c_t\}_{t\geq 0},\tau^*}\mathbb{E}_0\Bigg[\int_0^{\tau^{\min}}e^{-\rho t}u^e(c_t,h_t)\,dt+\mathbb{I}_{\{\tau^{\min}=\tau u\}}e^{-\rho\tau^u}\nu_{\tau^u}^{u_1}\left(a_{\tau^u},z_{\tau^u}\right) \tag{A4}$$

$$+\mathbb{I}_{\{\tau^{\min}=\tau^*\}}e^{-\rho\tau^*}\nu_{\tau^*}^{n_1}\left(a_{\tau^*},z_{\tau^*}\right)+\mathbb{I}_{\{\tau^{\min}=\tau^{n_0}\}}e^{-\rho\tau^{n_0}}\nu\tau_{\tau^{n_0}}^{n_0}\left(a_{\tau^{n_0}},z_{\tau^{n_0}}\right)\Bigg]$$

s.t.

$$c_t+\dot{a}_t=r_ta_t+(1-t)w_tz_th_t+\phi$$

$$a_t\geq 0.$$

Employed workers (e) can be displaced at rate λ_{zt}^{eu}, in which case they become eligible for UI benefits ($u=u_1$). Let τ^u be the first arrival time of this Poisson shock. At every instant τ^*, the employed worker can choose to quit the labor force ($\mathfrak{p}_t^e=0$).[50] In addition, an employed worker can exogenously switch to passive nonparticipant status at rate η^{en_0}, with τ^{n_0} being the first arrival time of this event.

50. Quitting into unemployment is never optimal, because the worker would not receive UI benefits, and would pay a higher disutility cost κ for the opportunity to be re-employed at the same wage.

Each problem (including the one for the active nonparticipant in the main text) can be expressed recursively as a Hamilton-Jacobi-Bellman quasi-variational inequality. This equation, in turn, can be discretized and solved.[51]

Appendix B. Problem of the Mutual Fund

The problem of the mutual fund, which takes prices as given, entails choosing the optimal portfolio composition between bonds and equity:

$$r_t A_t\left(X^m, B^m\right) = \max_{\dot{X}_t^m, \dot{B}_t^m} \Pi_t X_t^m - q_t \dot{X}_t^m + r_t^b B_t^m - \dot{B}_t^m + \partial_X A_t\left(X^m, B^m\right)\dot{X}^m + \partial_B A_t\left(X^m, B^m\right)\dot{B}_t^m + \partial_t A_t\left(X^m, B^m\right) \tag{B1}$$

with first-order conditions with respect to $\dot{X}_t^m$ and $\dot{B}_t^m$

$$q_t = \partial_X A_t\left(X^m, B^m\right)$$
$$1 = \partial_B A_t\left(X^m, B^m\right)$$

Substituting these first-order conditions into (B1) and exploiting the linear homogeneity of the problem, which implies that $A_t = q_t X_t^m + B_t^m$, we arrive at

$$r_t\left(q_t X_t^m + B_t^m\right) = \Pi_t X_t^m + r_t^b B_t^m + \dot{q}_t X^m.$$

By matching coefficients on equity and bonds, we obtain

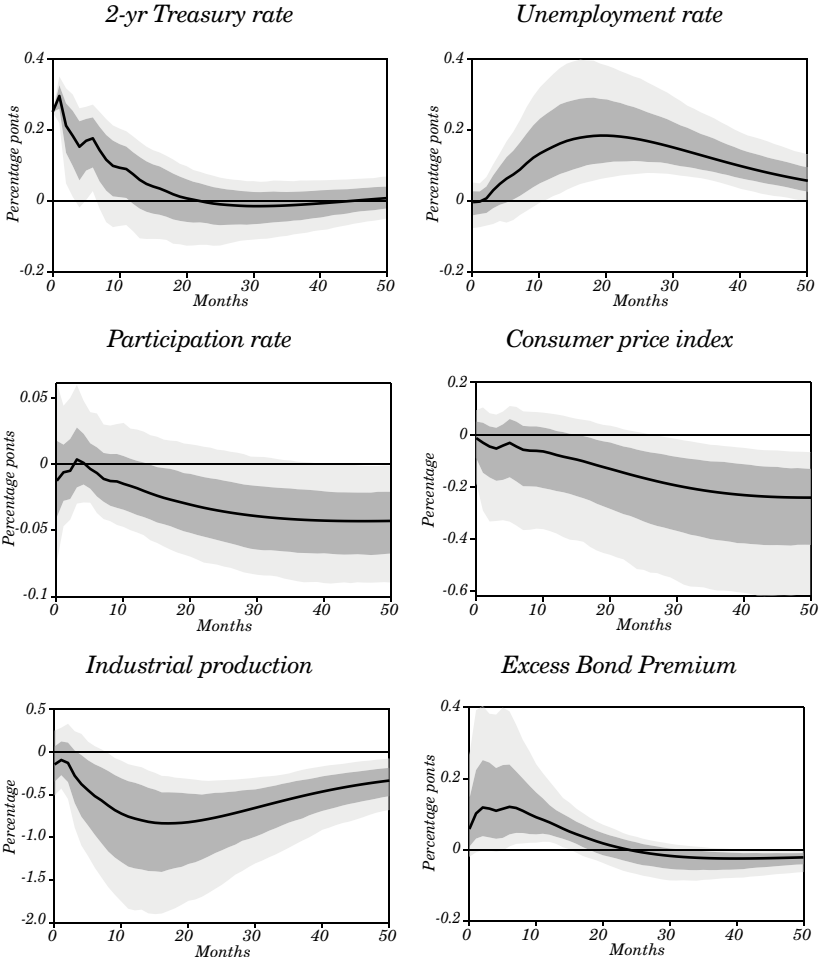
$$r_t = \frac{\Pi_t + \dot{q}_t}{q_t} = r_t^b, \tag{B2}$$

which determines the real return on the household financial asset a_t (wealth invested in the mutual fund) and establishes a no-arbitrage condition between government bonds and firm equity which holds at every t, except when a shock hits the economy, in which case the price q_t features a jump.

51. See Alves and Violante (2024) for details.

Appendix C. Empirical Estimates from Graves and others (2023)

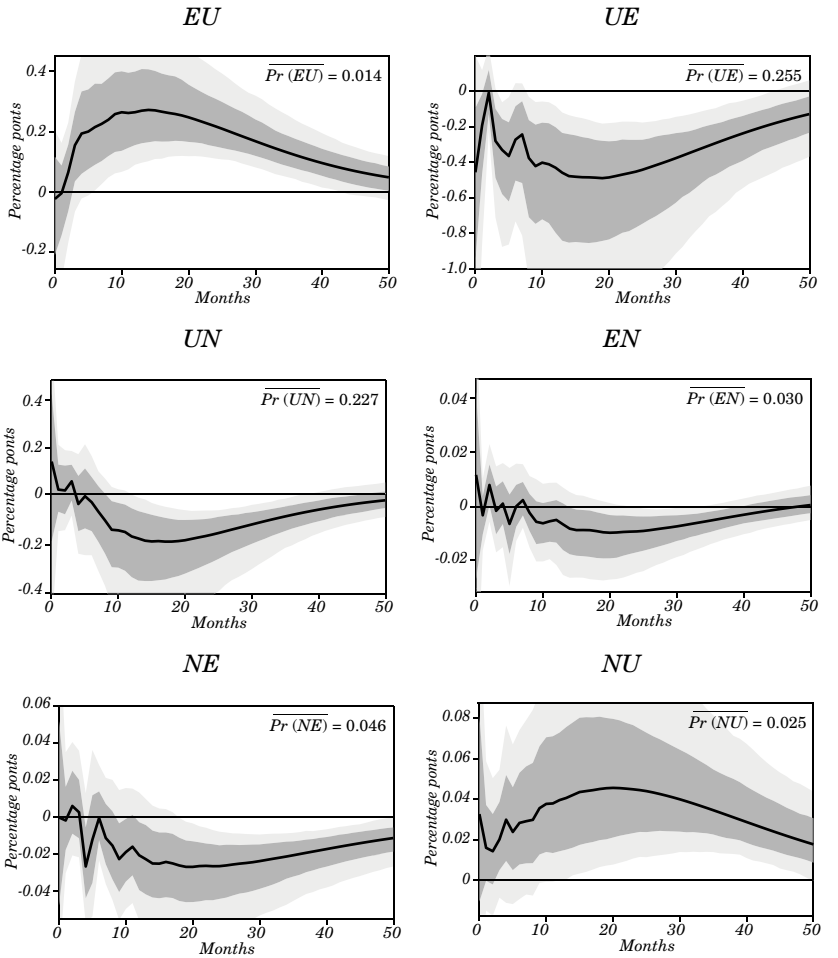
Figure C1. Response of Aggregate Variables to a Monetary Policy Shock



Source: Authors' calculations.

Note: Estimated impulse responses to a 25bp monetary policy tightening shock in the baseline VAR. Solid lines report IRFs, while light- and dark-shaded areas report bootstrapped 68% and 90% standard error bands. See Graves and others (2023) for more details.

Figure C2. Response of Labor-Market Flows to a Monetary Policy Shock



Source: Authors' calculations.

Note: Estimated impulse responses to a 25bp monetary policy tightening shock, computed by appending the given labor-market flow variable to the baseline VAR from figure C1. Solid black lines report IRFs while light- and dark-shaded areas report bootstrapped 68% and 90% standard error bands. See Graves and others (2023) for more details.

Appendix D. Log-Linear Approximation of the Labor Wedge

We start with the wage Phillips curve in equation (13), which we write as

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \frac{\epsilon}{\Theta}\left(H_t\right)\left[\psi h_t^{\frac{1}{\sigma}} - \left(\frac{\epsilon - 1}{\epsilon}\right)(1 - \mathrm{t})wZ_t^e\left(\tilde{C}_t^e\right)^{-1}\right],$$

where π_t is aggregate (wage and price) inflation rate, H_t aggregate hours, h_t average hours per worker, Z_t^e average productivity among the employed, and $\tilde{C}_t^e$ is the virtual aggregate consumption of the employed implicitly defined by the following equation

$$\frac{1}{\tilde{C}_t^e} = \int_{s_{it}=e} \frac{1}{c_{it}}\left(\frac{z_{it}}{\int_{s_{it}=e} z_{it}di}\right)di.$$

We now take a log-linear approximation around the steady state of the equation's right-hand side, obtaining

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \kappa^w\left[\sigma^{-1}\mathrm{dlog}\,h_t - \mathrm{dlog}\,Z_t^e + \mathrm{dlog}\,\tilde{C}_t^e\right], \tag{D1}$$

where $\kappa^w \equiv \frac{\epsilon}{\Theta}\left(H\right)\psi h^{\frac{1}{\sigma}}$. Using the aggregate production function (12), we can write

$$d\log Y_t = d\log h_t + d\log\left(1 - u_t\right) + d\log LFPR_t + d\log Z_t^e.$$

Under the assumption that the unemployment rate is proportional to average hours worked h_t, which is approximately true in our model given the imposed relation between labor-market frictions and h_t, we can write

$$d\log Y_t = \left(1 + \varepsilon_{e,h}\right)d\log h_t + d\log LFPR_t + d\log Z_t^e,$$

where $\varepsilon_{e,h}$ is the elasticity of the $(1 - u_t)$ to hours h_t. Using this to substitute out hours worked from (D1) and arrive at:

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \kappa^w \left[\frac{\sigma^{-1}}{1+\varepsilon_{e,h}} \left(d\log Y_t - d\log LFPR_t - d\log Z_t^e \right) \right. \tag{D2}$$
$$\left. - d\log Z_t^e + d\log \tilde{C}_t^e \right].$$

If we let

$$\xi \equiv \frac{\sigma^{-1}}{1+\varepsilon_{e,h}}$$

and collect terms, we can re-express (D2) as

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \kappa^w \left[\xi d\log Y_t - \xi d\log LFPR_t - \left(\xi + 1\right) d\log Z_t^e + d\log \tilde{C}_t^e \right].$$

Finally, add and subtract log deviations in aggregate consumption $d\log C_t$ to obtain

$$\rho\left(\pi_t - \pi^*\right) - \dot{\pi}_t = \kappa^w \left[\xi d\log Y_t + d\log C_t - \xi d\log LFPR_t - \left(\xi + 1\right) d\log Z_t^e, \right.$$
$$\left. + \left(d\log \tilde{C}_t^e - d\log C_t \right) \right]$$

which is equation (20) in the main text.

On the Optimal Use of Fiscal Stimulus Payments at the Zero Lower Bound

Alisdair McKay
Federal Reserve Bank of Minneapolis

Christian K. Wolf
Massachusetts Institute of Technology
National Bureau of Economic Research

For much of the decade following the Great Recession, central banks across the world remained constrained by a binding zero (or effective) lower bound (ZLB) on nominal interest rates. Much academic and policy interest thus centered on the question of how fiscal policy could be used to manage the economy instead.[1] A key takeaway from that literature is that fiscal instruments—either unconventional (e.g., consumption and labor subsidies) or more conventional (e.g., stimulus checks)—can in principle be used to replicate monetary policy's effects on aggregate demand, thus allowing policymakers to close aggregate output gaps and stabilize inflation even at the ZLB.

While very similar in their effects on aggregate demand and thus the economy as a whole, those instruments may however differ substantially in their distributional incidence. On the one hand, interest rate policy and consumption subsidies are likely to have broad-based effects: everyone tends to benefit, and so such policies tend to be stimulative across the distribution of households.[2] On the other hand, uniform stimulus checks—as seen frequently in the U.S.

The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. We thank Jordi Galí for a helpful discussion of our work.

1. See Eggertsson (2011); Christiano and others (2011); Correia and others (2013); Wolf (2021), among many others.

2. McKay and Wolf (2023a); Bachmann and others (2021).

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

over the past two decades—are much more progressive in their effects on household consumption: by construction, uniform transfers lead to a larger percentage change in income for low-income households—an effect that is only reinforced further by higher marginal propensities to consume at the bottom of the income and wealth distribution. If policymakers have distributional concerns in addition to their usual aggregate objectives, then this heterogeneous distributional incidence will shape optimal policy, including in particular at the ZLB.

The core contribution of this paper is to explore the optimal use of fiscal stabilization policy at the ZLB. Methodologically, doing so requires us to generalize our approach in McKay and Wolf (2023b) to environments subject to a binding ZLB constraint. Substantively, our core takeaway will be that—for canonical ZLB-type shocks, like a tightening in borrowing constraints or a distributional shock concentrated on low-income households—transfer stimulus payments are not just a mere substitute for classical unconstrained monetary policy; rather, they strictly improve upon it.

Environment. We consider a relatively standard heterogeneous-agent (HANK) model, rich enough to be consistent with the broad empirical patterns for the distributional incidence of monetary and fiscal stabilization policies. The economy is subject to a shock that disproportionately reduces the consumption of low-income households—a reduced-form stand-in for tighter borrowing constraints[3] or greater inequality.[4] The shock reduces aggregate demand and thus requires a policy response to stabilize the macroeconomy. We assume that the shock is large enough so that—in the presence of a ZLB on nominal interest rates—monetary policy alone is insufficient to stabilize aggregate demand.

Optimal Policy. We study the optimal policy problem of a policymaker that seeks to avoid cyclical changes in (i) the output gap, (ii) inflation, and (iii) the cross-sectional distribution of consumption. Such a loss function corresponds to a second-order approximation to a social welfare function where the Pareto weights are set so that the steady-state cross-sectional distribution of consumption is optimal.[5] "We refer to a policymaker with this particular loss function as the "Ramsey planner." We ask how such a policymaker uses three available tools—standard interest rate policy, unconventional fiscal policy à

3. For example, Eggertsson and Krugman (2012); Guerrieri and Lorenzoni (2017).
4. For example, Auclert and Rognlie (2018).
5. See McKay and Wolf (2023b).

la Correia and others (2013) (i.e., consumption and labor subsidies), and uniform stimulus checks—to stabilize the economy as well as possible.[6] It will also prove instructive to contrast those results with outcomes for an alternative policymaker that only cares about output and inflation—i.e., a conventional "dual-mandate" policymaker.

Our first results concern the use of conventional nominal interest rate policy. Without the ZLB constraint, a dual-mandate policymaker would lower nominal interest rates as far as needed to perfectly close the output gap and stabilize inflation.[7] Relative to this familiar dual-mandate benchmark, our Ramsey planner would additionally like to stabilize the cross-sectional consumption distribution. However, since interest rate cuts have broad-based stimulative effects across the consumption distribution, they do little to help the planner's distributional goals. Thus, if unconstrained, the Ramsey planner would cut interest rates in a manner similar to the usual dual-mandate outcome. With a binding ZLB, this interest rate cut is of course not feasible, and so now output and inflation gaps arise in addition to the cross-sectional consumption dispersion.

We next consider the use of unconventional fiscal policy—i.e., consumption subsidies to increase consumer demand, and labor taxes to offset the labor supply effects of the consumption subsidy. A dual-mandate policymaker could use these tools to perfectly stabilize aggregate output and inflation even with a binding ZLB, as discussed by Correia and others. We find, however, that such a policy is not particularly useful to the full Ramsey policymaker: unconventional fiscal policy again stimulates consumption across the entire cross-sectional income and wealth distribution and so—just like the infeasible interest rate cut—does little to address the inequality caused by the initial shock.

Finally we turn attention to conventional fiscal policy in the form of uniform stimulus payments. Consistent with the results in Wolf (2021), such uniform stimulus checks can also be used to perfectly stabilize aggregate output and inflation. Importantly, however, stimulus payments do so largely by boosting consumption of low-income households, directly counteracting the distributional incidence

6. Our analysis therefore takes a very particular perspective on the policy problem: the goal is to offset the effects of the business cycle without changing the long-run consumption distribution.

7. This is possible in our economy by the usual "divine coincidence" argument: our economy is subject to a demand shock, and monetary policy can in principle perfectly neutralize that shock's effects on aggregates.

of the original business-cycle shock. This delivers our headline result: for a Ramsey policymaker, at a binding ZLB caused by a distributional shock mostly hitting the poor, stimulus payments do not just substitute for conventional monetary policy—they strictly improve upon it.

Literature. A vast literature has studied macroeconomic stabilization policy at the ZLB—e.g., Krugman (1998); Eggertsson and Woodford (2003); Werning (2011). Our work in particular relates to the subset of that literature that has considered the interaction of inequality and the ZLB. As mentioned briefly above, Eggertsson and Krugman (2012) and Guerrieri and Lorenzoni (2017) study how deleveraging at the bottom end of the income distribution may act as a demand-type shock that pushes the economy towards the ZLB. The interaction between one classic monetary policy remedy to the ZLB—forward guidance—and inequality is analyzed in McKay and others (2016) and Farhi and Werning (2019). Closest to our focus on stimulus checks, Mehrotra (2018) and Wolf (2021) consider fiscal stimulus payments at the ZLB.

Outline. The paper is organized as follows. Section 1 introduces the HANK model and presents the optimal policy problem. The model calibration is described in section 2, and we there also discuss the distributional effects of our three policy instruments. The headline optimal policy results are presented in section 3. Section 4 concludes.

1. Model

For our optimal policy analysis, we rely on a relatively standard sticky-wage HANK model. The only nonstandard model feature is that it includes long-term bonds in addition to the usual short-term bonds. Importantly, the presence of such long-term bonds limits the extent of redistribution that occurs through changes in short-term interest rates, allowing our model to imply a realistic distributional incidence of monetary policy.

Time is discrete and runs forever, $t = 0,1,2,\ldots$ Consistent with our linear-quadratic framework in section 1.6, we will consider linearized perfect-foresight transition sequences. The perfect-foresight approach is in keeping with existing methods for analyzing business-cycle models with occasionally binding constraints on aggregate variables.[8] Throughout this section, boldface denotes time paths

8. See, e.g., Guerrieri and Iacoviello (2015); Holden (2016).

$\left(\text{so, e.g., } \boldsymbol{x} \equiv \left(x_0, x_1, x_2, \ldots\right)'\right)$, bars indicate the model's deterministic steady state $\bar{x}$, and hats denote (log–) deviations from the steady state $\hat{x}$.[9]

1.1 Households

The economy is populated by a unit continuum of ex-ante identical households indexed by $i \in [0,1]$. Household preferences are given by

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\frac{c_{it}^{1-\gamma} - 1}{1-\gamma} - \frac{\ell_{it}^{1+\phi}}{1+\phi} \right], \tag{1}$$

where c_{it} is the consumption of household i and ℓ_{it} is its labor supply.

