Wealth, Returns, and Economic Policy

Doctoral Thesis

Nicholas C. Hoffman

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy in Economics

Tepper School of Business
Carnegie Mellon University
Pittsburgh, Pennsylvania

April 21, 2025

Contents

Preface

Abstract

In the United States, households differ not only in their incomes but also in the rates of return that they earn on their investments. My dissertation studies how these differences in returns shape the economy's response to policy and influence the degree of social mobility. The first two chapters examine the implications of return and wealth heterogeneity for monetary and fiscal policy. The third chapter shows that accounting for differences in household returns improves our understanding of mobility—how households change places within the wealth distribution.

In *Chapter 1*, entitled "Redistribution and Reallocation: Monetary Policy with Return Heterogeneity," I study the aggregate and distributional effects of monetary policy when households face idiosyncratic return risk, and make levered investments. When high-return households make levered investments, a decrease in the policy rate redistributes assets towards these wealthier households. This redistributive channel causes monetary policy shocks to increase both aggregate Total Factor Productivity and wealth inequality, as they do in the data. The endogenous change in productivity amplifies the effect of the shock and flattens the Phillips curve, implying that the overall effect of a change in monetary policy is determined by the amount of redistribution that it induces. As a result of two countervailing forces, the power of monetary policy is hump-shaped in the degree of wealth inequality: monetary policy has small effects on output at low and high values of inequality, and larger effects for intermediate wealth inequality. Calibrating my model to the data suggests that the increase in wealth inequality from 1970-2022 can account for some of the decrease in the effect of monetary policy on output documented over the same period.

In *Chapter 2*, entitled "Optimal Taxation of Wealthy Individuals," Ali Shourideh and I characterize the optimal nonlinear taxation of capital income in an environment in which agents earn heterogeneous returns on their investments. Agents in our model can borrow and lend to one another at a common, risk-free rate, and invest in private business with idiosyncratic returns. In a static setting, we demonstrate that income streams from both sources should be taxed at positive, differential rates. Through the tax code, the government controls both the intensive and extensive margins of entrepreneurship, ensuring both that

the correct agents invest a positive amount in their private business, and that these agents invest the socially-optimal amount. For this reason, the optimal distortion on capital income is non-monotonic: the wedge rises for agents whom the government would like to discourage from entry into business ownership (extensive margin), and then falls for agents who do invest to ensure that they do so optimally (intensive margin). In the infinite-horizon extension, we take advantage of a homogeneity result to show that the distortions introduced by the optimal tax code are independent of an agent's history of shocks, and instead depend only on his *current* shock. Going forward, we are calibrating our dynamic economy to US data, in order to weigh in on the optimal level of long-run wealth inequality, as compared to its empirical counterpart.

In *Chapter 3*, entitled "Mobility," Daniel Carroll, Eric R. Young, and I study wealth mobility, the rate at which households change their position relative to one another in the wealth distribution. The US wealth data show a substantial amount of wealth mobility over short horizons. A standard heterogeneous-agents, incomplete markets model with labor income risk generates far less wealth mobility than in the data, even with the addition of standard augmentations to the income process that produce realistic wealth inequality. Agents facing income risk self-insure, accumulating assets to smooth consumption. This self-insurance motive slows the pace with which agents move through the wealth distribution. In the data, we find that families that make large moves through the wealth distribution over short time periods are more likely to receive shocks directly to their wealth, such as capital gains or losses from ownership of stocks or business. We find that incorporating idiosyncratic return risk produces mobility in line with the data. Across models that produce equal wealth inequality, the agents' preferred tax rate on capital income varies with the level of wealth mobility.

Committee Members

Ali Shourideh, *Chair*
Carnegie Mellon University

Ariel Zetlin-Jones
Carnegie Mellon University

Laurence Ales
Carnegie Mellon University

Christopher Sleet, *External Reader*
University of Rochester

Acknowledgements

There are a great many people without whom this dissertation, and the research therein, would never have come to be.

First and foremost, I wish to thank my committee members: Ali Shourideh, Laurence Ales, Ariel Zetlin-Jones, and Christopher Sleet. I thank them not only for providing me with innumerable hours of feedback, suggestions, corrections, and guidance, but also for pushing me to make this body of research the best it can be.

I also wish to extend my thanks to the teachers, colleagues, and friends at Carnegie Mellon who graciously volunteered their time and inputs to my research: Andre Sztutman, Liyan Shi, Kevin Mott, Stephen Spear, Rachel Childers, Martin Michellini, Shengxing Zhang, and the members of the Tepper Econ Reading group.

Prior to graduate school, Daniel Carroll and Ellis Tallman were my mentors at the Federal Reserve Bank of Cleveland. I thank them, as well as Edward Knotek, for giving me the opportunity to grow as a researcher and as a person. I also wish to acknowledge here Eric Young, Dionissi Aliprantis, and all of the other phenomenal economists with whom I was privileged to work at the FRBC.

I dedicate this dissertation to my family. To my wife, Geena: I am forever grateful for the love and support during this journey. To my parents Tom and Janet, my siblings Elliot and Mikaila: thank you for the constant encouragement and inspiration. And finally to Rose: I'm so proud and overjoyed to be your dad.

Dissertation Chapters

Chapter 1

Redistribution and Reallocation: Monetary Policy with Return Heterogeneity[1]

1.1 Introduction

This paper examines how monetary policy affects both aggregates and the distribution of wealth when agents earn heterogeneous rates of return on investments in which they may take leveraged positions. I construct a model wherein the distribution of wealth arises from financial frictions and idiosyncratic return risk, and nominal rigidities give the policy rate influence over this distribution in the short run. I find that a fall in the nominal interest rate redistributes productive assets from lower-return households, to wealthier households who earn higher returns. This redistribution amplifies the effect of monetary policy on the economy by increasing aggregate productivity. The increase in aggregate productivity endogenously flattens the Phillips curve, producing an inflation-output tradeoff more in line with empirical estimates. Changes in wealth inequality influence the efficacy of monetary policy through two channels: the *redistribution* channel, and the *reallocation* channel. These channels pull in opposite directions: the first implies that greater inequality *amplifies* policy, and the second that greater inequality *dampens* it. As a result, the extent of redistribution,

[1]This is my job market paper, and the first chapter of my PhD Thesis at Carnegie Mellon University. I am deeply indebted to Ali Shourideh, Laurence Ales, and Ariel Zetlin-Jones for invaluable guidance. I also extend my sincere gratitude to Andre Sztutman, Liyan Shi, Daniel Carroll, Eric Young, Kevin Mott, David Childers, and participants at the 2024 Midwest Macroeconomics Meetings, the 2024 Washington University in St. Louis Economics Graduate Student Conference, the University of Pittsburgh macroeconomics brownbag, the CMU Marvin Goodfriend Economics Lunch Seminar, and in the Carnegie Mellon Macroeconomics Reading Group for helpful discussions and feedback.

and thus of amplification, is hump-shaped in the degree of wealth inequality at the time of the policy change. In very *equal* and very *unequal* economies, this amplification is minimal; it is larger when wealth inequality takes an intermediate value. Thus, the rise in wealth inequality since the 1970s can partially explain the decrease in the potency of monetary policy observed over the same time period.

Empirical evidence points to leverage as an important channel by which changes in monetary policy manifest in changes in aggregate demand. Cloyne et al. (2020) show that consumption responses are driven by households with outstanding mortgage debt, and Kim and Lim (2020) and Flodén et al. (2021) find that as a result, monetary policy exerts greater effects on demand when households' leverage is higher. These findings indicate that household responses to interest rate policy are driven not by intertemporal substitution, but instead by levered households whose interest outlays change with the short-term policy rate. While mortgages are one reason why a household may have a leveraged portfolio, they are not the only one: here, I assume that households take on debt to make levered investments into an asset whose return is unique to them. With these forces present, the logic in Tobin (1982) is altered: rather than households with high marginal propensities to *consume*, borrowers are wealthy, high-return households with a high marginal propensity to *invest* in their business. As a result, the distributional consequences of monetary policy are aligned with the data: assets are redistributed towards wealthy households, raising their share of total wealth.[2] The presence of return heterogeneity and leverage gives rise to the *redistribution* channel in my model.

Modeling wealth inequality resulting from return heterogeneity, as opposed to labor income heterogeneity alone, provides a better fit to the data. It has long been recognized that models with income risk alone–on which the HANK literature has focused almost exclusively–are insufficient to generate the concentration of wealth seen in the data (see, for instance, the proof in Stachurski and Toda, 2019). As has been well-documented,[3] incomplete-markets models such as mine with *return* risk generate distributions of wealth more in line with their empirical counterparts. More importantly for the purposes of monetary policy, wealth inequality generated by heterogeneous returns to productive assets will imply a *supply-side* role for monetary policy. In my model, more productive investors are the beneficiaries of redistribution following a monetary shock. As an immediate result, *aggregate* productivity increases, following the change in the interest rate. Thus, my model rationalizes the well-known empirical fact that productivity rises in monetary expansions, first documented in Evans (1992)

[2]In the data, the wealth Gini and wealth share of the top 1% both rise in a monetary expansion, see the review in Colciago et al. (2019), or examples in Feilich (2021) and Medlin (2023).

[3]See, for instance, Quadrini (2000b) and Cagetti and De Nardi (2006) for computational examples, and theoretical arguments to this point in Benhabib et al. (2011) and Stachurski and Toda (2019).

and later reaffirmed by Christiano et al. (2005) and then Baqaee et al. (2021). The increase in aggregate productivity is important to capture because it alters the tradeoff between output and inflation, as captured by the Phillips curve: higher productivity in the aggregate implies lower marginal costs, reducing the amount of inflation that accompanies a given increase in output. Redistribution in my model amplifies the aggregate effect of a monetary shock, but in a novel way that is more in line with the data. To the extent that household heterogeneity amplifies monetary policy in existing Heterogeneous-Agent New Keynesian (HANK) models, it does so on the *demand* side: expansions increase the incomes of high-MPC households, resulting in more demand and thus more inflation following a policy change of a given size than in a representative-agent model. In my model, meanwhile, redistribution acts on the *supply* side, by shifting assets to more productive investors. This reallocation mutes the response of inflation, flattening the Phillips curve. It is well-documented that the Phillips curve in the data has indeed flattened over time—see, for instance, Del Negro et al. (2020) or Hazell et al. (2022). As such, my model better captures the policy-relevant tradeoff between output and inflation than existing HANK models. Quantitatively, my calibrated model allowing for redistribution flattens the slope of the output-inflation tradeoff by about 45% relative to the standard representative-producer framework employed in the HANK literature.

As I show, the redistributive channel on its own implies that the effect of a monetary policy shock is increasing in the degree of wealth inequality that is present when the shock hits. Intuitively, the larger is steady-state wealth inequality, the larger must be the scope for agents to maintain high returns over time, accumulating wealth out of repeated windfalls. If household shocks persist in this way in the long run, then the effect of saving also manifests itself in the short run. The initial policy shock generates increases in output and inflation, which as already noted, redistribute assets from poor households to wealthy. If returns are persistent, then the beneficiaries of this initial redistribution are likely to continue to have high-returns in the future, and thus they can save out of the initial redistribution. This savings behavior shifts the allocation of assets towards perennially high-productivity investors, which raises aggregate productivity. As steady-state wealth inequality increases, so too does the size of the amplification through redistribution: as returns increase in persistence and the scope for saving is broadened, the initial redistribution generates a larger and larger boost to productivity, which passes through to investment and output.

There is however a second channel present in my model, the *reallocation* channel, which dampens the effect of monetary policy as inequality increases. A key consequence of underpinning the wealth distribution with return heterogeneity is that the *aggregate* return on capital is now a wealth-weighted average of households' idiosyncratic returns. As wealth becomes more concentrated in the hands of high-return households in the steady state, and

the degree of misallocation decreases, the return on capital aligns with the idiosyncratic returns earned by high-productivity households. As such, following a monetary shock the growth rate of private wealth for productive households is closely aligned with the growth of total capital, and thus the response of the *share* of wealth held by high-return households is muted. The reduction in redistribution implies that monetary policy exerts less of an effect on productivity, and thus, less of an effect on aggregates.

This result in my model stands in contrast to findings elsewhere, and more accurately captures an important facet of the data. A recent group of papers has begun addressing the question of policy transmission with heterogeneous returns: for example, Melcangi and Sterk (2024) model heterogeneous stock market participation, Kekre and Lenel (2022) consider heterogeneities in risk tolerances with a risky asset, and González et al. (2024) and Matusche and Wacks (2021) consider entrepreneurs, as I do. However, in each the conclusion is the same: increasing concentration in the steady state (equivalently, less misallocation) always *amplifies* the effect of monetary policy on aggregates. While it is true that wealth concentration has increased since the 1970s, the amplification of monetary policy is not evident in the data: Boivin and Giannoni (2006) and Boivin et al. (2010) find consistently that monetary policy has become *less* effective since the 1970s and 1980s, a period over which wealth inequality has increased. My results are consistent with this observation: given the redistribution and reallocation channels present in my model, it is possible that an increase in wealth inequality may lead to a *decrease* in the efficacy of policy.

1.1.1 Related Literature

My paper bridges several strands in the literature. The first concerns the role of entrepreneurial and firm-level heterogeneity, in concert with financial market frictions, in determining aggregate activity and the response of the economy to policies. In these models, production is carried out by two or more firms who differ in their productivity.[4] Most closely related to my work is Baqaee et al. (2021), who augment a standard representative-agent New Keynesian model (as in, e.g., Galí, 2015) with firms of heterogeneous productivity. The mechanism at work here is the pass-through of marginal costs to prices: in response to a demand shock that raises marginal costs, high-productivity firms raise their prices by less than do low-productivity firms, leading to a shift in production towards efficient producers and a concomitant increase in overall TFP. My work complements theirs by demonstrating that, in addition to the reallocation of *labor*, monetary policy engenders a reallocation of *capital* between heterogeneous entrepreneurs. As a result, the wealth distribution is also a

[4]See Hopenhayn (2014), for instance, for a survey.

key determinant of the effect of these changes in interest rates.

The second strain of literature to which I contribute is the growing study of monetary policy in economies with household heterogeneity. The early papers in this literature focused on the role of *labor* income risk in altering the effects of a monetary shock relative to a representative-agent framework. McKay et al. (2016) and Kaplan et al. (2018) employ models with heterogeneous agents in the style of Aiyagari (1994b) to address inconsistencies between representative agent New Keynesian models and data, providing theoretical arguments that distributional concerns, precautionary savings motives, and household borrowing constraints are all relevant to a complete understanding the operation of monetary policy. These early "HANK" papers focused on the presence of borrowing-constrained agents, who have higher marginal propensities to consume out of current income shocks than do their unconstrained counterparts. This emphasis on the role of *poor* households has made tremendous strides in examining the role of heterogeneity in MPCs in the transmission of policy, but has left relatively under-explored the role of *wealthy* households in transmitting policy. My contribution is to fill in this gap, and to characterize the role played by rich entrepreneurs in translating a policy shock to aggregate output. I focus on heterogeneity in marginal propensities to *invest*, rather than to *consume*, in transmitting shocks.

My paper belongs to a growing subset of this literature, which aims to develop analytically tractable models combining household heterogeneity and monetary policy, in order to gain a better intuitive understanding of the mechanisms at play and the distinction between these models with heterogeneity and their representative-agent counterparts—for instance, Bilbiie (2021, 2020); Bilbiie et al. (2021); Acharya et al. (2020). As in the early HANK literature, these papers primarily focus on heterogeneity in labor income, and specifically, on the degree to which the income risk of various groups co-moves with the business cycle. I develop a complementary framework, demonstrating the importance of *return* heterogeneity in determining the effects of Central Bank policy.

Of course, my paper is not the first to study monetary policy with entrepreneurs. The celebrated "Financial Accelerator" literature—as in, for example, Carlstrom and Fuerst (1997), Bernanke et al. (1999), and Carlstrom and Fuerst (2001)—uses entrepreneurs as a mechanism to link aggregate activity to financial market fluctuations. The papers in this literature argue that shocks to financial markets affect the asset values of entrepreneurs, the activities of whose firms are linked to their personal wealth. In this way, adverse shocks which affect asset values are amplified, as these shocks reduce entrepreneurial investment, and thus output. Indeed, Kiyotaki (1998) mentions but does not study a simple version of the mechanism which lies at the heart of my paper: with entrepreneurs who are *ex-ante* heterogeneous, a monetary shock can potentially redistribute assets between entrepreneurs. I show, in a model

that retains much of the analytical tractability of these earlier papers, that this is indeed the case. As a result, my work extends and refines the results of this literature: where others have found that aggregate entrepreneurial wealth is an important determinant of monetary transmission, I go further and show that the *distribution* of wealth among entrepreneurs matters as well. Furthermore, relative to existing papers pursuing closed-form analyses of this transmission in the presence of heterogeneity, my approach does not rely on a degenerate distribution or a small, finite number of agents. Instead, my approach allows me to study monetary policy in a world where aggregates depend on the distribution of wealth.

Finally, a recent group of studies has focused on the role of monetary policy in reallocating resources across heterogeneous firms or investors. Ottonello and Winberry (2020) and Jeenas (2023) both study monetary policy with heterogeneous firms, finding that heterogeneity in firm balance sheets affects monetary transmission. In the former, accommodative shocks shift investment towards firms with lower default risk, and the latter that these shocks reallocate towards firms with more liquid balance sheets. I complement their analyses by showing that firm productivity is also a meaningful dimension along which monetary policy acts to reallocate assets. Kekre and Lenel (2022) show that with heterogeneity in risk tolerance, a decrease in interest rates decreases risk premia by shifting wealth to agents with higher willingness to invest in risky assets. Similarly, Melcangi and Sterk (2024) argue that monetary expansions increase wealth inequality by benefitting the small subset of the population active in the stock market, and conversely, have a larger effect when wealth is more concentrated in the hands of stockholders, as has increasingly become the case in the US data. Both of the latter two models would predict that, as wealth inequality increases, the efficacy of monetary policy should always increase. However, the data do not support this idea: over the past four decades, as wealth inequality has increased, what evidence there is of a change in the strength of monetary policy has shown that its potency has *decreased* (see e.g. Boivin et al., 2010; Boivin and Giannoni, 2002). While there are many ways of accounting for this potential change, my model provides a rationalization: beyond a certain point, increases in inequality dampen the redistributive channel of policy by making it harder for wealthy entrepreneurs to earn returns in excess of the aggregate return on capital. This channel is obscured when the risky asset in question evolves exogenously, rather than as a function of allocations.

The two papers most similar to mine are Matusche and Wacks (2021) and González et al. (2024). Matusche and Wacks (2021) construct a model of heterogeneous entrepreneurs who face diminishing returns, and show that an accommodative shock shifts wealth towards wealthier entrepreneurs, thereby increasing aggregate investment and amplifying the response of the economy to the monetary shift. They also demonstrate by means of a numer-

ical example that shifting wealth towards entrepreneurs in the steady state leads to larger effects of monetary policy. González et al. (2024) also demonstrate that an easing in monetary policy shifts resources towards firms with higher productivity, and derive a prescription for optimal monetary policy in this economy. They also provide empirical evidence using a representative sample of the universe of Spanish firms—both public and private—that heterogeneity in marginal returns to capital better explains the response of investment to monetary shocks, as opposed to other balance sheet, revenue, or productivity measures. My contribution, relative to these papers, is to study both the redistribution and reallocation channels by which inequality determines the power of monetary policy. The common conclusion among these papers is that further increases in wealth inequality increase the effect of a monetary policy change—that is, they study the channel that I term *redistribution*. My paper, however, points to the *reallocation* channel, by which increases in inequality imply a smaller effect of policy on aggregates. This second channel helps reconcile models such as this with the empirical evidence, outlined above, suggesting that the efficacy of monetary policy has decreased over the past 50 years as wealth inequality has widened.

My paper proceeds as follows. In section 1.2 I construct my model, and describe optimal behavior of all of the agents therein. I also study selected properties of the steady state in my model. I will argue that the wealth distribution in the steady state, and the assumptions underlying it, are crucial in shaping the transmission of monetary policy. As such, it is important to build an understanding of how the steady state varies across parameters. Section 1.3 contains my main results: responses of aggregate variables to an unanticipated change in monetary policy. In addition, Section 1.3 studies how these responses change in the wealth distribution at the time of the shock. Section 1.6 concludes.

1.2 Model

I consider a model with two types of agents: workers and entrepreneurs. It is not possible for a worker to become an entrepreneur, or vice versa. All workers are identical. Entrepreneurs are indexed by $i \in [0, 1]$. If entrepreneur i chooses to be active in period t, she hires workers on the spot labor market. The private firm i produces output according to

$$y_{it} = \max_{n_{it}^d} \left(z_{it} k_{it}\right)^{\alpha} \left(n_{it}^d\right)^{1-\alpha} \tag{1.1}$$

Here, k_{it} is the capital stock of household i, and n_{it}^d is the quantity of labor hired by firm i on the spot market. Entrepreneurs produce a homogeneous good y, which may be used for either consumption or capital investment. The entrepreneurial talent or productivity of

household i in period t is given by z_{it}.

1.2.1 Entrepreneurs' Problem and Collateral Constraints

I follow papers such as Buera et al. (2011) and Moll (2014) in assuming that entrepreneurs have the ability to save in one of two assets: capital used to run their firm, and risk-free nominal bonds.[5]Entrepreneurs maximize their lifetime expected utility

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t U^e \left(c_{it}\right) \tag{1.2}$$

subject to their nominal budget constraint,

$$P_t \left\{c_{it} + q_{it}\right\} = P_{tx} y_{it} - W_t l_{it} - \left(1 + i_t\right) D_{it} + D_{it+1} \tag{1.3}$$

where P_t is the aggregate price level, W_t the *nominal* wage (taken as given by the firm), and i_t the nominal interest rate between $t-1$ and t, set by the monetary authority at time $t-1$. D_{it+1} is the quantity of nominal bonds *issued* by firm i at time t; so $D_{it} < 0$ indicates that the household is a net lender, or purchaser of bonds. These agents split their total income between purchases of consumption goods c_{it} and investment (capital) goods q_{it}, which I assume are identical and hence share nominal price P_t. Entrepreneur i's capital stock evolves according to

$$k_{it+1} = \left(1 - \delta\right) k_{it} + q_{it} \tag{1.4}$$

where q_{it} is the quantity of investment goods purchased. Following Buera and Moll (2015), I assume that households are subject to a collateral constraint of the form

$$D_{it+1} \leq \theta P_t k_{it+1}, \qquad \theta \in [0, 1] \tag{1.5}$$

This collateral constraint implies that only a proportion θ of the nominal value of the next-period capital stock may be externally financed. As I will demonstrate, this framework is isomorphic to one in which entrepreneurs save only in risk-free bonds, and borrow their entire period-t capital stock from an intermediary. I also follow Buera and Moll (2015) in assuming that next-period productivity z_{it+1} is revealed to household at the end of period t, before they issue bonds D_{it+1}.

[5]As will become clear later, these bonds are risk-free in the sense that their *nominal* rate of return between periods t and $t+1$ is predetermined in period t. However, their *real* return is subject to risk in the event of unanticipated inflation between these two periods.

Each entrepreneurial household maximizes its nominal capital income:

$$P_t y_{it} = \max_{n_{it}^d} \; P_{tx} \left(z_{it} k_{it}\right)^{\alpha} \left(n_{it}^d\right)^{1-\alpha} - W_t l_{it} \tag{1.6}$$

The following lemma is a well-known result in problems such as (1.6):

Lemma 1. *Entrepreneurial labor demand is linear in capital:*

$$n_{it}^d = \left(\frac{1-\alpha}{W_t/P_t} \frac{P_{tx}}{P_t}\right)^{\frac{1}{\alpha}} z_{it} k_{it} \tag{1.7}$$

Lemma 1 is the result of the fact that the problem in (1.6) is static: entrepreneurs hire labor on the spot market to maximize their profit given their state (z_{it}, k_{it}) and the wage and price W_t and P_{tx}, which they take as given. Defining

$$\omega_t \equiv \alpha \left(\frac{1-\alpha}{w_t}\right)^{\frac{1-\alpha}{\alpha}} p_{tx}^{\frac{1}{\alpha}} \tag{1.8}$$

where w_t is the real wage and p_{tx} the real entrepreneurs' price, the budget constraint can be written as

$$P_t c_{it} + P_t k_{it+1} = P_t \left[\omega_t z_{it} + (1-\delta)\right] k_{it} - (1+i_t) D_{it} + D_{it+1} \tag{1.9}$$

It is useful to write the entrepreneurs' budget constraint in real terms. To do so, define real bond issuance as

$$d_{it+1} \equiv \frac{D_{it+1}}{P_t} \tag{1.10}$$

With this definition, the budget constraint for entrepreneurial household i in real terms is

$$c_{it} + k_{it+1} = \left[\omega_t z_{it} + (1-\delta)\right] k_{it} - (1+r_t) d_{it} + d_{it+1} \tag{1.11}$$

Here, r_t is the time-t ex-post real interest rate, defined by the Fisher equation:

$$1 + r_t = (1+i_t) \frac{P_{t-1}}{P_t} = \frac{1+i_t}{1+\pi_t} \tag{1.12}$$

Note that this interest rate depends on the realized inflation rate π_t. Define the real net worth as

$$a_{it} = k_{it} - d_{it} \tag{1.13}$$

With this definition, I can write the borrowing constraint in real terms:

$$d_{t+1} \leq \theta k_{t+1} \tag{1.14}$$

or substituting the definition of net worth a_t,

$$k_{t+1} \leq \lambda a_{t+1}, \qquad \lambda \equiv \frac{1}{1-\theta} \tag{1.15}$$

Then, I have the following lemma, similar to Moll (2014) and Buera and Moll (2015):

Lemma 2. *Entrepreneurs' capital and bond choices are corner solutions:*

$$k_{t+1} = \begin{cases} \lambda a_{t+1} & z_{t+1} \geq \underline{z}_{t+1} \\ 0 & z_{t+1} < \underline{z}_{t+1} \end{cases} \tag{1.16}$$

$$d_{t+1} = \begin{cases} (\lambda - 1)\, a_{t+1} & z_{t+1} \geq \underline{z}_{t+1} \\ -a_{t+1} & z_{t+1} < \underline{z}_{t+1} \end{cases} \tag{1.17}$$

where the cutoff $\underline{z}$ is such that

$$\omega_{t+1}\underline{z}_{t+1} = r_{t+1} + \delta \tag{1.18}$$

Lemma 2 shows that entrepreneurs are divided into two groups: those above the productivity threshold in (2), who are active, and those below it, who are inactive. Active entrepreneurs, who earn excess returns on their investment above the risk-free rate, borrow up to their limit and are thus bound by the collateral constraint. Inactive entrepreneurs save at the risk-free rate, lending to active entrepreneurs. Due to the linearity of the production technology, the cutoff productivity $\underline{z}_t$ is independent of wealth; instead, $\underline{z}_t$ is a linear function of the risk-free rate r_t and ω_t, which can be thought of as the private return per effective unit of capital zk.

1.2.2 Nominal Rigidities

To introduce nominal rigidities while maintaining tractability, I follow Bernanke et al. (1999) in assuming a three-tiered production structure. Entrepreneurs produce a homogenous good x_t, which is then sold to retailers. Retailers, a continuum of whom are indexed by $j \in [0, 1]$, in turn costlessly differentiate these goods. Retailers sell their output y_{tj} to a final good producer, who aggregates them using a CES technology:

$$Y_t = \left[\int_0^1 y_{tj}^{\frac{\varepsilon-1}{\varepsilon}} \right]^{\frac{\varepsilon}{\varepsilon-1}} \tag{1.19}$$

This assumption on the structure of production allows me to introduce price stickiness in a way that preserves the tractability of the entrepreneurs' problem. It is analytically convenient

to assume that entrepreneurs are price takers; otherwise, their investment and savings choices would be intermingled with a forward-looking pricing problem, which would complicate my model without providing any obvious upside.

Optimal behavior by the final good aggregator in (1.19) implies that the demand for variety j is

$$y_{t,j} = \left(\frac{P_{t,j}}{P_t}\right)^{-\varepsilon} Y_t \tag{1.20}$$

where

$$P_t = \left(\int P_{t,j}^{1-\varepsilon}\right)^{\frac{1}{1-\varepsilon}} \tag{1.21}$$

is the overall price level. Retailer j produces output y_{tj} according to

$$y_{tj} = x_{tj} \tag{1.22}$$

therefore, their marginal cost is $m_t = p_{tx}$. In addition, retailers incur Rotemberg (1982)-style quadratic adjustment costs to change their price:

$$\Theta\left(P_{tj}, P_{tj}^R\right) = \frac{\theta}{2}\left(\frac{P_{tj}}{P_{tj}^R} - 1\right)^2 Y_t \tag{1.23}$$

P_{tj}^R is the time-t reference price for retailer j. Typically, $P_{tj}^R = P_{t-1,j}$; that is, the firm incurs an adjustment cost when it wants to update its price relative to its own lagged price. As shown in lemma 3 below, I consider this case, as well as a "static" case chosen for additional gains in tractability.

Defining inflation as

$$\pi_t \equiv \frac{P_t}{P_{t-1}} - 1 \tag{1.24}$$

the following lemma describes the behavior of inflation over time:

Lemma 3. *Inflation evolves according to the New Keynesian Phillips Curve, which arises under optimal behavior by retailers:*

$$\pi_t = \frac{\varepsilon}{\theta}\left(p_{tx} - m^*\right) + \beta_f \mathbb{E}_t \pi_{t+1} \tag{1.25}$$

Here, $m^* = \varepsilon/(\varepsilon - 1)$ is the inverse of the optimal markup in the absence of price rigidities, and β_f is the rate at which retailers discount future profits. In the two cases that I consider,

$$\beta_f = \begin{cases} \beta & P_{tj}^R = P_{t-1,j} \\ 0 & P_{tj}^R = P_{t-1} \end{cases} \tag{1.26}$$

The intuition in Lemma 3 is standard. Iterating forward on equation (1.25) gives

$$\pi_t = \frac{\varepsilon}{\theta} \sum_{s=0}^{\infty} \beta_f^s \left[p_{t+s,x} - m^* \right] \tag{1.27}$$

In the presence of price stickiness, retailers raise their prices when they believe that future marginal costs will exceed their long-run optimal level—equivalently, retailer raise current prices when they believe that, in the future, markups will fall *below* their long-run optimal level. Additionally, the retailers' discount rate depends on the reference price on which their adjustment cost is based. Under the typical assumption that a retailer's adjustment costs are a function of the deviation between its own current and lagged prices ($P_{tj}^R = P_{t-1,j}$), retailers share a discount factor with the households by whom they are owned. If, on the other hand, the reference price for the firm is the lagged aggregate price, then the New Keynesian Phillips Curve is static, and current inflation depends only on the current deviation of marginal cost from its optimal long-run level. As noted in Bilbiie (2021), this assumption is empirically unrealistic; it implies that firms do not consider future profits in setting their current price. Nevertheless, it allows for a contemporaneous tradeoff between inflation and real output, and as such can offer a convenient alternative to the forward-looking assumption.

1.2.3 Equilibrium

There are two actors needed to close the model: workers, and a monetary authority. For expositional purposes, I assume for the time being that workers—who supply their labor to entrepreneurs at real wage w_t—cannot borrow or save, and are thus constrained to be hand-to-mouth. Workers are, however, free to adjust their labor supply in response to movements in the real wage. Worker households are all identical, and have preferences as in Greenwood et al. (1988):

$$U^w \left(C_t^w, N_t^w \right) = \frac{1}{1-\gamma} \left(C_t^w - \frac{\left(N_t^w \right)^{1+\eta}}{1+\eta} \right)^{1-\gamma} \tag{1.28}$$

In addition, worker households own the retailers, and receive the profits of these firms as real dividends T_t. Workers' budget constraint in real terms is simply $C_t = w_t N_t + T_t$. The labor supplied by the households is given by

$$N_t^w = w_t^{1/\eta} \tag{1.29}$$

I follow the literature in assuming that the monetary authority sets the nominal interest

rate i_t according to a Taylor rule:

$$i_{t+1} = \overline{r} + \phi_\pi \pi_t + \nu_t \tag{1.30}$$

Recall that i_t dictates the nominal cost that an entrepreneur pays for outside financing. In order to ensure notational consistency, I date all interest rates according to when they are earned, rather than when they are set. As such, the nominal rate i_{t+1} in (1.30) is set at time t, and dictates the interest rate on nominal debt issued in period t and maturing in period $t+1$. The term ν_t is an exogenous, stochastic innovation; I will use this shock to measure the impact of monetary policy in my model.

Definition 1. *An equilibrium is a sequence of prices* $\{P_t, P_{tx}, W_t\}$*, aggregates* $\{C_t^e, C_t^w, N_t, Y_t, K_t, Z_t\}$*, interest rates* $\{i_t, r_t\}$*, a path for inflation* π_t*, a sequence of aggregate shocks* ν_t*, and a sequence of distributions* $\{g_t(a,z)\}$ *over the idiosyncratic states for entrepreneurs such that:*

1. Entrepreneurs, workers, retailers, and the final good producer all maximize their respective objectives,
2. The monetary authority sets the nominal interest rate in accordance with the Taylor rule in (1.30), given an exogenous sequence for the shock ν_t,
3. Prices clear markets:

$$K_t = \int_0^{\overline{z}} \int_0^{\infty} a g_t(a,z)\, da dz \tag{1.31}$$

$$N_t^w = N_t^d \tag{1.32}$$

$$C_t^e + C_t^w + K_{t+1} + \Theta(\pi_t) = Y_t + (1-\delta) K_t \tag{1.33}$$

I will study the properties of the equilibrium defined above using *wealth shares*:

$$s_t(z) \equiv \frac{1}{K_t} \int_0^{\infty} a g_t(a,z)\, da \tag{1.34}$$

As in Moll (2014), among others, the wealth share $s_t(z)$ denotes the share of aggregate wealth held by agents of type z. There are a number of reasons why these objects are a convenient tool for studying the behavior of the model. First, the shares $s_t(z)$ can be thought of as a density: they are nonnegative for all z, and integrate to one: $\int_0^{\overline{z}} s_t(z)\, dz = 1$ for all t. As such, I can define the analogous cumulative share:

$$S_t(z) \equiv \int_0^{z} s_t(\hat{z})\, d\hat{z} \tag{1.35}$$

Second, note that because returns are linear in wealth, individual wealth follows a random

growth process.[6]As a result, the joint distribution $g_t(a, z)$ does *not* admit a stationary measure: the log of individual wealth a_t follows a random walk, and thus the cross-sectional variance of a_t grows without bound in t. However, it can be demonstrated that the *wealth shares* $s_t(z)$ do admit a stationary measure. This result is convenient: it allows me to study the long-run properties of my model without needing to augment the model with an assumption to deliver a stationary measure over wealth, such as random death and annuity markets (as in Gouin-Bonenfant and Toda, 2019) or a hard borrowing limit.

Aggregation

Using the definition of wealth shares in (1.34), aggregate quantities are easily derived:

Proposition 1. *Aggregate quantities satisfy*

$$Y_t = (Z_t K_t)^{\alpha} N_t^{1-\alpha} \tag{1.36}$$

$$K_{t+1} = \beta \left\{ \alpha p_{tx} Y_t + (1-\delta) K_t \right\} \tag{1.37}$$

Aggregate productivity is a function of the wealth distribution $s_t(z)$:

$$\begin{aligned} Z_t &= \frac{\int_{\underline{z}_t}^{\overline{z}} z s_t(z)\, dz}{\int_{\underline{z}_t}^{\overline{z}} s_t(z)\, dz} \\ &= \mathbb{E}_{s_t}\left[z | z > \underline{z}_t \right] \end{aligned} \tag{1.38}$$

Given a path for wealth shares $s_t(z)$, the cutoff productivity $\underline{z}_t$ is pinned down by capital market clearing:

$$1 = \lambda \left(1 - S_t(\underline{z}_t) \right) \tag{1.39}$$

Factor prices are

$$w_t = (1-\alpha) p_{tx} \left(\frac{Z_t K_t}{N_t} \right)^{\alpha} \tag{1.40}$$

$$\mathbb{E}_{t-1} r_t = \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t} \right)^{1-\alpha} \frac{\underline{z}_t}{Z_t} - \delta \tag{1.41}$$

[6]See Gabaix (2009) for a study of random growth processes in economics, and Benhabib et al. (2015b) for an example of how this process gives rise to wealth distributions in models that share the "fat-tailed" (Pareto) nature of their empirical counterparts.

Returns are given by

$$\omega_t = \alpha p_{tx} \left(\frac{N_t}{Z_t K_t} \right)^{1-\alpha} \tag{1.42}$$

$$R_{tK} = 1 - \delta + \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t} \right)^{1-\alpha} \tag{1.43}$$

The return to an entrepreneur of type z is given by

$$R_t(z) = 1 + i_{t-1} - \pi_t + \lambda \begin{cases} 0 & z < \underline{z}_t \\ \omega_t z - (i_{t-1} - \pi_t + \delta) & z > \underline{z}_t \end{cases} \tag{1.44}$$

Proposition 1 has largely the same interpretation as its counterparts in Moll (2014) and Buera and Moll (2015); I refer the reader there for excellent discussions. Equations (1.36) and (1.37) show that this economy behaves as one with a representative firm, with the key difference being that aggregate TFP is an endogenous result of the wealth distribution among entrepreneurs, as in (1.38). Per equation (1.40), the real wage is given by the real value of the aggregate marginal product of labor hired by entrepreneurs. The same is generally *not* true for the ex-ante expected risk-free real rate $\mathbb{E}r_t$: in equation (1.41), this object is equal to the aggregate marginal return to capital, weighted by $\underline{z}_t / Z_t$.[7] In this economy, capital market frictions prevent the return from investment to be equated with outside savings. As a corollary, the two are equated in the case that $\underline{z}_t = Z_t$, which is the case when capital markets are frictionless, $\lambda = \infty$. Note as well that both factor prices, as well as the returns R_{tK} and ω_t, move with the price paid to entrepreneurs by retailers for their goods.

In equation (1.43), the aggregate return to capital is derived from

$$R_{tK} = \int_0^{\overline{z}} R_t(z)\, s_t(z)\, dz \tag{1.45}$$

$$= \mathbb{E}_{s_t} \left[R_t(z) \right] \tag{1.46}$$

Thus, the aggregate return on capital is the average return across all entrepreneurs, weighted by their respective wealth shares. Finally, entrepreneurs' returns, per equation (1.44), exhibit a few key properties that will later drive my results. First, entrepreneurs with $z > \underline{z}_t$ are able to earn excess returns above the ex-ante risk-free rate $i_{t-1} - \pi_t$, due to their ability to make

[7] The expectations operator indicates that the ex-post real rate is subject to inflation risk. In the absence of nominal rigidities, Equation (1.41) would always hold. In order to study the redistributive effects of unanticipated inflation, I leave open the possibility that the ex-post real rates may differ from their ex-ante expectations. In the event that inflation is not equal to its ex-ante expectation, this equation will hold for the expected risk-free rate, upon which time-t contracts are based, but *not* for the ex-post rate r_{t+1}.

leveraged investments into their inside firm. In partial equilibrium, equation (1.44) previews the differential impact of a change in real rates on entrepreneurs of different productivities. Returns can also be written as

$$R_t(z) = \begin{cases} 1 + i_{t-1} - \pi_t & z \leq \underline{z}_t \\ 1 - (\lambda - 1)(i_{t-1} - \pi_t) + \lambda(\omega_t z - \delta) & z > \underline{z}_t \end{cases} \tag{1.47}$$

From (1.47) it is immediately obvious that a fall in the real rate $i_{t-1} - \pi_t$ lowers the returns of savers, who *earn* this rate, and raises that of active entrepreneurs, who *pay* this return to borrow capital. Additionally, the expression of returns in (1.47) makes clear the role of inflation in driving redistribution in this model: an unexpected increase in π_t reduces nominal obligations, redistributing from low types (lenders) to high types (borrowers). Crucially, π_t is realized one period *after* $\underline{z}_t$ has been determined: entrepreneurs operate their firms at time t with capital stock chosen at time $t-1$, and thus agents cannot switch from being inactive to active following unexpected inflation.

Before studying the properties of this equilibrium, I note again its generality in understanding the role of return heterogeneity in transmitting monetary policy. From the perspective of policy, the key features of this model are that (i) agents earn heterogeneous returns and (ii) take on leverage, (iii) the aggregate return to capital is a wealth-weighted average of returns above a cutoff for participation in the risky asset, and (iii) this cutoff is endogenously determined so as to clear the capital market. The most restrictive assumption that I make is that these firms are unable to issue equity, and thus must fund all investments above their existing wealth with debt. Nevertheless, these forces are likely to be present even when households have risky returns resulting from ownership of *public* firms as well.

1.2.4 Persistence in Returns

To study analytically the effects of a monetary shock in my model economy, I make the following assumption on individual entrepreneurs' productivities:

Assumption 1. *Individual entrepreneurial productivities are distributed according to some differentiable, time-invariant function $F(z)$. With probability p, an entrepreneur will maintain his productivity from one period to the next, $z_{t+1} = z_t$. With probability $1-p$, meanwhile, he draws his next-period productivity at random from the time-invariant distribution given by $F(z)$.*

Assumption 1 allows for gains in tractability while maintaining rich heterogeneity in the

model. With this assumption, the autocorrelation of z_t and z_{t+1} is parsimoniously given by

$$\rho(z_t, z_{t+1}) = p \tag{1.48}$$

This persistence is also incorporated in an appealing way: conditional on $z_{t+1} \neq z_t$, the distribution of z_{t+1} is independent of z_t. Thus, I maintain many of the desirable properties of an IID process while still allowing for a positive autocorrelation in returns. As pointed out by Moll (2014), persistence in returns is the empirically relevant case (see, e.g., DeBacker et al., 2023), and the case that leads to a correlation between entrepreneurs' wealth and their productivity. With autocorrelated productivities, more productive entrepreneurs will accumulate more wealth over time, using their own wealth as a complement to outside credit. I also assume that the distribution $F(z)$ has support on $\mathcal{Z} = [0, \overline{z}]$ with $\overline{z} < \infty$.

Under assumption 1, the behavior of the wealth shares $s_t(z)$ defined in (1.34) can be characterized in a clean and intuitive fashion:

Proposition 2. *The wealth share of type z, $s_t(z)$, evolves according to*

$$s_{t+1}(z) = p\frac{R_t(z)}{R_{tK}}s_t(z) + (1-p)f(z) \tag{1.49}$$

where R_{tK} is the aggregate (wealth-share weighted average) return to capital, as defined in Lemma 1.

Proposition (2) has an intuitive interpretation. There are two sources of change in the wealth share $s_{t+1}(z)$: entrepreneurs who retain their type ($z_{t+1} = z_t$), and entrepreneurs who transition to type z at $t+1$ from some other type ($z_t \neq z_{t+1}$). For each source of change, the sign of its contribution (whether it increases or decreases $s_{t+1}(z)$ relative to $s_t(z)$) depends on the returns of the agents in question, relative to the aggregate return on capital. For agents who retain their type: if the time-t return $R_t(z)$ is greater than the aggregate return to capital, then the wealth of agents of type z grows faster than the overall capital stock, and their share increases. Agents transitioning to type z from some other z' (the $(1-p)f(z)$) term in (1.49) on average earn, by definition, the aggregate return R_{tK}, hence the coefficient of 1 on this term.

1.2.5 Steady State

One of the primary questions of interest in my paper is: how does the distribution of wealth at the time of a policy shock, influence the economy's aggregate response to that shock? Here I analyze properties of the steady state that will be critical to answering this

question. I assume that prior to the unexpected change in monetary policy ν_0, the economy is in its *long-run, zero-inflation steady state*. Equation (1.34) implies that in the steady state, wealth shares are given by

$$s(z) = \frac{1-p}{1-\beta p R(z)} f(z) \tag{1.50}$$

Equation (1.50) uses the fact that in the steady state, the return to capital is pinned down by the entrepreneurs' discount factor:

$$R_K = 1 - \delta + \alpha p_x Z^{\alpha} (N/K)^{1-\alpha} = \frac{1}{\beta} \tag{1.51}$$

The wealth shares $s(z)$ are a mixture of the marginal distribution $f(z)$, which determines the mass of agents of type z, and the steady-state returns earned by type z, given by

$$R(z) = 1 + \alpha p_x Z^{\alpha} \left(\frac{N}{K}\right)^{1-\alpha} \frac{z_\lambda(z)}{Z} - \delta \tag{1.52}$$

where

$$z_\lambda(z) \equiv \begin{cases} z + (\lambda - 1)(z - \underline{z}) & z > \underline{z} \\ \underline{z} & z \leq \underline{z} \end{cases} \tag{1.53}$$

is the effective return to an entrepreneur of type z, taking into account leverage. The real interest rate is the effective productivity of the *marginal* entrepreneur:

$$r = \alpha p_x Z^{\alpha} \left(\frac{N}{K}\right)^{1-\alpha} \frac{\underline{z}}{Z} - \delta \tag{1.54}$$

The price for entrepreneurial goods is equal to its optimal level in the absence of nominal rigidities:

$$p_x = 1/\mathcal{M}^* = \frac{\varepsilon - 1}{\varepsilon} \tag{1.55}$$

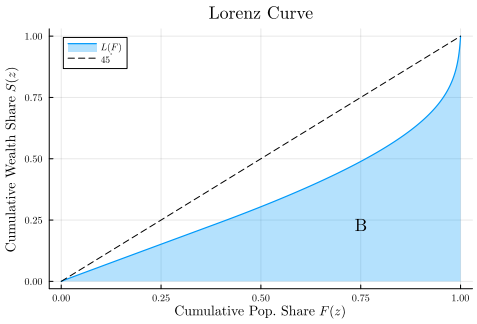
Aggregate productivity is determined by the allocation of wealth among entrepreneurs, as described by the shares $s(z)$:

$$Z = \lambda \int_{\underline{z}}^{\overline{z}} z s(z)\, dz \tag{1.56}$$

Equation (1.56) uses the fact that capital market clearing again implies $1 = \lambda(1 - S(\underline{z}))$.

Equations (1.50)-(1.52) show the forces that give rise to the steady state wealth distribution. The long-run wealth shares can be thought of as mixtures of the marginal density $f(z)$ and returns $R(z)$, with the weighting determined by p. Consider varying p from zero to one.

Figure 1.1: Lorenz Curve



At $p = 0$, productivity shocks are IID: each period, entrepreneurs draw a new z_{it+1} that is independent of their current productivity z_{it}. When this is the case, $s(z) = f(z)$; the wealth shares are equal to the marginal density. In this case, returns and wealth are uncorrelated: entrepreneurs who receive a high productivity shock today are unlikely to receive another high shock tomorrow, and are thus unable to accumulate wealth from a series of shocks. At the other extreme, $p = 1$, and entrepreneurs' productivities persist perfectly throughout time; heterogeneity is permanent. When this is the case, the wealth distribution will be $s(z) = \delta_z(\overline{z})$, where $\delta(\cdot)$ is the Dirac measure. In the long run, entrepreneurs of type $z$end up holding all of the wealth in the economy. As a result, the long-run risk-free rate and return to capital will equate with the returns of entrepreneurs of type $\overline{z}$,

$$R_K = R = R(\overline{z}) \tag{1.57}$$

For intermediate values of p, the wealth shares will be a transformation of the population weights f.

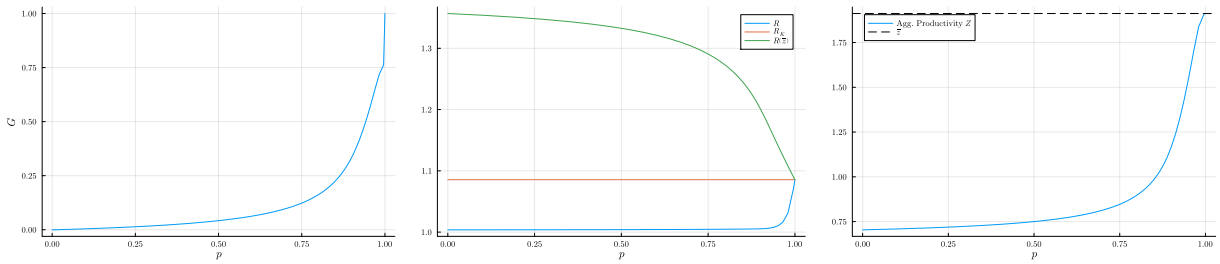
I introduce a notion of the Gini coefficient to measure inequality. The Gini coefficient at time t is defined by

$$G_t = 1 - 2\int_0^1 S_t\left[F^{-1}(x)\right] dx \tag{1.58}$$

In the context of this model, the Gini coefficient measures the discrepancy between $F(z)$, which determines the incidence of each type z in the population, and the wealth shares $S(z)$, which measure the distribution of wealth among these types. Visually, the Gini measures the area between the Lorenz curve, which is constructed by plotting the cumulative wealth shares S against the population shares F. Figure 1.1 illustrates the Lorenz curve; the Gini coefficient is $1 - 2B$, where B is the shaded blue area. This measure is convenient, as it can be calculated in the steady state and along transition paths using the wealth shares directly, without needing to calculate the underlying wealth distribution.

Figure 1.2: Steady State Comparative Statics in p

(a) Gini Coefficient (b) Returns (c) Agg. Productivity Z

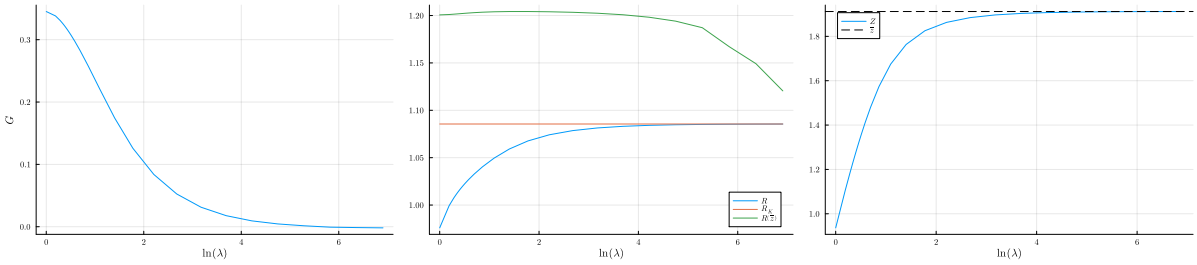


Clearly, the chief determinant of wealth inequality in this model is the persistence p of entrepreneurs' shocks. When $p = 0$ and shocks are IID, the Gini coefficient is equal to zero, corresponding to perfect equality; here, for instance, agents with z shocks in the bottom 25% of z values hold 25% of total wealth. By contrast, when $p = 1$, the highest types $\overline{z}$ hold all wealth, and the wealth shares thus put all mass here, implying that the area B in Figure 1.1 is equal to zero and the Gini coefficient is equal to one. At intermediate values of p, entrepreneurs' wealth accumulation decisions will lead to a Lorenz curve as in Figure 1.1 and a Gini coefficient between zero and one. Figure 1.2 shows how steady state inequality, returns, and productivity vary in the autocorrelation p. Panel 1.2a demonstrates that the relationship between inequality and persistence is convex in p. Panels 1.2b and 1.2c, meanwhile, shows how aggregate productivity and returns change in persistence. As $p \to 1$, a greater and greater share of aggregate wealth is held by the top type $\overline{z}$, and the risk-free rate R and the return $R(\overline{z})$ earned by these types converge to $R_K = \beta^{-1}$, while aggregate productivity converges to $\overline{z}$.

An increase in the collateral constraint, λ, will also increase productivity via a decrease in misallocation, but in a different way from p. Whereas an increase in persistence undoes financial frictions via wealth accumulation, an increase λ reduces misallocation by increasing the scope for *within-period* wealth transfers via the capital market. At the limit of $\lambda = \infty$, entrepreneurs are unconstrained in their ability to take on leverage, and can borrow as much as the market will bear. In this case, the real interest rate is again bid up to the marginal product of the highest-type entrepreneurs, $\overline{z}$, and the economy once again attains its first-best productivity. Figure 1.3, meanwhile, shows how productivity, returns, and wealth inequality are affected by an increase in the quality of financial markets, modeled as an increase in λ. Here again, Z is increasing in λ, and returns display a similar pattern, converging to the return of the high types $R(\overline{z})$. The wealth distributions underlying these changes, however, are markedly different. An increase in λ benefits low and high-productivity agents, shifting wealth away from the middle. Intuitively, a loosening of credit constraints allows for high-

Figure 1.3: Steady State Comparative Statics in λ

(a) Gini Coefficient (b) Returns (c) Agg. Productivity Z



type entrepreneurs to borrow more capital from inactive entrepreneurs ($z < \underline{z}$). This has the effect of increasing the risk-free real rate: recall that in the absence of capital market frictions ($\lambda = \infty$), the return to outside savings is equated with the marginal product of the highest types $\overline{z}$, who undertake all of the investment. The increase in the risk-free rate benefits types who do not invest, who now earn a higher return on their savings. The increase in λ also benefits high-z types, who still earn excess returns, but are now able to take on greater leverage to increase their capital income. Looser financial frictions shift wealth away from those with z in the middle of the support, near the cutoff. These types earn small excess returns, and to them, the benefit of looser capital markets is undone by the concomitant increase in the cost of external financing.

The discussion here presages the importance of the wealth distribution in determining the effects of monetary policy. As will become clear in Section 1.3, whether productivity is increased as a result of λ or p implies different responses to a policy change: it matters whether high productivity entrepreneurs accumulate capital over time, or borrow it on spot markets. The wealth distribution offers a way to disentangle whether aggregate returns and productivity are being driven by credit markets or by wealth accumulation, and thus what sort of response to policy we may expect.

1.3 Effect of a Monetary Shock

Here, I consider the effect of an unanticipated change in the stance of monetary policy. I assume that, prior to the shock, the economy is in its long-run, zero-inflation steady state as outlined in Section (1.2.5). Then, at time $t = 0$, there is an unanticipated innovation to the Taylor rule:

$$i_{t+1} = \overline{r} + \phi_\pi \pi_t + \nu_t \tag{1.59}$$

The shock $\nu_0 < 0$ decays at rate ρ_ν:

$$\nu_t = \rho_\nu \nu_{t-1} = \rho_\nu^t \nu_0 \tag{1.60}$$

Although agents do not anticipate the initial shock ν_0, they understand that it will decay according to the process above, and thus for $t > 0$ we return to a *perfect foresight* equilibrium, where there are no further aggregate shocks and agents perfectly anticipate the evolution of all aggregate variables. For a given variable X_t, I denote by $\hat{X}_t \equiv \ln X_t - \ln X$ its log-deviation from steady state. For all of the results in this section, I consider the model with the retailers' discount factor $\beta_f = 0$. as laid out in lemma 3. Though unrealistic, this assumption makes the models' mechanisms clearer. I confirm by simulation that the results hold qualitatively when $\beta_f > 0$.