Household Budget. We begin with the income side of the household budget. Households are endowed with stochastic idiosyncratic labor productivity e_{it} and have total labor earnings of $\left(1-\tau_{\ell,t}\right) w_t e_{it} \ell_{it}$, where w_t is the real wage per effective unit of labor and $\tau_{\ell,t}$ is the tax rate on labor income. We let ζ_{it} be a stochastic event that determines the labor productivity of household i at date t. The event ζ_{it} itself follows a stationary Markov process, and a canonical heterogeneous agent model would set $e_{it} = \zeta_{it}$. We will instead assume that there is a function Φ that maps ζ_{it} to e_{it},

$$e_{it} = \Phi\left(\zeta_{it}, d_t\right) \text{ with } \int e_{it} di = 1 \forall d_t.$$

This mapping depends on an exogenous distributional shock, d_t, which affects the dispersion of individual labor productivities. For the analysis in section 3, this shock d_t will be the shock that moves the economy towards the ZLB—a distributional shock that mostly affects low-income households.[10] As we will describe further below, household labor supply is determined by a labor market union, so hours worked ℓ_{it} are taken as given by an individual household. Households furthermore receive a time-varying lump-sum transfer $\tau_{x,t} + \tau_{e,t} e_{it}$. Here, the first component of the transfer, $\tau_{x,t}$, is the same for all households and will be manipulated as part of the optimal policy problem. We will thus refer to it as the "exogenous" component of fiscal

9. To be precise, we use log deviations for the variables $\{y, c, \ell, 1+r, 1+i, \varepsilon\}$ and level deviations for the variables $\{\pi, \tau_x, \tau_e, \tau_\ell, \tau_c, \mathcal{R}\}$.

10. While literally modeled here as a distributional shock, it is well-known that such a shock will have very similar effects to a tightening in borrowing constraints, as e.g., considered in Guerrieri and Lorenzoni (2017).

transfers, hence the subscript x, or "fiscal stimulus payment". The second component, $\tau_{e,t} e_{it}$, is the "endogenous" part of taxes, adjusting slowly over time to make sure that long-run fiscal budget balance is maintained.

Households use their income to consume and save. To consider unconventional fiscal stabilization policy, as in Correia and others (2013), we allow for time-varying consumption subsidies $\tau_{c,t}$. Following Correia and others, we furthermore assume that, for any given $\tau_{c,t}$, labor taxes $\tau_{\ell,t}$ adjust to offset the labor supply impact of the consumption subsidy. We thus treat $\tau_{c,t}$ as the single "unconventional" fiscal policy instrument.

Finally, households save through financial assets with expected real return r_t between periods t and $t + 1$, subject to an exogenous no-borrowing constraint. As we discuss later, households can save in multiple assets, with their returns linked by a no-arbitrage condition. In our perfect-foresight economy, all assets will earn exactly the same common realized return r_t at each date $t = 1,2....$ At date 0, however, the realized return may deviate from the ex-ante expected return, and in particular it may depend on the household's date-0 asset composition. For simplicity we assume that portfolios have the same composition everywhere in the cross-section of households, and we let $\mathcal{R}_t$ denote the common date-t revaluation factor of household portfolios—which, again, will only be nonzero at date 0.[11]

Putting all the pieces together, the household budget constraint is

$$\frac{1}{1+r_t} a_{it} + \left(1-\tau_{c,t}\right) c_{it} = a_{it-1}\left(1+\mathcal{R}_t\right) + \left(1-\tau_{\ell,t}\right) e_{it} w_t \ell_{it} + \tau_{x,t} + \tau_{e,t} e_{it} \qquad (2)$$

where a_{it} is the expected value of assets entering period $t + 1$.

Aggregate Consumption Function. The consumption-savings problem of an individual household i is to choose consumption $\boldsymbol{c}_j$ and savings $\boldsymbol{a}_i$ to maximize (1) subject to (2).

The solution is thus a mapping from paths of real wages $\boldsymbol{w}$, hours worked $\boldsymbol{\ell}$ expected real returns $\boldsymbol{r}$, transfers $\boldsymbol{\tau}_x$ and $\boldsymbol{\tau}_e$, prices $\boldsymbol{p}$, shocks $\boldsymbol{d}$, and date-0 revaluation effects $\mathcal{R}_0$ to that household's consumption $\boldsymbol{c}_i$.

11. We allow for heterogeneous household portfolios in McKay and Wolf (2023b). The conclusions of this paper are not affected by considering such a more complicated model variant.

Aggregating consumption decisions across all households, we thus obtain an aggregate consumption function $\mathcal{C}$

$$\boldsymbol{c} = \mathcal{C}\left(\boldsymbol{w}, \ell, \boldsymbol{r}, \tau_x, \tau_e, \tau_c, \tau_\ell, \boldsymbol{d}, \mathcal{R}_0\right) \tag{3}$$

Linearizing this consumption function around the deterministic steady state yields

$$\hat{\boldsymbol{c}} = \mathcal{C}_y\left(\hat{\boldsymbol{w}} + \hat{\ell}\right) + \mathcal{C}_r \hat{\boldsymbol{r}} + \mathcal{C}_{\tau x}\hat{\tau}_x + \mathcal{C}_{\tau e}\hat{\tau}_e + \mathcal{C}_{\tau c}\hat{\tau}_c + \mathcal{C}_{\tau \ell}\hat{\tau}_\ell + \mathcal{C}_d \hat{\boldsymbol{d}} + \mathcal{C}_{\mathcal{R}}\hat{R}_0, \tag{4}$$

where the derivative matrices $\mathcal{C}$ are evaluated at steady state and we have made use of the fact that it is only the product $w_t \ell_{it}$ that is relevant to the household.

1.2 Technology, Unions, and Firms

Labor supply is intermediated by a unit continuum of labor unions, and a competitive aggregate producer then packages union labor supply to produce the final good. Since this production model block is standard, we only state and briefly discuss the key relations here, with a detailed discussion relegated to Appendix A.

Union k demands ℓ_{ikt} units of labor from household i. The final good is sold at nominal price p_t and produced by aggregating the labor supply of all individual unions k, denoted $\ell_{kt} \equiv \int_0^1 e_{it}\ell_{ikt} di$. The aggregate production function takes a standard constant elasticity form, with elasticity of substitution between varieties ε. All unions satisfy labor demand by rationing labor equally across all households. This rationing rule together with marginal cost pricing ($W_t = p_t$) for the competitive producer implies that $e_{it}\ell_{it}\dfrac{W_t}{p_t} = e_{it}y_t$ for all i.

Each union sets its nominal wage in the usual Calvo fashion, with a probability 1 – θ of updating the wage each period. As usual, unions select their wages upon reset based on current and future marginal rates of substitution between leisure and consumption among its household members. Given that everyone supplies an equal amount of hours worked, and with our household preferences additively separable, it follows that all households share a common marginal disutility of labor. The marginal utility of consumption, however, need not be equalized. Following McKay and Wolf (2023b)—and similar to Auclert and others (2021)—, we assume that the union evaluates the benefits of higher after-tax income using the marginal utility of average consumption $\left(c_t^{-\gamma}\right)$ rather than the average of individual

household marginal utilities $\left(\int_0^1 c_{it}^{-\gamma} di\right)$. This assumption eliminates the impact of inequality on the supply side of the economy, and so we overall arrive at the following standard linearized perfect-foresight New Keynesian Phillips Curve (NKPC):

$$\hat{\pi}_t = \kappa \hat{y}_t + \beta \hat{\pi}_{t+1} \tag{5}$$

where $\kappa \equiv \left(\phi + \gamma\right)\frac{\left(1-\theta\right)\left(1-\beta\theta\right)}{\theta}$. In our derivation of (5), we allow for a (time-invariant) subsidy on union labor hiring, financed with lump-sum taxes also levied on the unions; this subsidy will yield the efficiency of the deterministic steady state needed to specify our optimal policy problem in a form consistent with a linear-quadratic analysis, as in Woodford (2003).

1.3 Asset Structure

There are two different assets in the economy: a short-term, nominal bond in zero net supply, and a long-term bond in positive net supply. By a no-arbitrage condition, both assets will provide the same expected returns along equilibrium transition paths (except possibly at $t = 0$), thus allowing us to consider a single asset in the household budget constraint (2). The realized return at date 0, however, will generally differ between the two assets. As mentioned above, the purpose of the long-term bond is to provide a more realistic description of the passthrough of monetary policy to household interest payments.

A unit of the short-term bond purchased at time t then returns $\frac{1+i_t}{1+\pi_{t+1}}$ units of the final good at time $t + 1$. For the long-term bond, at time t, households can purchase a unit of the bond for a real price of q_t (i.e., denominated in goods); at time $t + 1$, the household then receives a real "coupon" of $\left(\bar{r}+\delta\right)\left(1+\pi_{t+1}\right)^{-1}$ and furthermore retains a fraction $\left(1-\delta\right)\left(1+\pi_{t+1}\right)^{-1}$ of the initial asset position, now valued at $\left(1-\delta\right)\left(1+\pi_{t+1}\right)^{-1} q_{t+1}$ in units of goods. Note that the parameter δ controls the duration of the bond, with lower values of δ corresponding to higher duration. The coupon scaling factor ($\bar{r} + \delta$) is chosen to normalize the steady-state price of the bond to one. Finally, the presence of the inflation terms reflects the fact that the bond is nominal, so inflation reduces the real value of the current and future coupons, and so reduces the real value of the bond position.

Overall, it follows that the price of the long-term bond satisfies

$$q_t = \frac{(\bar{r}+\delta)(1+\pi_{t+1})^{-1} + (1-\delta)(1+\pi_{t+1})^{-1} q_{t+1}}{1+r_t}, \tag{6}$$

where r_t is the real interest rate between t and $t+1$. Real returns are furthermore linked to returns on the short-term rate via the standard Fisher relation

$$1 + r_t = \frac{1+i_t}{1+\pi_{t+1}}. \tag{7}$$

At date $t = 0$, the realized return on a household's portfolio will depend on the composition of its portfolio between the two assets. We assume that there are no existing gross positions in the short-term bond, so time-0 realized returns are simply those on the long-term bond, which implies that

$$1 + \mathcal{R}_0 = \frac{(\bar{r}+\delta)(1+\pi_0)^{-1} + (1-\delta)(1+\pi_0)^{-1} q_0}{1+\bar{r}}. \tag{8}$$

Note that this relation expresses the scaling factor $1+\mathcal{R}_0$ as the ratio of the actual realized return on the long-term bond (i.e., the numerator) to the expected return (i.e., the steady-state real rate in the denominator).

1.4 Government

The government collects tax revenue, pays out lump-sum transfers, sets the nominal rate on the short-term bond, and issues positive quantities of the long-term bond. Letting $a_t^g(1+\mathcal{R}_0)$ denote the value of claims on the government entering period t (inclusive of returns), the government budget constraint becomes

$$\frac{a_{t+1}^g}{1+r_t} = a_t^g (1+\mathcal{R}_t) + \tau_{x,t} + \tau_{e,t} + \tau_{c,t} c_t - \tau_{\ell,t} y_t. \tag{9}$$

Note that, when news arrives, the claims on the government are revalued in the exactly same way as previously discussed for the household sector's assets.

We consider the nominal rate of interest, i_t, the exogenous component of transfers, $\tau_{x,t}$, and the consumption subsidy, $\tau_{c,t}$, as the independent policy instruments of the government, used for business-cycle stabilization policy. The time-varying labor tax furthermore adjusts automatically so that the net consumption benefit of an hour of work is unaffected by the consumption subsidy, requiring that $1 - \tau_{\ell,t}$ be proportional to $1 - \tau_{c,t}$ at all times. Since all three policy instruments will generally have budgetary implications, it remains to specify how long-term budget balance is ensured. We will assume that the endogenous component of transfers $\tau_{e,t}$ adjusts gradually according to the rule

$$\tau_{e,t} = \left(\bar{r} + \delta\right)\left(a_t^g - \bar{a}^g\right) \tag{10}$$

where $\bar{a}^g$ is the real, steady-state value of government debt.

1.5 Equilibrium

Given paths of exogenous shocks $\{d_t\}_{t=0}^{\infty}$ and policy instruments $\left\{i_t, \tau_{x,t}, \tau_{c,t}\right\}_{t=0}^{\infty}$, a perfect-foresight equilibrium of our linearized economy is a set of sequences of endogenous aggregate variables $\left\{a_t^g, c_t, y_t, q_t, a_t, \pi_t, r_t, \ell_t, \tau_{e,t}, \tau_{\ell,t}\right\}_{t=0}^{\infty}$ and $\mathcal{R}_0$ that satisfy the following conditions:

1. The path of aggregate consumption $\{c_t\}_{t=0}^{\infty}$ is consistent with the linearized aggregate consumption function (4), and the path of household asset holdings $\{a_t\}_{t=0}^{\infty}$ is consistent with the budget constraint (2), aggregated across households.

2. The paths of $\{\ell_t, y_t\}_{t=0}^{\infty}$ satisfy the linearized aggregate production function $y_t = \ell_t$.

3. The paths $\{\pi_t, y_t\}_{t=0}^{\infty}$ are consistent with the Phillips curve (5).

4. The evolution of government debt a_t^g, the endogenous component of transfers $\tau_{e,t}$, and the labor income tax $\tau_{\ell,t}$ are consistent with the budget constraint (9), the law of motion (10), and the requirement that $\left(1 - \tau_{\ell,t}\right) / \left(1 - \tau_{c,t}\right) = \left(1 - \bar{\tau}_{\ell}\right) / \left(1 - \bar{\tau}_{c}\right)$.

5. The asset prices $\{q_t, r_t\}_{t=0}^{\infty}$ satisfy (6) and (7), and the revaluation effect $\mathcal{R}_0$ satisfies (8).

6. The output and asset markets clear, so $y_t = c_t$ and $a_t = a_t^g$.

Note that this definition of equilibrium takes the policy instruments $\left\{i_t, \tau_{x,t}, \tau_{c,t}\right\}_{t=0}^{\infty}$ as given. The paths for these will be determined by solving the optimal policy problem.

1.6 The Policy Problem

We consider a policymaker who wishes to both stabilize the aggregate economy and offset cyclical changes in consumption inequality.

Objective Function. To understand our formulation of the policymaker's objective, we begin by noting that households in our model are ex ante identical and only differ ex post due to different realizations of their idiosyncratic shocks. Households can therefore be indexed by the history of idiosyncratic shocks they have experienced, denoted $\zeta_i^t \equiv (\zeta_{it}, \zeta_{it-1}, \ldots)$. As the shocks ζ_{it} are drawn from a stationary process, the distribution of such histories is itself stationary. With this notation established, we write the policymaker objective as

$$\mathcal{L} \equiv \sum_{t=0}^{\infty} \beta^t \left[\hat{\pi}_t^2 + \frac{\kappa}{\varepsilon} \hat{y}_t^2 + \frac{\kappa}{\varepsilon} \frac{\gamma}{\gamma + \phi} \int \frac{\hat{\omega}_t(\zeta)^2}{\bar{\omega}(\zeta)} d\Gamma(\zeta) \right] \tag{11}$$

where ζ is an infinite history of idiosyncratic shocks, $\omega_t\left(\zeta_i^t\right) \equiv c_{it} / c_t$ is the consumption share of an individual with that history at date, t and Γ is the stationary distribution of household idiosyncratic histories. In McKay and Wolf (2023b), working with a very similar model, we derive the loss function (11) as a second-order approximation to a particular social welfare function—one that attaches Pareto weights to the welfare of individual households in exactly the right way to ensure that the policymaker does not wish to deviate from the steady-state distribution of household consumption.

Next, since household consumption and thus in particular the consumption shares $\omega_t\left(\zeta_i^t\right)$ are a function solely of the aggregate variables that influence the household's consumption-savings problem, it is straightforward to see that we can re-write (11) as[12]

12. A detailed argument—including details on how to compute Q—are provided in McKay and Wolf (2023b).

$$\mathcal{L} = \hat{x}' Q \hat{x},$$

where Q is a symmetric matrix and $\boldsymbol{x}$ stacks the paths of the various endogenous and exogenous variables entering the consumer problem,

$$x = \left(\boldsymbol{y}' \;\; \boldsymbol{\pi}' \;\; \boldsymbol{r}' \;\; \tau_e' \;\; \mathcal{R}' \;\; \boldsymbol{i}' \;\; \tau_x' \;\; \tau_c' \;\; \boldsymbol{d}' \right)'.$$

Constraints. We now turn to the constraints on the policy problem. Using sequence-space methods, we can compactly express the equilibrium of this economy as

$$\hat{\boldsymbol{x}} = \underbrace{\Theta_d \hat{\boldsymbol{d}}}_{\equiv \bar{\boldsymbol{x}}} + \Theta_i \hat{\boldsymbol{i}} + \Theta_x \hat{\boldsymbol{\tau}}_x + \Theta_c \hat{\boldsymbol{\tau}}_c \tag{12}$$

where the Θ's are general equilibrium impulse response matrices to the shock $\hat{d}$ and the policy instruments $\left\{ \hat{\boldsymbol{i}}_b, \hat{\tau}_x, \hat{\tau}_c \right\}$; and $\bar{\boldsymbol{x}}$ denotes outcomes if the policy instruments were not adjusted in response to the shock $\hat{\boldsymbol{d}}$.[13]

In addition, policy is constrained by a lower bound on the nominal interest rate. As we work with the model in deviations from a zero-inflation steady state, we express the ZLB constraint as $\hat{i} \geq \underline{i} = -\bar{r}$. We impose no constraints on the other two policy instruments.

Policy Problem. We can express the policy problem compactly by defining $\boldsymbol{p} \equiv \left(\hat{\boldsymbol{i}}, \hat{\boldsymbol{\tau}}_x', \hat{\boldsymbol{\tau}}_c' \right)'$ as the vector of policy instruments, letting $\underline{\boldsymbol{p}}$ denote the lower bounds on the instruments (which are $-\infty$ for the two fiscal instruments), and finally defining $\Theta_p \equiv \left(\Theta_i, \Theta_x, \Theta_c \right)$. We then solve the problem

$$\min_{p, \hat{x}} \hat{x}' Q \hat{x} \tag{13}$$

subject to

$$\hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \Theta_p \boldsymbol{p} \tag{14}$$

$$\boldsymbol{p} \geq \underline{\boldsymbol{p}}. \tag{15}$$

13. In practice, to compute the Θ's, we truncate the transition paths at some large (but finite) horizon T and assume the economy has returned to steady state by this time. As there are nine variables in x, each $\Theta \bullet$ is a $9T \times T$ matrix. See McKay and Wolf (2023b) for a discussion of how the Θ.'s are defined uniquely through policy shocks to a given baseline, determinacy-inducing monetary policy rule.

The policy problem therefore fits into a standard quadratic programming form.[14]

Finally, for reference, we will also find it useful to solve a simplified version of this problem for a dual-mandate policymaker—i.e., a policymaker with preferences as in (11), but ignoring the inequality-related term. This problem fits into (13) for a different (simpler) Q.

2. Model Parameterization

This section presents the model parameterization used for our analysis in section 3. We first discuss the calibration strategy in section 2.1 and then in section 2.2 focus on the model feature that matters most for our later results—the distributional incidence of policy.

2.1. Calibration Strategy

We provide a relatively brief sketch of our calibration strategy. A summary of the calibration is provided in table 1.

Table 1. Model Calibration

Parameter	*Description*	*Value*	*Calibration target*
γ	Relative risk aversion	1.2	Monetary shock effects
ϕ	Frisch elasticity	1	Standard
β	Discount factor	0.987	Asset market clearing
κ	Phillips curve slope	0.022	Monetary shock effects
ε	Labor Substitutability	6	Basu & Fernald (1997)
δ	Long-term bond duration	0.025	10-year maturity

Source: Authors' calculations.

14. For our numerical applications, we have found that guessing and verifying a horizon n over which the ZLB is binding to be a reliable computational strategy. In particular, given a candidate value of n, we first solve the simpler sub-problem in which the constraint binds as an equality constraint for n periods. We then verify the guess ex post. Appendix B provides details.