I begin by focusing on changes in the distribution of wealth, as measured by changes in the wealth shares $\hat{s}_t(z)$ across the space of types z. The linearized law of motion for the wealth share of type z is

$$\hat{s}_{t+1}(z) = p \underbrace{\beta R(z)}_{MPS(z)} \left\{ \underbrace{\hat{R}_t(z) - \hat{R}_{tK}}_{\widehat{\text{Excess Return}_t(z)}} + \hat{s}_t(z) \right\} \tag{1.61}$$

Equation (1.61) makes clear the sources of redistribution, as well as the role of persistence in determining these changes. As laid out in Section 1.2.4, for an entrepreneur to increase her wealth share, she must earn an idiosyncratic return in excess of the aggregate return to capital. Additionally, redistribution is impossible if $p = 0$ or $p = 1$. From equation (1.61), $\hat{s}_t(z) = 0$ for all t if $p = 0$, and if $p = 1$, then no entrepreneurs earn excess returns, and so here $\hat{s}_t(z)$ will be zero always as well. Near the steady state, returns to capital are approximately $\hat{R}_{tK} = r_K\left(\hat{\omega}_t + \hat{Z}_t\right)$, where $r_K \equiv R_K - 1$ is the net return to capital in the steady state. Excess returns can thus be written as

$$\hat{R}_t(z) - \hat{R}_{tK} = \begin{cases} \frac{1}{R(z)}\left\{(1-\lambda)\left(\hat{i}_t - \hat{\pi}_t\right) + \lambda\omega z\hat{\omega}_t\right\} - r_K\left(\hat{\omega}_t + \hat{Z}_t\right) & z > \underline{z}_t \\ \frac{1}{R}\left\{\hat{i}_t - \hat{\pi}_t\right\} - r_K\left(\hat{\omega}_t + \hat{Z}_t\right) & z < \underline{z}_t \end{cases} \tag{1.62}$$

Equation (1.62) also shows the way in which wealth shares revert back to their steady state values: as redistribution increases the aggregate return to capital, entrepreneurs' excess returns begin to shrink, which ultimately exerts a downward pull on their wealth shares.

Combining equations (1.61) and (1.62) gives the following Lemma:

Lemma 4. *The changes in wealth shares at $t = 1$ immediately following the shock are*

$$\hat{s}_1(z) = \begin{cases} p\beta R(z) \left[\frac{1}{R(z)} \left\{ (\lambda - 1)\pi_0 + \lambda\omega \cdot z\hat{\omega}_0 \right\} - \hat{R}_{0K} \right] & z > \underline{z} \\ -p\beta\pi_0 & z \leq \underline{z} \end{cases} \tag{1.63}$$

where $\hat{\omega}_0 = \hat{\omega}(\pi_0)$. Furthermore, $\hat{s}_1(z) > 0$ if $z > \underline{z}$, < 0 otherwise.

Both π_0 and $\hat{\omega}_0 = \hat{\omega}(\pi_0)$ are positive following an accommodative shock $\nu_0 < 0$: the increase in demand raises both the aggregate price level, and the real price that entrepreneurs earn for their goods. Equation (1.63) reveals two channels by which active entrepreneurs benefit from these higher prices. The first channel relates to the change in inflation π_0, and is known in the literature as the "Fisher" channel (see, e.g., Auclert, 2019). This channel is familiar: the rise in inflation devalues nominal debts, redistributing real assets from lenders to borrowers. The difference in my model is that this channel now benefits high-productivity entrepreneurs ($z > \underline{z}$), at the expense of lower-productivity. The second channel relates to rising profit margins, $\hat{\omega}(\pi_0)$. The increase in demand leads to a rise in aggregate profit margins $\hat{\omega}_0$, which benefits active entrepreneurs, who produce in the period of the shock ($t = 0$). Intuitively, the Fisher channel redistributes *existing* wealth (carried forward from the period before the shock), while the profit channel dictates that *newly created* wealth in the expansion is distributed in a pattern that favors higher-productivity entrepreneurs.

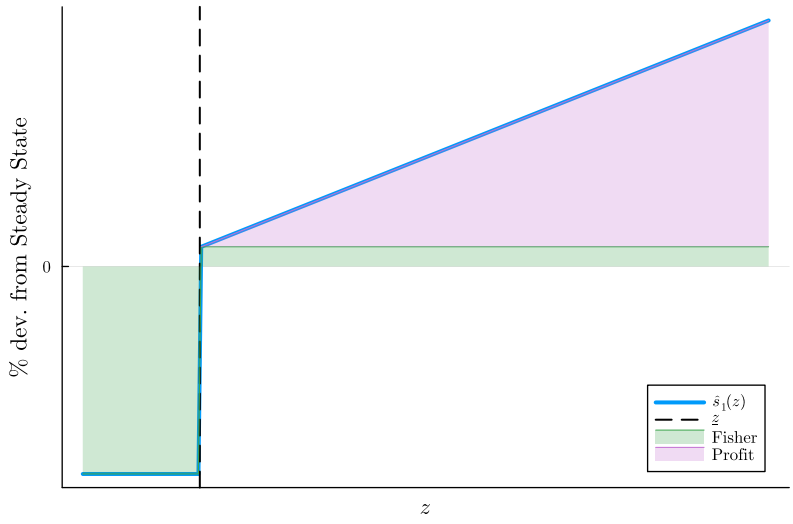
Figure (1.4) shows the pattern in $\hat{s}_1(z)$ across z: inactive types experience a uniform percentage decline in their wealth shares as a result of inflation, which devalues the nominal contracts written in the period before the shock. Active entrepreneurs benefit both from the inflationary shock, which devalues their *debt*, and from the increase in aggregate demand, visible in $\hat{\omega}_0$. As seen in Figure (1.4), this change benefits higher-z entrepreneurs more than lower.

This redistributive channel has the effect, of course, of raising aggregate productivity, further amplifying the effect on output. The redistribution of wealth toward high-types, who benefit from the shock, raises overall productivity. Linearizing the expression for TFP Z_t in Equation (1.38) and combining with capital market clearing (1.39), I have the following lemma:

Lemma 5. *The change in TFP following the shock ν_0 can be written as*

$$\begin{aligned} \hat{Z}_1 &= \mathbb{E}_s \left[(z - \underline{z})\, \hat{s}_1(z) \,|\, z > \underline{z} \right] \\ &\propto \int_{\underline{z}}^{\overline{z}} \underbrace{s(z)}_{\textit{S.S. dist.}} \times \underbrace{(z - \underline{z})}_{\textit{Excess prod}} \times \underbrace{\hat{s}_1(z)}_{\textit{Redistribution}} \, dz \end{aligned} \tag{1.64}$$

Figure 1.4: Wealth Share Change $\hat{s}_1(z)$, Following Shock



Equation (1.64) shows that the change in productivity depends on both the extent of redistribution, and the initial distribution of wealth. In particular, the immediate change in productivity $\hat{Z}_1$ as a result of the monetary shock is the conditional expectation of redistribution to entrepreneurs who are active in the steady state, weighted by their productivity net of the cutoff $\underline{z}$, evaluated according to the steady-state distribution $s(z)$. Because $\hat{s}_1(z) > 0$ for $z > \underline{z}$ by Lemma (4), $\hat{Z}_1 > 0$.

By capital market clearing, the cutoff $\underline{\hat{z}}_t$ must rise as well, in order to keep the mass of wealth above the cutoff constant. It is here that we see the importance of the wealth distribution in determining the overall effect of the change in policy: not only does investment increase (an effect which obtains even with identical entrepreneurs, or IID shocks), productivity also changes as the composition of investment is altered, with the additional investment being carried out by entrepreneurs who are more productive than average.

Turning to output, the linearized production function has the usual form:

$$\hat{Y}_t = \alpha\left(\hat{Z}_t + \hat{K}_t\right) + (1-\alpha)\hat{N}_t \tag{1.65}$$

This expression for output makes clear the amplification in this model, and the surrounding discussion. If shocks are IID, and wealth and productivity are uncorrelated in the steady state, then $\underline{\hat{z}}_t = \hat{s}_t(z) = 0$ for all t, and Equation (1.64) immediately implies $\hat{Z}_t = 0$ as well. In this case, Equation (1.65) becomes $\hat{Y}_t = \alpha\hat{K}_t + (1-\alpha)\hat{N}_t$, the standard form with fixed productivity. In this world, there is still amplification through investment: $\hat{K}_t$ will increase, as in e.g. Bernanke et al. (1999), which creates an elevated and long-lived response of output to the policy shift. In the case that $p > 0$, however, the results of Section (1.3) show that

the pattern of redistribution $\hat{s}_t(z)$ is such that $\hat{Z}_t > 0$ for a period following the shock. This creates further amplification of the shock, through an increase in productivity, which also occurs in the data (Christiano et al. 2005, Baqaee et al. 2021). Equations (1.61) and (1.64) then imply that this additional amplification also leads to longer-lasting effects: as the shock fades, entrepreneurs slowly spend down their wealth to return to steady state, leaving both investment and productivity elevated for some time after the shock has faded.

The effect of the steady-state wealth distribution on the economy's response to monetary policy can also be seen through the lens of inflation. As in standard New Keynesian models, the effect of the policy shock is to engender an increase in both real activity and inflation. The rise in inflation is driven by future inflation expectations. To see this, recall that solving the Phillips curve forward, as in Section (1.2.2), gives

$$\pi_t = \frac{\varepsilon}{\theta} \sum_{s=0}^{\infty} \beta^s \left\{ p_{sx} - \frac{\varepsilon - 1}{\varepsilon} \right\} \tag{1.66}$$

Thus, inflation is driven by expected future deviations of the price of entrepreneurial goods p_{tx}—the intermediate good for the retailers—from its long-run value. Anticipating this path of marginal costs, retailers preemptively raise prices beginning at the time of the shock. The increase in demand from intermediate retailers leads to an increase in labor demand by the entrepreneurs. This increase in labor demand also increases the aggregate return to effective capital ω_t and the overall return to capital R_{tK}, due to complementarity between capital and labor in production.

The linearized Phillips curve with $\beta_f = 0$ is

$$\pi_t = \kappa_p \hat{p}_{tx} \tag{1.67}$$

where $\kappa_p > 0$ determines the slope of the Phillips curve as a function of the elasticity of substitution and adjustment costs. Clearing in the markets for labor and entrepreneurial goods together imply that

$$\hat{p}_{tx} = \frac{\alpha + \eta}{1 - \alpha} \hat{Y}_t - \frac{\alpha(\eta + 1)}{1 - \alpha} \left(\hat{Z}_t + \hat{K}_t \right) \tag{1.68}$$

Combining the two gives

$$\pi_t = \kappa_p \left\{ \frac{\alpha + \eta}{1 - \alpha} \hat{Y}_t - \frac{\alpha(\eta + 1)}{1 - \alpha} \left(\hat{Z}_t + \hat{K}_t \right) \right\} \tag{1.69}$$

Above and beyond the impact of additional investment, the change in productivity $\hat{Z}_t$—itself

a function of reallocation, per Equation (1.64)—has the effect of *lowering* the Phillips curve, implying less inflation for a given pattern of economic activity. The size of this disinflationary force is driven by the response of $\hat{Z}_t$, and so it inherits from Equation (1.64) its dependence on the initial distribution. The larger are the gains in productivity from redistribution, the larger is the *downward* shift in the Phillips curve.

1.4 Redistribution and Reallocation: The Effect of Wealth Inequality on Monetary Policy

Section 1.3 lays out the fact that in this model, one effect of monetary policy is to redistribute wealth towards entrepreneurs, who have higher idiosyncratic productivities z, thereby increasing *aggregate* productivity Z. The increase in aggregate productivity amplifies the effect of the shock, flattening the Phillips curve and increasing the amount of output generated for a given change in capital and labor. The implication of Section 1.3, then, is this: the overall effect of a change in monetary policy depends on the degree to which this policy change redistributes assets among households. I show now that the extent of redistribution, and thus the overall effect of the monetary shock, depends on the distribution of wealth that is present at the time of the policy change.

1.4.1 Extreme Cases: Establishing the Bounds

In studying the effect of wealth inequality on monetary policy, it is clearest to start with the intermediate cases of zero inequality and perfect inequality, which correspond respectively to $p = 0$ and $p = 1$. In the IID case, with $p = 0$, the law of motion for the wealth shares $s_t(z)$ given in equation (1.34) implies that $s_t(z) = f(z)$ for all t and z; the wealth shares are fixed in time. Although entrepreneurs in this case do earn heterogeneous returns, they are just as likely in the next period to be low productivity as they are to be high, and thus wealth shares are unaffected by returns. As a direct result, aggregate productivity and the cutoff $\underline{z}$ are fixed in time as well, equal to

$$\underline{z} = F^{-1}\left(1 - \frac{1}{\lambda}\right) \tag{1.70}$$

$$Z = \lambda \int_{\underline{z}}^{\overline{z}} z dF(z) \tag{1.71}$$

At the other extreme of complete inequality ($G = 1$), which corresponds to perfect persistence $p = 1$, a similar result obtains. Because all of the wealth is held by the highest type, no

redistribution is possible, and $\underline{z} = Z = \overline{z}$: only the highest types produce, so aggregate productivity coincides with $\overline{z}$. The upshot is that in both cases, the economy acts as one with a representative producer whose productivity is fixed in time. In both cases, the equations governing the evolution of the economy are:

$$\hat{\omega}_t = \hat{p}_{tx} + (1-\alpha)\left(\hat{N}_t - \hat{K}_t\right) \tag{1.72}$$

$$\pi_t = \kappa_p \hat{p}_{tx} + \beta_f \mathbb{E}_t \pi_{t+1} \tag{1.73}$$

$$\hat{i}_{t+1} = \phi_\pi \pi_t + \nu_t \tag{1.74}$$

$$\hat{r}_t = \hat{i}_t - \pi_t \tag{1.75}$$

$$\hat{\omega}_{t+1} = \frac{1}{r+\delta}\hat{r}_{t+1} \tag{1.76}$$

$$\hat{N}_t = \frac{1}{\alpha+\eta}\hat{p}_{tx} + \frac{\alpha}{\alpha+\eta}\hat{K}_t \tag{1.77}$$

$$\hat{Y}_t = \alpha \hat{K}_t + (1-\alpha)\hat{N}_t \tag{1.78}$$

$$\hat{K}_{t+1} = [1-\beta(1-\delta)]\left(\hat{p}_{tx} + \hat{Y}_t\right) + \beta(1-\delta)\hat{K}_t \tag{1.79}$$

Equation (1.72) pins down the return to effective capital $\hat{\omega}_t$ as a function of aggregates; equation (1.76) requires that this return comove with the cost of capital, consistent with a constant cutoff $\underline{z}$. Equations (1.73) and (1.74) are the linearized Phillips Curve and Taylor rule, respectively. Equation (1.77) results from clearing in the labor market. Equations (1.78) and (1.79) are the linearized equivalents of the production function and capital accumulation equation.

This system can be simplified down and solved via the method of undetermined coefficients (see, e.g. Chapter 3 of Galí 2015). However, the key property of its behavior following an accommodative shock ($\nu_0 < 0$) can be inferred without any such solution. Note that when linearized about the steady state, the only aggregate steady-state object which appears is r, the steady state real interest rate. As such, all economies featuring constant wealth shares and a common interest rate r will behave *identically* following a monetary shock, regardless of any differences in their steady states. In particular: assuming that both economies are calibrated to match the US economy, such that they replicate the real interest rate observed in the data[8], the two economies will respond in the same way to a monetary shock, despite displaying polar opposite levels of wealth inequality. Even without re-calibration, the transition paths in these economies will be very similar, as evidenced by the system in (1.72)-(1.79). This equivalence between the two extreme economies lays the foundation for the hump-shaped relationship between inequality and the effect of monetary policy: because

[8]In the case of IID shocks, this can be done by choice of $F(z)$ or λ.

the response is the same at the polar extremes of inequality, to the extent that interest rate changes redistribute in the intermediate cases, the effect will be larger there. I turn to these cases next.

1.4.2 Intermediate Cases: Countervailing Forces

For intermediate values of wealth inequality, the tractability of my model allows for further insights as to the role of wealth inequality in determining the effects of monetary policy. Recall that the extent of redistribution, measured as the change in wealth shares $\hat{s}_1(z)$ immediately following the shock, can be written as

$$\hat{s}_1(z) = p\beta R(z)\left\{\hat{R}_0(z) - \hat{R}_{0K}\right\} \tag{1.80}$$

The question that I ask now is how this initial redistribution depends on the persistence p. To do so, I fix a value of π_0, to study how redistribution alters the response of the economy to monetary policy for a given path of prices. Later, I solve for the full response of the economy, including that of prices, computationally.

The channels of redistribution and reallocation are laid out in the following Proposition:

Proposition 3. *Given a price change π_0, for $z > \underline{z}$*

$$\frac{d}{dp}\hat{s}_1(z) = \underbrace{\beta R(z)\left(\hat{R}_0(z) - \hat{R}_{0K}\right)}_{>0} \qquad ①$$

$$+ p\beta \times \underbrace{\frac{d}{dp}\left\{R(z)\cdot\left(\hat{R}_0(z) - \hat{R}_{0K}\right)\right\}}_{<0} \qquad ②$$

Term ①, which is positive for $z > \underline{z}$, is redistribution. Term ②, which is negative for $z < \underline{z}$, is reallocation.

The proof of Propsition 3 can be found in the appendix. This proposition highlights the countervailing forces that affect redistribution as wealth inequality, driven by changes in persistence, increases. The first term, redistribution, is positive. To understand the intuition behind the redistribution channel, fix a type z and a level of excess returns $\widehat{ER}_0(z) = \hat{R}_0(z) - \hat{R}_{0K}$. Term ① measures the extent to which this change in excess returns filters into wealth shares: the more persistent are returns, the greater is the scope for agents who earn returns in excess of the return on capital following the shock to *accumulate* from those returns. Term ②, however, is negative for $z < \underline{z}$, and captures the reallocation channel. This term captures, for a given level of persistence (and thus scope for accumulation), what

level of excess returns the agents can expect to earn following the shock. To understand this term, recall the comparative statics on steady-state returns in Figure 1.2b. As wealth concentration increases, the idiosyncratic returns $R(z)$ for active entrepreneurs ($z > \underline{z}$) align with the return on capital. Proposition 3 shows that this property of the steady state, also holds along the transition path following a monetary shock: as $p \to 1$, idiosyncratic returns $\hat{R}_0(z)$ align with the return on capital $\hat{R}_{0K}$, and thus $\widehat{ER}_0(z)$ declines in p.

Figure 1.5 illustrates the results of Proposition 3. In each panel, I plot the responses of the variables in question again given a fixed price response π_0. Panel 1.5a shows the initial change in the wealth share of the highest type, $\hat{s}_1(\overline{z})$, following the shock across persistence p. As illustrated in Lemma 4, the changes for all of the remaining active entrepreneurs will be bounded above by $\hat{s}_1(\overline{z})$. Panel 1.5a shows the hump-shaped response of wealth shares, as a result of the redistribution and reallocation channels. Panel 1.5b shows the derivative $d\hat{s}_1(\overline{z})/dp$ in green, and the two terms in blue and orange respectively. As in Proposition 3, the first term is positive, reflecting the benefit of accumulation that comes from higher persistence. The second term is negative, reflecting the logic of the reallocation channel: further wealth concentration in the steady state reduces the scope for active entrepreneurs to earn returns in excess of the return on capital. Panel 1.5c illustrates this point directly: analogously to Figure 1.2b, Panel 1.5c shows the idiosyncratic return of type-$\overline{z}$ agents immediately following the shock, $\hat{R}_0(\overline{z})$, as well as the return on capital $\hat{R}_{0K}$, for fixed π_0 across p. Note that for a given change in prices, the response of the return on capital $\hat{R}_{0K}$ is fixed and independent of steady-state wealth concentration. As in Figure 1.2b, we see that as p increases, the return earned by these top entrepreneurs converges to the return on capital, and their difference $\widehat{ER}_0(\overline{z})$ in green converges to zero. Do note that each of the panels in Figure 1.5 shows a discontinuity that is characteristic of the steady state in this model: for any p arbitrarily close to zero, credit markets will be active, and entrepreneurs will borrow up to their debt limit. At $p = 1$, however, the top types hold all of the wealth, and the credit markets vanish—inactive entrepreneurs no longer hold any wealth to lend to the highest types. The discontinuity reflects this fact: for any p near but less than one, active entrepreneurs still benefit from the Fisher channel described in Section 1.3, and thus still enjoy a small amount of redistribution from inflation. At $p = 1$, meanwhile, credit markets are inactive, and the idiosyncratic return of the top types is by definition equal to the return on capital, thus shrinking redistribution to zero.

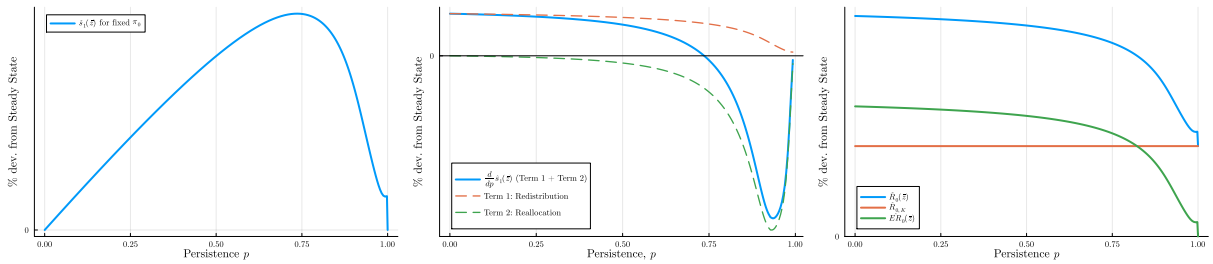
With the results of Proposition 3 established, the connection between wealth inequality and monetary policy is immediate. Because the magnitude of redistribution changes with wealth inequality, so too does the magnitude of the response of productivity to the shock. The amplification of the shock, and therefore its overall affect on aggregates, then depends

Figure 1.5: Redistribution and Reallocation

(a) Change in Top Wealth Share $\hat{s}_1(\overline{z})$, fixed π_0

(b) Derivative $\frac{d}{dp}\hat{s}_1(\overline{z})$

(c) Alignment of Returns in p



on how unequally wealth is distributed to begin with. To pin down this relationship, the following Corollary calculates the derivative of the initial change in TFP $\hat{Z}_1$ in p:

Corollary 1. *Given a price change* $\pi_0 > 0$,

$$\frac{d}{dp}\hat{Z}_1 = \underbrace{\mathbb{E}_s\left[(z-\underline{z})\times\frac{d}{dp}\hat{s}_1(z)\,|z>\underline{z}\right]}_{\lesseqgtr 0} \qquad ①$$

$$+\underbrace{\mathbb{E}_s\left[\frac{1}{s(z)}\left\{\frac{d}{dp}s(z)\right\}(z-\underline{z})\,\hat{s}_1(z)\,|z>\underline{z}\right]}_{>0} \qquad ②$$

$$+\underbrace{\mathbb{E}_s\left[\left(-\frac{d\underline{z}}{dp}\right)\hat{s}_1(z)\,|z>\underline{z}\right]}_{<0} \qquad ③$$

The main message of Corollary 1 is that the response of aggregate TFP to the monetary shock is *also* subject to countervailing forces in the persistence of entrepreneurs' returns. First and foremost, whether $\hat{Z}_1$ is increasing or decreasing in wealth inequality depends on the response of the wealth shares, as in Propostion 3. From the lens of aggregate productivity, higher persistence implies that the initial redistribution—which benefits agents with high z types in the period of the shock—is *targeted*, in the sense that it benefits agents who will have high productivity *tomorrow*. All else being equal, additional persistence implies that redistribution has a larger effect on productivity, as the effect of the shock will be to redistribute wealth to entrepreneurs of higher productivity in the period following the shock, thereby increasing aggregate productivity. Of course, as p increases the size of redistribution $\hat{s}_1(z)$ for $z > \underline{z}$ shrinks with these entrepreneurs' excess returns, as discussed above.

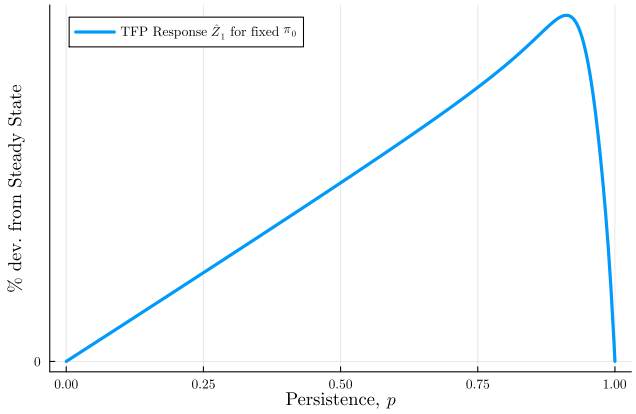
There are two additional forces which determine the magnitude of $\hat{Z}_1$ in Corollary 1. Tern ② captures the fact that increases in p also alter the steady-state distribution $s(z)$, which

determines how changes in wealth shares post-shock $\hat{s}_1$ are translated into productivity gains. All else equal, a given pattern of redistribution $\hat{s}_1$ will have a greater effect on productivity if its beneficiaries have greater wealth to begin with, as this gives them more weight in the expectation that determines $\hat{Z}_1$. As discussed in Section 1.2.5, increases in p shift wealth toward high-z entrepreneurs, which increases the response $\hat{Z}_1$ holding the pattern of redistribution $\hat{s}_1$ fixed. However, Term ③ captures a downward effect similar to that in Propostion 3: as p increases, the cutoff productivity $\underline{z}$ increases as well. This shrinks the response of $\hat{Z}_1$ to redistribution $\hat{s}_1(z)$, for a similar reason as before: as $\underline{z}$ increases, the excess productivity $z - \underline{z}$ shrinks for all active z. Effectively, with greater wealth concentration the beneficiaries of redistribution have productivities closer to the cutoff, and so the productivity gains to giving these entrepreneurs a bit more wealth shink.

Figure 1.6 illustrates the response of $\hat{Z}_1$ that results from Corollary 1, across p and again for fixed π_0. Here again, the hump-shaped response in $\hat{Z}_1$ is evident: aggregate TFP does not respond to a monetary shock at $p = 0$ and $p = 1$, and the magnitude of its response peaks between these two values. Notably, the response of $\hat{Z}_1$ peaks at a higher value of p than does the response $\hat{s}_1(\overline{z})$ in Figure 1.5a. This result is due to Tern ② in Corollary 1: higher persistence shifts steady-state wealth shares towards agents with higher z, which increases the response of $\hat{Z}_1$ to any given pattern of redistribution, $\hat{s}_1(z)$. Crucially, it implies that even as the magnitude of $\hat{s}_1(\overline{z})$ begins to decline in p, the productivity response continues to rise in p: even though redistribution is smaller, it is targeted towards households who exert a greater effect on aggregate productivity due to their higher shares of aggregate weatlh. This is the benefit of higher wealth inequality among firms in Baqaee et al. (2021), and households in Colciago et al. (2019)—essentially, this feature results from the redistribution channel, which their results highlight. However, Figure 1.6 demonstrates that these benefits fade away for high values of wealth inequality: as wealth concentration nears its maximum, the response of TFP to monetary policy begins to fade to zero.