Households. We begin with preferences. We set the coefficient of relative risk aversion to 1.2, allowing us to match the empirically measured sensitivity of aggregate consumption to monetary policy shocks. The elasticity of labor supply is set to a standard value of 1. Next, the discount factor β is calibrated to match the steady-state level of aggregate assets in the economy. We set this asset supply to 1.4 times GDP, as in McKay and others (2016), with the implicit interpretation that assets in our model correspond to liquid assets. Turning to the idiosyncratic income process, we associate ζ_{it} with the persistent AR(1) process in the estimates of Floden and Lindé (2001) adapted to a quarterly frequency, which results in a persistence of 0.978 and an innovation variance of 0.0114. The function Φ is then given by

$$\log e_{it} = \log \zeta_{it} (1 + d_t) - \bar{e}_t,$$

where d_t is the exogenous distributional shock with $\overline{d} = 0$, and $\bar{e}_t$ is a normalization constant so that the cross-sectional average of e_{it} is always 1. Notice that an increase in d_t amplifies the dispersion in labor productivity by amplifying the differences in ζ_{it}—that is, it is an inequality shock that redistributes from the poor to the rich.

Assets and Government. We assume that households save in long-term bonds with a maturity of ten years, which corresponds to $\delta = 0.025$. The steady-state real interest rate is set to 2.4 percent per annum. Steady-state consumption subsidies are zero, and the steady-state tax rate on labor income $\overline{\tau}_\ell$ is then determined endogenously to satisfy the government budget constraint.

Supply Block. We calibrate the slope of the Phillips curve to 0.022 in order to match the magnitude of the response of inflation to a monetary policy shock.[15] Finally, the elasticity of substitution between varieties of intermediate goods is set to 6, following Basu and Fernald (1997).

2.2 The Distributional Implications of Policy

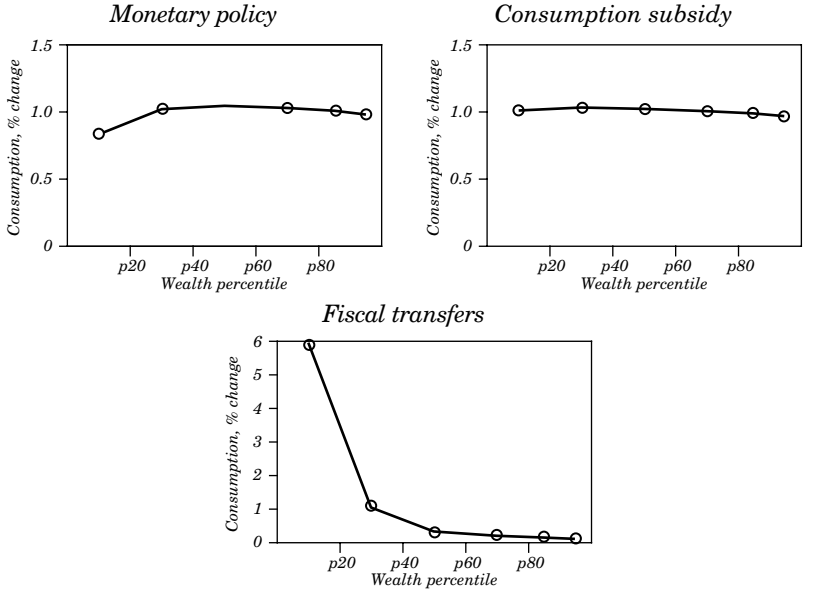
As established in prior work,[16] the three policy instruments available to our policymaker are equivalent in their effects on macroeconomic aggregates—they all equally flexibly perturb aggregate net excess demand. For optimal Ramsey policy, however, their distributional effects

15. See McKay and Wolf (2023b).
16. Correia and others (2013); Wolf (2021).

also matter. We here show that stimulus payments have very different distributional incidence from standard monetary and unconventional fiscal policy. The results are displayed in figure 1.

Monetary Policy. The top-left panel of figure 1 reveals that monetary policy in our model is broadly distributionally neutral: an interest rate cut stimulates consumption across the entire wealth distribution. This feature of our model is broadly consistent with prior empirical evidence on the heterogeneous effects of monetary policy, as for example reviewed in McKay and Wolf (2023a). We conclude that monetary policy is unlikely to have a material impact on the inequality term in the policymaker loss function (11).[17]

Figure 1. Effects of Policy Instruments on Consumption Inequality



Source: Authors' calculations.

Note: The figures show the initial ($t = 0$) change in consumption following a policy stimulus. In the top-left panel, we consider an expansionary monetary policy shock that induces a one-percent increase in aggregate output on impact. Thereafter, aggregate output decays with a persistence of 0.7. In the top-right panel, we consider an unconventional fiscal policy and, in the bottom panel, we study transfer payments, with both chosen to induce a response of output of the same size and persistence as the monetary policy shock.

17. More precisely, if monetary policy were to move all households exactly up and down in tandem (and at all horizons), then monetary policy would not affect consumption shares at all, and so Ramsey policy would be identical to dual-mandate policy.

Unconventional Fiscal Policy. The top-right panel of figure 1 shows the response of consumption to a consumption subsidy, where that subsidy is chosen to replicate the aggregate output effects of the conventional monetary policy studied previously. We see that the effects on inequality are very similar to the nominal interest rate cut: households across the entire net worth distribution increase their consumption in response to the consumption subsidy. Empirically, this feature of our model is consistent with prior evidence.[18] Theoretically, the close agreement between the top-left and top-right panels follows the arguments in Seidl and Seyrich (2023).

Stimulus Checks. Finally, the bottom panel of figure 1 shows how the cross-section of consumption responds to a stimulus payment policy. The stimulative effects on consumption are now not broad-based: the consumption of the poor is disproportionately stimulated, mainly reflecting (i) the fact that the initial transfer is a larger share of their steady-state income, and (ii) their higher marginal propensities to consume. At the top end of the distribution, consumption rises mainly because, in general equilibrium, higher inflation leads to a decline in real rates, thus inducing intertemporal substitution. The differences in distributional incidence across the three policy instruments documented in figure 1 will be key to understand our optimal policy results in the next section.

3. Optimal Policy Results

This section presents our headline results on optimal stabilization policy at the ZLB. We proceed in three steps, with one subsection for each of the three policy instruments: monetary policy in section 3.1, unconventional fiscal policy in section 3.2, and fiscal stimulus payments in section 3.3. Throughout we consider an economy subject to a contractionary distributional demand shock d_t, where that shock is large enough so that the ZLB constraint becomes binding for conventional monetary policy.

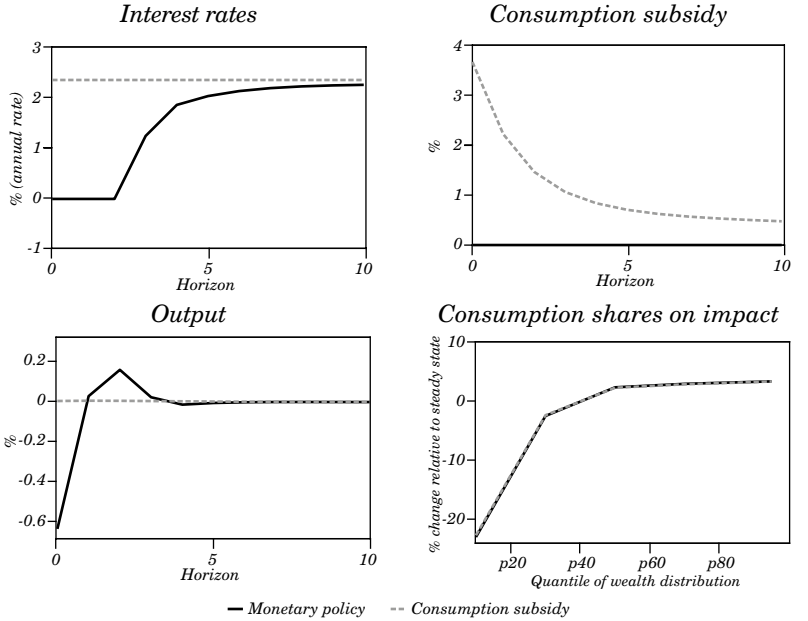
3.1 Monetary Policy

We begin with conventional nominal interest rate policy. Results for optimal Ramsey monetary policy subject to the ZLB are displayed

18. Bachmann and others (2021).

as the solid lines in figure 2. By construction, the inequality shock is sufficiently large so that, given the ZLB constraint, monetary policy is unable to stabilize aggregate output. We see that nominal interest rates are cut as much as possible and remain at zero for a couple of quarters (top right), leading to an initial decline and then an overshoot in output (bottom left). The overshooting of output reflects the usual "low-for-longer" logic of optimal monetary policy at the ZLB.[19] Perhaps most importantly, monetary policy fails to counteract the distributional implications of the shock—consumption drops sharply for low-income households while remaining relatively stable for the rich (bottom right).

Figure 2. Optimal Monetary and Unconventional Fiscal Policies



Source: Authors' calculations.

Note: Impulse responses of nominal interest rates, the consumption subsidy, output, and consumption shares (at $t = 0$) to the inequality shock under optimal (Ramsey) monetary policy subject to the ZLB (solid line) and unconventional fiscal policy (dashed).

19. See, e.g., Eggertsson and Woodford (2003).

In Appendix C we show what happens, first, in the counterfactual absence of a ZLB constraint, and second, under optimal dual-mandate monetary policy. Naturally, optimal Ramsey policy without the ZLB constraint would lower interest rates more aggressively, thus stabilizing output. Importantly, however, this additional interest rate cut does little to counteract the distributional implications of the initial shock, so consumption shares still decline significantly at the bottom end of the income and wealth distribution. This reflects the same logic as figure 1—monetary policy has small effects on the shape of the consumption distribution. In light of this, it is furthermore also not surprising that the optimal Ramsey policy looks rather similar to optimal dual-mandate policy. As monetary policy has relatively little power to moderate the effects of the initial demand shock on inequality, even the Ramsey policy is essentially only concerned with aggregate stabilization.

3.2 Unconventional Fiscal Policy

We next consider unconventional fiscal policy, as analyzed in Correia and others (2013). Results for the optimal Ramsey unconventional fiscal policy are displayed as the dashed lines in figure 2. The top-right panel shows that, as expected, the policymaker finds it optimal to subsidize consumption, thus spurring aggregate demand and almost perfectly stabilizing the macro-economy as a whole. However, as we see in the bottom-right panel, this policy does relatively little to offset the distributional tilt of the original inequality shock, with consumption shares of low-income households still dropping substantially. This is again exactly what was expected in light of figure 1: unconventional fiscal policy has broad-based stimulative effects, and so—just like conventional monetary policy—it is relatively ill-suited to deal with explicitly distributional shocks.

We note that our numerical findings in figure 2 are consistent with analytical results in Seidl and Seyrich (2023). These authors show that, for a particular mix of unconventional fiscal policy and government debt issuance, household-by-household outcomes are exactly identical to monetary stimulus. In our case the equivalence is not exact (as we consider a somewhat different debt issuance policy), but the broad intuition remains: both interest rate and consumption subsidy policy affect households in similar ways, and in particular—at least in our model calibration—those effects are rather uniform cross-sectionally.

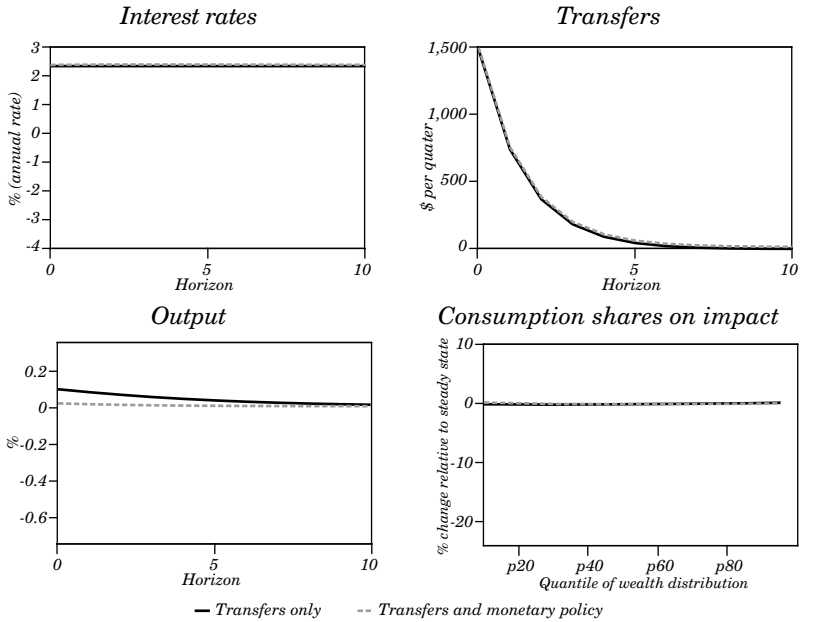
3.3 Stimulus Payments

Finally we consider our main policy alternative: uniform stimulus payments. Results for optimal Ramsey stimulus transfer policy are displayed as solid lines in figure 3. We see that the policymaker decides to optimally pay out a relatively large positive transfer (top right), thus almost perfectly stabilizing—in fact even slightly overshooting—aggregate output (bottom left). Crucially, however, and differently from the results for monetary and unconventional fiscal policy in figure 2, this stimulus to demand at the same time also stabilizes the cross-sectional consumption distribution (bottom right). Intuitively, stimulus checks increase aggregate demand precisely by boosting the spending of those households that were hardest hit by the initial contractionary demand shock. From the point of view of our Ramsey planner with loss function (11), such transfer payments are thus the ideal tool: with one instrument, they can stabilize all three terms of their loss function. Indeed, as we show in Appendix C, the Ramsey loss for a policymaker that only uses stimulus checks is an order of magnitude lower than the loss for a counterfactual monetary policy Ramsey planner, even without the ZLB constraint. Transfers thus do not merely substitute for—but in fact very much improve upon—conventional monetary policy, at least in response to the kind of distributional demand shock that we consider here.[20]

For comparison, the dashed lines in figure 3 display optimal joint monetary-fiscal policy, which sets both stimulus payments τ_x as well as interest rates i optimally. To stabilize cross-sectional consumption shares, the stimulus payment-only policy induced a slightly excessive aggregate demand boom, overheating the economy. The optimal joint policy thus features a very mild increase in interest rates, thus closing the output gap while also keeping consumption shares stable. Overall, however, the difference in loss between transfer-only and joint optimal policy is minimal, in particular compared to the large losses that occurred under monetary-only or unconventional fiscal policy.

20. Naturally, for more broad-based initial demand shocks, conventional monetary policy or unconventional fiscal policy would again emerge as more suitable tools. However, we view explicitly distributional shocks as relevant empirically: tightening borrowing constraints were particularly important in the Great Recession, and the COVID-19 recession mostly impacted the bottom of the income distribution.

Figure 3. Optimal Stimulus Payments



Source: Authors' calculations.

Note: Impulse responses of nominal interest, uniform stimulus payments, output, and consumption shares (at $t = 0$) to the inequality shock under optimal (Ramsey) stimulus payment policy (solid) and optimal joint monetary-transfer policy (dashed).

4. Conclusion

How should policymakers stabilize the macro-economy when conventional monetary policy is constrained by a zero lower bound on nominal rates? In particular, what policy options are most attractive if—as seems empirically plausible—the economy was subject to a negative shock that mostly impacted low-income households? Building on our prior work in McKay and Wolf (2023b), we here tried to answer those questions through the lens of a textbook heterogeneous-household model. Our headline result was that stimulus checks are more than a substitute for conventional monetary policy; in fact, since they are much more well-adapted to the distributional incidence of the shock, they are strictly preferable as a tool for cyclical stabilization, and so the ZLB poses no meaningful constraint on the policymaker.

We emphasize two important qualifiers of our results. First, our findings are necessarily sensitive to a key feature of our model—the distributional neutrality of monetary policy interventions. While this model feature is consistent with prior work,[21] further empirical investigation would be welcome. Second, our conclusions apply to particular, empirically relevant kinds of demand shocks—those mostly affecting low-income households. Conclusions may be different for other types of demand disturbances, e.g., those directly affecting firms rather than households.

21. See, e.g., McKay and Wolf (2023a) and the references therein.

References

Auclert, A. and M. Rognlie. 2018. "Inequality and Aggregate Demand." Technical Report, National Bureau of Economic Research.

Auclert, A., M. Rognlie, M. Souchier, and L. Straub. 2021. "Exchange Rates and Monetary Policy with Heterogeneous Agents: Sizing up the Real Income Channel." Technical Report, National Bureau of Economic Research.

Bachmann, R., B. Born, O. Goldfayn-Frank, G. Kocharkov, R. Luetticke, and M. Weber. 2021. "A Temporary VAT Cut as Unconventional Fiscal Policy." Technical Report, National Bureau of Economic Research.

Basu, S. and J.G. Fernald. 1997. "Returns to Scale in U.S. Production: Estimates and Implications." *Journal of Political Economy* 105(2): 249–83.

Christiano, L., M. Eichenbaum, and S. Rebelo. 2011. "When is the Government Spending Multiplier Large?" *Journal of Political Economy* 119(1): 78–121.

Correia, I., E. Farhi, J.P. Nicolini, and P. Teles, P. 2013. "Unconventional Fiscal Policy at the Zero Bound." *American Economic Review* 103(4): 1172–211.

Eggertsson, G.B. 2011. "What Fiscal Policy is Effective at Zero Interest Rates?" *NBER Macroeconomics Annual* 25: 59–112.

Eggertsson, G.B. and P. Krugman. 2012. "Debt, Deleveraging, and the Liquidity Trap: A Fisher-Minsky-Koo Approach." *Quarterly Journal of Economics* 127(3): 1469–513.

Eggertsson, G.B. and M. Woodford. 2003. "Zero Bound on Interest Rates and Optimal Monetary Policy." *Brookings Papers on Economic Activity* (1): 139–233.

Farhi, E. and I. Werning. 2019. "Monetary Policy, Bounded Rationality, and Incomplete Markets." *American Economic Review* 109(11): 3887–928.

Floden, M. and J. Lindé. 2001. "Idiosyncratic Risk in the United States and Sweden: Is There a Role for Government Insurance?" *Review of Economic Dynamics* 4(2): 406–37.

Guerrieri, L. and M. Iacoviello. 2015. "Occbin: A Toolkit for Solving Dynamic Models with Occasionally Binding Constraints Easily." *Journal of Monetary Economics* 70: 22–38.

Guerrieri, V. and G. Lorenzoni. 2017. "Credit Crises, Precautionary Savings, and the Liquidity Trap." *Quarterly Journal of Economics* 132(3): 1427–67.

Holden, T.D. 2016. "Computation of Solutions to Dynamic Models with Occasionally Binding Constraints." EconStor Preprints No. 144569.

Krugman, P.R. 1998. "It's Baaack!: Japan's Slump and the Return of the Liquidity Trap." *Brookings Papers on Economic Activity* 29(2): 137–206.

McKay, A., E. Nakamura, and J. Steinsson. 2016. "The Power of Forward Guidance Revisited." *American Economic Review* 106(10): 3133–58.

McKay, A. and C.K. Wolf. 2023a. "Monetary Policy and Inequality." *Journal of Economic Perspectives* 37(1): 121–44.

McKay, A. and C.K. Wolf. 2023b. "Optimal Policy Rules in HANK." Working Paper, FRB Minneapolis.

Mehrotra, N. R. 2018. "Fiscal Policy Stabilization: Purchases or Transfers?" *International Journal of Central Banking* 14(2): 1–50.

Seidl, H. and F. Seyrich. 2023. "Unconventional Fiscal Policy in HANK." DIW Berlin Discussion Paper.

Werning, I. 2011. "Managing a Liquidity Trap: Monetary and Fiscal Policy." Technical Report, National Bureau of Economic Research.

Wolf, C.K. 2021. "Interest Rate Cuts vs. Stimulus Payments: An Equivalence Result." Working Paper No. 29193, National Bureau of Economic Research.

Woodford, M. 2003. *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton, NJ: Princeton University Press.

APPENDICES

Appendix A. Supply Side and Phillips Curve Derivation

We here provide further details for the production side of our economy, as sketched in section 1.2. We begin by specifying the details of the economy's production technology and then derive our Phillips curve (5).

Technology. A unit continuum of unions, indexed by $k \in [0,1]$, differentiates labor into distinct tasks. Union k aggregates efficiency units into the union-specific task $\ell_{kt} = \int e_{it} \ell_{ikt} di$, where ℓ_{ikt} are the hours worked supplied by household i to union k. A competitive final goods producer then packages these tasks using the technology

$$y_t = \left(\int_k \ell_{kt}^{\frac{\varepsilon-1}{\varepsilon}} dk \right)^{\frac{\varepsilon}{\varepsilon-1}}.$$

The price index of a unit of the overall labor aggregate is

$$W_t = \left(\int W_{kt}^{1-\varepsilon} dk \right)^{1/(1-\varepsilon)},$$

where W_{kt} is the price of the task supplied by union k. Marginal cost pricing by final goods producers requires $p_t = W_t$. The resulting demand for labor from union k is

$$\ell_{kt} = \left(\frac{W_{kt}}{W_t} \right)^{-\varepsilon} y_t. \tag{A.1}$$

Integrating both sides across k yields the aggregate production

$$y_t \int \left(\frac{W_{kt}}{W_t} \right)^{-\varepsilon} dk = \ell_t,$$

with ℓ_t denoting total effective hours supplied by households and the integral term capturing the efficiency losses due to price dispersion. The dispersion term disappears in a first-order approximation to the dynamics of the model.

From Union Problem to NKPC. We assume that union wage payments to households are subsidized at gross rate $\frac{\varepsilon}{(\varepsilon-1)(1-\bar{\tau}_\ell)}$,

where $\bar{\tau}_\ell$ is the steady-state labor income tax. The union's problem is to choose the reset wage W^* and ℓ_{kt} to maximize

$$\sum_{s\geq 0}\beta^s\theta^s\left[u_c\left(c_{t+s}\right)\frac{1-\tau_{\ell,t}}{1-\tau_{c,t}}\frac{\varepsilon}{(\varepsilon-1)(1-\bar{\tau}_\ell)}\frac{W^*}{p_{t+s}}\ell_{kt}-\nu_\ell\left(\ell_{t+s}\right)\ell_{kt}\right]$$

subject to (A.1) and taking c_{t+s} and ℓ_{t+s} as given (since the individual labor union is atomistic). The first-order condition is

$$\sum_{s\geq 0}\beta^s\theta^s\nu_\ell\left(\ell_{t+s}\right)y_{t+s}\epsilon_{t+s}\left(\frac{p_{t+s}}{p_t}\right)^{\varepsilon}=\frac{\varepsilon}{\varepsilon-1}\sum_{s\geq 0}\beta^s\theta^s u_c\left(c_{t+s}\right)(\varepsilon-1)\frac{W_t^*}{p_t}\left(\frac{p_{t+s}}{p_t}\right)^{\varepsilon-1}y_{t+s}, \quad \text{(A.2)}$$

where w_t^* is the optimal reset wage chosen at date t, and we have used the fact that $(1-\tau_{\ell,t})/(1-\tau_{c,t})$ is constant and equal to $1-\bar{\tau}_\ell$. Log-linearizing the first-order condition around a zero-inflation steady state:

$$\sum_{s\geq 0}\beta^s\theta^s\left(\phi\hat{y}_{t+s}+\varepsilon\left(\hat{p}_{t+s}-\hat{p}_t\right)-\hat{W}_t^*+\varepsilon\hat{p}_t-(\varepsilon-1)\hat{p}_{t+s}+\gamma\hat{y}_{t+s}\right)=0,$$

where $\phi\equiv\frac{\nu_{\ell\ell}(\ell)\ell}{\nu_\ell(\ell)}$ and we have used the fact $\hat{\ell}_t=\hat{y}_t$ in a first-order approximation of the dynamics. Rearranging

$$\hat{W}_t^*-\hat{p}_t=(1-\beta\theta)\sum_{s\geq 0}\beta^s\theta^s\left((\phi+\gamma)\hat{y}_{t+s}+\hat{p}_{t+s}-\hat{p}_t\right).$$

Next, from the definition of the price index, we have

$$1+\pi_t\equiv\frac{p_t}{p_{t-1}}=\left(\theta^{-1}-\frac{1-\theta}{\theta}\left(\frac{W_t^*}{p_t}\right)^{1-\varepsilon}\right)^{\frac{1}{\varepsilon-1}}. \quad \text{(A.3)}$$

Log-linearizing around a zero inflation steady state, this gives

$$\hat{\pi}_t=\hat{p}_t-\hat{p}_{t-1}=\frac{1-\theta}{\theta}\left(\hat{W}_t^*-\hat{p}_t\right).$$

Eliminating $\hat{W}_t^*-\hat{p}_t$ and simplifying, we get

$$\hat{\pi}_t=\kappa\hat{y}_t+\beta\hat{\pi}_{t+1},$$

where $\kappa=\frac{(1-\theta)(1-\beta\theta)(\phi+\gamma)}{\theta}$.

Appendix B. Computation of Optimal Policy

Here we will describe how to solve the problem for an application for which the ZLB binds for the first n periods after the shock occurs. We partition $\boldsymbol{p} = (\boldsymbol{p}_1' \, \boldsymbol{p}_2')'$, where the lower bound binds on $\boldsymbol{p}_1$ and not on $\boldsymbol{p}_2$. We can then rewrite the policy problem as

$$\min_{\boldsymbol{p}_2, \hat{\boldsymbol{x}}} \frac{1}{2} \hat{\boldsymbol{x}}' Q \hat{\boldsymbol{x}}$$
$$\text{s.t.} \hat{\boldsymbol{x}} = \bar{\boldsymbol{x}} + \Theta_{p,1} \underline{\boldsymbol{p}}_1 + \Theta_{p,2} \boldsymbol{p}_2,$$

where we have partitioned Θp to correspond to the partition of $\boldsymbol{p}$. The first-order conditions of this problem yield

$$\Theta_{p,2}' Q \left(\bar{\boldsymbol{x}} + \Theta_{p,1} \underline{\boldsymbol{p}}_1 + \Theta_{p,2} \boldsymbol{p}_2 \right) = 0,$$

which we can easily solve for $\boldsymbol{p}_2$.

To solve the full problem, we perform the above calculation for all possible values of n (within reason). For each candidate n, we solve for $\boldsymbol{p}_2$ as above and then check if it violates the constraint $\boldsymbol{p}$. If so, we discard this candidate. If not, we compute and store the objective value $\hat{\boldsymbol{x}}' Q \hat{\boldsymbol{x}}$. After evaluating all the candidate values of n, we select the one that yields the lowest objective value.

This procedure is a simple and robust method for typical macroeconomic shocks that mean revert, resulting in a binding ZLB only for the first n periods. For more complicated ZLB episodes, one could use more sophisticated quadratic programming methods.

Appendix C. Further Optimal Policy Results

This appendix presents two additional sets of optimal policy results. First, figure C.1 shows optimal monetary policy for the dual-mandate policymaker and in the absence of the ZLB constraint. Second, table C.1 shows the loss function values achieved by policymakers using different policy tools.

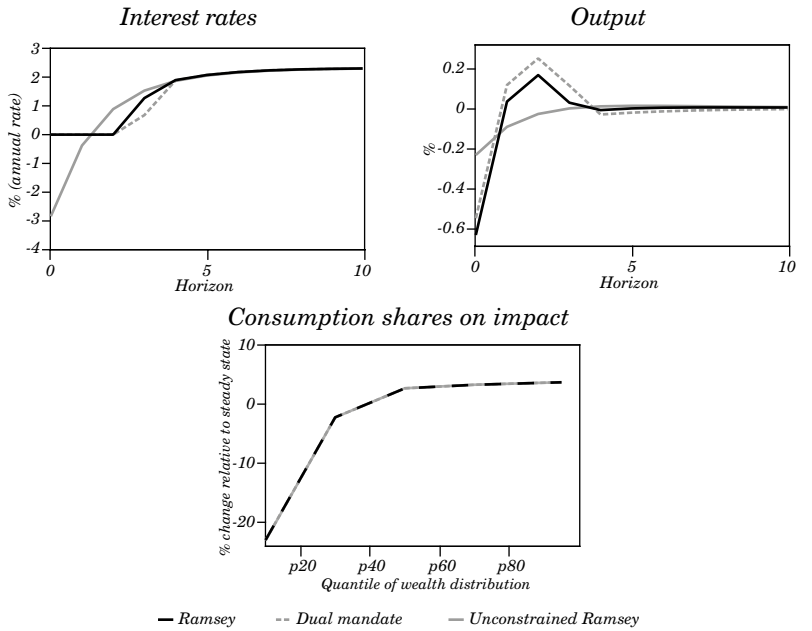
Table C.1 Ramsey Loss Achieved Relative to Monetary Policy

Policy instrument	*Relative loss*
Monetary policy (ZLB)	1.00
Monetary policy (unconstrained)	0.99
Unconventional fiscal policy	0.95
Fiscal stimulus payments	0.05
Joint monetary-transfer policy	0.04

Source: Authors' calculations.

Note: The table shows the policymaker loss under optimal policy for different policy tools. All values are reported relative to the loss achieved by ZLB-constrained Ramsey monetary policy.

Figure C.1 Optimal Ramsey and Dual-Mandate Monetary Policies



Source: Authors' calculations.

Note: Impulse responses of interest rates, output, and consumption shares (at $t = 0$) to the inequality shock under optimal monetary policy for the Ramsey planner subject to the ZLB (black), the dual-mandate policymaker subject to the ZLB (dashed grey), and the unconstrained Ramsey planner (solid grey).

The Role of Progressivity on the Economic Impact of Fiscal Transfers: a HANK for Chile

Benjamín García
Central Bank of Chile

Mario Giarda
Central Bank of Chile

Carlos Lizama
Central Bank of Chile

During the Covid-19 pandemic, the Chilean government provided unprecedented economic assistance to households. Direct fiscal transfers through stimulus checks amounted to nine percent of the country's GDP. Aditionally, three times during the period, policymakers allowed for the possibility of withdrawing up to ten percent of the workers' individual pension accounts savings. This policy provided households with access to additional resources equivalent to 19 percent of GDP. Overall, the extra liquidity provided amounted to 28 percent of GDP, thus becoming Chile's most extensive support package in recent history.[1]

The magnitude of these measures highlights how important it is to understand the impact of fiscal transfers on economic activity. However,

The opinions and mistakes are our exclusive responsibility and do not necessarily represent the opinion of the Central Bank of Chile or its board. We thank Gastón Navarro and attendees at the XXV Annual Conference of the Central Bank of Chile for fruitful comments and Giancarlo Acevedo, Javiera Azócar, Ignacio Rojas, and Valentina Vásquez for superb research assistance.

1. To put these numbers in context, before the Covid-19 pandemic, the Chilean government's total spending in subsidies and direct transfers—including education and health—was about 11% of GDP.

Heterogeneity in Macroeconomics: Implications for Monetary Policy, edited by Sofía Bauducco, Andrés Fernández, and Giovanni L. Violante, Santiago, Chile.

our paper does not specifically focus on the Covid-19 experience.[2] Instead, we aim to make a more general point by emphasizing the significance of policy design progressivity in achieving the expected effects on aggregate outcomes, building on the findings of Céspedes and others (2013).

Throughout our analyzed sample period, from 2018 to 2022, we document significant heterogeneity in the scope and progressivity of twelve programs. This heterogeneity allows us to study the differential impact on macroeconomic outcomes of policies with different degrees of progressivity. We start by empirically studying the macroeconomic effects of fiscal transfers. First, we estimate a Bayesian structural vector autoregressive model (BSVAR) to show that fiscal transfers significantly impact economic activity. Second, we document that some policies were mainly flat along the income distribution, while others displayed significant progressivity, thus showing how households with different marginal propensities to consume (MPCs) were affected by the transfers varied across time and policies. Third, with the help of micro data on credit- and debit-card transactions at the municipal level,[3] we study whether fiscal transfers with different degrees of progressivity showed a differentiated impact on household card purchases. To do so, we estimate a local projection-like equation of the dynamic effects of different policies and find that, while all of them show significant effects on this proxy for consumption, the impact of progressive transfers was significantly larger than their nonprogressive counterparts. In other words, these results show that, per unit of help, progressive fiscal transfers, by stimulating purchases the most, were more effective in increasing aggregate consumption. These results support the view that the Chilean economy displays strong non-Ricardian elements, which motivate the use of models that depart from the permanent income hypothesis.

To study to what extent (and under what conditions) transfers progressivity has a role at the aggregate level, we build an heterogeneous agents New Keynesian (HANK) model for the Chilean

2. Vaskov and others (2022) present a comprehensive analysis of the macroeconomic effects of the different fiscal programs implemented by the Chilean government during the Covid-19 pandemic.

3. Administratively, Chile is subdivided into 346 municipalities, also called communes. Wikipedia defines them as "the smallest administrative subdivision in Chile. It may contain cities, towns, villages, hamlets, and rural areas. A conurbation may be broken into several communes in highly populated areas, such as Santiago, Valparaíso, and Concepción." See https: //en.wikipedia.org/wiki/Communes_of_Chile

economy featuring progressive and nonprogressive transfer policies. Both policies are modeled as lump-sum transfers to households. Our model follows Auclert and others (2018), who develop a general equilibrium model with heterogeneous agents and nominal rigidities to study the macroeconomic effects of fiscal policy in the United States. We extend their analysis by considering two features we find essential for the case of Chile: unemployment—with search and matching (SAM) frictions—and progressivity of fiscal transfers. The model also features capital adjustment costs and a government that can finance its spending through taxes and debt accumulation.

Following a strategy similar to Kaplan and others' (2018), we calibrate the model to the Chilean economy by matching the share of hand-to-mouth (HtM) households as documented in household wealth surveys. We also use highly granular administrative data (from the Social Security Administration) on labor income quarterly to calibrate the household's income risk and consumption profiles.

To fix ideas, we propose a statistic we dub the "policy slack", that summarizes to what extent the policy undertaken is expansionary. We define the policy slack as the excess transfer delivered to households due to fluctuations in income. For instance, a positive slack in a downturn means transfers are more generous than needed to offset the household's income loss. We show that a positive policy slack is present in some of the policies implemented in Chile. Furthermore, the slack is heterogeneous across different households and policies. We also show that, under certain conditions, we can summarize the effects of policies on consumption by the relationship between the slack and the households' MPCs. In particular, we decompose the fluctuations in consumption into an average effect, which summarizes how averages fluctuate, and a distributional effect, which summarizes how the distribution of the slack affects the evolution of aggregate consumption. Moreover, we show that the distributional component is significant for all calibrations. It then follows that, when evaluating the effects of fiscal policies, it is crucial to consider not only the magnitude of the policy itself but also how far the policy took each household away from their 'normal' income. We then show that the progressivity of the transfers considerably affects the macroeconomic impact of the programs in that the more concentrated on high MPCs they are, the higher the response of aggregate variables. This result is especially marked when the government finances its spending with debt instead of taxes, so tax-paying households do not contemporaneously pay the additional government expenses. We also find that the aggregate

effect (as it is common in this literature) depends on how investment responds. We find, however, that this dependence is mostly orthogonal to the progressivity of the policy. Therefore, it does not affect the differential impact between high and low progressivity transfers. Furthermore, we show that more progressive transfers, as they affect the economy more, have more substantial general equilibrium effects than the less progressive ones.

Related Literature. A relevant part of the HANK literature emphasizes the role of fiscal policy and how it relates to non-Ricardian agents in the economy. Oh and Reis (2012) study the role of targeted transfers in the context of the Great Recession of 2008–2009, and point out the need for models that account for the positive effects of transfers; McKay and Reis (2016) study the role of progressive fiscal policies to show quantitatively that unemployment benefits and progressive taxes generate an attenuation of the business cycle because of their role as automatic stabilizers; Ferriere and Navarro (2020) study the role of tax progressivity for the transmission of government spending, and show that in times where spending is progressively financed, the fiscal multiplier was higher in the U.S. than in times where taxes were less progressive; Hagedorn and others (2019) dissect the transmission of government spending and transfers into the aggregate economy in HANK models; Auclert and others (2018) show that HANK models feature a Keynesian multiplier that gives rise to a Keynesian cross that amplifies the effects of fiscal policies; Kaplan and Violante (2018) argue that HANK models feature stronger nonequivalence than their representative agent counterparts, showing that the inclusion of heterogeneous agents changes both the transmission mechanism and the aggregate effect of fiscal shocks. This paper also relates to the literature on HANK with SAM frictions. We closely follow Gornemann and others (2016), who study the role of SAM in the transmission of monetary policy with heterogeneous agents, and Ravn and Sterk (2020), who show analytically how HANK and SAM frictions interact.

Finally, this paper is related to the empirical analysis of the effects of fiscal transfers on consumption. We follow Johnson and others (2006) and Parker and others (2013), who study the effects of the 2001 and 2008 fiscal rebates on consumption to estimate MPCs in the U.S. by using the Consumer Expenditure Survey. Another relevant paper is Misra and Surico (2014), who estimate the heterogeneous effects of these rebates. We study the dynamic effects of fiscal transfers as in a local projection analysis following the literature on the estimation of MPCs.

We contribute to this literature in four dimensions. First, we show suggestive evidence that the progressivity of transfers matters for the transmission of these policies, i.e., more progressive transfers have stronger effects on aggregates. Second, we extend the theoretical analysis to the labor market to study how unemployment affects the transmission mechanisms of fiscal transfers.[4] Third, we show that the effects of policies can be decomposed into an average effect and a distributional effect (extending Patterson, 2019), and that the way the policy is distributed across households with different MPCs is crucial. Finally, we show that a relevant part of the transfers' second-round general equilibrium effects is driven by the presence of frictional unemployment.

The remainder of the paper is organized as follows. Section 1 presents the empirical evidence we use to motivate this paper. Section 2 describes the model. Section 3 discusses the calibration. Section 4 describes what we call the policy slack—a statistic that summarizes the expected effect of the shocks on aggregate consumption. Section 5 shows the quantitative results from the model. Section 6 concludes.

1. Fiscal Support Measures in Chile: Stylized Facts and Macroeconomic Implications

In this section, we document some stylized facts about the magnitude and implementation of the fiscal transfers given to Chilean households between 2018 and 2022 and perform some empirical estimations showing the macroeconomic impact of the policies. We start by showing some key macroeconomic aggregates to contextualize the scope of the implemented policies. Then, we describe the amounts involved, both in aggregate and by quintiles of the income distribution. Finally, we show suggestive evidence that the effects these measures have on household expenditure are statistically and economically significant and related to the progressivity of the transfers, motivating our further study on the theoretical channels that may generate the observed heterogeneous impact of the different policies on macroeconomic aggregates.

4. Guerra-Salas and others (2021) emphasize the importance of including unemployment in the analysis of the dynamics of the Chilean business cycle, where variation along the extensive margin of the labor supply is particularly relevant.

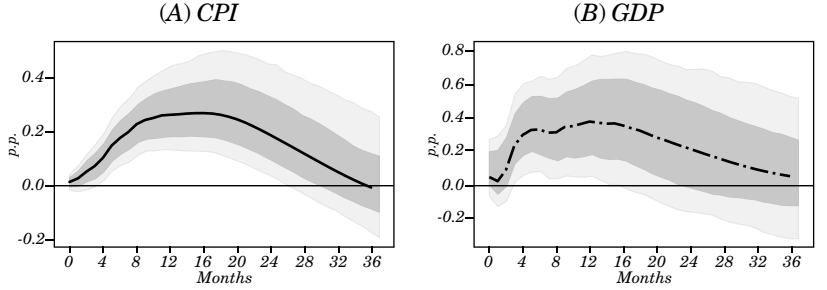
1.1 Fiscal Transfers Stimulate Economic Activity

To study how fiscal transfers affect macroeconomic aggregates, we update the estimates from Céspedes and others (2013) by running a fiscal structural VAR at monthly frequency. We follow Blanchard and Perotti (2002) approach by using a Cholesky identification. Due to our short sample, we estimate the VAR with Bayesian methods. The BSVAR includes fiscal transfers from aggregate fiscal accounts (as a share of GDP), fiscal income, CPI, and industrial production—in that order. The sample spans from January 2005 to August 2022. We consider twelve lags, detrend and seasonally adjust the series, and assume the usual normal-Wishart priors.

The impulse response functions from figure 1 show the response of the log of CPI and the log of GDP to a one-percent of GDP increase in transfers and subsidies, with the corresponding 90 and 68 percent confidence intervals. The results are both statistically and economically significant: a one-percent increase in the transfers-to-GDP ratio generates a 0.4 percent increase in GDP.

Notice that government transfers amounted to about ten percent of GDP during the Covid-19 pandemic, a greater order of magnitude than the exercise in figure 1 so that the effects of the policies undertaken during the crisis would have a substantial impact on the aggregates. This evidence suggests an important non-Ricardian component in the Chilean economy, showing that, as households see their disposable income increase after receiving fiscal transfers, they spend a significant part of this inflow in the subsequent periods, and this leads to substantial short-run effects on industrial production. Also, there is a significant rise in CPI inflation after these shocks.