As a result of the hump-shaped pattern of the TFP response $\hat{Z}_1$ in Figure 1.6, the *overall* effect of monetary policy will be be hump-shaped in wealth inequality. In essence, the degree of wealth inequality present at the time of a change in the policy rate determines how much redistribution the policy rate creates, and thus how much the response is amplified by the resulting change in aggregate productivity. The results here also imply that beyond a certain point, increases in wealth inequality *dampen* the efficacy of monetary policy.

Figure 1.6: Productivity Response $\hat{Z}_1$ (fixed π_0)



1.5 Quantitative Results

In Sections 1.3 and 1.4, I study theoretically the response of my economy to monetary policy, and measure how that response depends on wealth inequality. Here, I calibrate my model to match moments of the joint distribution of wealth and returns in the US economy as of 2022, to assess the quantitative relevance of the channels studied above. To begin, I discuss my strategy for calibrating my model to the data. Then, with the calibrated model in hand, I ask a few questions about how the model speaks to the data. First, I show that this model captures the response of the economy to monetary policy fairly well; in particular, it captures movements in TFP and inequality resulting from monetary easings. Then, I show that the degree of amplification resulting from the redistribution in assets is substantial, implying that this is a meaningful channel of monetary policy to capture. My calibrated model also implies a markedly different relationship between output and inflation than in standard RANK and HANK models, again arguing that heterogeneity in household returns is an important source to capture. Finally, I show that my calibration suggests that the increase in wealth inequality since the 1970s can partially account for the decrease in monetary policy that has been documented in the empirical literature over the same time period.

1.5.1 Calibration

The time period in my model is one quarter. My parameters are summarized in Table 1.1. To begin, in my model the wealth-weighted average return to wealth in the steady state is equal to $1/\beta$, and so I set this to match the average return in Fagereng et al. (2020), approximately 0.9 percent quarterly. Depreciation is then set so that the quarterly capital

Table 1.1: Quarterly Calibration, 2022

Parameter	Value	Source
β	0.9911	Steady-state real interest rate of 0%
δ	0.021	Capital stock-output of 12
η	2	Labor elasticity (standard)
λ	1.43	Corp. Debt/Assets (FRB)
ε	10	Avg. Markup 11%
θ	90	Slope of Phillips Curve 0.1
α	0.4	Capital share $\alpha p_x = 0.36$ (standard)
ϕ_π	1.5	Standard value
ρ_ν	0.6	Standard value
μ_w	0.4	Lorenz Curve
p	0.995	Wealth Gini (SCF)
$F(z)$	$\Gamma(1.11, 1.26)$	Cross-section of returns (see text)

stock to output ratio is 12. Note that in my model, this ratio is given by

$$\frac{K}{Y} = \frac{\alpha p_x}{\beta^{-1} - (1-\delta)} \tag{1.81}$$

As a result of my assumption on workers' preferences, workers supply labor with elasticity $1/\eta$. McClelland and Mok (2012) report an estimate of this elasticity of about 0.5, implying $\eta = 2$. These households hold no wealth, and so I assume that they make up 40% of the population ($\mu_w = 0.4$), corresponding to the fact that in the SCF, the bottom 40% of households by wealth cumulatively hold approximately zero net worth.

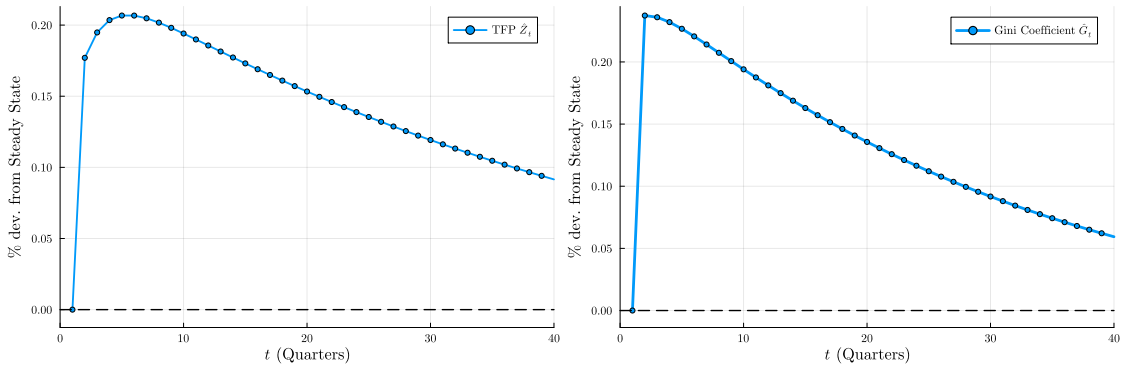
For λ: entrepreneurs in my model borrow a proportion $\lambda - 1$ of their assets in debt, i.e. $d/a = \lambda - 1$ if $z > \underline{z}$. I therefore set λ so that firms in equilibrium have a debt to net worth ratio of 0.43, as in the Financial Accounts data from the Federal Reserve Board of Governors. I choose this particular leverage ratio because assets and debt in my model are tied to businesses, and so I only want to model leverage in the economy as resulting from this particular form of equity, rather than forms such as student debt or mortgages.

The Phillips curve parameters ε and θ follow Kaplan et al. (2018), and imply a Phillips curve slope of 0.1, as in Schorfheide (2008). The capital loading factor α is then set so that the capital share αp_x is 0.36, a standard value. The weight on inflation in the Taylor Rule ϕ_π and the persistence $\rho_\nu$are assigned standard values from the literature.

What remains are the persistence of entrepreneurs returns p and the distribution of types $F(z)$. The degree of persistence p matches a wealth Gini of 0.85, corresponding to the Gini in the SCF. The distribution $F(z)$, meanwhile, targets the wealth-weighted cross sectional standard deviation of returns to wealth as documented in Fagereng et al.

Figure 1.7: Impulse Responses to 100bp monetary shock

(a) Total Factor Productivity $\hat{Z}_t$ (b) Gini Coefficient $\hat{G}_t$



(2020). In particular, I require that the standard deviation in my model be one third of their documented standard deviation of 1.95% quarterly, consistent with their estimate that about a third of the variation in returns is attributable to ex-ante heterogeneity (as opposed to ex-post risk), which is the only source of return variation present in my model.

1.5.2 Impulse Responses in the Calibrated Model

Here, I discuss the response of my economy to a 100bp monetary shock. To calculate the full impulse response, I use a computational strategy, which I outline in Appendix 1.7.4. To begin, Figure 1.7 shows the responses of two variables of particular interest in my model: aggregate TFP $\hat{Z}_t$, and the wealth Gini coefficient $\hat{G}_t$.

At its peak, TFP $\hat{Z}_t$ in Panel 1.7a increases by about 0.21% relative to the steady state. This effect is about one-fifth the size in the data (e.g. Baqaee et al., 2021). Although the model cannot explain the totality of the increase in total factor productivity, it can nevertheless capture this channel in ways that a standard RANK or HANK model cannot. As I will explain later, this exercise may in fact sell short the ability of this model to capture movements in TFP. On inequality $\hat{G}_t$, the model fits the data somewhat better: the Gini coefficient increases at its peak by about 0.25%, which is about sixty percent of the size of the increase in this statistic documented in Medlin (2023).

Figure 1.8 shows the responses of output and inflation in the calibrated model. Unsurprisingly, both increase, and investment by more than output, as is the case in the data. Figure 1.9, meanwhile, shows the remaining impulse responses. In particular, Panel 1.9a shows the response of the overall price level π_t, as well as the response of entrepreneurs' price $\hat{p}_{tx}$ and profit margin $\hat{\omega}_t$. Here, both the price that entrepreneurs earn on their goods $\hat{p}_{tx}$ and their profit margin per unit of effective capital $\hat{\omega}_t$ both rise in response to the increase

Figure 1.8: Impulse Responses to 100bp monetary shock

(a) Output $\hat{Y}_t$ (b) Investment $\hat{I}_t$

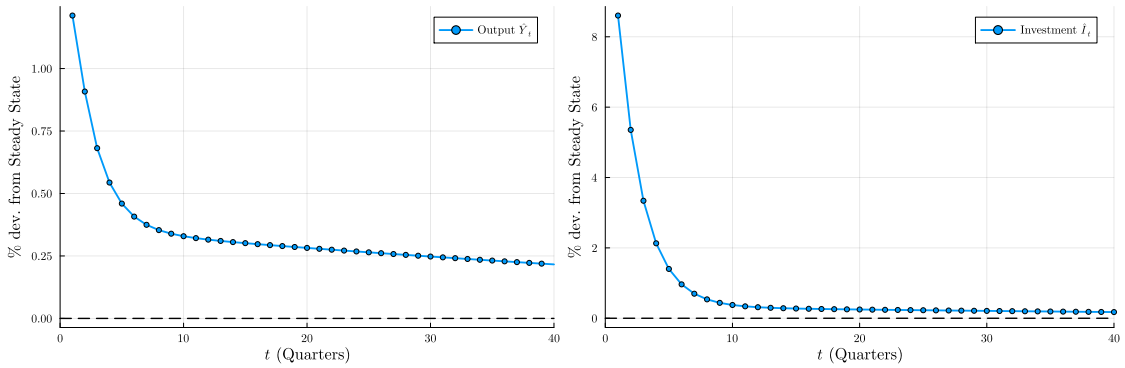
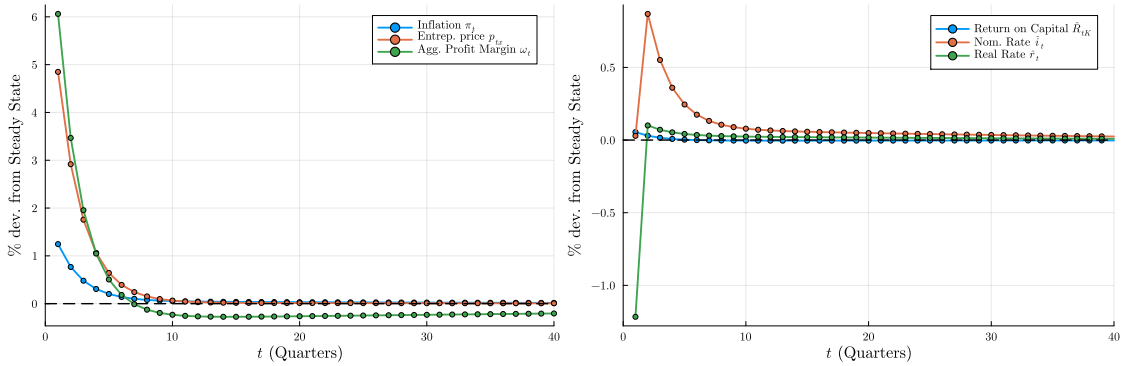


Figure 1.9: Impulse Responses to 100bp monetary shock

(a) Prices (b) Rates



in demand. As a result, then, the overall price level rises as well. Panel 1.9b, meanwhile, shows the impulse responses of the nominal and real interest rates $\hat{i}_t$ and $\hat{r}_t$ respectively, as well as the response of the aggregate return on capital $\hat{R}_{tK}$. The real interest rate falls as the monetary authority loosens, as intended, and then rises following the subsequent tightening that the policymaker enacts to counter inflation. As a result of the rise in prices and returns, as well as the redistribution that follows, the return on capital $\hat{R}_{tK}$ rises as well.

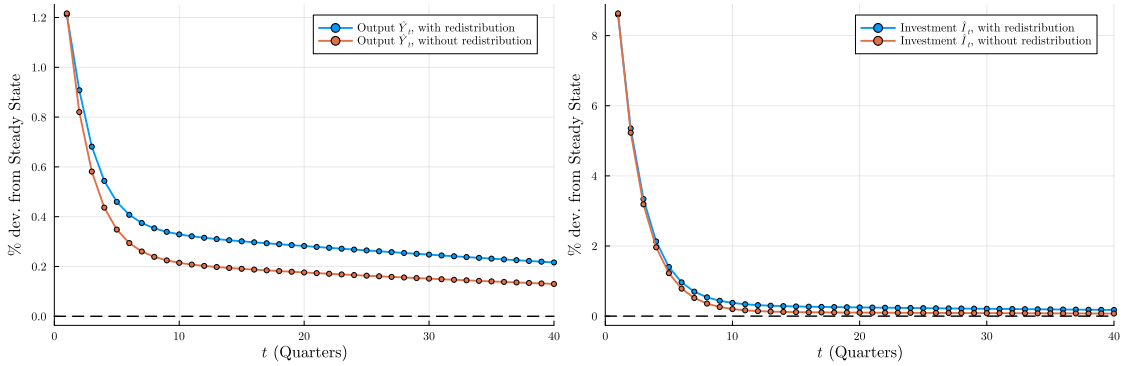
1.5.3 Amplification

Here, I provide two arguments as to the empirical significance of the redistributive channel of monetary policy. To begin, Figure 1.10 assesses the effect of redistribution on output $\hat{Y}_t$ and investment $\hat{I}_t$. For each variable, I plot two impulse responses to the same 100bp monetary shock as above. In blue, I plot the full impulse response, allowing for redistribution, which is identical to that in Figure 1.8. In orange, meanwhile, I plot the impulse response of

Figure 1.10: Amplification through Redistribution

(a) Impulse Response to 100bp shock: Output $\hat{Y}_t$

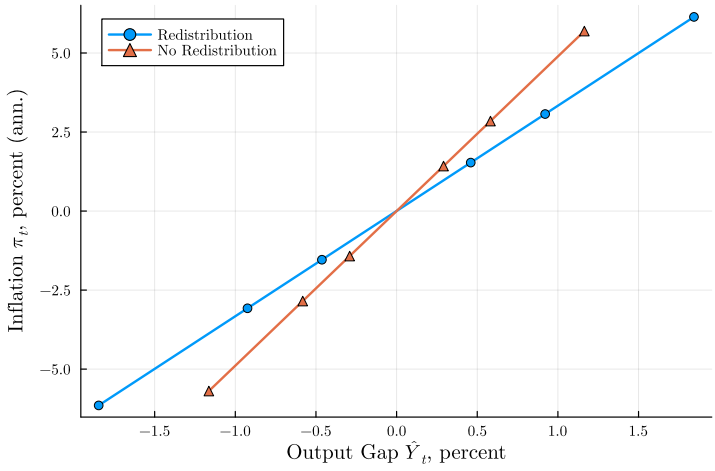
(b) Impulse Response to 100bp shock: Investment $\hat{I}_t$



an identical economy with redistribution shut down: I begin from the same steady state, but fix wealth shares $s_t(z)$ following the shock. Clearly, the redistributive channel exerts a strong influence on output in particular: the increase in productivity that results from shifting assets towards higher-producitivty entrepreneurs implies that over the life of the shock, the cumulative change in output is about fifty percent higher than the world without redistribution.

Figure 1.11 shows the flattening of the Phillips curve described theoretically in Section 1.3. For this experiment, I consider a range of monetary shocks ν_0 ranging from -200bp to 200bp. For each, I calculate the response of output $\hat{Y}_1$ and inflation π_1 in response to the shock, both in the period *after* the shock in order to account for any potential redistributive effects. I plot the former against the latter in Figure 1.11. I repeat this experiment for the two economies described immediately above: one with redistribution, and the other without, both starting from the same initial conditions. As demonstrated by Figure 1.11, allowing for changes in productivity resulting from redistribution meaningfully alters this relationship. For the theoretical reasons discussed in Section 1.3, the rise in productivity implies that the marginal cost to producing an additional unit of output falls, which implies less inflation for a given increase in output. Figure 1.11 shows this directly: the slope of the output-inflation line is meaningfully smaller than in the economy without redistribution. Quantitatively, allowing for redistribution flattens the slope of the Phillips curve by about 45%. This is a relevant comparison for two reasons, one normative and one positive. On the positive side, standard HANK models assume that production is carried out by a representative producer, thereby shutting down the redistributive, supply-side channel in my model—analogous to the orange line in Figure 1.11. Furthermore, when evaluating *optimal* monetary policy, the tradeoff between output and inflation is the fundamental choice faced by the policymaker.

Figure 1.11: Phillips Curve Slope



As a result, the optimal response of monetary policy to a given shock will be altered in the presence of redistribution. González et al. (2024) confirm that this is indeed the case.

1.5.4 Increase in Inequality

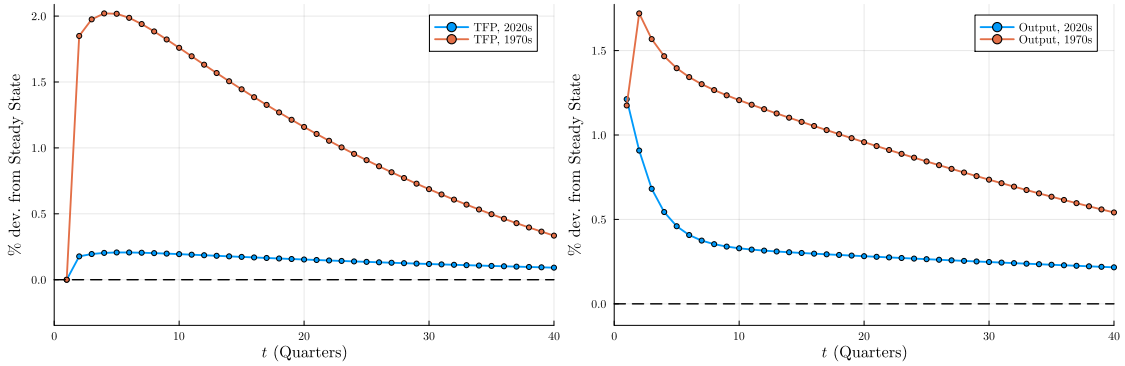
Since the early 1970s, two trends are present in the data. First, and famously, wealth inequality has increased: I measure using historical SCF data that the wealth Gini coefficient for the US has increased from 0.7 in 1971, to 0.85 in 2022. Concurrently, the efficacy of monetary policy shocks has decreased; for example, Boivin et al. (2010) estimate that the effect of an identified monetary shock on output is about half as large now as in the 1960s and early 1970s. Because my model suggests that increases in inequality can dampen the effects of monetary policy, it offers a natural lens with which to explore whether these two trends are related.

I study this question in the following way. I take the 2022 calibration presented above, and adjust the persistence parameter p to match the wealth Gini of 0.7 in 1971. I then calculate two sets of impulse responses to a 100bp shock: one beginning from the 1970s economy, and the other beginning from the 2022 economy. The level of inequality in the 1970s implies a somewhat lower value of persistence: whereas $p_{2022} = 0.996$, I find that $p_{1981} = 0.981$. The impulse responses in the two economies, presented in Figure 1.12, are markedly different. As compared to the 2022 economy, a monetary shock to the 1970s economy generates a much larger increase in productivity, as seen in Panel 1.12a. Per the results in Section 1.4, this change in the producvitivy suggests that as inequality has increased over the past fifty years, the *reallocation* channel has dominated, and the efficacy of policy has been muted as the productivity channel has shut down. Because the productivity response is smaller

Figure 1.12: Impulse Responses: 1970 and 2022

(a) Productivity (TFP) Impulse Responses, 1970 and 2022

(b) Output Impulse Responses, 1970 and 2022



in the 2022 economy, so too is the response of output, as shown in Panel 1.12b. Thus, my calibration suggests that the increase in wealth inequality since the 1970s is indeed partially responsible for the decrease in the efficacy of policy over the same period.

I note two further observations from Figure 1.12. First, the data suggest that following a 100bp monetary shock, aggregate TFP increases by about one percent (Baqaee et al., 2021). These estimates typically use data that covers the entirety of the period in time considered here. In Panel 1.12a, the "true" response of TFP lies between the responses of the two economies, suggesting that a calibration which averages the two would accurately capture the response of TFP to monetary policy. Additionally, one finding in the literature on the changing efficacy of monetary policy over time is that, relative to the 1960s-1970s, the response of output in the post-1970s period to monetary policy peaks at a lower level, but persists for longer. This is exactly the pattern of output in Panel 1.12b: following the shock, output in the 2022 economy with higher wealth inequality peaks at a lower level, but remains elevated for longer than in the 1970 economy.

1.6 Conclusion

I argue here two key points concerning the effect of monetary policy on economies with unequal wealth distributions generated by entrepreneurs who earn persistently different returns on their businesses. First, in this framework, redistribution of wealth among entrepreneurs is a key component of the transmission of monetary policy: in particular, a reduction in interest rates ultimately redistributes from low-productivity entrepreneurs to those with higher productivity. Second, the size and duration of the economy's response to monetary policy is determined by the wealth distribution, and the process for entrepreneurs' productivity that

generates it. The more persistent are entrepreneurs' idiosyncratic shocks, the more concentrated the wealth distribution will be prior to a shock, and the larger and longer-lasting the response of aggregates to this shock will be.

This paper reconciles two findings in the data: that expansionary monetary policy increases productivity, and that economies are more responsive to monetary policy when wealth inequality is greater. It also has important implications for the *optimal* conduct of monetary policy (see González et al. 2024 for further discussion in a similar framework). My model can also make some headway in explaining empirical evidence that suggests the effects of monetary shocks have changed over time (e.g. Canova and Gambetti, 2009; Boivin et al., 2010), as in the US the concentration of wealth in the hands of the most successful entrepreneurs has increased. The most important implication, in my opinion, is this: my paper contributes to a growing notion in the literature on monetary policy that measuring and predicting responses to changes in interest rates cannot be done by observing aggregates alone, and that distributions play an equally important role. Where many early papers in this strain emphasize the importance of heterogeneities in marginal propensities to *consume*, I stress that marginal propensities to *invest*, and their correlation with wealth, are of equal importance.

1.7 Appendix

1.7.1 Proof of Proposition (1)

Proof. I begin with aggregate output. Given the optimal choice of labor in Lemma 1, the output of an entrepreneur with productivity $z$and capital k_t is

$$y_t(z, k_t) = \frac{\omega_t}{\alpha} z k_t$$

Integrating over entrepreneurs using the stationary distribution $g_t(a, z)$, and using the fact that $k_t(a, z) = \lambda a$ if $z > \underline{z}_t$ and 0 otherwise, gives

$$\begin{aligned} Y_t &= \frac{\omega_t}{\alpha} \lambda \int_{\underline{z}}^{\infty} \int_0^{\infty} a z g_t(a, z)\, da dz \\ &= \frac{\omega_t}{\alpha} \lambda K_t \int_{\underline{z}}^{\infty} z s_t(z)\, dz \\ &= \frac{\omega_t}{\alpha} \lambda K_t X_t \end{aligned}$$

where

$$X_t = \int_{\underline{z}}^{\infty} z s_t(z)\, dz$$

From the labor market, we have

$$\begin{aligned} N_t &= \left(\frac{\omega_t}{\alpha}\right)^{\frac{1}{1-\alpha}} \lambda \int_{\underline{z}}^{\infty} \int_0^{\infty} a z g_t(a, z)\, da dz \\ &= \left(\frac{\omega_t}{\alpha}\right)^{\frac{1}{1-\alpha}} \lambda K_t X_t \end{aligned}$$

which implies

$$\omega_t = \alpha \left(\frac{N_t}{\lambda K_t X_t}\right)^{1-\alpha}$$

so production is

$$\begin{aligned} Y_t &= \frac{\omega_t}{\alpha} \lambda K_t X_t \\ &= \left(\frac{N_t}{\lambda K_t X_t}\right)^{1-\alpha} \lambda K_t X_t \\ &= (\lambda X_t K_t)^{\alpha} N_t^{1-\alpha} \end{aligned}$$

In order to eliminate λ, note that capital market clearing requires

$$\begin{aligned} K_t &= \int_0^\infty \int_0^\infty k_t(a,z)\, g_t(a,z)\, dadz \\ &= \int_{\underline{z}}^\infty \int_0^\infty \lambda a g_t(a,z)\, dadz \\ &\downarrow \\ 1 &= \lambda \int_{\underline{z}}^\infty s_t(z)\, dz \end{aligned}$$

and thus

$$\lambda = \frac{1}{\int_{\underline{z}}^\infty s_t(z)\, dz}$$

Replacing this into production gives

$$Y_t = (Z_t K_t)^\alpha N_t^{1-\alpha}$$

where

$$\begin{aligned} Z_t &= \lambda X_t \\ &= \frac{\int_{\underline{z}}^\infty z s_t(z)\, dz}{\int_{\underline{z}}^\infty s_t(z)\, dz} \\ &= \mathbb{E}_\omega [z | z > \underline{z}] \end{aligned}$$

Now, the law of motion for the aggregate capital stock is

$$\begin{aligned} K_{t+1} &= \int\int a_{t+1}(a,z)\, g_t(a,z)\, dadz \\ &= \int\int \beta R_t(z)\, a g_t(a,z)\, dadz \\ &= \beta K_t \int R_t(z)\, s_t(z)\, dz \end{aligned}$$

Recall that

$$R(z_t) = 1 + r_t + \lambda \max\{\omega_t z_t - r_t - \delta, 0\}$$

and so I can write this as

$$\begin{aligned}
K_{t+1} &= \beta K_t \int \left[1 + r_t + \lambda \max\left\{\omega_t z_t - r_t - \delta, 0\right\}\right] s_t\left(z\right) dz \\
&= \beta K_t \left(1 + r_t + \lambda \int_{\underline{z}}^{\infty} \left(\omega_t z - r_t - \delta\right) s_t dz\right) \\
&= \beta K_t \left(1 + r_t + \lambda \int_{\underline{z}}^{\infty} z\omega_t s_t\left(z\right) dz - \lambda\left(r_t + \delta\right) \int_{\underline{z}}^{\infty} s_t dz\right)
\end{aligned}$$

The two integrals are

$$\begin{aligned}
\omega_t \lambda \int_{\underline{z}}^{\infty} z s_t\left(z\right) dz &= \omega_t X_t \lambda \\
&= \omega_t Z_t
\end{aligned}$$

and

$$\begin{aligned}
\lambda \int_{\underline{z}}^{\infty} s_t dz &= \lambda\left(1 - S\left(\underline{z}\right)\right) \\
&= 1
\end{aligned}$$

and so the LoM is

$$K_{t+1} = \beta K_t \left(1 + \omega_t Z_t - \delta\right)$$

From labor market clearing,

$$\omega_t Z_t = \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha}$$

and so the LoM becomes

$$\begin{aligned}
K_{t+1} &= \alpha\beta p_{tx} \left(Z_t K_t\right)^{\alpha} N_t^{1-\alpha} + \beta\left(1 - \delta\right) K_t \\
&= \alpha\beta p_{tx} Y_t + \beta\left(1 - \delta\right) K_t
\end{aligned}$$

The return on capital equals the average return across all entrepreneurs, calculated above:

$$\begin{aligned}
R_{tK} &= \int_0^{\overline{z}} R_t\left(z\right) s_t\left(z\right) \\
&= 1 - \delta + \omega_t Z_t \\
&= 1 - \delta + \alpha p_{tx} Z_t^{\alpha} \left(\frac{N_t}{K_t}\right)^{1-\alpha}
\end{aligned}$$

Finally, the factor prices. The wage can be calculated from labor market clearing: recall that

$$N_t = \left(\frac{1-\alpha}{w_t} p_{tx}\right)^{\frac{1}{\alpha}} \lambda K_t X_t$$

Rearranging gives

$$w_t = (1-\alpha)\, p_{tx} \left(\frac{Z_t K_t}{N_t}\right)^{\alpha}$$

as in the text. The net real interest rate comes from the definition of the cutoff $\underline{z}_t$:

$$r_t = \omega_t \underline{z}_t - \delta$$

Substituting the definition of ω_t from labor market clearing gives the form in Equation (1.41). Note that this only holds in expectation, as nominal debt contracts are negotiated at the end of period t, and the ex-post real return r_t depends on the realization of inflation. □

1.7.2 Proof of Proposition (2)

Proof. By definition, the law of motion for the *cumulative* wealth share $S_t(z)$ is

$$\begin{aligned} S_{t+1}(z) = &\frac{1}{K_{t+1}} p \int_0^z \int_0^\infty a'(a,\hat{z})\, g_t(a,\hat{z})\, da d\hat{z} + \\ &\frac{1}{K_{t+1}} (1-p) F(z) \int_0^\infty \int_0^\infty a'(a,\hat{z})\, g_t(a,\hat{z})\, da d\hat{z} \end{aligned}$$

Differentiating with respect to z gives

$$\begin{aligned} S'_{t+1}(z) = &\frac{1}{K_{t+1}} p \int_0^\infty a'(a,z)\, g_t(a,z)\, da + \\ &\frac{1}{K_{t+1}} (1-p) f(z) \int_0^\infty \int_0^\infty a'(a,\hat{z})\, g_t(a,\hat{z})\, da d\hat{z} \end{aligned}$$

With the policy functions:

$$\begin{aligned} s_{t+1}(z) = &\frac{1}{K_{t+1}} p \int_0^\infty \beta R(z)\, a g_t(a,z)\, da + \\ &\frac{1}{K_{t+1}} (1-p) f(z) \int_0^\infty \int_0^\infty \beta R(\hat{z})\, a g_t(a,\hat{z})\, da d\hat{z} \\ = &\frac{K_t}{K_{t+1}} p \beta R(z)\, s_t(z) + \\ &\frac{K_t}{K_{t+1}} (1-p) f(z) \int_0^\infty \beta R(\hat{z})\, s_t(\hat{z})\, d\hat{z} \end{aligned}$$

Leibniz:

$$\frac{d}{dz}\int_0^z \int_0^\infty \beta R\left(\hat{z}\right) a g_t\left(a,\hat{z}\right) da d\hat{z} = \int_0^\infty \beta R\left(z\right) a g_t\left(a,z\right) da$$

Furthermore,

$$\int R\left(z\right) s_t\left(z\right) dz = 1 - \delta + \alpha p_{tx} Z_t^\alpha \left(\frac{N_t}{K_t}\right)^{1-\alpha}$$

as derived above, i.e. the RoK. So the LoM is

$$s_{t+1}\left(z\right) = \frac{K_t}{K_{t+1}}\beta\left[pR\left(z\right)s_t\left(z\right) + \left(1-p\right)f\left(z\right)\left(1 - \delta + \alpha p_{tx} Z_t^\alpha \left(\frac{N_t}{K_t}\right)^{1-\alpha}\right)\right] \tag{1.82}$$

Recall that from the law of motion for capital,

$$\frac{K_{t+1}}{\beta K_t} = 1 - \delta + \alpha p_{tx} Z_t^\alpha \left(\frac{N_t}{K_t}\right)^{1-\alpha} = R_{tK}$$

Substituting this into (1.82) gives Equation (1.49) in the text. □

1.7.3 Full Log-Linearized System

The full system is

$$\hat{s}_{t+1}(z) = p\frac{R(z)}{R_K}\left\{\hat{R}_t(z) - \hat{R}_{tK} + \hat{s}_t(z)\right\} \tag{1.83}$$

$$\hat{R}_t(z) = \begin{cases} \frac{1}{R(z)}\left\{(1-\lambda)\left(\hat{i}_t - \hat{\pi}_t\right) + \lambda\omega z\hat{\omega}_t\right\} & z > \underline{z}_t \\ \frac{1}{R}\left\{\hat{i}_t - \hat{\pi}_t\right\} & z < \underline{z}_t \end{cases} \tag{1.84}$$

$$\hat{R}_{tK} = r_K\left\{\hat{\omega}_t - \hat{Z}_t\right\} \tag{1.85}$$

$$\hat{\omega}_t = \hat{p}_{tx} + (1-\alpha)\left(\hat{N}_t - \hat{Z}_t - \hat{K}_t\right) \tag{1.86}$$

$$\hat{w}_t = \eta\hat{N}_t \tag{1.87}$$

$$\pi_t = \kappa_p\hat{p}_{tx} + \beta_f\mathbb{E}_t\pi_{t+1} \tag{1.88}$$

$$\hat{i}_t = \phi_\pi\pi_t + \nu_t \tag{1.89}$$

$$\hat{r}_t = \hat{i}_{t-1} - \pi_t \tag{1.90}$$

$$\hat{\underline{z}}_{t+1} = \frac{1}{r+\delta}\hat{r}_{t+1} - \hat{\omega}_{t+1} \tag{1.91}$$

$$\hat{N}_t = \frac{1}{\alpha+\eta}\hat{p}_{tx} + \frac{\alpha}{\alpha+\eta}\left(\hat{Z}_t + \hat{K}_t\right) \tag{1.92}$$

$$\hat{Y}_t = \alpha\left(\hat{Z}_t + \hat{K}_t\right) + (1-\alpha)\hat{N}_t \tag{1.93}$$

$$\hat{K}_{t+1} = [1-\beta(1-\delta)]\left(\hat{p}_{tx} + \hat{Y}_t\right) + \beta(1-\delta)\hat{K}_t \tag{1.94}$$

$$\hat{Z}_{t+1} = \lambda\left\{\underline{z}s(\underline{z})\left(1 - \frac{\underline{z}}{Z}\right)\hat{\underline{z}}_{t+1} + \int_{\underline{z}}^{\overline{z}}\left(\frac{z}{Z} - 1\right)s(z)\,\hat{s}_{t+1}(z)\,dz\right\} \tag{1.95}$$

$$0 = \underline{z}s(\underline{z})\,\hat{\underline{z}}_t + \int_0^{\overline{z}} s(z)\,\hat{s}_t(z)\,dz \tag{1.96}$$

1.7.4 Computational Strategy

Steady State

The equations pinning down the steady state are as follows:

$$s(z) = \frac{1-p}{1-p\beta R(z)} f(z) \tag{1.97}$$

$$R(z) = \begin{cases} 1 + r + \lambda(\omega z - r - \delta) & z > \underline{z} \\ 1 + r & z \leq \underline{z} \end{cases} \tag{1.98}$$

$$\omega = \alpha p_x \left(\frac{N}{ZK}\right)^{1-\alpha} \tag{1.99}$$

$$p_x = \frac{\varepsilon - 1}{\varepsilon} \tag{1.100}$$

$$\frac{1}{\beta} = 1 + \alpha p_x Z^{\alpha} \left(\frac{N}{K}\right)^{1-\alpha} \tag{1.101}$$

$$N = \left[(1-\alpha) p_x\right]^{\frac{1}{\alpha+\eta}} (ZK)^{\frac{\alpha}{\alpha+\eta}} \tag{1.102}$$

$$\underline{z} = \frac{r+\delta}{\omega} \tag{1.103}$$

$$Z = \lambda \int_{\underline{z}}^{\overline{z}} z s(z)\, dz \tag{1.104}$$

$$1 = \lambda \int_{\underline{z}}^{\overline{z}} s(z)\, dz \tag{1.105}$$

My assumption on the process for z_t allows the computation of the steady state to be reduced to a system of two equations in two unknowns. I guess a pair (ω^0, r^0). From there, I calculate

$$\underline{z}^0 = \frac{r^0 + \delta}{\omega^0}$$
$$Z^0 = \frac{1/\beta - (1-\delta)}{\omega^0}$$

where the second follows from stationarity of the capital stock. Aggregate productivity then pins down the capital-labor ratio:

$$\frac{K^0}{N^0} = \left(\frac{\alpha p_x (Z^0)^{\alpha}}{\frac{1}{\beta} - (1-\delta)}\right)^{\frac{1}{1-\alpha}}$$
$$\equiv \xi^0$$

and then labor market clearing pins down equilibrium labor N^0:

$$N^0 = \left[(1-\alpha)\, p_x \left(Z^0 \xi^0\right)^{\alpha}\right]^{1/\eta}$$

and then $K^0 = \xi^0 N^0$. Now, I check errors on the implied definitions of Z^0 and $\underline{z}^0$. I calculate $\underline{z}^1$ using capital market clearing: given r^0 and ω^0, I can calculate the wealth shares $s^0(z)$, and then $\underline{z}^1$ solves

$$1 - \frac{1}{\lambda} = \int_{\underline{z}^1}^{\overline{z}} s^0(z)\, dz$$

and Z^1 is given by

$$Z^1 = \int_{\underline{z}^0}^{\overline{z}} z s^0(z)\, dz$$

I define a discrete grid for z on $[0, \overline{z}]$, and compute the above integrals using the Trapezoidal rule. As a starting guess, I set

$$r^0 = r^{FB} = \frac{1}{\beta}$$

and

$$\omega^0 = \omega^{FB} = \frac{1/\beta - (1-\delta)}{\overline{z}}$$

equal to their values under the first-best equilibrium.

Transitions

I use the following algorithm to compute impulse responses to a monetary shock in my linearized model:

1. Start with $\hat{K}_0, \hat{s}_0(z), \hat{\underline{z}}_0, \hat{Z}_0, \hat{\imath}_0$ all equal to zero.
2. Guess a path for $\left\{\hat{\imath}_{t+1}\right\}_{t=0}^{T-1}$. For boundary conditions, we have $\hat{\imath}_0 = \hat{\imath}_{T+1} = 0$.
3. Calculate the path for π_t from the Taylor Rule:

$$\pi_t = \frac{\hat{\imath}_{t+1} - \nu_t}{\phi_t}$$

4. Using π_t and π_{t+1}, calculate $\hat{p}_{tx}$ from the Phillips curve:

$$\pi_t = \kappa_p \hat{p}_{tx} + \beta_f \mathbb{E}_t \pi_{t+1}$$

$$\hat{p}_{tx} = \frac{1}{\kappa_p}\left\{\pi_t - \beta \pi_{t+1}\right\}$$

5. Calculate $\hat{N}_t$:

$$\hat{N}_t = \frac{1}{\alpha+\eta}\hat{p}_{tx} + \frac{\alpha}{\alpha+\eta}\left(\hat{Z}_t + \hat{K}_t\right)$$

6. Calculate $\hat{\omega}_t, \hat{r}_t, \hat{R}_{tK}, \hat{R}_t(z)$:

$$\begin{aligned}
\hat{R}_t(z) &= \begin{cases} \frac{1}{R(z)}\left\{(1-\lambda)\left(\hat{i}_t - \hat{\pi}_t\right) + \lambda\omega z\hat{\omega}_t\right\} & z > \underline{z}_t \\ \frac{1}{R}\left\{\hat{i}_t - \hat{\pi}_t\right\} & z < \underline{z}_t \end{cases} \\
\hat{R}_{tK} &= r_K\left\{\hat{\omega}_t - \hat{Z}_t\right\} \\
\hat{\omega}_t &= \hat{p}_{tx} + (1-\alpha)\left(\hat{N}_t - \hat{Z}_t - \hat{K}_t\right) \\
\hat{r}_t &= \hat{i}_t - \pi_t
\end{aligned}$$

   - Note the notation in $R_t(z)$: it's important that we not let them switch ex-post to investing after the initial unanticipated π_0.

7. Using the returns, update the wealth shares:

$$\hat{s}_{t+1}(z) = p\frac{R(z)}{R_K}\left\{\hat{R}_t(z) - \hat{R}_{tK} + \hat{s}_t(z)\right\}$$

8. Using the continuation shares $\hat{s}_{t+1}(z)$, find $\hat{\underline{z}}_{t+1}$ from capital market clearing:

$$\hat{\underline{z}}_{t+1} = \frac{1}{s(\underline{z})}\int_{\underline{z}}^{\overline{z}} \hat{s}_{t+1}(z)\, dz$$

9. Repeat steps 3-7 for all t. Then, evaluate the $\underline{z}_{t+1}$ values implied by capital market clearing against their definition: for $t = 0, ..., T-1$:

$$\hat{\underline{z}}_{t+1} \stackrel{?}{=} \frac{1}{r+\delta}\hat{r}_{t+1} - \hat{\omega}_{t+1}$$

If these are all satisfied, stop. Otherwise, update the path $\left\{\hat{i}_t\right\}$ and return to step 2.

This section contains the full suite of impulse responses to the monetary shock, as discussed in Section 1.3.

Chapter 2

Optimal Taxation of Wealthy Individuals

Joint with Ali Shourideh[1]

2.1 Introduction

It is well-established that capital income is the proximate force in shaping the wealth distribution in advanced economies—see, for instance, Benhabib et al. (2011) or Benhabib et al. (2019). The fortunes of those on the top rung of the economic ladder are primarily composed of high-risk, high-return assets, such as ownership of businesses and shares of stock. Given these forces behind the wealth distribution, however, it is less clear what sort of tax and transfer system is optimal to both satisfy the redistributive motives of the government and encourage investors to undertake risky projects with upside benefits to society. Although heterogeneous rates of return are crucial to modeling the distribution of wealth, the vast majority of the literature in optimal taxation has assumed that all agents face the same schedule of returns on their savings. This paper addresses this gap by studying the optimal taxation of capital income in an environment in which agents earn heterogeneous returns on their investments.

The Dynamic Public Finance literature, which began with the seminal work of Mirrlees (1971), offers us a framework in which to weigh the concerns of redistribution and efficiency in designing an optimal tax scheme. What distinguishes this literature from prior work in optimal taxation ("Ramsey" taxation) is that no exogenous restrictions are imposed on the tax schedule. Instead, the government can implement any type of tax that it wishes, subject to revenue requirements and informational frictions. Given the nonlinear tax schedules and

[1]Carnegie Mellon University, ashourid@andrew.cmu.edu

informational frictions present in reality, we view this as an intuitively appealing setting in which to study capital taxation. Following this literature, we first state the problem of a government choosing the optimal nonlinear tax schedule. We then recast this problem as one of mechanism design under direct revelation: instead of a government levying taxes, we solve the problem in terms of a planner who collects potentially erroneous reports of agents' types, and then allocates to them consumption, savings, and investment. As such, we hereafter discuss the solution to our problem in terms of the allocations induced by a tax code, rather than the tax code itself.

In order to study capital taxation when returns are heterogeneous, we construct a model in which agents have access to two assets: a bond with common return across the population, and a private technology with idiosyncratic, constant returns. The crucial assumption is that while individuals are privately aware of their returns (types), the government is not, and thus cannot levy taxes based on types. Instead, the government must levy taxes based on ex-post income from both sources of saving. To differentiate between the two, we henceforth refer to allocating capital to one's idiosyncratic technology as "investing," and to zero net-supply borrowing and lending as "saving."

The primary differentiation of our paper from others in the Mirrleesian tradition is that we assume that the factors of production are fully mobile. In combination with our assumption of constant (rather than decreasing) returns to scale for entrepreneurs, full mobility of factors of production opens up the possibility that the planner may want to allocate all risky investment to agents of the highest "type." To formalize this intuition: if agents are indexed by $\theta \in [\underline{\theta}, \bar{\theta}]$, with expected returns increasing in θ, it may be optimal for the tax system to be designed such that only agents with $\theta = \bar{\theta}$ invest, and all other types simply lend to these entrepreneurs. One goal of this paper is to establish conditions under which such an allocation is *not* optimal, and instead, investment is divided among a range of entrepreneurs. We refer to such an allocation as "endogenous span of control," after Lucas (1978).

We begin by studying a static, two-period version of our model, in which entrepreneurs face a one-time decision between consumption, risk-free savings, and risky investing. In this setting, we vary our assumptions on information observable to the planner in order to develop a set of criteria under which the model exhibits an endogenous span of control. Here we demonstrate that, in order to have multiple types invest under the optimal allocation, it is crucial that the highest possible expected excess return not be too high relative to a measure of its riskiness, and that utility be bounded below. Furthermore, we derive an analytic condition that determines whether, under the fully-private information case, an agent of a given type invests. Whether an agent is allocated investment depends not only on his expected return, but also on all of the expected returns in the population *above* his, due

to informational rents. This condition leads directly to the primary result in this section: not all agents with expected returns above the risk-free rate are called upon to invest. In order to allocate investment to a given type, the planner must compensate all types who may wish to *mimic* that type, and for agents with sufficiently low (but positive) expected excess returns, this cost outweighs the benefit of the additional output that the agent may have produced.

We then turn to a dynamic extension of our model economy. Here, we exploit the homogeneity of the planner's problem to simplify the computation of optimal allocations. We also show that this homogeneity implies that optimal distortions in the dynamic context are independent of an agent's history of shocks or income, and instead depend only on his *current* expected returns. We consider two cases: one in which the planner can observe expected returns but not investment ex-ante (hidden action), and one in which neither investment nor expected returns are known to the planner (hidden action and type). Going forward, we are working to calibrate our dynamic model economy to US data on aggregate output and capital stock, as well as wealth inequality (as measured by the Pareto tail coefficient), in order to study the optimal level of long-run wealth inequality.

In Section 2.2, we discuss the existing literature on optimal taxation of capital income, and the ways in which our paper contributes to this. In Section 2.3, we analyze a static version of our model, where agents face a single investment choice. Section 2.4 extends our model to an infinite horizon. Finally, Section 2.5 concludes by discussing the remaining work to be done on this paper, as well as potential directions for future research.

2.2 Literature Review

A canonical result in the Ramsey taxation literature, in which the government seeks to raise its revenue using the most non-distortionary linear taxes possible, is that the tax on capital should be set to zero,[2] as any distortions introduced to the savings decisions of agents in the economy will decrease the future capital stock. This result is underpinned by the assumptions that all agents earn a common rate on their savings, and that there is no tax-relevant information about these individuals that the government cannot observe. In a "Mirrleesian" setting with informational frictions, by contrast, it is often the case that capital taxes are positive at the optimum—for example, as demonstrated in Golosov et al. (2003). With unobservable labor productivity, the government taxes capital in order to prevent agents from self-insuring against idiosyncratic income shocks, ensuring that they will continue to exert labor effort. Kocherlakota (2005) shows that this result holds in an

[2]See for instance Atkinson and Stiglitz (1976), Chamley (1986), and Judd (1982).

economy subject to aggregate shocks. He also shows that the intertemporal wedge can be implemented by a tax system that is nonlinear in capital gains and linear in current wealth, and that in such a system, the tax on wealth is zero in the aggregate and raises no revenue. Albanesi and Sleet (2006) construct a dynamic economy in which agents are subject to idiosyncratic shocks to their disutility of labor, and show that optimal allocations can be implemented in a market economy using a simple tax schedule conditioned on wealth and current labor income. In this decentralization, the tax on capital may be nonzero, depending on how the tax system incorporates an agent's labor income history.

One common feature of these papers is that the intertemporal rate of return on savings is constant across individuals, even if it may vary over time. Such a process of accumulation, however, fails to capture the way in which wealth is built in reality. Benhabib et al. (2011) show that the main force behind the thick tails in the wealth distribution–as can be seen in the US data–is variation in individual rates of return. To see intuitively why this is the case, consider two favorable shocks to an agent: one affects his income, and the other affects the rate of return. The shock to income allows him to save more, increasing his wealth additively. The shock to his rate of return, however, *multiplies* his wealth. It is this second type of shock that fills in the thick upper tail in the stationary distribution of wealth.

Because variable rates of return can help models capture the thick tail in the empirical wealth distribution, more recent literature has attempted to characterize optimal taxation in settings where agents earn heterogeneous returns. Albanesi (2006) considers the optimal taxation of entrepreneurial capital in a model where entrepreneurs are ex-post heterogeneous in the rates of return that they earn. She finds that a wedge on risky entrepreneurial capital is optimal, and that the sign on the wedge depends on the risk aversion of entrepreneurs. Additionally, when considering a market structure similar to ours, she finds that it is optimal to tax risk-free bonds and risky capital at different rates, with a positive wedge on risk-free saving serving as a deterrent from entrepreneurs self-insuring against the possibility of adverse shocks. As we will demonstrate, this differential schedule arises in our model as well. By contrast, though, we obtain this result in an environment in which entrepreneurial returns differ as a result of ex-ante heterogeneity, and do not depend on entrepreneurial effort. This latter assumption is crucial, as our positive wedge arises in a setting wherein there is no way to separate the utility from consumption and the disutility of effort (or in our model, investment), as is the case in many models in this literature. Scheuer (2014) similarly considers an environment with workers and entrepreneurs, and notes that the optimal tax code depends on whether taxes can be conditioned on occupation (i.e., entrepreneur or worker) separately from income. His main result is that, if the two occupations can be taxed separately, it is Pareto-optimal for the government to achieve its redistributive goals through

taxation of entrepreneurs and transfers to workers, rather than to rely on a lower tax on entrepreneurs to "trickle down" to workers in the form of higher wages.

Guvenen et al. (2019) study the welfare gains from wealth and capital income taxation in an OLG model similar to ours, wherein individuals have idiosyncratic productivity in their entrepreneurial ventures, and sell the outputs of these ventures to a final-goods producer who aggregates them using a CES production function. A key distinction between our work and theirs, though, is the notion of optimality: Guvenen et al. (2019) restrict attention to linear taxes, and consider the welfare gains from replacing capital income taxes with revenue-equivalent taxes on wealth. As such, they are able to show that implementing their wealth tax generates a welfare *improvement*, but not whether this tax is welfare-*optimal*. Our paper, by contrast, identifies a fully welfare-optimal tax code in the presence of informational frictions, a tax code which may be (and in fact is) very much nonlinear. Phelan (2021) similarly restricts attention to the optimal *linear* taxation of business owners, whose idiosyncratic productivity evolves over time, in part as a result of effort exerted in their business.

2.3 Static Model

2.3.1 Model

Our static economy takes place over two time periods, $t \in 0, 1$. The economy is populated by a continuum of households, $i \in [0, 1]$, who differ with respect to their privately-known types θ. We assume that $\theta \in \Theta = \left[\underline{\theta}, \overline{\theta}\right]$, and that θ is distributed according to the twice-differentiable CDF $F(\theta)$. In the first period, all agents are endowed with initial wealth w, which they allocate to consumption and savings. Agents have two methods of saving: investments k, and borrowing or lending b. All agents earn a common return of R on their borrowing and lending. Meanwhile, an agent of type θ who at $t = 1$ invests k in his private technology produces $y = zk$ at $t = 1$, where z is drawn randomly from a distribution with CDF $G(z|\theta)$. We denote by $\mathcal{Y}(\theta)$ the set of possible values of capital income for type θ, and $H(y|\theta)$ the implied distribution for capital income. For simplicity, we assume that agents have quasilinear utility:

$$U\left(c_0, c_1\right) = c_0 + \beta \mathbb{E}_0\left[u\left(c_1\right)|\theta\right]$$

The government chooses a tax function $T(zk, Rb)$ to maximize total welfare:

$$\max \int_{\Theta} U\left(\theta\right) dF\left(\theta\right) \tag{2.1}$$

where $U\left(\theta\right)$ is the utility to agents of type θ induced by the tax schedule. A household of

type θ maximizes the objective in (2.3.1), subject to

$$c_0 + k + b = w \tag{2.2}$$

$$c_1(z) = zk + Rb - T(zk, Rb) \tag{2.3}$$

As such, we can define the *wedges* induced by the tax function T as distortions in the household's optimality conditions:

$$\tau_k \equiv T_1(zk, Rb) = 1 - \frac{1}{\beta \mathbb{E}\left[zu'\left(c_1(z)\right)\right]} \tag{2.4}$$

$$\tau_b \equiv T_2(zk, Rb) = 1 - \frac{1}{\beta R \mathbb{E} u'\left(c_1(z)\right)} \tag{2.5}$$

As noted by Mirrlees (1971), the government's problem can be recast as a mechanism design problem, and by the Revelation Principle, we can focus on the direct mechanism. In this setting, the planner collects reports from households of type θ and chooses allocations

$$c_0(\theta), b(\theta), k(\theta), \{c_1(\theta, y)\}_{y \in \mathcal{Y}(\theta)} \tag{2.6}$$

in order to maximize the social welfare function in (2.1). The planner faces the following promise-keeping constraint:

$$U(\theta) = c_0(\theta) + \beta \mathbb{E}\left[c_1(\theta, y)\right] \tag{2.7}$$

and feasibility constraints:

$$\int \left[c_0(\theta) + k(\theta)\right] dF(\theta) = w \tag{2.8}$$

$$\int \int c_1(\theta, y)\, dH(y|\theta)\, dF(\theta) = \int \int z(\theta)\, k(\theta)\, dG(z|\theta)\, dF(\theta) \tag{2.9}$$

The planner's allocations may also be governed by incentive constraints, depending on what information we assume to be available to her. We assume that consumption and investment goods are identical and can be costlessly substituted for one another. Hereafter, we will refer to the allocations that solve the planner's problem as "constrained-efficient." In order to make the connection between constrained-efficient allocations and the original tax problem we can analyze the wedges $\tau_b(\theta)$ and $\tau_k(\theta)$, which are defined exactly as in (2.5) and (2.4), with the constrained-efficient allocations c_1 substituted in.

2.3.2 Simplification and Dual Problem

For ease of exposition, we assume that z takes on two possible values:

$$z(\theta) = \begin{cases} \bar{z}(\theta) & \text{with probability } p(\theta) \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

We assume further that $p : \Theta \to [0,1]$ is a differentiable function, and that $p'(\theta) < 0$. In addition, we focus on the dual to the problem in (2.1)-(2.9): the planner minimizes the cost of delivering a minimum level of utility $\underline{U}$ via the allocations in (2.6), subject to the promise-keeping constraint in (2.7) and potential additional incentive constraints. Because there are two possible values of capital income in the second period, the planner chooses two levels of second-period consumption:

$$c_1(\theta, 0) \equiv c_{1,L}(\theta)$$
$$c_1(\theta, y) \equiv c_{1,H}(\theta)$$

In order to study the planner's problem, we consider three separate assumptions on information available to the her. To understand the distinction between these cases, note that there are two dimensions along which we can introduce informational frictions: the reporting of type θ, and the choice of investment $k(\theta)$. As such, the three cases that we consider are as follows:

1. **First-best**: the planner can observe θ, *and* investment $k(\theta)$ (full information)
2. **Second-best**: the planner can observe type θ, but *cannot* observe investment $k(\theta)$; that is; she only observes capital income zk
3. **Third-best**: the planner can observe neither θ nor $k(\theta)$

Under first-best, the planner faces no informational frictions: she observes each agent's type θ, and can punish him if he invests anything other than his allocated $k(\theta)$. In particular, recall that the agent either produces θk (with probability $p(\theta)$), or nothing at all. Under first-best, we assume that the planner can distinguish an agent who produces no output due to poor luck, and one who produces no output because he did not invest at $t = 0$. In this scenario, we can characterize the solution to the planner's problem, which we do in Lemma 1.