Figure 1. CPI and GDP Response to a One-Percent of GDP Rise in Government Transfers



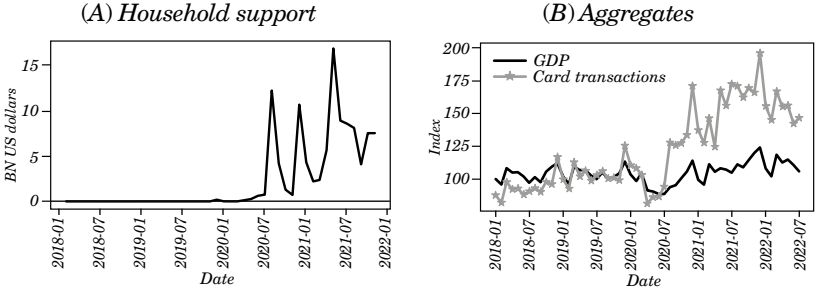
Source: Authors' calculations.

1.2 Not All Support Is the Same

In this section, we characterize the household support measures implemented in Chile in the form of direct transfers from January 2018 to November 2021 and study to what extent these policies affected consumption differently depending on the level of progressivity they displayed.

We consider twelve programs featuring different sizes, timings, cyclicality, and progressivity.[5] The data on the different programs come from the Ministry of Social Security and from the Pensions System Regulator. While these data are available at the individual level, for the empirical analysis performed in this section, we aggregate them at a municipal level as this allows us to draw a direct comparison with our measure for consumption, only available up to that level of aggregation. Figure 2 shows the programs' size and its relationship with economic activity. The left panel depicts the total amount of additional liquidity households obtain thanks to these measures. The right panel, on the other hand, shows how the path of these policies correlated with the evolution of aggregate demand during the period.

Figure 2. Total Household Support and Aggregate Outcomes



Source: Authors' calculations.

5. The twelve programs are i. Family help check; ii. Family base check; iii. Christmas Covid-19 check; iv. School homework check; v. Child homework check; vi. Covid-19 emergency check; vii. Protection check; viii. Emergency Income Covid-19; ix. Emergency Covid-19 2020; x. Guaranteed Minimum Income; xi. Universal Covid-19 check; xii. Pension Funds Withdrawals. In this paper, we consider the latter as fiscal transfers, since pension funds in Chile are fully illiquid accounts in the short run, hence, they are most likely perceived as extra income.

Although all of the features mentioned earlier may play a role in the effectiveness of the different programs, in what follows, we concentrate on only one dimension—the progressivity of the policies. To do so, we define progressivity (conceptually) as the way the government distributes these transfers among households of different incomes. To compute each policy's progressivity, we use the ratio between the absolute amount of liquidity provided to the first and fifth quintiles (Q_1/Q_5). Then, a unitary value for our progressivity score means that all quintiles receive the same amount. That is the relevant threshold since, in the model below, we define MPC as the response of households to a unitary increase in income where this additional amount is the same for everyone.[6] To build the index, we start by classifying each municipality into an income distribution quintile. We then build a per quintile population-weighted transfer measure for all twelve policies and then compute the ratio Q_1/Q_5 for every period for each policy. Finally, we assign each of the twelve programs into two categories: progressive and nonprogressive. As the same program may have different progressivity scores at different periods, we label a program as progressive if the policy has $Q_1/Q_5 > 1$ every month during its implementation and nonprogressive otherwise.

Figure 3 shows the evolution of the average progressivity of both types of policies. We can see that progressivity levels have been falling steadily since early 2020, suggesting a shift towards high-coverage fiscal transfers.[7]

We now analyze the differentiated impact of progressive and nonprogressive policies on consumption. In particular, we study the effect of the policies per unit of additional liquidity provided to the households. To carry out the analysis, we use several data sources, including data on credit- and debit-card transactions at the municipal level as a proxy of consumption obtained from Transbank, a private firm that processes most of the credit and debit transactions in Chile; data on labor income at the municipal level as a control (to account for heterogeneous fluctuations in income) obtained from the Chilean Unemployment Insurance Administration Agency; per municipality

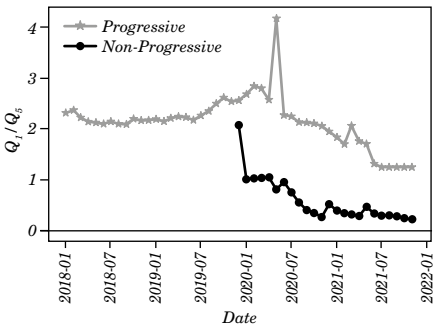
6. This is an absolute measure of progressivity, as opposed to alternative relative progressivity metrics that consider transfers as a share of the household's income or how the transfer helped increase the income of different households. Moreover, a policy with progressivity index 1 (same lump-sum transfer for everyone) is, in fact, progressive in relative terms.

7. In the Appendix, Figure 15 shows the progressivity scores for all of the analyzed programs.

total amounts given by the different programs obtained from the Ministry of Social Security and the Pensions System Regulator; finally, as additional controls, we use data on GDP, CPI, and exchange rates available from the Central Bank and the National Statistics Institute.

Our credit- and debit-card transaction data are available at the municipal level and distinguish between in-person and online purchases. We use the former, as the latter is harder to associate with the buyer's residence. Using these data as a proxy for aggregate consumption has a few shortcomings. First, it only considers card transactions and hence only represents a fraction of the aggregate consumption in the economy, not including cash purchases. Second, although we have access to the firm and place where the transactions were made, we do not know the individual who made the purchase. Due to these restrictions, we carry out our analysis at the municipal level.[8] In a companion paper,[9] we show that card transactions track national accounts data well and that municipalities in Chile are a good approximation of their inhabitants.

Figure 3. Progressivity of Household Support

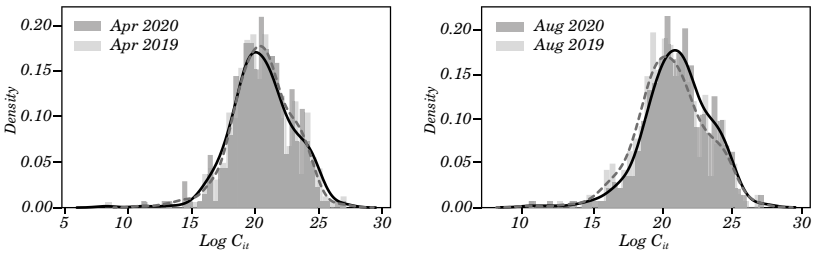


Source: Authors' calculations.

8. The geographical approach is used, for instance, by Mian and Sufi (2009) and Mian and others (2013) to study the effects of wealth on consumption. This approach is also extensively discussed by Guren and others (2020) to disentangle general equilibrium from the partial equilibrium effects of these estimates.

9. García and others (2023b).

Figure 4. Histograms of Consumption at Municipal Level, Selected Dates



Source: Authors' calculations.
Note: Solid line: 2020. Dashed line: 2019.

The aggregation of the more granular fiscal support data down to the municipality level is, as mentioned above, a compromise due to the availability of consumption data. Still, its level of aggregation is appropriate for our analysis, given the observed heterogeneity across municipalities in all the dimensions we are studying: consumption, income, and fiscal support. Figure 4 helps us visualize this by showing the cross-sectional distribution of consumption at the municipal level on selected dates. The figure allows us to point out some relevant facts. First, there is considerable heterogeneity with significant dispersion. Second, the distributions are not static, as they seem to evolve: In April 2020, we observed a tightening of the distribution with respect to 2019; perhaps even more importantly, we observed a rightwards shift in consumption in August 2020, the date of the first pension funds withdrawal, where households received a significant liquidity influx. An outlier does not drive that month's shift, as we observe that in almost all municipalities, consumption rose. These facts give us confidence that aggregating at a municipal level allows for a good representation of the heterogeneity we want to exploit in our analysis.

We study the differential effects of progressive and nonprogressive policies by exploiting the abovementioned heterogeneity. We follow the specification by Misra and Surico (2014), who estimate the effects of the 2001 and 2008 rebates in the United States by using the Consumer Expenditure Survey.[10] To be able to analyze not only

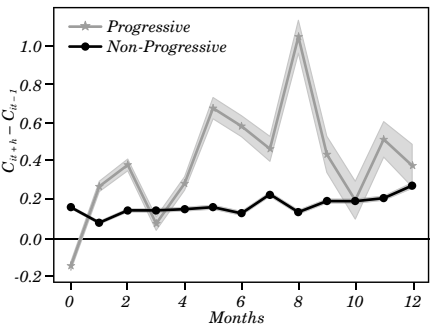
10. Misra and Surico (2014) further study the heterogeneous effects of those rebates following Johnson and others (2006) and Parker and others (2013). A similar approach is also used by Fuster and others (2020), who use surveys from experiments to study the effects on consumption of raising households' income.

the contemporaneous response of consumption to fiscal transfer shocks but also their dynamics, we estimate the following local projection-like regression:

$$C_{it+k} - C_{it-1} = \alpha_k + \beta_k T_{it}^{p} + \delta_k T_{it}^{np} + \Gamma_k' X_{it} + \epsilon_i + \psi_t + \varepsilon_{ikt}, \text{ for } k = 0, \ldots, K \quad (1)$$

where C_{it+k} is total credit- and debit-card purchases in municipality i in period $t + k$; α_k is a constant for projection k; T_{it}^{p} and T_{it}^{pn} denote the total amount of progressive and nonprogressive policies given to a municipality i in time t; ϵ_i and ψ_t are respectively a municipality and a time fixed-effect; and X_{it} is a vector of controls that include two lags of income growth and of a mobility index at a municipal level, as well as two lags of T_{it}^{p}, and T_{it}^{pn}.[11] The estimated coefficients β_k and δ_k denote the consumption response up to period $t + k$ after the household support given in period t.[12]

Figure 5. Response of Consumption to Progressive and Nonprogressive Policies



Source: Authors' calculations.

11. We control for an index of mobility which varies along municipality and over time. In Chile, the lockdowns during Covid-19 were at a municipal level, with their degrees varying from 1 (the most restrictive) to 5 (the least restrictive).

12. Robustness exercises with four and eight lags yield qualitatively similar results.

Figure 5 shows the results from estimating Equation (1). The figure presents the effect on household card spending after both types of policies. Several results are worth commenting on. First, transfers have a positive and significant impact on consumption. The regression results show that municipalities that received more transfers saw a more pronounced increase in their consumption. Second, there is a significant differential effect on consumption between progressive and nonprogressive policies. On the one hand, the peak effect on consumption after a progressive transfer is almost five times higher than after a nonprogressive. However, on the other hand, the response to progressive policies is more front-loaded than nonprogressive policies, appearing much more evenly distributed over time. In the remainder of the paper, with the help of a HANK model, we will study the theoretical reasons behind these results.

There are a few essential points to address. First, the observed consumption responses may only partially reflect the reactions to exogenous fiscal transfers, even when considering factors such as income and employment at the municipal level. This could be due to consumption decisions being influenced by increased transfer expectations. While the short interval between policy announcement and implementation (one month) suggests the possibility of exogeneity, definitive causal claims cannot be made. Second, due to the aggregation of individuals up to the municipal level, these results can be interpreted neither as fully partial equilibrium MPCs (as they include potential spillovers from the municipal aggregate consumption to the individual household) nor fully general equilibrium aggregate effects (as it does not consider the GE effects that an increase in aggregate consumption has on a municipality's consumption spending). This issue, common to all estimates using cross-sectional data, arises due to what is called the missing intercept problem,[13] where we cannot be sure of the total effect of a shock on the aggregate economy, and we can only infer the differential effect of the more exposed individuals versus the average, which in this case would mean the differences in consumption between municipalities that received more transfers than others. Finally, while the debit- and credit-card transaction data closely resemble aggregate consumption patterns, it is important to note that it might not capture total household consumption. Biases could arise due to self-selection in debit-/credit-card usage. For instance, households without prior card usage might adopt it after the transfers, mainly since many programs

13. See Wolf (2023) and Nakamura and Steinsson (2014).

require a bank account for more accessible receipt of financial aid. These biases are assumed to be evenly spread across municipalities and affect both policies equally, allowing an unbiased comparison of differential effects between the two policy types.

2. A HANK Model with Heterogeneity in Transfers Progressivity

To rationalize the facts presented in the previous section and study the different policies' roles, we build a HANK model calibrated for Chile. We closely follow the approach—and methods—presented by Auclert and others (2021). The model is a HANK with unemployment risk[14] with liquid and illiquid assets.[15]

We extend the model to include unemployment risk, as it has been shown that the extensive margin of the labor supply is a fundamental driver of the income risk and employment fluctuations in Chile.[16] This feature is especially relevant for households at the bottom of the distribution and that depend crucially on labor income.

In the model, the government is able to provide transfers in different amounts to households of different income levels. In addition, the government can finance its spending by issuing debt or raising taxes. Finally, the model has the usual features of New Keynesian models: price rigidities, monopolistic competition in intermediates, and capital adjustment costs. Since we use the methods developed by Auclert and others (2021) to solve the model, which relies on economies with aggregate shocks but without uncertainty, we omit the expectation operator over time in the model's description. In particular, the method applies a linearization of the sequence space, which relies on unexpected shocks but with a known future path.

2.1 Households

The economy is populated by a continuum of households of measure one. Households are heterogeneous in their assets, productivity, and employment state. Households deliver utility from consumption and leisure. They maximize the time-separable utility function $\mathbb{E}\left\{\sum_{k=0}^{\infty}\beta^{k}u\left(c_{t+k},h_{t+k}\right)\right\}$, where $u(c,h)$ is of the usual CRRA

14. As in Ravn and Sterk (2020) or Gornemann and others (2016).
15. As in Auclert and others (2018).
16. See Guerra-Salas and others (2021).

form $\frac{c^{1-\gamma}}{1-\gamma} - \psi \frac{h^{1+\varphi}}{1+\varphi} \mathbb{I}_e$ and $\mathbb{E}$ is the expectation operator over labor productivity and employment uncertainty. $\mathbb{I}_w = 1$ if working and zero otherwise (extensive margin), and h is hours worked (intensive margin). There are N_z possible idiosyncratic states in the productivity dimension where the probability of transitioning between states z and z' is given by $\Pi(z, z')$.

Agents can be employed or unemployed. If employed, they supply H_t hours common to all workers due to labor market frictions we explain below (thus, $h_t = H_t$ for all households). Workers earn $(1 - \tau_t) w_t H_t z_t$, where w_t is the wage per efficient hour and τ_t is a proportional labor income tax. If unemployed, households receive an unemployment benefit denoted by ω, distributed in proportion to agents' productivity times wages $w_t z_t$. Following Diamond-Mortesen-Pissarides' framework, we denote by δ the separation rate and $f(\theta)$ job-finding rate of transitioning between the states w and u such that $s = [w,u]$. Hence, $\Pi(z, z', s, s')$ is the transition matrix considering both unemployment and income risk. Consequently, income becomes $y_t(z_t, s)$ with $y_t(z_t, .) = [(1 - \tau_t) w_t H_t z_t, z_t w_t \omega]$.

Agents can trade in two assets, i.e., $\boldsymbol{a} \equiv \{a_1, a_2\}$. These assets pay an interest rate $r_{ht}(h = \{1,2\})$ and are subject to a non-borrowing constraint. The value function of an agent in the states $(z, \boldsymbol{a}, s)$ at time t is, therefore[17]

$$V_t(z, \boldsymbol{a}, s) = \max_{c, \boldsymbol{a}} u(c) + \beta \sum_{z,s} \Pi(z, z', s, s') V_{t+1}(z', \boldsymbol{a}', s')$$

$$\text{s.t. } c + \sum_h a'_h = \sum_h (1 + r_{ht}) a_h + y(z, s) + f_t(z) + d_t(z)$$

$$\boldsymbol{a} \geq 0.$$

Households receive a fiscal transfer which is a function of household productivity $f_t(z)$; i.e., it depends on the household type. We determine this function in the calibration below. $d_t(z)$ are individual firms' dividends received by households. For the structure of assets of households, we take the approach by Auclert and others (2018), who assume there is a fully liquid (government bonds) and a fully illiquid

17. In Appendix B we present the value functions and first-order conditions of this problem.

asset (capital and equity). The illiquid assets returns are accrued in the liquid account. These assumptions allow us to match the high MPC (through the high share of HtM) and a high level of aggregate wealth while keeping the model tractable.[18]

Given optimal policies $c_t^{\star}(z,\boldsymbol{a},s), a_t'^{\star}(z,\boldsymbol{a},s), b_t'^{\star}(z,\boldsymbol{a},s)$ and denoting $\Psi(z,\boldsymbol{a},s) = \Pr(z_t = z, a_{t-1} \in A, s_t = s)$ the probability of that combination of states at the start of date t, the distribution Ψ_t has a law of motion

$$\Psi_{t+1}(z',\boldsymbol{a}',s') = \sum_{z,s} \Psi_t\left(z',\boldsymbol{a}'^{\star -1},s'\right)\Pi(z,z',s,s'), \tag{2}$$

where $\boldsymbol{a}'^{\star -1}$ are the inverse of the optimal policies of $\boldsymbol{a}$. For simplicity, we summarize in an index i, the combination of possible states, i.e., $i = (z,\boldsymbol{a},s)$. Therefore, in what follows, $\Psi(z,\boldsymbol{a},s) = \Psi(i)$, and the aggregate of a variable $x_t(i)$ is given by $\int x_t(i)\Psi(i)di = X_t$. However, we use the long notation when needed.

With the distribution and the optimal allocations we compute the aggregates $C_t = \int c_t(i)\Psi(i)di$ and the stock of liquid assets, $B_t = \int b_t(i)\Psi(i)di$ with counterpart in the government budget constraint.

2.2 Government

Fiscal policy is one of the main ingredients in our model. The government, in our setting, allocates its spending between government consumption G_t, fiscal transfers to households $f_t(z)$, and unemployment benefits ω. Transfers are heterogeneous across households and can be progressive ($f_t'(z) < 0$), regressive ($f_t'(z) > 0$), or flat $f_t'(z) = 0$. How transfers are distributed across households satisfies $\int f_t(z)\Psi(i)di = T_t$ where T_t denotes the aggregate amount of transfers. The government finances its spending by issuing real-denominated debt B_t^g and by charging proportional taxes on labor income. Government debt is held by households in their liquid account and pays a real return r_t. Transfers are lump-sum in the sense that households take these as given and do not enter their first-order conditions. However, they affect optimal decisions due to market incompleteness. The government's budget constraint is then given by

$$B_{t+1}^g = T_t + G_t + \omega w_t U_t - \tau_t w_t H_t N_t + (1 + r_t)B_t^g.$$

18. As in Auclert and others (2018), we assume the fact shown by Fagereng and others (2021) that households do not change their illiquid assets in response to income shocks.

The evolution of the fiscal balance depends on a smoothing parameter ρ_T, which determines to what extent additional spending is financed with debt according to:

$$dB_t^g = \rho_T \left(dB_{t-1}^g + dT_t \right).$$

This fiscal balance rule captures the fact that governments do not necessarily raise taxes contemporaneously to finance additional spending, as they can also issue more debt. As we will see below, the government financing strategy is key for characterizing consumption dynamics in response to fiscal transfers in general equilibrium.