Lemma 1 (First-Best Solution)**.** *Assume that the planner can observe θ and k. The alloca-*

tions that solve the planner's problem are as follows:

$$c_0(\theta) : \textit{Free}$$
$$c_{1,L} = c_{1,H} = \overline{c}$$
$$k(\theta) = \begin{cases} K & \theta = \overline{\theta} \\ 0 & o/w \end{cases}$$

for some K large (possibly infinite).

In the First-Best scenario, it is optimal to allocate all investment to those with the highest type $\bar{\theta}$. Note that this implicitly assumes that there is no constraint on borrowing, and the planner can frictionlessly allocate all of the wealth that is not consumed at $t = 0$ to the highest types through the market for borrowing and lending. For the remaining allocations, because utility is quasilinear, agents' marginal utilities are independent of their consumption c_0, and thus the planner is indifferent between all distributions of $c_0(\theta)$. In the second period, meanwhile, consumption is equated across agents, and equal to output $\bar{\theta}p\left(\bar{\theta}\right)K$. For the rest of our analysis of the static model, we aim to study conditions under which this allocation cannot be implemented with informational frictions.

In the Second-Best scenario, we continue to assume that the planner can observe θ, but we assume that she cannot observe their investment k; instead, she can only see output $y = zk$. This setup is similar to the familiar moral hazard problems in, for example, Holmström (1979) and Mirrlees (1999). This informational friction allows for one type of potential deviation: an agent of type θ, who receives first-period allocations $c_0(\theta)$ and $k(\theta)$, can consume $c_0(\theta)+k(\theta)$ in the first period, thereby guaranteeing that he will produce no output at $t = 1$. In the second period, then, he can claim to have been unlucky, and thus receive $c_{1,L}(\theta)$. This type of deviation requires that the planner's allocations satisfy the following incentive constraint (suppressing dependence on θ):

$$c_0 + \beta\left[pu\left(c_{1,H}\right) + (1-p)\,u\left(c_{1,L}\right)\right] \geq c_0 + k + \beta u\left(c_{1,L}\right) \tag{2.11}$$

Constraint (2.11) requires that the agent be no worse off if he abides by the planner's prescription for investment and consumption in the first period, than if he follows the deviation strategy described above.[3] Because the objective is strictly decreasing in $k(\theta)$, (2.11) will

[3]As is standard in the mechanism design literature, we assume that if the agent is indifferent between cooperating with the planner's recommendation and deviating, that he will cooperate.

hold with equality, and thus this equation pins down investment $k(\theta)$:

$$k = \beta p \left[u\left(c_{1,H}\right) - u\left(c_{1,L}\right) \right] \tag{2.12}$$

Under this assumption, we have the following proposition:

Proposition 1 (Second-Best allocations)**.** *Suppose that the utility of second-period consumption is CES,*

$$u\left(c\right) = \frac{c^{1-\gamma}}{1-\gamma} \tag{2.13}$$

with inverse intertemoporal elasticity of substitution $\gamma < 1$. Then, the first-best allocation cannot be implemented with unobservable k.

Proposition 1 is our first result on *endogenous span of control*: if the elasticity of intertemoporal substitution is greater than one, the planner cannot implement the first-best allocations, under which all investment is allocated to agents of the highest type $\bar{\theta}$. The intuition is as follows: under second-best, the planner must now provide sufficient incentive to the highest types to invest $k\left(\bar{\theta}\right)$—if she fails to do so, they will simply eat this amount in the first period and claim in the second to have been unlucky. By the incentive constraint in (2.12), the incentives for investment are delivered through the spread between $c_{1,H}\left(\bar{\theta}\right)$ and $c_{1,L}\left(\bar{\theta}\right)$. Because consumption cannot be negative, the planner in this case will set $c_{1,L}\left(\bar{\theta}\right) = 0$, and all incentives will be provided through $c_{1,H}\left(\bar{\theta}\right)$. The remainder of the proof is a limiting argument: we assume that $k(\theta) > 0$ for $\theta \in \left[\bar{\theta} - \varepsilon, \bar{\theta}\right]$ and send $\varepsilon \to 0$. As $\varepsilon \to 0$, the investment allocated to the agents of type $\bar{\theta}$ diverges to infinity. With an elasticity of intertemoporal substitution is greater than one, these increases in $k(\bar{\theta})$ must be compensated more than one-for-one with increases in $c_{1,H}\left(\bar{\theta}\right)$, and thus the cost to the planner of implementing the first-best allocations (Lemma 1) is infinite, and thus these allocations cannot solve the planner's problem under second-best informational assumptions.

We now turn to the third-best case, wherein the planner does not observe type θ or investment k. This case allows for the agents to deviate from the planner's allocations in two manners: they can misreport their type, invest a deviant amount, or both. Such possible

deviations in turn necessitate two sets of incentive constraints: for all $\theta, \hat{\theta} \in \Theta$,

$$U(\theta) \geq c_0\left(\hat{\theta}\right) + k\left(\hat{\theta}\right) - \frac{\overline{z}\left(\hat{\theta}\right)}{\overline{z}(\theta)} k\left(\hat{\theta}\right) + \beta\left[p(\theta)\, u\left(c_{1,H}\left(\hat{\theta}\right)\right) + (1 - p(\theta))\, u\left(c_{1,L}\left(\hat{\theta}\right)\right)\right] \tag{2.14}$$

$$U(\theta) \geq c_0\left(\hat{\theta}\right) + k\left(\hat{\theta}\right) + \beta u\left(c_{1,L}\left(\hat{\theta}\right)\right) \tag{2.15}$$

Constraint (2.14) handles one potential deviation strategy: an agent of type θ can misreport his type, instead reporting $\hat{\theta}$, and then invest an amount such that he mimics the capital income of type $\hat{\theta}$ upon a successful investment. Do note, however, that the probability of such a successful investment still depends on his true type, θ. Constraint (2.15), meanwhile, obviates the second possible deviation: an agent of type θ can claim to be of type $\hat{\theta}$, and then invest nothing. In this scenario, he consumes $c_0\left(\hat{\theta}\right) + k\left(\hat{\theta}\right)$ in the first period, produces nothing with certainty, and then receives consumption $c_{1,L}\left(\hat{\theta}\right)$ in the second period.

Similar to Second-Best case (Proposition 1), we can characterize one assumption that guarantees endogenous span-of-control:

Proposition 2 (Third-Best)**.** *If utility is CES as in equation* (2.13) *with* $\gamma > 1$*, then the first-best allocations* ***can*** *be implemented with private types* θ *and investment* k*.*

The intuition for Proposition 2 is the same as for Proposition 1. If $\gamma > 1$, then utility is unbounded below ($\lim_{c \to 0} u(c) = 0$) and the intertemoporal elasticity of substitution $1/\gamma$ is less than one. In such a scenario, the planner can implement the First-Best allocations in Lemma 1 by allocating all capital to the agents of type $\bar{\theta}$, and then punishing them with infinite disutility upon the realization of zero output. The infinite disutility discourages other types from reporting $\bar{\theta}$, and the IES of less than one ensures that rewarding the $\bar{\theta}$ types with high $c_{1,H}$ does not become infinitely costly.

Assuming that utility is CES with $\gamma < 1$, and thus that a range of types will invest a positive amount (endogenous span-of-control), we can demonstrate additional properties of the solution to the planner's problem under Third-Best information constraints:

Proposition 3. *Under the optimal third-best allocations,*

$$\frac{d}{d\theta}\left(\Phi(\theta)\, p(\theta)\right) < 0 \implies k(\theta) > 0 \tag{2.16}$$

where

$$\Phi(\theta) = \int_{\theta}^{\bar{\theta}} \left\{ \left[\frac{p(t)\,\overline{z}(t)}{R} - 1 \right] \frac{\overline{z}(\theta)}{\overline{z}(t)} \right\} f(t)\,dt \tag{2.17}$$

Proposition 3 characterizes the role of the tax code in determining entry into entrepreneurship, defined as $k(\theta) > 0$. The function $\Phi(\theta)p(\theta)$ can be thought of as capturing the main tradeoffs that the planner faces when allocating investment k to agents of a given type θ. From (2.17), we see that Φ can be thought of as a weighted average of excess returns for types *above* θ, where the weights are the relative returns $\bar{z}(t)/\bar{z}(\theta)$. Thus, Φ captures the fact that, by allocating investment to type θ, the planner must adjust the investment-consumption schedule of types $\hat{\theta} > \theta$, so that these types do not *mimic* θ. As such, Φ can be thought of as a measure of informational rents associated with allocating $k(\theta)$. In order to allocate $k(\theta) > 0$, it must be that the benefits to the planner (in the form of returns $p(\theta)\bar{z}(\theta)$) are equal to the costs of information rents and insurance to type θ in the case of an unsuccessful project.

One important implication of proposition 3 is that the condition in equation (2.16) is not equivalent to requiring simply that $\mathbb{E}[z|\theta] \geq R$. That an agent's expected return on investment is greater than the risk-free rate is *not* sufficient to guarantee that the planner will find it optimal for that agent to invest. Indeed, in the numerical example in Section 2.3.3, we present an example in which there exists a range of θ values such that $\mathbb{E}[z|\theta] \geq R$, but $k(\theta) = 0$ under the optimal allocations. This property of the solution results from the costs to the planner incurred by allocating investment, as discussed above. For the planner to allocate $k(\theta) > 0$, it must be that the benefit of type θ's expected returns is equal to the costs implied by the incentive constraints (2.14) and (2.15). For types with $\mathbb{E}[z|\theta]$ above but near the risk-free rate R, the benefits of their potential investments cannot outweigh these costs, and as a result, these types are discouraged from investing.

2.3.3 Numerical Example

To help visualize optimal allocations and wedges in our model, we turn to a numerical example. We assume that $\beta = 0.95$, and $R = 1/\beta$. In keeping with Proposition 2, we assume that utility is CES with $\gamma = 0.4$. We assume that types are drawn according to a truncated Normal distribution on $\Theta = [1, 3]$, with mean $\frac{1}{2}(\underline{\theta} + \bar{\theta})$ and standard deviation $\sigma_\theta = 1/2$. The probability of a successful investment is determined by

$$p(\theta) = e^{-a\theta} \tag{2.18}$$

with $a = 0.2$. Finally, we assume that $\bar{z}(\theta) = \theta$, so that upon the realization of a successful project, and agent of type θ who invested k in the first period produces $y = \theta k$ in the second. With these assumptions on returns, the wedges in (2.4) and (2.5) become

$$\tau_b(\theta) = 1 - (\beta R)^{-1} \left[p u'(c_{1,H}(\theta)) + (1-p) u'(c_{1,L}(\theta)) \right]^{-1} \tag{2.19}$$

$$\tau_k(\theta) = 1 - \left[\beta \theta p(\theta) u'(c_{1,H}(\theta)) \right]^{-1} \tag{2.20}$$

Following Proposition 3, Figure 2.1 shows $p(\theta)$, $\Phi(\theta)$, and their produce $p(\theta)\Phi(\theta)$. As presaged by the discussion of Proposition 3, the produce $p\Phi$ is hump-shaped over Θ, first rising and then falling as $\theta \to \bar{\theta}$. Recall that for a type to invest, it must be that $p(\theta)\Phi(\theta)$ is declining at θ.

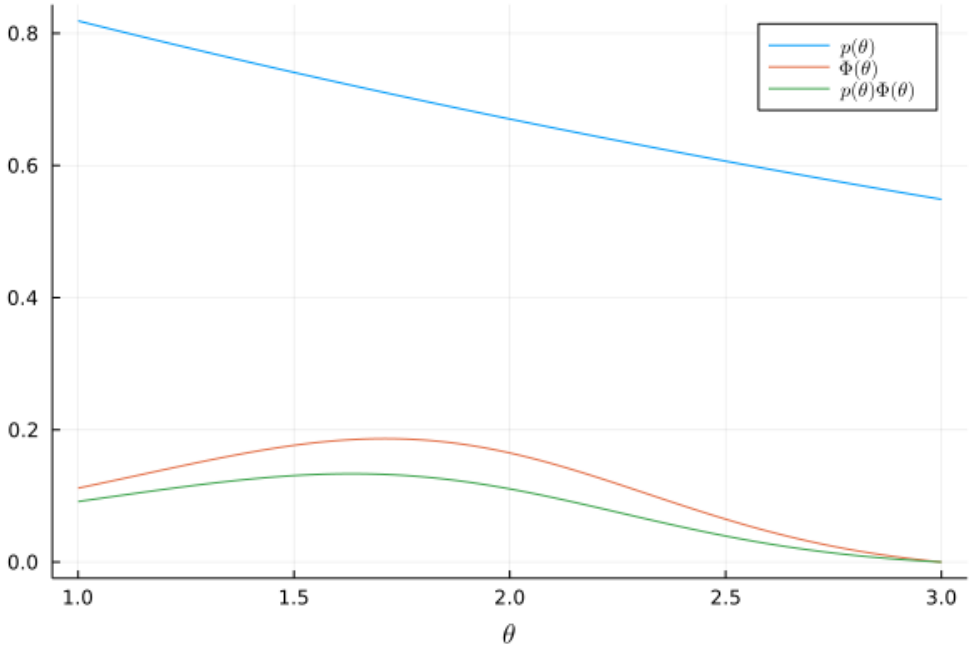


Figure 2.1: Determinants of entry

Figure 2.2 illustrates the remaining results for our numerical example. The top left panel shows the first-period allocations $c_0(\theta)$ and $k(\theta)$, with the dashed vertical line indicating the lowest θ value for which $k > 0$. The top right panel, meanwhile, shows $c_{1,H}$ and $c_{1,L}$, the second-period values of consumption following a successful and unsuccessful project respectively. For values of θ to the left of the dashed vertical line, $k(\theta) = 0$; these types do not invest, and thus their second-period consumption c_1 is independent of capital income. For types who *do* invest, meanwhile, the spread between $c_{1,H}$ and $c_{1,L}$ incentivizes the proper investment k. As k increases, so too does the spread $c_{1,H} - c_{1,L}$. To understand this spread, recall the incentive constraints (2.14) and (2.15). As in Mirrlees (1971), constraint (2.14)

requires that utility be increasing in θ in order to ensure that each type invests the efficient amount. Constraint (2.15), meanwhile, implies that $c_{1,L}$ decline in θ as $k(\theta)$ increases. The decline in $c_{1,L}$ ensures that as agents are allocated higher levels of k, agents with lower values of θ are discouraged from claiming to be of type θ, eating the entire $c_0(\theta)+k(\theta)$, and claiming to be unlucky.

The bottom right panel of Figure 2.2 shows the wedges on investing and saving, τ_k and τ_b. This panel and the bottom left have two vertical dashed lines: the first indicating the point at which $\mathbb{E}[z|\theta] = R$, and the second indicating the lowest value of θ for which $k(\theta) > 0$. Beginning with the wedge on investing τ_k, notice that this distortion is nonmonotonic: it increases up to the point at which investment becomes positive, at which point it begins to decrease. Note as well that at the point where $\mathbb{E}[z|\theta] = R$, τ_k becomes positive, and is negative below this point. This pattern of τ_k points to the dual role that the tax code plays in regulating entrepreneurship along both the intensive and extensive margins. The types in between the dashed vertical lines–those for whom τ_k is both positive and increasing–are the types who *would* select into entrepreneurship (invest $k > 0$) in the absence of tax distortions, but do not under the optimal allocation. This result is key: recalling the discussion of Proposition 3, although these types have expected returns in excess of the risk-free rate, their returns are not sufficient to overcome the costs that the planner would incur by allocating to them $k > 0$. For agents for whom $k > 0$, the wedge is *regressive*, declining towards zero as θ increases. Note, though, that this wedge is positive: investment income is taxed at a positive rate, in order to satisfy the planner's redistributive objective.

The wedge on risk-free borrowing and lending τ_b, meanwhile, is monotonic over Θ. Note that this wedge is zero for types who do not invest—these agents and the planner agree on their consumption-savings choice. For agents who do invest, the wedge is progressive, which further ensures efficient investment. Because investment is risky, agents who choose $k > 0$ would also like to save a positive amount ($b > 0$), in order to self-insure against a failed project. Under the constrained-efficient allocation, however, the planner would like these types to *borrow* ($b < 0$), thereby shifting economy-wide resources towards investment undertaken by the most productive types. The positive wedge on savings income nudges the higher-θ types towards borrowing and away from saving, so that capital is efficiently deployed.

2.4 Wealth and Taxes in the Infinite Horizon

While our static model is informative regarding optimal distortions to investment, it cannot distinguish between capital income and wealth. To address questions relating to the

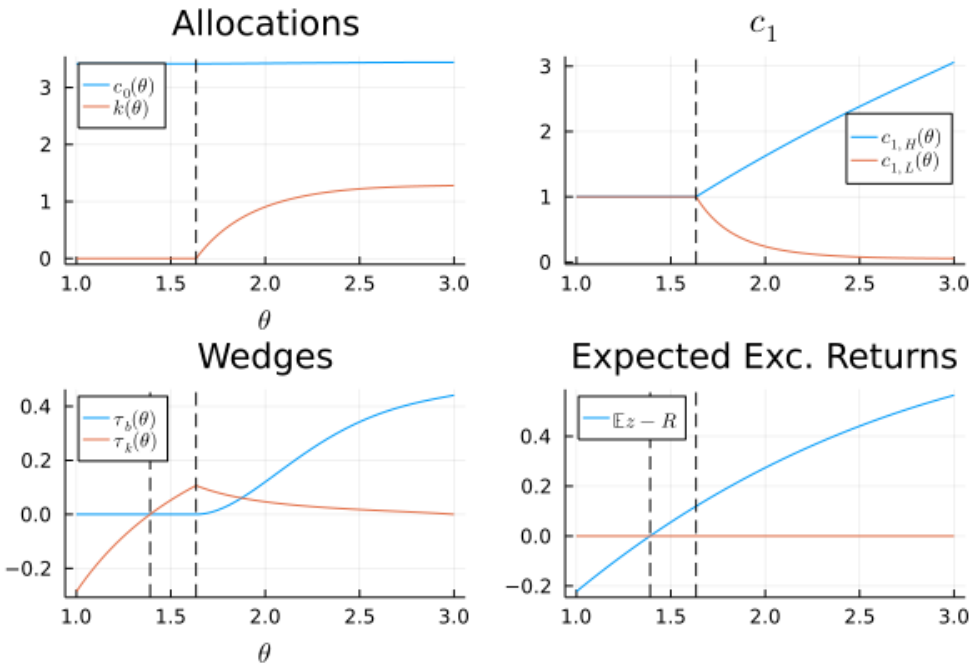


Figure 2.2: Solution to Planner's Problem

long-run wealth distribution, we now turn to a dynamic extension of the model in Section 2.3. Time is discrete, $t \in 0, 1, \ldots$ The economy is populated by a unit continuum of agents, who draw a new type θ_t in each period. For simplicity, we assume that these draws are IID across agents and time periods.

2.4.1 Household Problem

We follow Angeletos (2007) in assuming that each household has the opportunity to become and entrepreneur in each period, and operate a private business with idiosyncratic productivity z_{it}. Households also have access to a risk-free bond, which we assume to be in zero net supply. Finally, households also supply labor to the entrepreneurial sector, for which they are paid a common wage. While we microfound return heterogeneity using the assumption of private firms, we wish to stress that our results apply to return heterogeneity generally—for example, holders of public firms are also subject to similar risk to their returns.

We assume that households are taxed on their income from operating their business, if any, as well as their interest income from bonds. Households are *not* taxed on their labor income; because labor is supplied inelastically, a tax on this would amount to a lump-sum tax, which would be optimal. As in Section 2.3, we refer to a household choosing capital to run their business as "investing," and choosing bonds as "saving."

Formally, the household solves

$$\max_{\{c_t,k_{t+1},b_{t+1}\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \left(\hat{\beta}\zeta\right)^t u\left(c_t\right) \tag{2.21}$$

given an initial wealth $\bar{a}$ and type θ_0. We assume that households pass away with constant probability $1-\zeta$, and are replaced upon death with a mass of households of wealth $\bar{a}$ and θ_0 drawn at random. Due to the perpetual-youth framework, the households' effective discount rate is $\beta = \hat{\beta}\zeta$, the product of their "pure" discount factor $\hat{\beta}$ and their survival probability ζ.

Each period begins with production: given the savings choices k_t and $\hat{b}_t$ made in the prior period, the household realizes its productivity shock z_t and earns capital, interest, and wage income. At this point, the household pays taxes and receives transfers based on capital and interest income. At this point, the household also draws its next-period type θ_t, which governs the distribution for the next-period productivity shock z_{t+1}. For simplicity, we assume that types are IID, and each household draws a new θ_t at random in each period from a time-invariant distribution with CDF $F(\theta)$. Given the resulting post-tax cash on hand, and their new type, the household then makes its consumption, savings, and investment decisions for the following period. The household budget constraints are:

$$c_t + k_{t+1} + \hat{b}_{t+1} = \pi_t\left(z_t, k_t, w_t\right) + R_t\hat{b}_t + w_t + T_t\left(\pi_t, R_t\hat{b}_t\right) \tag{2.22}$$

The household divides its current cash on hand between consumption c_t and savings, which it further allocates between capital k_t and bonds $\hat{b}_t$. The household supplies one unit of labor inelastically, for which it earns wage w_t. Households earn income from two other sources. The first is from saving in bonds $\hat{b}_t$, which pay gross interest rate R_t. We allow for borrowing, up to a natural borrowing limit:

$$\hat{b}_{t+1} \geq -h_t \tag{2.23}$$

$$h_t = \sum_{s=1}^{\infty} \frac{w_{t+s} + T_{t+s}}{\hat{R}_{t+1}...\hat{R}_{t+s}} \tag{2.24}$$

The term h_t in equation (2.24) is *human wealth*: the present discounted value of all future income streams.

The other source of income for the household is capital income from operating their private business, denoted as $\pi(z_t, k_t; w_t)$. If a household chose to carry forward capital k_t from the previous period, it produces output by combining this capital with labor n_t hired on

the spot market at wage rate w_t, subject to idiosyncratic productivity z_t. Functionally, we assume that the shock z_t affects the entire capital stock. As such, household capital income, inclusive of its nondepreciated capital stock, is

$$y\left(z_t, k_t; w_t\right) = \left(z_t k_t\right)^{\alpha} n_t^{1-\alpha} + (1-\delta) z_t k_t \tag{2.25}$$

The household's profit, then, is

$$\pi\left(z_t, k_t; w_t\right) = \max_{n_t} \left(z_t k_t\right)^{\alpha} n_t^{1-\alpha} + (1-\delta) z_t k_t - w_t n_t \tag{2.26}$$

Note that the production function in (2.25) exhibits constant returns to scale in capital and labor. As a result, the optimal labor choice n_t of the entrepreneur, as well as their output y_t and profit π_t will be linear in *effective* capital $z_t k_t$:

$$n\left(z_t, k_t; w_t\right) = \left(\frac{\omega_t}{\alpha}\right)^{\frac{1}{1-\alpha}} z_t k_t \tag{2.27}$$

$$y\left(z_t, k_t; w_t\right) = \left[\frac{\omega_t}{\alpha} + (1-\delta)\right] z_t k_t \tag{2.28}$$

$$\pi\left(z_t, k_t; w_t\right) = \left[\omega_t + (1-\delta)\right] z_t k_t \tag{2.29}$$

where

$$\omega \equiv \alpha \left(\frac{1-\alpha}{w_t}\right)^{\frac{1-\alpha}{\alpha}} \tag{2.30}$$

denotes the aggregate return to a unit of effective capital zk. As a result of this linearity, we can write the household budget constraints as

$$c_t + k_{t+1} + \hat{b}_{t+1} = \hat{a}_t + w_t + T_t \tag{2.31}$$

$$\hat{a}_{t+1} = \left(1 - \tau_{t,k}\right) \pi_{t+1}\left(z_{t+1}, k_t, w_t\right) + \left(1 - \tau_{t,b}\right) \hat{R}_{t+1} \hat{b}_t \tag{2.32}$$

$$k_{t+1} \geq 0 \tag{2.33}$$

$$\hat{b}_{t+1} \geq -h_t \tag{2.34}$$

With this formulation, the household problem now involves homothetic preferences and linear budget constraints, and thus the household's value function is *homogeneous* in wealth a_t, and policy function will be linear in wealth. The same property results in Angeletos (2007), and simplifies the solution of the problem.

Finally, the function T_t denotes net taxes and transfers that the household receives, based

on its capital and interest income. For the purposes of calibration, we assume that taxes on both are linear, and thus the tax function takes the form

$$T_t\left(\pi_t, R_t\hat{b}_t\right) = \text{Transfer}_t - \tau_{t,k}\pi_t - \tau_{t,b}R_t b_t \tag{2.35}$$

were Transfer_t denotes the lump-sum transfer paid to the households from the proceeds of taxation, the size of which is determined in equilibrium.

2.4.2 Equilibrium

A household's state consists of its assets a, inclusive of human wealth, and its current type θ. Denote the time-t distribution over these types as $\Psi_t(a,\theta)$.

Definition 1 (Decentralized Equilibrium.)**.** *Given fiscal policies $\{\tau_{k,t}, \tau_{b,t}, T_t\}$, a decentralized equilibrium is a sequence of prices $\{w_t, R_t\}_{t=0}^{\infty}$ and allocations*

$$\{c_t(a,\theta), k_{t+1}(a,\theta), b_{t+1}(a,\theta)\}_{t=0}^{\infty}$$

such that:

1. *Given sequences of shocks $\{\theta_t\}_{t=0}^{\infty}$ and $\{z_t\}_{t=0}^{\infty}$ choices solve the household problem in* (2.21) *and* (2.22)*.*
2. *Aggregates represent the sum of household choices:*

$$C_t = \int c_t(a,\theta)\, d\Psi_t(a,\theta) \tag{2.36}$$

$$K_t = \int k_t(a,\theta)\, d\Psi_t(a,\theta) \tag{2.37}$$

$$B_t = \int b_t(a,\theta)\, d\Psi_t(a,\theta) \tag{2.38}$$

$$N_t = \int n_t(a,\theta)\, d\Psi_t(a,\theta) \tag{2.39}$$

$$Y_t = \int y_t(a,\theta)\, d\Psi_t(a,\theta) \tag{2.40}$$

$$\hat{K}_t = \int \theta k_t(a,\theta)\, d\Psi_t(a,\theta) \tag{2.41}$$

3. Markets clear:

$$C_t + K_{t+1} = Y_t + (1-\delta)\hat{K}_t \qquad \textit{Goods} \tag{2.42}$$

$$N_t = 1 \qquad \textit{Labor} \tag{2.43}$$

$$B_t = 0 \qquad \textit{Bonds} \tag{2.44}$$

The computation of the equilibrium in Definition 1 is simplified using a homogeneity property as in Angeletos (2007). First, define the following:

$$a_t = \hat{a}_t + h_t \tag{2.45}$$

$$b_t = \hat{b}_t + h_t \tag{2.46}$$

Now, a_t represents total wealth, the sum of financial wealth and the present discounted value of future labor income ("human" wealth). We then have the following result:

Proposition 4. *The household problem is solved by the following policy functions:*

$$c(\theta, a) = \psi(\theta)\, a \tag{2.47}$$

$$k'(\theta, a) = [1 - \psi(\theta)]\, \phi(\theta)\, a \tag{2.48}$$

$$b'(\theta, a) = [1 - \psi(\theta)]\, [1 - \phi(\theta)]\, a \tag{2.49}$$

where $\psi(\theta)$ and $\phi(\theta)$ are functions of R and ω.

There is more that we can say about the long-run steady state equilibrium. First: because types θ_t are IID, wealth and θ will be independent, and their joint distribution $\Psi(a, \theta)$ will be the product of the marginal distributions $F(\theta)$ and $\mu(a)$. As such, we can integrate separately; for instance, we have

$$\hat{K} = \int \theta k(a, \theta)\, d\Psi(a, \theta) = \int [1 - \psi(\theta)]\, \phi(\theta)\, dF(\theta) \int a\mu(a)\, da$$

Furthermore, wealth a follows a *random growth process*:

$$\frac{a'}{a} = (1 - \psi(\theta))\, [\phi(\theta)\, (\omega + 1 - \delta)\, z + (1 - \phi(\theta))\, R] \tag{2.50}$$

In particular, the growth rate of wealth is independent of wealth itself.[4] This result, combined with the assumption that agents die and are replaced by new agents of average wealth, implies

[4] In other words, the process for wealth satisfies Gibrat's law.

that the long-run distribution of wealth $\mu(a)$ will have a *Pareto tail*:

$$\lim_{a\to\infty} \Pr(x > a) = 1 - d \cdot a^{-\kappa}$$

for some d, κ. The value κ is referred to as the tail index or the Pareto parameter; the lower is $@k$, the greater is the concentration of wealth in the tail and the higher is inequality.

2.4.3 Calibration

We calibrate the model above to match features of the United States economy, both in the aggregate and in the cross-section of household returns. We assume a time period of one year. To begin, we assume that θ_t is distributed according to a standard lognormal:

$$\ln\theta \sim \mathcal{N}\left(0, \sigma_\theta^2\right) \tag{2.51}$$

We solve the model on a truncated support for θ: $\Theta = [-3\sigma_\theta, 3\sigma_\theta]$, which captures 99.7% of all θ values. For the ex-post shocks, we assume that for a type θ,

$$z \sim \Gamma\left(a(\theta), b(\theta)\right) \tag{2.52}$$

where

$$a(\theta) = \left(\chi\theta^\rho\right)^{-2} \tag{2.53}$$
$$b(\theta) = \chi^2\theta^{1+2\rho} \tag{2.54}$$

for some parameters χ, ρ. This formulation ensures that

$$\mathbb{E}[z|\theta] = \theta \tag{2.55}$$
$$\mathrm{Sd}[z|\theta] = \chi\theta^{1+\rho} \tag{2.56}$$

Furthermore, the coefficient of variation is

$$\frac{\mathrm{Sd}[z|\theta]}{\mathbb{E}[z|\theta]} = \chi\theta^\rho \tag{2.57}$$

This formulation therefore ensures that expected returns are increasing in θ, as are the size of the shocks relative to their mean, provided $\rho > 0$.

Table 2.1 lists the parameters that are calibrated externally. The parameter ζ, which is the probability of surviving from one period to the next, pins down a working life of

$(1-\zeta)^{-1} = 40$ years. The value α determines the relative factor shares, and targets a labor share of 60%, approximately equal to the value in the US data. Finally, the flat tax rates τ_k and τ_b are taken from McDaniel (2007), who computes average tax rates across the population by dividing aggregate tax receipts by the corresponding base. She calculates the average tax rate on capital income to be approximately 25%, as of 2003. In the US, interest income from risk-free assets is taxed at the individual's corresponding labor income tax rate. As such, for the tax on risk-free savings τ_b I use the average tax rate on labor income, which McDaniel (2007) calculates to be approximately 20%. There are five parameters remaining:

Table 2.1: Externally Calibrated Parameters

Parameter	Value	Target
ζ	0.975	Working life of 40 years
α	0.4	Labor share 60%
$\mathbb{E}\ln\theta$	0	Normalization
$\ln\underline{\theta}, \ln\overline{\theta}$	$\pm 3\sigma_\theta$	Capture 99.7% of all θ values
τ_k	0.25	Capital income tax rate
τ_b	0.2	Savings income tax rate

the "pure" rate of time preference $\hat{\beta}$, depreciation δ, the Variance of types σ_θ^2, and χ and ρ, which govern the distribution of z conditional on θ. With these, we target the ratios of the capital stock and consumption to GDP, using standard values from the National Accounts:

$$\frac{K}{Y} = 3 \qquad \frac{C}{Y} = \frac{2}{3} \tag{2.58}$$

Additionally, we target a value for the Pareto tail parameter of the wealth distribution $\kappa = 1.5$, roughly equal to its value in the US data (Vermeulen, 2018). Finally, we target the first and second moments of returns to wealth as documented in Fagereng et al. (2020). For a household i, these authors define the return to wealth $r_{i,t}^n$ as

$$r_{it}^n = \frac{y_{it}^f + y_{it}^r - y_{it}^b}{w_{it}^g + \frac{F_{it}^g}{2}} \tag{2.59}$$

where y_{it}^f represents income from financial assets such as stocks and bonds, y_{it}^r income from real assets including private businesses, and y_{it}^b debt expenses. Net worth is w_{it}^g, and the term F_{it}^g captures flows to and from net worth during the period to avoid overestimating returns. The authors use pre-tax incomes to measure returns. Furthermore, income from businesses is measured as the household's share of retained profits in the business, before taxes and depreciation; that is, they use EBITDA. The analogous return to wealth in our

model is

$$r_a(\theta, z) = (1 - \psi(\theta)) [\phi(\theta) \omega z + (1 - \phi(\theta)) r] \tag{2.60}$$

Note that this return r_a excludes the principal on bond income and any non-depreciated capital, and is taken pre-tax. Note also that due to the linearity of policy functions, the return to wealth is independent of wealth itself.

Fagereng et al. (2020) weight by wealth, and find that the mean return to wealth is 3.65% per year, and the standard deviation of these returns is about 7.81%. Therefore, our final two targets are

$$\mathbb{E}[r_a] = 0.0365 \tag{2.61}$$

$$\text{Sd}[r_a] = 0.0781 \tag{2.62}$$

Where the expectation is taken over all (θ, z).

2.4.4 Planner's Problem

We now consider the problem of the planner. As in Section 2.3.2, we work in the space of allocations, and solve the problem of a planner who collects reports of types θ_t and observes capital incomes y_t, inclusive of depreciation. We consider again the cost minimization problem for a planner assigned at time zero a minimum level of utility U^*, and must minimize the cost of delivering this level of average lifetime utility. When agents pass away, they are replaced with new agents with type θ_0 drawn at random from $F(\theta)$, and common promise utility $\bar{w}$.

Formally, for a variable x_t, denote the time-t history as $x^t \equiv \{x_0, x_1, ..., x_t\}$. The planner observes histories (θ^t, y^t) and chooses allocations

$$\left\{c_t\left(\theta^t, y^t\right), k_{t+1}\left(\theta^t, y^t\right), n_{t+1}\left(\theta^t, y^t\right)\right\} \tag{2.63}$$

where n_{t+1} denotes the labor hired by the household to run its private firm, if any. Before defining the planning problem, note that because production takes place using the technology in (2.25), optimal labor demand is again linear in effective capital $z_t k_t$. We assume that labor hired is observable to the planner, and thus the planner is able to observe labor productivity:

$$\begin{aligned} y_t &= (z_t k_t)^\alpha n_t^{1-\alpha} \\ &= \hat{y}_t n_t^{1-\alpha} \end{aligned}$$

From the planner's perspective, $\hat{y} = (zk)^\alpha$ is labor productivity. Conditional on a level of $\hat{y}$,

the planner directs the household to hire labor until it is equated with its marginal product:

$$w_t = (1-\alpha)\,\hat{y}_t n_t^{-\alpha} \tag{2.64}$$

Replacing for labor demand in the production function implies that the output produced by a household with capital k_t, conditional on their productivity shock z_t, is once again $y_t = (\omega_t + 1 - \delta)\, z_t k_t$ where ω_t is once again the aggregate return to a unit of effective capital zk, defined as in (2.30).

Let $\mu_t(\theta^t, y^t)$ denote the time-t joint distribution over histories. By selecting the allocations in (2.63), the planner solves

$$\begin{aligned}\min \sum_{t=0}^{\infty}\left(\prod_{s=0}^{t-1} R_s\right)^{-1}\Bigg\{H_t + \int\left[c_t\left(\theta^t, y^t\right) + k_{t+1}\left(\theta^t, y^t\right)\right] d\mu_t - \\ \int\left(\int\left(\frac{\omega}{\alpha} + 1 - \delta\right) z_t k_t\left(\theta^{t-1}, y^{t-1}\right) dG\left(z_t|\theta_{t-1}\right)\right) d\mu_{t-1}\Bigg\}\end{aligned} \tag{2.65}$$

such that her allocations deliver the given utility:

$$U^* \geq \sum_{t=0}^{\infty}\beta^t u\left(c_t\left(\theta^t, y^t\right)\right) d\mu_t \tag{2.66}$$

and satisfy incentive constraints. To formulate the incentive constraints, we define promise utility as

$$w_{t+1}\left(\theta^t, y^{t+1}\right) = \sum_{s=t+1}^{\infty}\beta^{s-t-1}\int u\left(c_s\left(\theta^s, y^s\right)\right) d\mu_s\left(\theta^s, y^s|\theta^t, y^t\right) \tag{2.67}$$

Note that promise utility w_{t+1} is contingent upon the realization of y_{t+1}. As in Section 2.3, we consider two cases: **Second-best**, wherein the planner can observe θ_t but not k_{t+1}, and **Third-best**, wherein the planner can observe neither θ_t nor k_{t+1}. To formulate both sets of incentive constraints, we denote with $H\left(y_{t+1}|\theta, k\right)$ the distribution for next-period output conditional on type θ and investment k. Under **second-best**, the planner's allocations must satisfy

$$\begin{aligned}u\left(c_t\left(\theta^t, y^t\right)\right) + \beta\mathbb{E}_t w_{t+1}\left(\theta^t, y^{t+1}\right) \geq \max_{\hat{k}} u\left(c_t\left(\theta^t, y^t\right) + k\left(\theta^t, y^t\right) - \hat{k}\right) + \\ \beta\int w_{t+1}\left(\theta^t, \left\{y^t, y_{t+1}\right\}\right) dH\left(y_{t+1}|\theta_t, \hat{k}\right)\end{aligned} \tag{2.68}$$

The constraint in (2.68) has the same interpretation as in the static model: given that θ is observable, the planner must ensure that the agent is no worse off following the prescribed

level of investment k_{t+1} than he is deviating, investing some other amount $\hat{k}$. The incentive constraint under **third-best** is similar:

$$u\left(c_t\left(\theta^t, y^t\right)\right) + \beta \mathbb{E}_t w_{t+1}\left(\theta^t, y^{t+1}\right) \geq \max_{\hat{\theta}, \hat{k}} u\left(c_t\left(\left\{\theta^{t-1}, \hat{\theta}\right\}, y^t\right) + k\left(\left\{\theta^{t-1}, \hat{\theta}\right\}, y^t\right) - \hat{k}\right) + \beta \int w_{t+1}\left(\left\{\theta^{t-1}, \hat{\theta}\right\}, \left\{y^t, y_{t+1}\right\}\right) dH\left(y_{t+1} | \theta_t, \hat{k}\right) \tag{2.69}$$

Here, the planner once again faces a double-continuum of constraints: the agent can lie about his type and report some other type $\hat{\theta}$, *and* invest a deviant amount $\hat{k}$.

2.4.5 Recursive Formulation and Homogeneity

Because types θ_t are IID, promised utility $w_t\left(\theta^{t-1}, z^t\right)$ is a sufficient statistic for the history $\left(\theta^{t-1}, z^t\right)$. Hence, we can further simplify the problem by focusing on the problem of a *component planner*, as in Albanesi and Sleet (2006), who is tasked with delivering promised utility w_t to all agents with this value for the state. To do so, the component planner chooses allocations $\left\{c_t\left(w, \theta\right), k_{t+1}\left(w, \theta\right)\right\}$ along with a schedule of promise utility values $\left\{w_{t+1}\left(w, \theta, y\right)\right\}_{y \in \mathbb{R}_+}$ contingent upon the realization of next-period output. These allocations solve a version of (2.65) from time t onward. Furthermore, we focus on the problem of a component planner in the steady state, where aggregate prices and quantities are constant. This problem admits a recursive formulation, which is as follows:

$$C\left(w\right) = \min \int \left[c\left(\theta\right) + k\left(\theta\right) - p - \left(\frac{\omega}{\alpha} + 1 - \delta\right) \int z k\left(\theta\right) dG\left(z|\theta\right) + \frac{1}{R} \int C\left(w'\left(\theta, z\right)\right) dG\left(z|\theta\right)\right] dF\left(\theta\right) \tag{2.70}$$

The promise-keeping constraint is

$$w = \int U\left(\theta\right) dF\left(\theta\right) \tag{2.71}$$

$$U\left(\theta\right) = u\left(c\left(\theta\right)\right) + \beta \int w'\left(\theta, z\right) dG\left(z|\theta\right) \tag{2.72}$$

Focusing on the **second-best** case, where θ is observable to the planner, we derive a recursive formulation for (2.68). To simplify this continuum of constraints, we enforce a *local* incentive

constraint as in Jewitt (1988):

$$u'\left(c\left(\theta\right)\right)k\left(\theta\right)=\beta\int w'\left(\theta,z\right)\left[\frac{-g\left(z|\theta\right)-zg_{z}\left(z|\theta\right)}{g\left(z|\theta\right)}\right]dG\left(z|\theta\right) \tag{2.73}$$

The constraint in (2.73) requires that compliant investment be at a stationary point in the choice of $\hat{k}$ in (2.68), rather than a global optimum. With our assumption on $g(z|\theta)$ in (2.54)-(2.56), (2.73) becomes

$$u'\left(c\left(\theta\right)\right)k\left(\theta\right)=\beta\int w'\left(\theta,z\right)\left(\frac{z-\theta}{b\left(\theta\right)}\right)dG\left(z|\theta\right) \tag{2.74}$$

2.4.6 Allocations and Wedges in the Infinite Horizon

As in the decentralized model, the planner's problem in (2.65)-(2.73) is homogeneous in promise utility w: an agent entering the period with double the promise utility will be allocated twice the *current* utility. Formally, we have the following proposition:

Proposition 5. *Assume that utility is CES:* $u(c)=c^{1-\gamma}/(1-\gamma)$ *with* $\gamma\neq1$. *The steady-state allocations solve*

$$\begin{aligned} C\left(w\right)&=A\left[\left(1-\gamma\right)w\right]^{\frac{1}{1-\gamma}}-H & \qquad U\left(\theta,w\right)&=U\left(\theta\right)\left(1-\gamma\right)w\\ k\left(\theta,w\right)&=k\left(\theta\right)\left[\left(1-\gamma\right)w\right]^{\frac{1}{1-\gamma}} & \qquad w'\left(\theta,z,w\right)&=w'\left(\theta,z\right)\left(1-\gamma\right)w\\ c\left(\theta,w\right)&=c\left(\theta\right)\left[\left(1-\gamma\right)w\right]^{\frac{1}{1-\gamma}} \end{aligned}$$

for some functions $k\left(\theta\right),c\left(\theta\right),U\left(\theta\right),w'\left(\theta,z\right)$, *where* $H=Rp/(R-1)$ *is the present discounted value of all future labor income, including that in the current period.*

As an immediate corollary to Proposition 5, it can be shown that the optimal steady-state wedges are independent of history (as summarized by promise utility w):

Corollary 1. *The wedges implied by the constrained-efficient allocations in the component planner's problem in* (2.65)-(2.73) *are independent of history* w, *and instead depend only on the current type* θ.

Another result of Proposition 5 is that, like wealth in the decentralized economy, promised utility follows a random growth process:

$$\frac{w'\left(\theta,z,w\right)}{w}=w'\left(\theta,z\right)\left(1-\gamma\right)$$

with the growth rate independent of w. As a result, the long-run distribution of promised utility $\hat{\mu}(w)$ will also have a Pareto tail. One question which we address with our numerical model is how the two tails compare, which gives insight as to whether the planner ultimately prefers more or less inequality than under the status quo.

We define an equilibrium in a similar way as in Definition (1):

Definition 2 (Constrained-Efficient Equilibrium)**.** *A constrained-efficient, steady-state equilibrium is a set of allocations*

$$\{c(\theta, w), k(\theta, w), U(\theta, w), w'(\theta, w, z')\}$$

such that

1. *The planner's objective in* (2.70) *is maximized, subject to the constraints in* (2.72) *and* (2.73)*.*
2. *Aggregates are consistent with allocations and the respective distributions of types θ and promise utilities w:*

$$K = \int k(\theta)\, dF(\theta)\, W \tag{2.75}$$

$$\hat{K} = \int \theta k(\theta)\, dF(\theta)\, W \tag{2.76}$$

$$C = \int c(\theta)\, dF(\theta)\, W \tag{2.77}$$

$$Y = \frac{\omega}{\alpha} \int \theta k(\theta)\, dF(\theta)\, W \tag{2.78}$$

$$L = \left(\frac{\omega}{\alpha}\right)^{\frac{1}{1-\alpha}} \int \theta k(\theta)\, dF(\theta)\, W \tag{2.79}$$

where

$$W = \begin{cases} \int e^{(1-\beta)w} d\mu(w) & \gamma = 1 \\ \int [(1-\gamma)w]\, d\mu(w) & \gamma \neq 1 \end{cases} \tag{2.80}$$

and $\mu(w)$ is the stationary distribution of promise utilities.

3. *Markets clear:*

$$C + K = Y + (1-\delta)\hat{K} \tag{2.81}$$

$$L = 1 \tag{2.82}$$

2.5 Conclusion

This paper has studied the optimal taxation of capital income when agents are heterogeneous in the rates of return on their investments. In a static setting, we have demonstrated that the optimality of positive capital income taxation established in Golosov et al. (2003) is preserved when agents earn heterogeneous returns, rather than a common rate. Furthermore, we demonstrate that informational frictions create a disagreement between households and the planner, wherein under constrained efficiency the planner does not find it optimal to allocate investment to all households with expected returns above the risk-free rate. Therefore, in this setting, the tax code acts on entrepreneurship along both the intensive and extensive margins: through a combination of distortions on both investment and risk-free savings, the government ensures that the correct types select into entrepreneurship (invest a positive amount), and that these entrepreneurs invest the socially-optimal level into their business. These forces give rise to a wedge on investing that is nonmonotonic over the space of types, rising up to the point of socially-optimal entry to ensure that types who would invest absent distortions instead abide by the planner's recommendation to lend their capital to more productive entrepreneurs. For those agents who *do* invest, a decreasing pattern of distortions on doing so ensures that more productive entrepreneurs invest higher levels. The wedge on borrowing and lending, meanwhile, is zero up to the point of entry, and increasing thereafter. This pattern takes advantage of the full mobility of capital in our economy, ensuring that less-productive agents lend their capital to those with higher productivity so that capital is allocated efficiently. We also derive conditions under which informational frictions in our model produce what we refer to as "endogenous span-of-control," a constrained-efficient allocation under which entrepreneurs other than the most productive type invest. We demonstrate that a crucial condition for such an allocation is to enforce a lower bound on utility, as allocating all investment to the most productive agents require that these agents be punished with arbitrarily low utility upon the realization of an unsuccessful project.

In the dynamic context, we exploit the homogeneity of the planner's value function in order to simplify the component planning problem, which enables us to solve the model by simply solving the problem for a single history. We show in this context that wedges are independent of history, and instead depend only on the agents' current types. As in the static context, the tax code on capital income from both risk-free savings and risky investment ensures optimal selection into entrepreneurship, and optimal investment thereupon. We show that in the dynamic context, informational frictions exert a strong downward force on investment. Finally, the long-run wedges largely mirror their static counterparts: the distortion to capital income acts on both the intensive and extensive margins of entrepreneurship, ensur-

ing that the correct households select into operating private businesses, and that those who do enter invest the correct amount. The tax on outside saving, meanwhile, incentivizes investor households to take on leverage to increase their investment, rather than to self-insure against unfavorable ex-post returns.

Moving forward, there are a number of extensions and refinements that we aim to add to our model. First, while our infinite-horizon model is informative regarding the qualitative ways in which the tax code operates on investment, it is not quantitatively realistic. Our current work is focused on calibrating an economy in the style of Angeletos (2007) to the US data on aggregate output and wealth inequality, and studying the optimal tax code in such a model. In addition to information on the optimal long-run tax code in the US economy, we can also use such a model to compare the optimal level of long-run wealth inequality to its empirical counterpart. In addition, we wish to consider the case where returns are persistent, using the techniques from Farhi and Werning (2013). Though the assumption of autocorrelation in returns adds complexity to the analysis of our model, we feel that it is a more natural assumption—it seems plausible that entrepreneurial talent should be at least somewhat persistent. Additionally, we will attempt to construct a tax and transfer scheme that implements the constrained-efficient allocations in a competitive equilibrium.

Finally, we would like to add a competitive labor market to the model, consisting of workers who are employed by entrepreneurs. As shown by Scheuer (2014), in the presence of both workers and entrepreneurs, the tax code plays a key role in the decision to become an entrepreneur, in addition to its obvious objective of redistribution from high-income capitalists to lower-income workers. In particular, because the workers are employed by the entrepreneurs, there is a question as to whether the government is better off taxing entrepreneurs at a higher rate, and redistributing the proceeds to workers, or taxing business income at a lower rate and relying on market forces to redistribute this windfall to workers (the "trickle-down" approach). In the framework of Scheuer (2014), the former approach is optimal. Importantly, though, entrepreneurs in his model are heterogeneous not in their rates of return, but in the utility cost of running their own business. We would like to study whether the superiority of direct taxation and transfers, rather than a "trickle-down" scheme, remains when entrepreneurial businesses do in fact earn different returns. Adding this dimension to our model would help us gain a better understanding of how capital income ought to be taxed in a more realistic environment, in which the lower portion of the wealth distribution is populated by laborers, and the thick upper tail by the entrepreneurs by whom they are employed.

Chapter 3

Mobility

***Joint with Daniel Carroll*[1] *and Eric R. Young*[2]**

3.1 Introduction

This paper examines wealth mobility in both the US data, and in a simple dynamic stochastic general equilibrium model with incomplete markets in the spirit of Bewley (1986), Aiyagari (1994a), and Huggett (1993). This model, with its many variations, has become the workhorse model of macroeconomics in large part because it generates an endogenous distribution of agents across income and wealth. This endogenous distribution is ideal for studying the effects of policy on inequality. Very little is understood in this environment regarding wealth mobility—the frequency with which agents "switch places" in the wealth distribution. Our paper fills the gap: we document the inability of these standard models to generate realistic short-run wealth mobility, and explore the data to find evidence for augmentations to the standard model to rectify this discrepancy.

Mobility is distinct from inequality. Of course, inequality is a necessary condition for mobility—if everyone holds the same wealth, it makes no sense to talk about households switching places in any distribution—but inequality can arise in the absence of mobility as well. For example, without risk, inequality can be not only present but permanent, depending on how savings choices vary in the population, while mobility may be zero as agents remain frozen in their ordering within the wealth distribution. Thus, inequality on its own paints an incomplete picture of the social opportunities that households face. In order to fully assess how equally resources are distributed we need to look beyond a "snapshot" of the distribution at one point in time; we also need to evaluate the frequency with which households move

[1]Federal Reserve Bank of Cleveland, daniel.carroll@clev.frb.org
[2]University of Virginia, ey2d@virginia.edu

within that distribution.

In addition to a better understanding of opportunity, thie question of mobility has concrete policy implications. In Guvenen et al. (2015a), the optimality of wealth taxes relative to capital income taxes depends on the mobility of agents. In their environment, a switch from a capital income tax to a wealth tax induces a gain in output by shifting the tax burden from low to high-return investors; this gain in efficiency depends on the ability of younger households to supplant the incumbent wealthy at the top of the distribution. In Carroll et al. (2016), voting with commitment to future taxes implies that agents who expect to transit quickly are reluctant to impose high taxes on capital income and high transfers, since they expect to rapidly end up on the wrong side of the wealth transfer distribution (with a skewed distribution the mean wealth type loses wealth to the median type under majority voting).

In this paper, we measure the degree of wealth mobility present in the US data, and compare this to the wealth mobility implied in the steady-state of an incomplete markets model. To this end, we first introduce a consistent, interpretable way in which to measure mobility. Generally speaking, our process for measuring mobility is as follows. We take the starting and ending wealth distributions, order households by wealth, and divide them into bins, or quantiles. For each household, we then note the quantile in which it begins (its place in the starting distribution), and the quantile in which it ends up after the prespecified amount of time (ending distribution). We use this information to construct a Markov transition matrix, where entry (i, j) is the probability that a household beginning the sample period in quantile i will find itself in quantile j by the end of the time horizon. Finally, we several measures of mobility found in the literature to summarize the resulting transition matrix into a single number, which captures some aspect of how much mobility the matrix exhibits. Due to the information loss inherent in reducing a matrix of multiple rows and columns to a single number, we consider several complementary measures, each of which captures different degrees of movement that make up overall mobility. To keep our results consistent across data and model results, we focus on *quintiles*, dividing the wealth distribution into five bins, each containing twenty percent of households.

We begin by studying wealth mobility in the data, using household-level panel data on wealth collected by the Panel Study of Income Dynamics (PSID). Our focus is on short-run wealth mobility, which we define as movements over a five or six-year span.[3] Here we find that even over this relatively short time horizon, households in the US exhibit a significant amount of wealth mobility. The middle of the distribution, from the twentieth to eightieth percentiles, exhibits a substantial degree of mixing: consistently across surveys going back to 1984, households in this range are more likely to depart their starting quintile than to remain

[3]Whether we consider a horizon of five or six years depends on the availability of PSID data.

in it. Mobility is present at the extremes as well; for instance, across all of our samples, a household beginning a period in the top quintile of wealth has at least a 29% chance to move down withing five to six years. Furthermore, the mobility matrices calculated from the PSID data show statistically significant proportions of "jumps," in which a household moves two or more quintiles in a given direction.

We then turn to a standard heterogeneous-agents, incomplete markets model. Using a reasonable calibration for the income process (taken from Floden and Lindé, 2001) with a high persistence of shocks, we find that the benchmark model delivers too little short-run mobility relative to the data (over five-year horizons). Specifically, we find that the model implies far too little mobility overall, but in particular it fails to deliver the high mobility observed in the lowest and highest quintiles; in the model, households stay in these quintiles on average 38 and 63 years, in contrast to values in the data closer to 15 and 17 years, respectively. Furthermore, households in the model also stay in their initial quintile too frequently, and when they move they move only one quintile at a time; in the data wealth moves more rapidly, with significant numbers of households switching more than one quintile in either direction.