2.3 Firms

There is a continuum of identical firms (indexed by $j \in [0,1]$) that produce differentiated goods using capital and labor, combining them with a Cobb-Douglas function $y_{jt} = Z_t k_{jt-1}^{\alpha} n_{jt}^{1-\alpha}$, with Z_t denoting an aggregate productivity level. Although identical, these intermediate firms are in monopolistic competition and set prices taking into account the demand for their variety. Varieties are aggregated with a Dixit-Stiglitz aggregator with a price elasticity equal to $\frac{\mu_p}{\mu_p - 1}$, with μ_p being the steady state markup charged by these firms. Price setting is subject to quadratic Rotemberg adjustment costs, with the cost given by $\Theta_{jt}^{\pi} = \frac{\mu_p}{\mu_p - 1} \frac{1}{2\kappa_p} \left[\log\left(1 + \pi_{jt}\right) \right]^2 y_{jt}$, with κ_p being the adjustment cost parameter that is also the slope of the Phillips curve. Intermediate firms solve:

$$J_t\left(p_{jt-1}\right) = \max_{y_{jt}, p_{jt}, k_{jt}, n_{jt}} \left\{ \frac{p_{jt}}{p_t} y_{jt} - w_t h_{jt} n_{jt} - r_t^k k_{jt-1} - \Theta_{jt}^{\pi} + \frac{J\left(p_{jt}\right)}{1 + r_{t+1}^a} \right\}$$

s.t.

$$y_{jt} = Z_t k_{jt-1}^{\alpha} \left(h_{jt} n_{jt} \right)^{1-\alpha},$$

$$y_{jt} = \left(\frac{p_{jt}}{p_t} \right)^{-\frac{\mu_p}{\mu_p - 1}} Y_t.$$

The first-order conditions, after symmetry, read

$$\log\left(1+\pi_t\right) = \kappa_p\left(mc_t - \frac{1}{\mu_p}\right) + \frac{1}{1+r^a_{t+1}}\frac{Y_{t+1}}{Y_t}\log\left(1+\pi_{t+1}\right)$$

$$mpl_t = \left(1-\alpha\right)mc_t\frac{Y_t}{N_t}$$

$$r^k_t = \alpha mc_t\frac{Y_t}{K_{t-1}},$$

where mc_t is the marginal cost. The aggregate amount of profits generated each period by intermediate firms is given by

$$\Pi^y_t = \left(1-mc_t\right)Y_t - \Theta^\pi_t.$$

2.4 Labor Markets

There is a union that determines hours worked (the intensive margin) by aggregating households' preferences, solving the individual problem at an aggregate level. This maximization procedure generates the following labor supply, which is given by the average marginal rate of substitution equal wages:

$$\psi H^\varphi_t = \mathcal{U}' w_t,$$

with $\mathcal{U}' = \int\left(1-\tau\left(z_t\right)\right)z_t u'\left(c_t\left(i\right)\right)\Psi_t\left(z,a,s=e\right)di.$

To account for fluctuations in unemployment and unemployment risk, we consider a labor market with search frictions as in Ravn and Sterk (2020) and Gornemann and others (2016). The model is a canonical Diamond-Mortensen-Pissarides model. We assume there is a Cobb-Douglas matching function $M\left(U_t,V_t\right) = m_t U^\gamma_t V^{1-\gamma}_t$, which leads to a job-finding probability $f_t\left(\theta_t\right) = m_t\theta^{1-\gamma}_t$ and a job-filling probability $q\left(\theta_t\right) = m_t\theta^{-\gamma}_t$, where $\theta_t = \frac{V_t}{U_t}$ is the market tightness. U_t is the measure of unemployed workers with $U_t = \int d\Psi\left(z_t,b,a,s=u\right)$, and the level of employment is given by $N_t = 1 - U_t$. The probability of becoming unemployed while working is given by an exogenous separation probability δ.

We assume that households cannot individually supply—and set—labor. Instead, there is an intermediary for each type who hires and sells labor services. This firm's value of a worker with productivity z_t is

$$J(z_t) = (mpl_t - w_t)z_t + (1-\delta)\frac{1}{1+r_{t+1}}\mathbb{E}_z\left[J(z_{t+1}|z_t)\right],$$

where mpl_t is the marginal product of labor. The free-entry condition for these intermediaries is

$$\frac{c_v}{q(\theta_t)} = \frac{1}{1+r_{t+1}}\int_{z_t}\mathbb{E}_z\left[J(z_{t+1}|z_t)\right]d\Phi(z_t, b, a, s=u).$$

Additionally, we use a Nash-inspired wage rule

$$w_t = (1-\eta)\omega + \eta(mpl_t + c_v\theta_t),$$

where η is workers' wage bargaining power.

Finally, the intermediary generates profits from the difference between the marginal productivity of labor and the real wage given by

$$\Pi_t^w = mpl_t - w_t.$$

These profits are delivered to households in the same way monopolistic profits are.

2.5 Capital

We assume there is an investment fund that produces capital. The investment fund owns the economy's capital stock K_t. The fund makes the economy's investment decision subject to an adjustment cost $\Gamma_t(K_{t+1}, K_t)$, solving the problem

$$\max_{K_{s+1}, I_s}\sum_{s=0}^{\infty}\left(\frac{1}{1+r_s}\right)\left[r_t^k K_t - I_t - \Gamma(K_{s+1}, K_s)\right]$$

s.t.

$$K_{s+1} = (1-\delta_K)K_s + I_s,$$

where $\Gamma(K_t, K_{t+1}) = \frac{1}{2\delta_K \epsilon_I}\left(\frac{K_{t+1}-K_t}{K_t}\right)^2 K_t$. The first-order conditions are:

$$(1+r_{t+1})q_t^k = r_{t+1}^k - \left[\frac{K_{t+1}}{K_t} - (1-\delta_K) + \frac{1}{\delta_K \epsilon_I}\left(\frac{K_{t+1}-K_t}{K_t}\right)^2\right] + \frac{K_{t+1}}{K_t} q_{t+1}^k$$

$$q_t^k = 1 + \frac{1}{\delta_K \epsilon_I}\left(\frac{K_{t+1}-K_t}{K_t}\right),$$

equations that reduce to the Tobin's-Q solution.

2.6 Dividends

Dividends in this economy are given by the sum of the return to capital, profits from intermediate producers, and profits from the labor intermediary. Therefore, it can be shown that dividends are given by

$$\text{Div}_t = Y_t - w_t H_t N_t - \Theta_t^{\pi} - c_v V_t - I_t - \Gamma(K_{t+1}, K_t).$$

These dividends are delivered with an ad-hoc rule similar to Kaplan and others (2018), in proportion to household productivity.

2.7 Monetary Authority

In the presence of nominal rigidities, the real interest rate r_t is determined by monetary policy, which sets the nominal interest rate i_t according to a Taylor rule that responds to inflation and unemployment:

$$i_t = i^* + \phi_\pi(\pi_t - \overline{\pi}) + \phi_U \frac{(U_t - U)}{U}.$$

We denote by $\phi_\pi > 0$ and $\phi_U < 0$ the preference parameters for inflation and unemployment respectively. Monetary authorities seek a nominal interest rate target in steady state given by i^*. Given inflation and the nominal interest rate, the real rate is determined by the Fisher equation $(1+r_t) = \frac{(1+i_t)}{1+\pi_{t+1}}$.

2.8 Aggregation

Total consumption expenditure is given by

$$C_t = \int c(i)\Psi(i)di. \tag{4}$$

Goods market clearing implies

$$Y_t = C_t + I_t + G_t + \Theta_t^{\pi} + c_v V_t + \Gamma\left(K_{t+1}, K_t\right),$$

and the market for bonds closes:

$$B_t^g = \int b\,\Psi(i)\,di.$$

3. Calibration

3.1 Households

Households' Assets. We follow Kaplan and others (2018) to develop our aggregated two-asset (liquid-illiquid) structure. For this purpose, we use a mix of data from the Chilean Financial Regulator (CMF) and the Chilean Household Financial Survey (EFH). This latter survey is the Chilean counterpart of the Survey of Consumer Finances. We consider this mix to have a reasonable estimate of the aggregates (from CMF) and the distribution of assets in the Chilean economy.

We closely follow the taxonomy proposed by Kaplan and Violante (2014), which is given by the following components (summarized in table 1). On the side of liquid assets, revolving debt corresponds to bank credit cards, lines of credit, bank or financial consumer loans, credit cards from nonbanking institutions, consumer loans in commercial houses (cash advances), credits in savings banks, cooperatives, educational loans, and other nonmortgage debts. Deposits are the total amount households keep in their checking or sight accounts. We also include equity in the liquid account from the data, which is the sum of investment in shares, mutual funds, participation in investment funds, and investment in other equity instruments (options, futures, swaps, among others). Finally, fixed income is the total amount households have invested in different instruments such as time deposits, bonds, savings accounts, and insurance with savings.

We consider three illiquid assets: net housing, defined as the value households assign to their primary home or other real estate they own, discounting the present value of the mortgage loan debt; net durables which correspond to the value of automotive assets, such as cars or trucks, motorcycles, vans or utility vehicles, and other motorized vehicles (boats, planes, helicopters, etc.), as well as other assets such as agricultural or industrial machinery, animals, works of art, etc., discounted from the debt in auto loans.

Table 1 summarizes this taxonomy as a fraction of the 2017 annual GDP. When considering aggregates, we obtain figures not so far from the ones shown in Kaplan and others (2018) for the United States. Liquid assets are a small fraction of total wealth, and housing is the largest fraction of wealth. This means that in Chile is also appropriate to use the liquid-illiquid split when considering the assets' structure.

Regarding the shape of the distribution of assets, we use the EFH to build these distributions. However, unlike Kaplan and others (2018), we only focus on the share of HtM of Chilean households, which is a key target in our calibration. Table 2 shows the shares of HtM of Chilean households. We define an HtM household as one that holds up to five percent of their quarterly income in liquid assets (in absolute value). We find that for Chile, the total share of HtM is about 39 percent of households. This figure is considerably higher than that of the United States, which is about 30 percent. Another difference that we find with respect to the U.S. is that in Chile the share of wealthy HtM households is 31 percent, while in the U.S it is six percent. The poor's HtM, though, is 8 percent, i.e., lower than the 20 percent the U.S. has. These differences are interesting, but in this paper, we only use the total share of HtM to calibrate our model.[19]

Table 1. Taxonomy of Households' Assets in Chile in 2017. Values as a Percentage of GDP

Liquid (*B*)		*Iliquid* (*A*)		*Total*
Revolving consumer debt	-0.12	Net housing	1.93	
Deposits	0.05	Net durables	0.13	
Fixed income	0.12			
Equity	0.12			
Total	0.17		2.06	2.23

Source: Commission for the Financial Market (CMF) and Internal Revenue Service.

Table 2. Share of Wealthy and Poor Hand-to-Mouth Households (Relative to the Total Population)

	Data	
Poor	Frac. With B≈0 and A=0	0.08
Wealthy	Frac. With B≈0 and A>0	0.31

Source: Authors' calculations.

19. We study the effects of these features for Chile in García and others (2024).

Income distribution and income risk. Empirically, the challenge in estimating the frequency of earnings is that almost all high-quality panel earnings data are available only at an annual (or lower) frequency. We overcome this issue by employing a confidential dataset from the Chilean Pension Regulator.[20] We calculate the empirical moments of the distribution of income fluctuations to obtain a discretized process for income risk. In particular, following Guvenen and others (2019), we consider fluctuations in income at different frequencies. We consider from the second to the fourth standardized moments (variance, skewness, and kurtosis), which, as has been shown in previous literature,[21] can be essential for aggregate fluctuations and wealth accumulation.

We assume idiosyncratic income (in logs) is given by the sum of two processes z_{1t} and z_{2t}:

$$y_t = z_{1t} + z_{2t}, \tag{5}$$

where z_{it} follows

$$z_{it} = \rho_i z_{it-1} + \sigma_i \varepsilon_{it}$$

$$\varepsilon_{it} = \begin{cases} \mu_{it} \geq p_i \sim \mathcal{N}(0,1) \\ \mu_{it} < p_i \end{cases}$$

$$\mu_{it} \sim U[0,1].$$

Therefore, we estimate parameters $\{\rho_1, \rho_2, \sigma_1, \sigma_2, p_1, p_2\}$. As noted by the previous literature, the combination of these two processes returns high kurtosis (given by a $p_i \neq 0$) and can match the moments of the growth in income at lower frequencies.

To match the moments of the empirical distribution with the income process in Equation (5), we approximate z_1 and z_2 using a discretization method first proposed by Farmer and Toda (2017) and Tanaka and Toda (2013, 2015). This method is based on matching conditional moments of the discrete approximation with the moments of the true continuous-state process. This is similar to the Rouwenhorst method proposed by Kopecky and Suen (2010), extended for nonlinear, non-Gaussian Markovian processes. Therefore, our job is to pin down the parameters that describe the processes z_i, namely ρ_i, σ_i, p_i to

20. See appendix A for a description of this database.
21. See Kaplan and others, 2018 and McKay (2017).

match the moments observed in the data and then apply the method by Farmer and Toda (2017) to obtain the discretized version that we feed into the model. We find the parameters by minimizing a loss function that takes a proposed set of parameters and computes how far we are from the desired moments.

Table 3 shows the moments of quarterly labor income for one-quarter and twenty-quarters log-change in labor income and the variance of the log of income ($\log(y_t)$). We compare the empirical moments with the ones we obtain with our discretization method.[22] What we observe here is that, naturally, the variance increases with the lag of the difference, and these distributions have a high kurtosis, which decreases with the lag of the change. Although decreasing, the kurtosis is still higher than that of a normal distribution for the twenty-period change. Table 3 shows that our model matches the empirical moments well.

We show the estimated process in table 4. We estimate a permanent process with high persistence with a half-life of around 43 years (a career shock) and a low probability of occurrence: workers receive these shocks every 3.5 years. The other shock is less persistent but more likely. Households receive it almost every quarter, while its half-life is about 0.4 quarters. With these parameters, we build the transition matrix to discretize these processes, and we consider three points for the persistent component and eleven for the transitory component.[23]

Table 3. Empirical and Estimated Moments of Labor Earnings in Chile at a Quarterly Frequency

Moment	*Data*	*Model*
Var $\log(y_t)$	0.719	0.714
Var $\Delta\log(y_t)$	0.195	0.226
Var $\Delta_{20}\log(y_t)$	0.463	0.448
Kur $\Delta_{\log}(y_t)$	11.589	11.617
Kur $\Delta_{20}\log(y_t)$	6.143	6.076

Source: Unemployment Fund Administration, Chile.

22. In García and others (2024), we study the role of all these features in Chile. In particular, we compare Chile's moments to those observed in the United States. We show that Chile has a higher variance than the United States but a lower risk.

23. This process suggests that in Chile, income risk is higher than what we observe in the United States. A reason for this high risk is the high worker turnover in Chile. Albagli and others (2017) conclude that, turnover in Chile is higher than all of the OECD countries.

Table 4. Parameter Estimates for Idiosyncratic Income Process

ρ_1	ρ_2	σ_1	σ_2	p_1	p_2
0.996	0.145	0.511	0.382	0.071	0.958

Source: Authors' calculations.

3.2 Labor Markets and Firms

Labor Markets. We use the same targets as in the quantitative model of the Central Bank of Chile.[24] We calibrate unemployment in steady state at eight percent, the vacancy filling probability $q(\theta) = 0.8$, and the separation rate to $\delta = 0.04$. In steady state, the job-finding probability is given by

$$u = \frac{\delta}{\delta + p(\theta)} \Rightarrow p(\theta) = \delta \cdot \frac{1-u}{u} = 0.46.$$

The Nash bargaining parameter is set to $\eta = 0.5$.[25] We set $\alpha = 0.5$ (Hosios condition). We calibrate the productivity of the matching function to satisfy the previous conditions, with $m = \frac{p(\theta)}{\theta^{1-\alpha}}$. Finally, we set the Frisch elasticity of labor supply $1/\varphi$ equal to one, and we calibrate the parameter of disutility of labor to match $H_t = 1$.

Firms. We assume in the steady state a capital level of 2.01 as a share of GDP (8.04 quarterly) to match the value of illiquid assets in steady state in table 1. The capital share α_k is equal to 1/3. Productivity Z in steady state is set to have GDP in steady state equal to one ($Y = 1$). The depreciation rate is equal to 0.01,[26] and in the baseline calibration, the capital adjustment cost parameter is set to $\epsilon_I = 2$. Finally, we assume markups are $\mu_p = 1.1$, and the slope of the price Phillips curve is set to 0.1.

3.3 Government

We set the Taylor rule parameters to $\phi_\pi = 1.25$ and $\phi_U = -1$ in the baseline calibration. We set the level of government spending and fiscal transfers equal to ten percent of GDP each. Fiscal transfers

24. García and others (2019).
25. As in García and others (2019) and Mortensen and Pissarides (1994).
26. From García and others (2019).

have two components, a progressive and a nonprogressive transfer. We set both to five percent of GDP. Individual transfers are defined by a nonlinear function $f(z) = T_t z^{-\aleph_f} f_0$, where f_0 is a scalar which ensures $\int f(z)\Psi(i)di = T_t$ and $\aleph_f$ is the level of progressivity. We solve the model with two transfers which only differ in the progressivity level $\aleph_f$. In the next sections, we introduce two types of policies simultaneously, a progressive and a nonprogressive one, to match the distribution of two selected policies delivered in 2020. These parameters are $\aleph_p = -1.1$ $\aleph_{np} = 0.4$ in the progressive and the nonprogressive policies respectively. We explain how we set these parameters in the next section. Finally, we set the tax rate on dividends equal to 25 percent, and we show results for different ways of government financing, ρ_T.

3.4 Solution Method

To solve this heterogeneous-agent model with borrowing constraints, we follow Auclert and others (2021). To solve the value function we use Carroll's (2006) endogenous grid method, which is a fast and accurate algorithm to solve these kinds of problems. Then, we use a Newton method to solve the steady state of this economy. And finally, to solve the model with aggregate shocks, we follow Auclert and others (2021) as well, who propose to write the model in its sequence space and linearize around that system of equations. The method relies on the fact that any model without aggregate uncertainty can be written as a sequence of equations in the transition. This is, if we assume shocks are one-time and unexpected, we can write the system as a sequence of equations in the transitional dynamics. This system of equations, which is given by $T \times M$, with T standing for the horizon of the transition and M the number of equations to solve, can be linearized around the steady state. This linearization leads to jacobians of all variables with respect to others, and the impulse-responses can be obtained by a composition of these jacobians. This method, based on Boppart and others (2018), is faster, more accurate, and more robust than methods like the ones that follow Reiter (2009). We refer the reader to the paper for more details on the method.

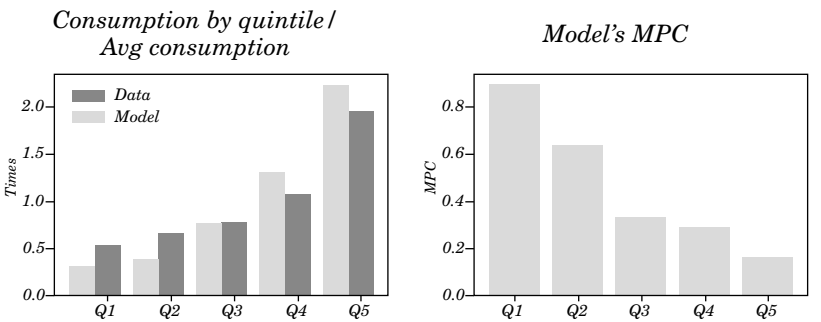
3.5 Calibration in the Steady-State and Micro Fit

To solve the steady state we leave free the disutility of labor (ψ), the discount factor (β), the level of labor income taxes (τ_w), the vacancy cost (c_v), and aggregate bond holdings (B or B^g). As targets, we set

an interest rate of five percent yearly, a HtM share of 0.39, hours at one; the unemployment rate is determined implicitly by satisfying the free-entry condition in the labor market, and τ^w by satisfying the government budget constraint. After this calibration procedure, we obtain $\beta = 0.95$, $\psi = 0.51$, $c_v = 0.19$ which leads to 0.8 percent of GDP in vacancy costs, a tax rate equal to $\tau^w = 0.08$, and aggregate bond holdings equal to 0.18 as a share of annual GDP. Finally, we set the elasticity of intertemporal substitution equal to one ($\gamma = 1$).

Additionally, our goal is to characterize the consumption distribution well. In figure 6 we show the distribution of consumption in steady state and this distribution in the data. The left-hand panel shows consumption with respect to the mean by quintile in the model and the data. Our calibration overestimates inequality: in the model, consumption of the first quintile is lower than the data, and consumption of the fifth quintile is larger. This may be problematic if we are interested in inequality itself. However, as we are interested in the response of each quintile, and there is a fall in MPCs along the income distribution, we argue that this feature of our model underestimates the effects of progressivity. That, because a more progressive policy gives money to households with high MPCs, which in our model are weighted lower than in the data. This is, if the distribution of consumption was as in the data, the response to progressive transfers would be larger than in our results. The right-hand panel shows the MPCs by quintile in the income distribution. In our calibration, the MPCs are decreasing in the quintile of income. The consumption-weighted MPC in our model is 0.31 quarterly. These values are larger than in the US,[27] as expected.

Figure 6. Distribution of Consumption and MPCs



Source: Authors' calculations.

27. As reported by Kaplan and others (2018), 0.16.

4. The Policy Slack

Often, fiscal transfers occur in response to exogenous aggregate shocks affecting households' income. When fiscal support is larger than the drop in household income, there is a gap we call the policy slack, which for household i we denote by $\chi_t(i)$ and satisfies the following identity:

$$dT_t(i) = d\chi_t(i) - dy_t(i) \tag{6}$$

with $dT_t(i)$ being the change in transfer and $dy_t(i)$ the change in income of household i. Equation (6) means that the policy slack is a measure of extra resources taken or given to a household with respect to a perfect compensation to the fall in income, where this perfect compensation is the response of transfers that keeps consumption of most consumers constant at their steady-state levels.

The policy slack is empirically observable. Take, for example, the policies undertaken during the Covid-19 pandemic. It was a combination of progressive and nonprogressive programs with different policy slacks. Table 5 shows how both policies allocated resources differently for each quintile of the income distribution. In this case, the more progressive policies showed a markedly decreasing pattern along the income distribution: the fifth quintile received less than one percent of their income, whereas the first quintile received close to 20 percent. A second group of less progressive policies was much less targeted towards low-income households. In those programs, high-income households received about the same as low-income households as a share of income. Transfers were one of many sources of policy slack. Also to be considered is the drop in income, which is also very heterogeneous across households. While the first quintile saw their income fall by about 19 percent, the income of a typical household from the fifth quintile remained practically unchanged. The combination of fiscal programs and Covid-related drops in income resulted in very heterogeneous policy slacks across quintiles. Due to the relatively low amounts given by the average progressive policy, it generated a negative policy slack. On the other hand, the more generous nonprogressive ones generated an overcompensation in the income fall.

Table 5. Household Support Measures in 2020

$Quintile$	$\frac{T_t^p(q)}{y_t(q)}$	$\frac{T_t^{np}(q)}{y_t(q)}$	$dy_t(q)$	$d\chi_t^p(q)$	$d\chi_t^{np}(q)$	$d\chi_t^{\text{tot}}(q)$
Q1	0.20	0.25	-0.19	0.01	0.06	0.26
Q2	0.09	0.31	-0.24	-0.15	0.07	0.16
Q3	0.04	0.32	-0.27	-0.23	0.05	0.09
Q4	0.02	0.28	-0.19	-0.17	0.11	0.11
Q5	0.003	0.24	0.00	0.003	0.24	0.243

Source: Authors' calculations.
Notes: Total annual labor income by quintile (q) in 2019 $y_t(q)$ obtained from the Social Security Administration (AFC), $T_t^p(q)$ is total progressive transfers by quintile in 2020, and $T_t^{np}(q)$ are nonprogressive transfers by quintile in 2020. $d(y_t(q))$ denotes the change in income of households at a given quintile (q) between 2020 and 2019.

This policy slack can be an important statistic because it helps us to evaluate the policies and has a direct effect on consumption. Moreover, in models with inequality, not only does the size of the policy slack matter, but also its distribution, which is directly related to the progressivity of the policy and interacts with the MPCs of households. To explain this, denote the household's i MPCs with $M_{t,s}(i)$, which is the response of consumption in t to an income windfall on s with $s = [0,\ldots,T-1]$. Therefore, a matrix $M_t(i)$ summarizes the intertemporal MPCs and is a $T \times T$ matrix for every i where each row is the response in period t to a shock in period s. Hence, the response of household consumption in t is the multiplication of the $M_t(i)$, the row of the matrix $M(i)$ for the period t, and the whole path of future policy slacks $d\chi(i)$, with $d\chi(i)$ being a column vector. Hence, the response of consumption in period t, assuming a constant interest rate, is

$$dC_t = \int M_t(i)\, d\chi(i)\, di, \tag{7}$$

which can be rewritten as

$$dC_t = \underbrace{\bar{M}_t\, d\bar{\chi}}_{\text{Average Effects}} + \underbrace{COV_i\left(M_t(i), d\chi(i)\right)}_{\text{Distributional Effects}}. \tag{8}$$

Equation (8) decomposes consumption fluctuations into two components: the average effect and the distributional effect of the policy slack. The first component represents the responses to the size of the policy, and the second one represents the response of consumption to the progressivity of the policy by the relationship between households' MPCs $M_t(i)$ and the policy slack $\chi(i)$. This implies

that given the same average MPCs and a given path in the policy slack, there are effects from how fluctuations in income and transfers are distributed among households. These decompositions have recently been popular in the HANK literature.[28]

We can decompose consumption further by separating 'direct' effects from the slack and 'indirect' effects[29] to analyze if the covariance object fluctuates more from partial or general equilibrium effects:

$$dC_t = \underbrace{\bar{M}_t \, d\bar{T} + COV_i\left(\bar{M}_t(i), \bar{d}T(i)\right)}_{\text{Direct}} + \underbrace{\bar{M}_t \, d\bar{y} + COV_i\left(M_t(i), dy(i)\right)}_{\text{Indirect}}. \quad (9)$$

Next, we apply this decomposition to the calibrated model. To do so, we first solve the model in the baseline calibration, assuming a constant real interest rate and calibrating the progressivity of the policy to match the second and third columns of table 5, which requires $\aleph_p = 0.4$ and $\aleph_{np} = -1.1$. After solving the model, we compute the paths for the average and distributional effect of a one percent of GDP increase in transfers (with a persistence of 0.5). In this case, we show the results for the decomposition of consumption in figure 7. Consumption increases in response to both shocks. However, the progressive transfer is twice as effective as the nonprogressive.

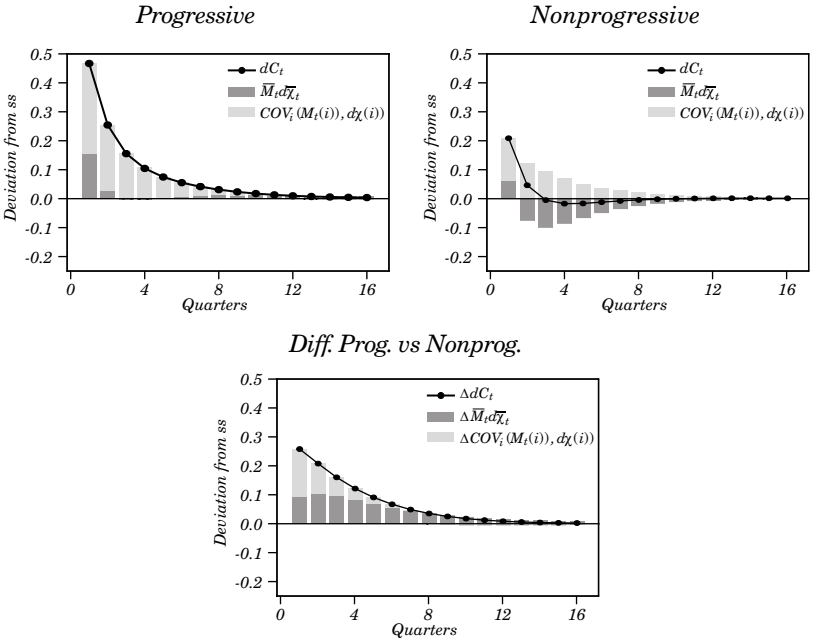
We find that the progressive policy propagates through both channels in the whole horizon. This is, the progressive policy is able to generate a positive response through the average and the distributional channels. However, this is not the case in the nonprogressive policy, where the bigger share of the fiscal transfers given to the wealthier households leads to an average channel that partially reverses the effects generated from redistribution in general equilibrium.

In figure 8, we show the decomposition described in Equation (9), separating both the distributional and the average components into their direct and indirect effects. Since MPCs and the path for the transfer are the same in both cases, the differences arise from the covariances and the general equilibrium effects.

Figure 8 shows different effects on consumption from progressive and nonprogressive transfers. In the former, the component COV $(M_t(i), dT_t(i))$ is positive, contributing to the increase in consumption. In the latter, however, the component COV $(M_t(i), dT_t(i))$ is negative and hence, counteracts the initial impulse of the transfer.

28. See Patterson (2019).
29. As in Kaplan and others (2018) or Auclert (2019).

Figure 7. Consumption Decomposition in Average and Distributional Effects. Constant r and $\rho_T = 0.5$

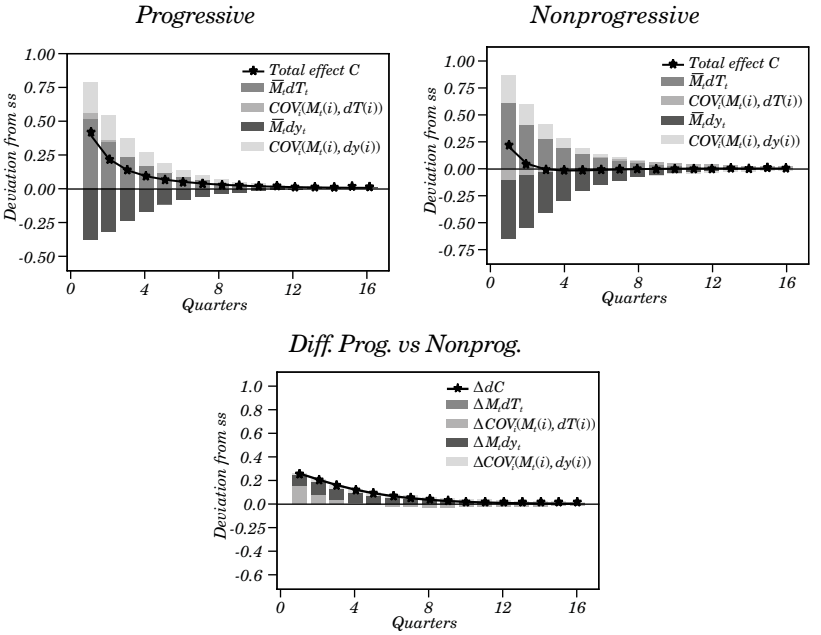


Source: Authors' calculations.

On top of those effects, we have the general equilibrium effects (or indirect effects) through fluctuations in income. For our calibration, we find that the average effect is negative for both policies but stronger for the nonprogressive policy. Moreover, these indirect effects seem to be distributed unevenly among the distribution of MPCs: the covariance term associated with that channel is positive, counteracting the negative response of the average. This result is due mainly to the countercyclical dividends our model features, which is the main driver of the negative responses in the average indirect effect. Since we distribute these dividends increasingly in productivity (and hence, on MPCs), we observe a positive COV $(M_t(i), dy_t(i))$.[30]

30. Aldunate and others (2023) find that labor income in the lowest quintiles responds more strongly than in the highest quintiles to foreign shocks, which would generate an additional source of positive COV $(M_t(i), dy_t(i))$ and hence, that would deliver more amplification in our setup.

Figure 8. Consumption Decomposition in Average and Distributional, Indirect, and Direct Effects. Constant r and $\rho_T = 0.5$



Source: Authors' calculations.

While in this exercise we study the role of these channels in the response of consumption to fiscal transfers, the decompositions from Equations (8) and (9) can be used to study the effects of a broad range of policies, like the ones described in table 5.

5. Quantitative Analysis

In this section, we explore the aggregate effects of transfers when we relax the assumption of fixed real interest rate and let monetary policy have a more active role over the business cycle. In addition to that, we show the role of government financing rules on the expected effect of the transfers.

In particular, in the exercises that follow, we show the responses of macroeconomic variables to a rise in fiscal transfers of one percent of GDP. We assume a persistence of 0.5, halving the impulse every

quarter. For each of the exercises, we show two figures. First, we show the responses to transfers, with the effect of progressive transfers on the top panels and the nonprogressive on the bottom panels. We show the response of macroeconomic aggregates, labor market variables, and prices. Second, we show a decomposition of the policies' effect on consumption by separating the total effect on consumption between 'direct' and 'indirect' effects.[31]

Baseline $\rho_T = 0.5$ and tight monetary policy. Figure 9 shows the response of macroeconomic variables to a rise in fiscal transfers of one percent of GDP. In this case, the 'baseline' monetary policy reacts to inflation and unemployment ($\phi_\pi = 1.25$ and $\phi_U = -1$). This figure shows that, quantitatively, fiscal transfers impact all the macro variables, triggering a boom on impact with a subsequent bust in both cases, progressive and nonprogressive. However, in both cases, transfers have a low total effect on consumption due to the endogenous response of labor income taxes (to finance the transfer partially) and unemployment due to an endogenous response with feedback from consumption and output. Additionally, the endogenous response of the nominal interest rate contributes to the downturn after the shock.

Figure 10 shows, on the other hand, the decomposition of the response of consumption between the direct (that from changes in transfers) and indirect (the other variables). The direct and indirect effects are different. In particular, the direct effect in the progressive case is about 40 percent on impact more significant than in the nonprogressive. Consistent with the evidence in the previous section, the indirect effect becomes more negative in the nonprogressive than the progressive. This latter result is significant because it is evidence of the transfer's large impact and that general equilibrium effects operate in the transfer's transmission.

Loose monetary policy and $\rho_T = 0.5$. Figure 11 shows the same exercise but in a case where the monetary authority does not respond to inflation or unemployment rate. In this case, we assume monetary policy 'coordinates' with the fiscal policy in stimulating the economy by not responding to the fiscal impulse. The consumption response in the progressive case is about twice as large as in the nonprogressive policy. This result is because, as the nominal interest rate does not adjust, the real interest rate falls (due to the rise in inflation and the Fisher equation). This substantial fall in the real interest rate also mutes the response of the tax rate since there is lower debt servicing

31. As in Kaplan and others (2018).

during these periods. The third reason for a significant consumption surge is the fall in the unemployment rate, which we did not observe in the previous case. This result shows the significant general equilibrium effects of having a progressive transfer, which is paid by itself because taxes go down.[32]

Figure 12 shows the decomposition of consumption into direct and indirect effects. We observe that the indirect effects are significant in both cases. The indirect effect of the progressive policy is larger than that of the less progressive one. These results imply that progressive policies have stronger impacts through targeting high MPCs than nonprogressive ones and through the general equilibrium effects.

Figures 11 and 12 provide evidence that the effect of these kinds of policies depends on the monetary policy stance. Therefore, to maximize the response to government transfers, policies must target households with high MPCs, and monetary policy must be loose. Conversely, when monetary policy counteracts these impulses, fiscal policy may become contractionary. These results are present in any New Keynesian model.[33] Finally, having a monetary policy stance that does not entirely counteract the fiscal impulse is not unrealistic, at least in the short run. We observed this policy coordination in times of Covid-19.

Tight monetary policy and Tax-Financed Transfers, $\rho_T = 0$. Figure 13 shows the previous exercises when government finances transfers with taxes $\rho_T = 0$. Even though the responses to the transfer are lower than in the previous exercises, the differences between progressive and nonprogressive transfer are significant. At least on impact, the response of the progressive case is positive, and the nonprogressive is negative. The response of the progressive one is about 0.4 i.e., 30 percent lower than the partially financed transfers. This result arises from the increases in labor income taxes, unemployment, and real interest rate (due to the rise in inflation).

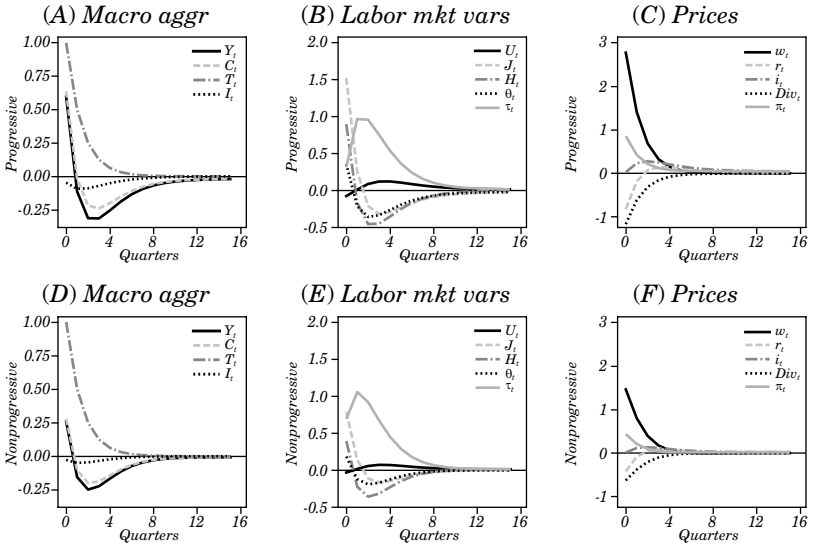
The decomposition in figure 14 shows that, in this case, the indirect effect is negative in both cases, and the direct effect is about the same as the one in the previous cases. However, the general equilibrium effect is less negative for the progressive transfer than for the nonprogressive one.

More Results. In the Appendix, we show other combinations of these exercises. In particular, we find that consumption's response to transfers is the largest in extreme cases of debt-financed transfers

32. This result is also stressed by Angeletos and others (2023).
33. See Woodford (2011).

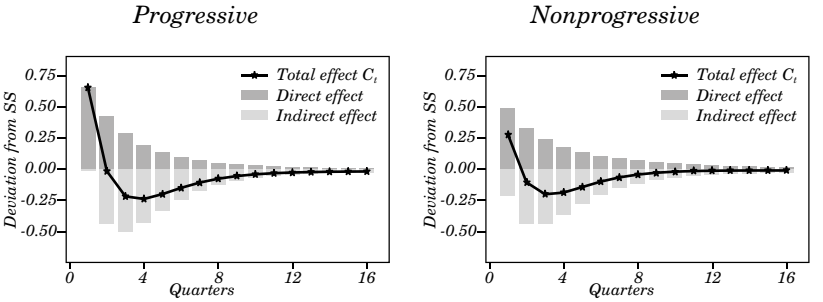
and loose monetary policy. However, in that case, the contribution of progressivity is lower than what we showed above. We also study the effect of muting investment and do not find significant differences between the cases with and without it.

Figure 9. Responses of Aggregate Variables to a 1% Rise in Fiscal Transfers. Tight Monetary Policy and $\rho_T = 0.5$



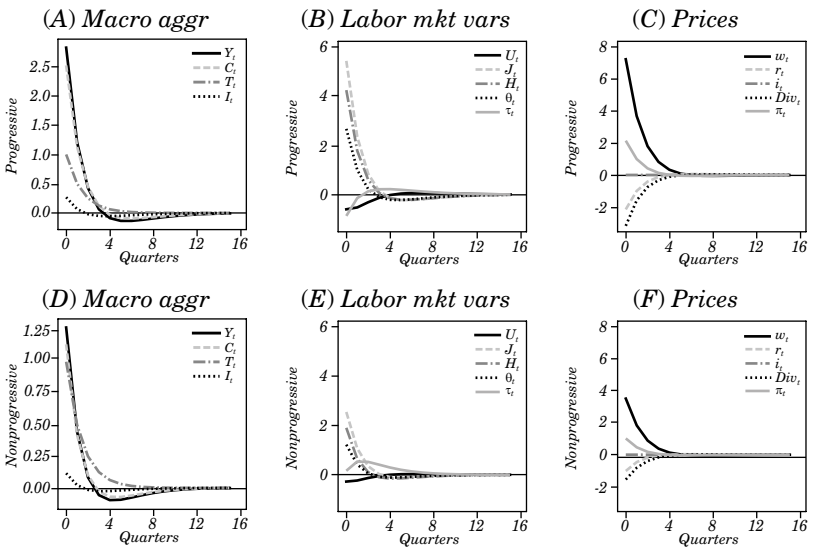
Source: Authors' calculations.

Figure 10. Decomposition of Consumption into Direct and Indirect Effects in Response to Fiscal Transfers. Tight Monetary Policy and $\rho_T = 0.5$



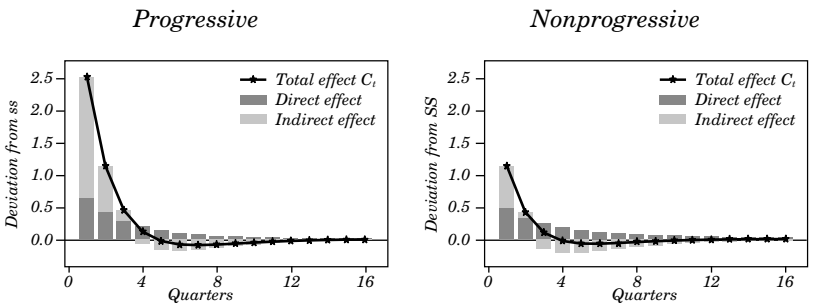
Source: Authors' calculations.

Figure 11. Responses of Aggregate Variables to a 1% Rise in Fiscal Transfers. Loose Monetary Policy and $\rho_T = 0.5$



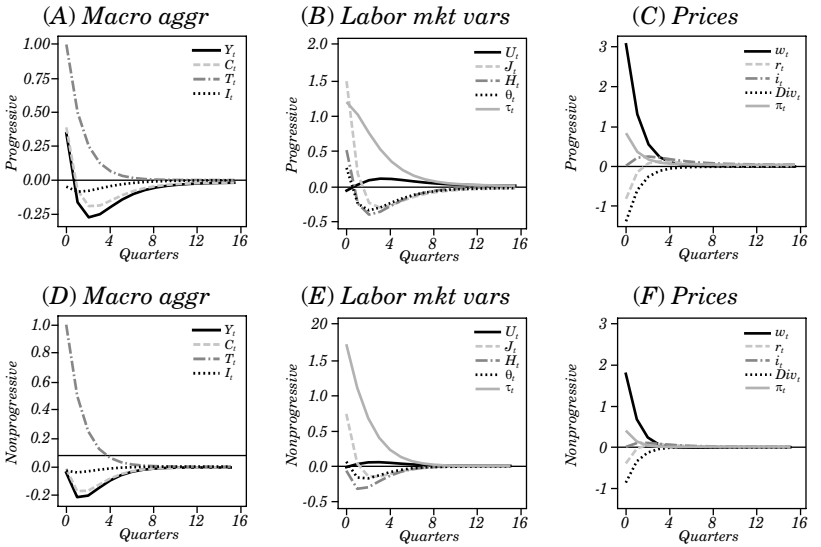
Source: Authors' calculations.

Figure 12. Decomposition of Consumption into Direct and Indirect Effects in Response to Fiscal Transfers. Loose Monetary Policy and $\rho_T = 0.5$



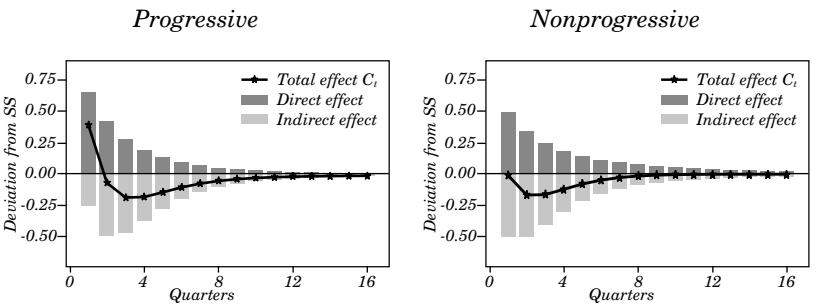
Source: Authors' calculations.

Figure 13. Responses of Aggregate Variables to a 1% Rise in Fiscal Transfers. Tight Monetary Policy and $\rho_T = 0$



Source: Authors' calculations.

Figure 14. Decomposition of Consumption into Direct and Indirect Effects in Response to Fiscal Transfers. Tight Monetary Policy and $\rho_T = 0$



Source: Authors' calculations.

6. Conclusion

In this paper, we build an heterogeneous agents New Keynesian model calibrated for Chile. We test the model implications by comparing its results to empirical facts regarding the effects of fiscal transfers on real activity. These facts derive from two separate estimations. First, fiscal transfers significantly impact GDP and inflation by running a fiscal SVAR as in Blanchard and Perotti (2002).

Second, at a municipal level, we analyze the impact of different fiscal programs between 2018 and 2022. By combining receipts of credit- and debit-card transactions with data on household income and fiscal support, we show that consumption in Chile responds more strongly to policies classified as progressive, suggesting a considerable non-Ricardian behavior of Chilean households.

Our calibrated model can replicate these empirical findings and several other key moments of the Chilean economy. We show that more progressive transfers, associated with higher covariance between allocated funds and household's MPCs, have stronger effects on consumption than less progressive policies. We also show that the magnitude of this differential impact depends crucially on how the government finances its policies and the monetary policy response to the shock. Finally, we decompose the shock's impact between an average and a distributional effect with a statistic that we call the policy slack. We show that a higher transfer progressivity is associated with a higher share of its effect attributed to the distributional channel. A stronger second-round general equilibrium effect compounds the higher direct effect in more progressive policies.

References

Albagli, E., A. Chovar, E. Luttini, C. Madeira, A. Naudon, and M. Tapia. 2023. "Labor Market Flows: Evidence for Chile Using Microdata from Administrative Tax Records." *Latin American Journal of Central Banking*, 4(4): 100102.

Aldunate, R., A. Blanco, A. Fernández, M. Giarda, and G. Navarro. 2023. "The Cross Sectional Labor Market Dynamics After a Foreign Shock." Manuscript, Central Bank of Chile.

Angeletos, G.-M., C. Lian, and C.K. Wolf. 2023. "Can Deficits Finance Themselves?" NBER Working Paper No. 31185.

Auclert, A. 2019. "Monetary Policy and the Redistribution Channel." *American Economic Review* 109(6): 2333–67.

Auclert, A., B. Bardóczy, M. Rognlie, and L. Straub. 2021. "Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models." *Econometrica* 89(5): 2375–408.

Auclert, A., M. Rognlie, and L. Straub. 2018. "The Intertemporal Keynesian Cross." NBER Working Paper No. 25020.

Blanchard, O. and R. Perotti. 2002. "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output." *Quarterly Journal of Economics* 117(4): 1329–68.

Boppart, T., P. Krusell, and K. Mitman. 2018. "Exploiting MIT Shocks in Heterogeneous-Agent Economies: The Impulse Response as a Numerical Derivative." *Journal of Economic Dynamics and Control* 89: 68–92.

Carroll, C.D. 2006. "The Method of Endogenous Gridpoints for Solving Dynamic Stochastic Optimization Problems." *Economics Letters* 91(3): 312–20.

Céspedes, L.F., J. Fornero, and J. Galí. 2013. "Non-Ricardian Aspects of Fiscal Policy in Chile." In *Fiscal Policy and Macroeconomic Performance,* edited by L.F. Céspedes and J. Galí. Santiago, Chile: Central Bank of Chile.

Fagereng, A., M. B. Holm, and G.J. Natvik. 2021. "MPC Heterogeneity and Household Balance Sheets." *American Economic Journal: Macroeconomics* 13(4): 1–54.

Farmer, L.E. and A.A. Toda. 2017. "Discretizing Nonlinear, Non-Gaussian Markov Processes with Exact Conditional Moments." *Quantitative Economics* 8(2): 651–83.

Ferriere, A. and G. Navarro. 2020. "The Heterogeneous Effects of Government Spending: It's All About Taxes." Manuscript, Federal Reserve Board.

Fuster, A., G. Kaplan, and B. Zafar. 2020. "What Would You Do with $500? Spending Responses to Gains, Losses, News, and Loans." *Review of Economic Studies* 88(4): 1760–95.

García, B., S. Guarda, M. Kirchner, and R. TranamiL. 2019. "XMAS: An Extended Model for Analysis and Simulations." Working Papers No. 833, Central Bank of Chile.

García, B., M. Giarda, C. Lizama, and I. Rojas. 2024. "Transmission Mechanisms In HANK: an Application to Chile." *Latin American Journal of Central Banking*.

García, B., M. Giarda, C. Lizama, and D. Romero. 2023. "Time-Varying Expenditure Shares and Macroeconomic Fluctuations." Manuscript, Central Bank of Chile.

Gornemann, N., K. Kuester, and M. Nakajima. 2016. "Doves for the Rich, Hawks for the Poor? Distributional Consequences of Monetary Policy." International Finance Discussion Papers No. 1167, U.S. Board of Governors of the Federal Reserve System.

Guerra-Salas, J., M. Kirchner, and R. Tranamil-Vidal. 2021. "Search Frictions and the Business Cycle in a Small Open Economy DSGE Model." *Review of Economic Dynamics* 39: 258–79.

Guren, A., A. McKay, E. Nakamura, and J. Steinsson. 2020. "What Do We Learn from Cross-Regional Empirical Estimates in Macroeconomics?" NBER Macro Annual No. 35.

Guvenen, F., F. Karahan, O. Serdar, and J. Song. 2019. "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Dynamics?" Manuscript, University of Minnesota.

Hadzi-Vaskov, M., E. Luttini, K. Kuester, and L. RicCi. 2022. "Consumption during Covid-19 in Chile." Manuscript, Central Bank of Chile.

Hagedorn, M., I. Manovskii, and K. Mitman. 2019. "The Fiscal Multiplier." NBER Working Paper No. 25571.

Johnson, D.S., J.A. Parker, and N.S. Souleles. 2006. "Household Expenditure and the Income Tax Rebates of 2001." *American Economic Review* 96(5): 1589–610.

Kaplan, G., B. Moll, and G.L. Violante. 2018. "Monetary Policy According to HANK." *American Economic Review* 108(3): 697–743.

Kaplan, G. and G.L. Violante. 2014. "A Model of the Consumption Response to Fiscal Stimulus Payments." *Econometrica* 82(4): 1199–239.

Kaplan, G. and G.L. Violante. 2018. "Microeconomic Heterogeneity and Macroeconomic Shocks." *Journal of Economic Perspectives* 32(3): 167–94.

Kopecky, K.A. and R.M. Suen. 2010. "Finite State Markov-Chain Approximations to Highly Persistent Processes." *Review of Economic Dynamics* 13(3): 701–14.

McKay, A. 2017. "Time-Varying Idiosyncratic Risk and Aggregate Consumption Dynamics." *Journal of Monetary Economics* 88: 1–14.

McKay, A. and R. Reis. 2016. "The Role of Automatic Stabilizers in the U.S. Business Cycle." *Econometrica* 84(1): 141–94.

Mian, A. and A. Sufi. 2009. "The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis." *Quarterly Journal of Economics* 124: 1449–96.

Mian, A., K. Rao, and A. Sufi. 2013. "Household Balance Sheets, Consumption, and the Economic Slump." *Quarterly Journal of Economics* 128(4): 1687–726.

Misra, K. and P. Surico. 2014. "Consumption, Income Changes, and Heterogeneity: Evidence from Two Fiscal Stimulus Programs." *American Economic Journal: Macroeconomics* 6(4): 84–106.

Mortensen, D.T. and C.A. Pissarides. 1994. "Job Creation and Job Destruction in the Theory of Unemployment." *Review of Economic Studies* 61(3): 397–415.

Nakamura, E. and J. Steinsson. 2014. "Fiscal Stimulus in a Monetary Union: Evidence from US Regions." *American Economic Review* 104(3): 753–92.

Oh, H. and R. Reis. 2012. "Targeted Transfers and the Fiscal Response to the Great Recession." *Journal of Monetary Economics* 59: S50-S64, supplement issue:October 15–16 2010 Research Conference on 'Directions for Macroeconomics: What Did We Learn from the Economic Crises?' Sponsored by the Swiss National Bank.

Parker, J.A., N.S. Souleles, D.S. Johnson, and R. McClelland. 2013. "Consumer Spending and the Economic Stimulus Payments of 2008." *American Economic Review* 103(6): 2530–53.

Patterson, C. 2023. "The Matching Multiplier and the Amplification of Recessions." *American Economic Review* 113(4): 982–1012.

Ravn, M.O. and V. Sterk. 2020. "Macroeconomic Fluctuations with HANK & SAM: An Analytical Approach." *Journal of the European Economic Association* 19(2): 1162–202.

Reiter, M. 2009. "Solving Heterogeneous-Agent Models by Projection and Perturbation." *Journal of Economic Dynamics and Control* 33(3): 649–65.

Tanaka, K. and A.A. Toda. 2013. "Discrete Approximations of Continuous Distributions by Maximum Entropy." *Economics Letters* 118: 445–50.

Tanaka, K. and A.A. Toda. 2015. "Discretizing Distributions with Exact Moments: Error Estimate and Convergence Analysis." *SIAM Journal on Numerical Analysis* 53(5): 2158–77.

Wolf, C.K. 2023. "The Missing Intercept: A Demand Equivalence Approach." *American Economic Review* 113(8): 2232–69.

Woodford, M. 2011. "Simple Analytics of the Government Expenditure Multiplier." *American Economic Journal: Macroeconomics* 3(1): 1–35.

Appendices

Appendix A. Data on Fiscal Aid and Pension Fund Withdrawals

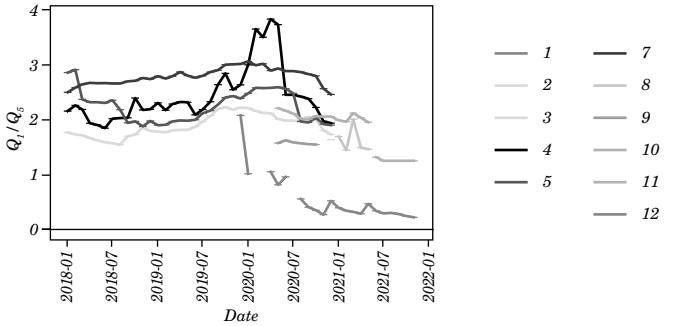
The data on pension fund withdrawals are obtained from the regulator of pension funds. The database is administrative, and we have access to the universe of withdrawals. The database includes the dates of the delivery of the withdrawal, the amount, and an individual identification number.

Until October 2022 there were 11,108,917 requests on the first withdrawal. The average disbursed is 1,422,919 (close to USD 1500). In dollars, the total given amounts to 16.14 billion. The second withdrawal had 9,310,312 requests. The average disbursed was 1460955 pesos (about USD 1500 as well) and the total amounted to USD 13.81 billion. The third withdrawal had 8,866,610 requests in which the average was about USD 1500 as well. The total amount in the third withdrawal was USD 13.05 billion. Therefore, the total amount in withdrawals was USD 43 billion.[34]

The transfer programs available for this study are of different types, sizes, and progressivities. These programs usually target different types of households, focused mainly on poorer ones. We list them as follows: 1. Family help check; 2. Family base check; 3. Christmas Covid-19 check; 4. School homework check; 5. Child homework check; 6. Covid-19 emergency check; 7. Protection check; 8. Covid-19 emergency income; 9. Covid-19 2020 emergency; 10. Guaranteed minimum income; 11. Universal Covid-19 check. These policies have been available since January 2018. These are all direct transfers to individuals, which may be conditional (like homework checks) and unconditional, like Universal Covid-19 checks. These are all targeted to households somehow, as we can observe in figure A.1.

34. Source: Chile' Superintendency of Pensions.

Figure A1. Progressivity of Household Support



Source: Authors' calculations.
Notes: 1. Family help check; 2. Family base check; 3. Christmas Covid-19 check; 4. School homework check; 5. Child homework check; 6. Covid-19 emergency check; 7. Protection check; 8. Covid-19 emergency Income; 9. Covid-19 2020 emergency; 10. Guaranteed minimum income; 11. Universal Covid-19 check; 12. Pension Funds Withdrawals. We exclude policy 6 from the graph because it goes off the chart.

Appendix B. Household Problem

$$V(u_t, z_t, b_{t-1}) = \max_{c_t, b_t} u(c_t) + \beta\big[\, p(\theta_t) V(e_{t+1}, z_{t+1}, b_t) + (1 - p(\theta_t)) V(u_{t+1}, z_{t+1}, b_t)\big]$$

$$\text{s.t. } c_t + b_t = (1 + r_t)\, b_{t-1} + \omega z_t - \tau_t \tau(z_t) + d_t d(z_t)\, b_t \geq 0$$

$$V(e_t, z_t, b_{t-1}) = \max_{c_t, b_t} u(c_t) + \beta\big[(1 - \delta) V(e_{t+1}, z_{t+1}, b_t) + \delta V(u_{t+1}, z_{t+1}, b_t)\big]$$

$$\text{s.t. } c_t + b_t = (1 + r_t)\, b_{t-1} + w_t z_t - \tau_t \tau(z_t) + d_t d(z_t)\, b_t \geq 0.$$

Series on Central Banking, Analysis, and Economic Policies

The Book Series on "Central Banking, Analysis, and Economic Policies" of the Central Bank of Chile publishes new research on central banking and economics in general, with special emphasis on issues and fields that are relevant to economic policies in developing economies. Policy usefulness, high-quality research, and relevance to Chile and other open economies are the main criteria for publishing books. Most research published by the Series has been conducted in or sponsored by the Central Bank of Chile.

Volumes in the series:

1. Análisis empírico del ahorro en Chile
   Felipe Morandé and Rodrigo Vergara, editors
2. Indexation, Inflation, and Monetary Policy
   Fernando Lefort and Klaus Schmidt–Hebbel, editors
3. Banking, Financial Integration, and International Crises
   Leonardo Hernández and Klaus Schmidt–Hebbel, editors
4. Monetary Policy: Rules and Transmission Mechanisms
   Norman Loayza and Klaus Schmidt–Hebbel, editors
5. Inflation Targeting: Design, Performance, Challenges
   Norman Loayza and Raimundo Soto, editors
6. Economic Growth: Sources, Trends, and Cycles
   Norman Loayza and Raimundo Soto, editors
7. Banking Market Structure and Monetary Policy
   Luis Antonio Ahumada and J. Rodrigo Fuentes, editors
8. Labor Markets and Institutions
   Jorge Enrique Restrepo and Andrea Tokman R., editors
9. General Equilibrium Models for the Chilean Economy
   Rómulo Chumacero and Klaus Schmidt–Hebbel, editors
10. External Vulnerability and Preventive Policies
    Ricardo J. Caballero, César Calderón, and Luis Felipe Céspedes, editors
11. Monetary Policy under Inflation Targeting
    Frederic S. Mishkin and Klaus Schmidt–Hebbel, editors
12. Current Account and External Financing
    Kevin Cowan, Sebastián Edwards, and Rodrigo Valdés, editors
13. Monetary Policy under Uncertainty and Learning
    Klaus Schmidt–Hebbel and Carl E. Walsh, editors
14. Banco Central de Chile 1925-1964, Una Historia Institucional
    Camilo Carrasco, editor
15. Financial Stability, Monetary Policy, and Central Banking
    Rodrigo A. Alfaro, editor
16. Monetary Policy under Financial Turbulence
    Luis Felipe Céspedes, Roberto Chang, and Diego Saravia, editors
17. Fiscal Policy and Macroeconomic Performance
    Luis Felipe Céspedes and Jordi Galí, editors
18. Capital Mobility and Monetary Policy
    Miguel Fuentes D., Claudio E. Raddatz, and Carmen M. Reinhart, editors
19. Macroeconomic and Financial Stability:
    Challenges for Monetary Policy
    Sofía Bauducco, Lawrence Christiano, and Claudio Raddatz, editors
20. Global Liquidity, Spillovers to Emerging Markets and Policy Responses
    Claudio Raddatz, Diego Saravia, and Jaume Ventura, editors
21. Economic Policies in Emerging-Market Economies
    Festschrift in Honor of Vittorio Corbo
    Ricardo J. Caballero and Klaus Schmidt–Hebbel, editors
22. Commodity Prices and Macroeconomic Policy
    Rodrigo Caputo and Roberto Chang, editors
23. 25 Años de Autonomía del Banco Central de Chile
    Alberto Naudon D. and Luis Álvarez V., editors
24. Monetary Policy through Asset Markets: Lessons
    from Unconventional Measures and Implications
    for an Integrated World
    Elías Albagli, Diego Saravia, and Michael Woodford, editors
25. Monetary Policy and Global Spillovers: Mechanisms,
    Effects, and Policy Measures
    Enrique G. Mendoza, Ernesto Pastén, and Diego Saravia, editors
26. Monetary Policy and Financial Stability: Transmission Mechanisms
    and Policy Implications
    Álvaro Aguirre, Markus Brunnermeier, and Diego Saravia, editors
27. Changing Inflation Dynamics, Evolving Monetary Policy
    Gonzalo Castex, Jordi Galí, and Diego Saravia, editors
28. Independence, Credibility, and Communication of Central Banking
    Ernesto Pastén and Ricardo Reis, editors
29. Credibility of Emerging Markets, Foreign Investors' Risk Perceptions, and
    Capital Flows
    Álvaro Aguirre, Andrés Fernández, and Şebnem Kalemli–Özcan, editors

Heterogeneity in Macroeconomics: Implications for Monetary Policy

The HANK revolution has been one of the most important developments in modern macroeconomics. This brilliant set of papers from a brilliant set of authors spells out the implications for monetary and fiscal policy. If you want to catch up to the frontier of this literature, you will want this volume.

Mark Gertler
New York University

The contributors to this book are the leading researchers on HANK models and their implications for the monetary policy transmission. Every student and central banker will want to read this outstanding collection of articles.

Monika Piazzesi
Stanford University

This collection of papers showcases both how the HANK machinery already can be used very productively in concrete applications and that it still offers much room for breaking new conceptual ground.

Per Krusell
Stockholm University

"Heterogeneity in Macroeconomics: Implications for Monetary Policy" provides an impressive overview of recent developments in this fast-growing literature, with eight insightful articles written by some of its key contributors. The volume's introduction as well as Sargent's keynote paper both masterfully place HANK models in the history of macroeconomic thought and should be required reading for both academic macroeconomists and practitioners at central banks.

Benjamin Moll
Sir John Hicks Professor of Economics, London School of Economics

banco central
Chile