Digging deeper, we uncover the forces which prevent the workhorse model from replicating wealth mobility in the data. We find that overall wealth mobility is "hump-shaped" in the persistence of income shocks: mobility is low for near-IID and near-permanent income shocks, and higher for intermediate values. To explain this result, we decompose mobility into two components: *structural* mobility, resulting from changes in the wealth distribution, and *exchange* mobility, resulting from changes in agents' savings following changes in income. We further decompose exchange mobility into a *behavioral* component, which measures pure changes in savings behavior, and a *luck* component, which captures mobility resulting from agents' experiencing different ex-post shocks. We find that as the persistence of income changes, structural mobility remains relatively fixed; agents' movements scale with the overall distribution of wealth. However, we find that the behavioral and luck components act as countervailing forces: as income shocks become more persistent, agents experience longer strings of shocks in one direction or another, which increases mobility through the "luck" channel. However, this increase in mobility through luck is offset through the behavioral channel: the Permanent Income Hypothesis implies that when agents' incomes are more persistent, they adjust their consumption by more (and thus savings by less) in response to a change in income. These countervailing forces prevent the standard model from producing empirically realistic wealth mobility.

We then move to consider augmentations used elsewhere in the literature to allow the baseline model to account for wealth inequality. We first examine the Krusell and Smith

(1998) modification that introduces stochastic movements in discount factors that are highly persistent. The discount factor model improves a small amount by increasing mobility at the lower end of the wealth distribution, but actually reduces it at the high end; the reason the model gets high wealth concentration is that it essentially "freezes" rich households in the top quintile, since high discount factor types will save a significant amount whether their income is high or not, and these discount factor states must be very persistent if they are to match wealth inequality.

We then examine a "rockstar" model as in Castaneda et al. (2003), in which the earnings process has a rare and transitory state with very high income and a relatively high probability of dropping to the lowest state. The rockstar model works relatively well, as it increases mobility across the board and introduces some households that shift across more than one quintile; nevertheless, mobility at the high end is still substantially too low as households do not choose to let their wealth fall fast enough. This failure can be understood as the result of standard buffer-stock behavior combined with decreasing absolute risk aversion: with high temporary income, households save rapidly to move away from the borrowing constraint but dissave slowly. Furthermore, the rockstar model requires an earnings process unlike anything in the data (see Guvenen et al., 2015bb).

Having shown that the basic model does not replicate the wealth mobility statistics particularly well, we next drill deeper into the facts: can we learn anything about why these families move up and down the wealth distribution? We run probits to study the determinants of the probability that a family makes a "jump" (a movement of more than two quintiles) over a five-year horizon. First, we see that families that make one jump are significantly more likely to make another jump; that is, some families are just more mobile than others. Second, we find that the portfolio of the family is critical for these jumps: families with stocks are more likely to move up and less likely to move down, and families with private business income are more likely to move up and down as well as more likely to jump up or down.[4] Third, families that experience a marriage, divorce, or inheritance are also more likely to move and jump (in the obvious directions).

We find that the most straightforward way to introduce direct wealth shocks into the model is to allow for heterogeneity in agents' rates of return, as in Benhabib et al. (2019). Persistent heterogeneity in returns allows our model to replicate the level of inequalty and mobility present in the data. These shocks are particularly useful for generating realistic *downward* wealth mobility: households who find themselves locked into a string of successive poor returns will see their wealth mechanically eroded by these returns, and will save less in

[4]Quadrini (1999) and Quadrini (2000a) also point out the connection between private business activity and mobility, but only over the period 1984-1989.

the face of these returns as well. This dual force allows us to reconcile our model with the large downward movements seen in the data, and to match the propensity for agents in the highest quintiles to depart.

Finally, we study households' preferred tax policy in our model, and argue that models of policy should aim to replicate the wealth mobility in the data. To do so, we fix a parameterization of our model with return shocks that produces realistic inequality and mobility in wealth, and compute within this model the linear tax rate on capital income that households prefer, given that the proceeds are rebated to these households lump-sum. We then perform the same experiment in a number of different models that replicate inequality, but *not* mobility. For this comparison we consider both alternative return processes in our model with return shocks, and a model of discount factor heterogeneity following Krusell and Smith (1998). We find that the optimal capital tax rate varies across economies different levels of mobility, even though all exhibit the same level of inequality. The reason for this difference is that in our preferred model, agents face idiosyncratic, uninsurable risk to the return on their savings. All else being equal, then, agents prefer a safer savings vehicle for consumption smoothing, and the tax-and-transfer scheme provides this insurance.

3.1.1 Related Literature

3.2 Measuring Mobility

The literature on measuring income mobility with transition matrices dates back to at least as early as Prais (1955), who examined transitions between occupational classes in England. There is no standardized measure in part because there are many aspects to mobility.[5] In this paper, we are primarily interested in so-called *relative mobility*. Relative mobility measures how likely it is that a household in wealth quantile n_1 at time s will be in some other quantile n_2 at time $s+t$, where t is a fixed number of periods in the future. Here we review the process by which we measure mobility to allow for comparisons across steady states, models, and time periods.

Formally, represent by $x(\Gamma)$ the distribution over each of N wealth quantiles (i.e., $x = [\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}]$) and by $q(\Gamma)$ the wealth values defining the quantiles. That is,

$$q(\Gamma) = [q_1, q_2, ..., q_N]$$

where $q_1 = \underline{a}$ and $q_i = a_i : \sum_{j=1}^{J} \int d\Gamma(a, \varepsilon_j) \mathbf{1}_{\{q_{i-1} \leq a < q_i\}} = \frac{1}{5}$, for $i = \{1, ..., N\}$. The q_i

[5] For a broad overview of the literature, see Fields and Ok (1999).

values define the cutoff wealth values for entering the i^{th} quantile (the lowest wealth value in the quantile). Further, denote by $Q_i = \{a : a \in [q_i, q_{i+1})\}_{i=1,N-1}$ and $Q_N = \{a : a \geq q_N\}$; these sets define the wealth levels that constitute a given quantile. Finally, let $M_{NxN}(\Gamma)$ be a regular transition matrix induced by Γ with the element m_{ij} indicating the probability that a household in quantile Q_i will be in quantile Q_j after some fixed number of periods.[6]

We will consider four measures from the literature, discussed at length in Dardanoni (1993). In particular, we highlight how each measure captures somewhat different aspects of mobility; due to the loss of information generated by moving from a matrix to a scalar, we find it important to consider multiple measures.

Shorrocks Measure The Shorrocks (1978) measure of mobility focuses on the probability weight along the diagonal of M. One interpretation of the measure is that it reports the "stickiness" of initial conditions. Formally, the Shorrocks measure is

$$\mu_S(M) = \frac{N - \text{trace}(M)}{N - 1}.$$

The Shorrocks measure takes values between 0 and 1, with smaller values indicating a lower likelihood that a household will escape its initial quantile. Importantly, the measure is unaffected by a reallocation of mass along off-diagonal elements. The Shorrocks measure makes no distinction between economies where households move immediately from rags to riches and those where the poor become only slightly less poor.

Bartholomew's Immobility Measure In contrast to the Shorrocks measure, Bartholomew and Bartholomew (1967) deals exclusively with the off-diagonal elements:

$$\mu_B(M) = \frac{1}{N-1} \sum_{i=1}^{N} \sum_{j=1}^{N} m_{ij} |i - j|$$

is the expected number of quantiles a household would cross into each period. The measure puts positive weight only on the off-diagonal probabilities. The term $|i - j|$, the absolute number of quantiles crossed into, places more weight on transitions that cross multiple quantiles; a transition matrix with more probability mass further from the diagonal has

[6] According to Theorem 4.1.2 in Kemeny and Snell (1976)), a transition matrix is regular if and only if for some $t > 0$, M^t has no zero entries. Regularity guarantees that starting from any state in the Markov chain any other state can be visited in a finite amount of periods (that is, all states communicate). This condition is related to the "monotone mixing condition" (see Hopenhayn and Prescott (1992)) used to prove the existence of a stationary distribution Γ, which Ríos-Rull (1998) labels "the American Dream and the American Nightmare" condition. This condition is a long-run mobility requirement, whereas we are interested in short-run effects.

greater mobility (like Π_B in the previous subsection). Fields and Ok (1999) point out that Bartholomew's measure can be thought of as capturing total movement; economies in which households oscillate between being very rich and very poor would be measured as much more mobile than those where households transition more slowly through adjacent quantiles, even if the former involved fewer such transitions. In Appendix 3.10.1, we show examples which illustrate the different aspects of mobility that the Bartholomew and Shorrocks measures capture.

Second Largest Eigenvalue The second largest eigenvalue of a stochastic matrix governs the mixing rate of a Markov chain process, where a larger eigenvalue implyies a slower mixing rate. Let $\lambda_i(M)$ be the i^{th} largest eigenvalue of M. A natural measure of mobility is $\mu_{2E}(M) = 1 - |\lambda_2(M)|$. Because M is regular $\lambda_1 = 1$, and $\lambda_i < 1$ for all $i > 1$. Sommers and Conlisk (1979) show that $\mu_{2E}(M)$ measures the total deviation of M from a matrix with perfect mobility.[7] To understand why this measure captures mobility, we show in Appendix 3.10.1 for a two-state Markov chain that the second highest eigenvalue is equal to the autocorrelation of the chain. This result also holds if we confine ourselves to only Markov chains generated using the Rouwenhorst method, which preserves the autocorrelation of the two-state process as additional states are introduced.

Mean First Passage Time The mean first passing matrix $T(M)$ measures the expected number of periods until a household initially in quintile i first arrives in quintile j. Mean first passage time (MFP), then, is the expected number of periods before one household enters the quintile of another household when both are drawn at random from the wealth distribution Γ. For ease of comparison to the other measures, we define

$$\mu_{MFP}(M) = \frac{N}{MFP}.$$

thereby normalizing $\mu_{MFP}(M)$ between zero and one. This measure can therefore be conceived of as a measure of the "speed" of agents' transitions: because N is the number of quantiles, and MFP loosely speaking is the expected periods to transition from one quantile to another, $\mu_{MFP}(M)$ is denoted in units of quintiles per period. If M is "perfectly mobile" $\mu_{MFP} = 1$; agents can expect to move one quantile per period, on average. As the diagonal elements of M approach one, $\mu_M FP \to 0$.[8]

[7] Perfect mobility for an NxN matrix is one with all elements equal to $1/N$. This concept is related to "origin independence:" the probability that an agent ends up in a given quintile is independent of their starting position.

[8] μ_{MFP} cannot exceed 1 if M is *monotone* (i.e., each row is stochastically dominated by the one below it). Huggett (1993) proves the monotonicity of M in Bewley models with positively autocorrelated shocks.

So far we have defined these measures generally for any set of evenly spaced quantiles. In the remainder of this paper, we will restrict our attention to quintiles, that is, $\dim(M) = 5$.

3.2.1 Components of Mobility

Structural and Exchange Mobility We are concerned with how quickly and to what extent agents change their ordering within the stochastic stationary distribution of wealth (known as *relative mobility*). In the steady state, households' wealth positions change, but the wealth distribution itself is time-invariant. It would be intuitive to presume that relative mobility is just a simple function of the rates at which agents accumulate wealth and that greater relative mobility implies that households transition more quickly through quintiles, by rising and falling over a shorter time span. This *exchange* or *pure* mobility, however, is only one component of relative mobility. Differences in relative mobility can also arise from changes in the shape of the wealth distribution, even if individual savings behavior is the same. This concept is called *structural mobility*, and it can appear in the data when wealth inequality changes over time. In the stochastic steady state of a Bewley model, wealth inequality does not change over time. Nevertheless, structural mobility must still be taken into account when comparing the steady states from two models. Because of general equilibrium effects, changes in the model environment induce changes in the shape of the stationary distribution as well and are likely to alter the cutoffs defining wealth quintiles.

To illustrate, consider two distributions of wealth, Γ_1 and Γ_2, and let Γ_2 be a mean-preserving spread of Γ_1. Take a household from each distribution and label them according to their distribution of origin. Because there is more wealth inequality in Γ_2 than in Γ_1, the cutoffs that define the quantiles will be spread more apart. Even if households 1 and 2 begin with the same initial wealth, have the same optimal saving policies, and experience identical realizations of labor productivity, household 2 will transition across quantiles less frequently over the same amount of time, and so our measures of mobility would rank Γ_2 as less mobile than Γ_1. Figure 3.1 plots the cutoffs for entering each quintile as defined by the distribution of wealth from our experiments. Notice that there is not much change in the cutoffs until ρ exceeds 0.7. Beyond that, as the productivity process becomes more persistent, the distribution spreads out, and the cutoffs become further apart. We will detail how we identify exchange mobility from structural mobility in both the data and the model.

Exchange Mobility: Behavior vs. Luck Once the movement of households through the distribution has been isolated from movements in the distribution itself, exchange mobility can be further separated into changes due to differences in the productivity shock process and changes in household behavior. Consider two households A and B with two Π matrices.

Let $\rho_A = 0$ and $\rho_B = 0.5$ so

$$\Pi_A = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

and

$$\Pi_B = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}.$$

One might initially suppose that household A will have greater mobility than household B. After all, according to any one of the above measures, the earnings mobility of A is considerably greater than that of B. This fact, however, does not necessarily translate to greater wealth mobility. The reason is that randomness in household earnings does not wholly determine a household's wealth. Because household utility is strictly concave, households try to smooth consumption over time. Since shocks for A are less persistent, the optimal response of household A to a switch in productivity is to adjust savings. The more persistent the shocks, the more closely earnings resemble permanent income and the less savings adjusts.

3.3 Wealth Mobility in the Data

3.3.1 Data

There has been relatively little empirical work on the intragenerational evolution of wealth.[9] We study eight waves of wealth supplements from the Panel Study of Income Dynamics (PSID) from 1984-2015 to measure wealth mobility. Following Hurst et al. (1998), we use identifiers from the family and individual files in the PSID to link families in the wealth supplements.[10] Then, using the population weights from the family files, we construct the distribution of wealth in each year and divide each distribution into quintiles. Finally, we measure the fraction of households that transition between quintile i and quintile j for $i, j \in \{1, ..., 5\}$, between the starting and ending years.

We study three time horizons: short, medium, and long. We define the short horizon as 5-6 years, the medium horizon as 9-10 years, and the long horizon as 19-21 years.[11] Table

[9] Several studies on intragenerational wealth mobility have been conducted using a small number of waves from the PSID. See Castaneda et al. (2003), Hurst et al. (1998), and Díaz-Giménez et al. (2011).

[10] We include only families that have the same head at the beginning and end of the sample period. This would exclude cases where the head becomes deceased or institutionalized. In the case of a divorce, our methodology retains the head, but the non-head spouse is discarded. On average, this restriction removes 8-9 percent of any sample.

[11] While it would be ideal to have a fixed length for each horizon, the irregular timing of PSID releases does not permit it.

3.1 reports the short-, medium-, and long-horizon wealth mobility matrices obtained from the PSID data.

Table 3.1: Mobility Matrices

Short Horizon

1984-1989

$$\begin{bmatrix} 0.70 & 0.23 & 0.05 & 0.02 & 0.00 \\ 0.25 & 0.45 & 0.22 & 0.06 & 0.02 \\ 0.06 & 0.24 & 0.44 & 0.19 & 0.06 \\ 0.02 & 0.06 & 0.22 & 0.47 & 0.23 \\ 0.01 & 0.01 & 0.06 & 0.22 & 0.70 \end{bmatrix}$$

1989-1994

$$\begin{bmatrix} 0.66 & 0.24 & 0.07 & 0.02 & 0.01 \\ 0.27 & 0.45 & 0.18 & 0.07 & 0.02 \\ 0.08 & 0.25 & 0.42 & 0.19 & 0.06 \\ 0.03 & 0.06 & 0.27 & 0.42 & 0.21 \\ 0.01 & 0.03 & 0.05 & 0.24 & 0.66 \end{bmatrix}$$

1994-1999

$$\begin{bmatrix} 0.64 & 0.26 & 0.07 & 0.02 & 0.01 \\ 0.25 & 0.47 & 0.21 & 0.05 & 0.02 \\ 0.10 & 0.22 & 0.43 & 0.21 & 0.04 \\ 0.04 & 0.07 & 0.24 & 0.44 & 0.21 \\ 0.01 & 0.03 & 0.06 & 0.20 & 0.70 \end{bmatrix}$$

2001-2007

$$\begin{bmatrix} 0.62 & 0.25 & 0.10 & 0.03 & 0.01 \\ 0.27 & 0.43 & 0.22 & 0.07 & 0.02 \\ 0.09 & 0.28 & 0.37 & 0.21 & 0.05 \\ 0.03 & 0.08 & 0.24 & 0.43 & 0.22 \\ 0.01 & 0.03 & 0.05 & 0.22 & 0.69 \end{bmatrix}$$

2007-2013

$$\begin{bmatrix} 0.60 & 0.30 & 0.08 & 0.02 & 0.00 \\ 0.26 & 0.44 & 0.24 & 0.06 & 0.01 \\ 0.14 & 0.18 & 0.43 & 0.21 & 0.04 \\ 0.07 & 0.08 & 0.20 & 0.47 & 0.19 \\ 0.02 & 0.02 & 0.05 & 0.21 & 0.71 \end{bmatrix}$$

Medium Horizon

1984-1994

$$\begin{bmatrix} 0.63 & 0.24 & 0.09 & 0.03 & 0.02 \\ 0.23 & 0.41 & 0.21 & 0.10 & 0.05 \\ 0.10 & 0.28 & 0.33 & 0.21 & 0.09 \\ 0.05 & 0.08 & 0.26 & 0.37 & 0.23 \\ 0.02 & 0.03 & 0.09 & 0.25 & 0.61 \end{bmatrix}$$

1994-2003

$$\begin{bmatrix} 0.61 & 0.26 & 0.09 & 0.03 & 0.02 \\ 0.24 & 0.44 & 0.23 & 0.06 & 0.03 \\ 0.11 & 0.25 & 0.35 & 0.23 & 0.06 \\ 0.06 & 0.09 & 0.24 & 0.39 & 0.22 \\ 0.03 & 0.04 & 0.08 & 0.21 & 0.65 \end{bmatrix}$$

2003-2013

$$\begin{bmatrix} 0.57 & 0.29 & 0.10 & 0.03 & 0.01 \\ 0.27 & 0.41 & 0.23 & 0.07 & 0.02 \\ 0.14 & 0.21 & 0.39 & 0.20 & 0.05 \\ 0.06 & 0.08 & 0.21 & 0.44 & 0.22 \\ 0.02 & 0.02 & 0.05 & 0.23 & 0.68 \end{bmatrix}$$

Long Horizon

1984-2003

$$\begin{bmatrix} 0.58 & 0.25 & 0.11 & 0.05 & 0.02 \\ 0.26 & 0.35 & 0.22 & 0.12 & 0.05 \\ 0.09 & 0.29 & 0.27 & 0.22 & 0.13 \\ 0.05 & 0.11 & 0.27 & 0.32 & 0.26 \\ 0.03 & 0.06 & 0.11 & 0.26 & 0.55 \end{bmatrix}$$

1989-2009

$$\begin{bmatrix} 0.56 & 0.28 & 0.10 & 0.04 & 0.03 \\ 0.27 & 0.37 & 0.20 & 0.12 & 0.05 \\ 0.12 & 0.25 & 0.29 & 0.22 & 0.12 \\ 0.08 & 0.11 & 0.29 & 0.32 & 0.20 \\ 0.02 & 0.05 & 0.09 & 0.25 & 0.60 \end{bmatrix}$$

1994-2015

$$\begin{bmatrix} 0.58 & 0.24 & 0.11 & 0.04 & 0.03 \\ 0.28 & 0.38 & 0.20 & 0.10 & 0.04 \\ 0.13 & 0.25 & 0.32 & 0.21 & 0.08 \\ 0.07 & 0.11 & 0.24 & 0.34 & 0.25 \\ 0.03 & 0.05 & 0.09 & 0.25 & 0.58 \end{bmatrix}$$

The wealth data, shown in Table 3.1, display quite a bit of mobility. Particularly, we see that although the first and fifth quintiles are the most persistent, families that begin in these quintiles have at least a 30 percent chance of ending elsewhere, at all time horizons. Additionally, families in the middle three quintiles are, in every period and at all horizons, more likely to leave their starting quintile than they are to stay. Finally, a non-trivial fraction of families make large transitions, crossing multiple quintiles over a single period. For example, among families that began in the first quintile in 2001, about 3 percent end in the fourth quintile, and about 1 percent end in the fifth after 10 years. We also see large movements in the opposite direction: over the same period, about 3 percent of families that began in the fifth quintile finished in the second, and about the 1 percent ended in the first quintile.

Figures 3.17 through 3.19 show the evolution of our mobility measures over time for each horizon. There is no apparent trend at the short horizon. However, over the medium and long horizons, our measures show a decline in wealth mobility since 1984.

3.3.2 Confidence Intervals for Mobility

As with any empirical measurement, there is some uncertainty around with these mobility matrices and the measures applied to them. For example, Table 3.1 shows that the proportion of families that begin in the first quintile and end in the fifth fell from 1.2 percent between 1994 and 1999, to about 0.7 percent between 2001 and 2007. Similarly, the Shorrocks measure of mobility rose between these two periods. Can we say that these changes are statistically significant?

To address these questions, we estimate standard errors associated with wealth mobility using a bootstrapping procedure. In each iteration, we draw a new panel sample with replacement, using the same number of observations as our original sample. For each such sample, we re-estimate the wealth distributions in the starting and ending years, and construct a mobility matrix using the same procedure as before. Using this procedure, we construct 95 percent confidence intervals around each entry in the mobility matrix, as well as the same intervals around the mobility exhibited by these matrices, as measured by each of the aforementioned statistics. In Table 3.2 we show an example of the results of this process for the period 1994-1999. The large entries in the matrix represent our point estimates for the transition rates over this period, as reported in Table 3.1. The smaller parenthetical entries below represent a 95 percent confidence interval for each element of the matrix.

Table 3.2: Wealth Transition Matrix with 95% Confidence Intervals

1994-1999

0.64	0.26	0.07	0.02	0.01
(0.61,0.67)	(0.24,0.29)	(0.05,0.08)	(0.01,0.02)	(0.01,0.02)
0.25	0.47	0.21	0.05	0.02
(0.22,0.28)	(0.44,0.50)	(0.19,0.24)	(0.04,0.07)	(0.01,0.03)
0.10	0.22	0.43	0.21	0.04
(0.08,0.12)	(0.19,0.24)	(0.40,0.46)	(0.18,0.23)	(0.03,0.06)
0.04	0.07	0.24	0.44	0.21
(0.02,0.05)	(0.05,0.09)	(0.21,0.27)	(0.41,0.47)	(0.18,0.24)
0.01	0.03	0.06	0.20	0.70
(0.00,0.02)	(0.02,0.05)	(0.04,0.08)	(0.17,0.22)	(0.67,0.73)

The answer to the above questions is "No." We cannot conclude that the proportion of families transitioning from the first to the fifth quintile did not fall significantly between the periods 1994-1999 and 2001-2007. Between 61 and 67 percent of families in the first quintile in 1994 ended in the first quintile in 2003, while between 1 and 2 percent of these families transitioned to the fifth quintile. The interval for the latter figure over the period 2001 to 2007 is about 0.2 percent to 1.2 percent.

Our bootstrapping procedure also gives us the opportunity to reassess time trends in mobility. In each bootstrapping iteration, we apply our four measures of mobility to the resultant matrix. By doing so, we can compute bootstrapped standard errors for measures of wealth mobility for each sample. Figures 3.20 through 3.22 show the measures of short, medium, and long-horizon wealth mobility as reported before, with shading indicating 95 percent confidence intervals. Broadly, we see that the size of these intervals depends on both the time horizon (which influences the number of observations available) and the measure used. Once again, it is difficult to extrapolate a trend in the measures of short-horizon wealth mobility.[12] At the medium horizon, we find that, at the 5 percent level, only one measure rates any given period as having significantly less mobility than the preceding period. However, by three of the four measures used, wealth mobility from 2003-2013 was significantly lower than from 1984-1994. Thus, we can still safely say that medium-horizon wealth mobility has significantly declined over our entire sample period. At the long horizon, we cannot say that mobility has declined significantly since 1984.

3.3.3 Structural and Exchange Mobility in the Data

As was mentioned in Section 3.1, wealth mobility arises from two sources: *structural* mobility and *exchange* mobility. Again, structural mobility refers to mobility that arises from changes in the shape of the wealth distribution, while exchange mobility is mobility arising from households changing their wealth position relative to other households. We aim to decompose mobility in the PSID wealth data into structural and exchange mobility. Using the same samples that we use to measure overall mobility, we estimate exchange mobility by recalculating the mobility matrix for each period, holding fixed the cutoff values for wealth quintiles. That is, for each sample we divide the families into quintiles based on their starting wealth as before, and keep track of the wealth values that demarcate those quintiles. We then take that family's wealth in the ending year and record the quintile in which that wealth value would have fallen, using the quintile cutoffs from the starting year. In this way, we hold the distribution fixed, and any mobility is purely the result of households changing their

[12] Although, to answer a question posed earlier, the increase in mobility from 1994-1999 to 2001-2007 was in fact significant at the 5 percent level.

relative wealth. In order to estimate structural mobility, we then subtract exchange mobility from total mobility over the sample period, as calculated by the procedure outlined above.

Figures 3.23 through 3.25 show the time patterns in structural, exchange, and total mobility over our sample period. In most cases, total mobility lies between exchange and structural mobility, and the contribution of structural mobility to total mobility is negative. This observation is consistent with the well-documented fact that wealth inequality in the US has increased over the past 30 years. That being said, in most cases, the mobility lost to structural changes in the wealth distribution is minimal. This decomposition shows that the large majority of empirical wealth mobility is *exchange* mobility, resulting from differences among household income and savings rather than changes in the wealth distribution.

3.4 Model

As a starting point, we study the long run properties of Aiyagari (1994a) with no borrowing.[13] There is a unit measure of *ex ante* identical households. Every period, each household receives an idiosyncratic labor productivity shock, ε, from a finite set $\mathcal{E} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_J]$ with $\varepsilon_1 < \varepsilon_2 < ... < \varepsilon_J$. The process for productivity shocks be Markov with stochastic transition matrix $\Pi = \Pr(\varepsilon_j|\varepsilon_i)$ for $j, i \in 1, .., J$. Every household supplies the same fixed number of hours, $\overline{h}$, and earns total labor income equal to $\omega\overline{h}\varepsilon$, where ω is a market-wide wage. Because the wage and hours supplied do not change across periods, labor productivity shocks are equivalent to random labor income endowments. As in the standard incomplete-markets model, there is only one asset, a, which is a claim to the capital stock K. Because no state contingent claims exist, households have a motive to self-insure through precautionary savings.

A stand-in firm combines capital and effective labor through a constant-returns-to-scale production technology $F : \Re^+ \times \Re^+ \to \Re^+$ to produce a final good which may be consumed or invested in capital for next period. The firm manages the capital stock from household's saving, pays an interest rate r on assets, hires labor, and invests in new capital. Capital depreciates at a constant rate δ each period. We assume that the firm behaves competitively. Letting F be Cobb-Douglas, the optimal choice of the firm implies that each factor is paid its marginal product:

$$\omega = (1 - \alpha)\left(\frac{K}{N}\right)^{\alpha}$$

[13] Because we are concerned with mobility in the stochastic steady state, we omit time subscripts.

and

$$r = \alpha \left(\frac{K}{N} \right)^{\alpha - 1} - \delta.$$

The state vector of the household has two elements: current wealth, a, and current labor productivity, ε. Let period utility be represented by a continuous, strictly concave function $u : \Re^{+} \to \Re$, and assume that u is continuously differentiable as many times as necessary. The household problem in recursive form is

$$V(a, \varepsilon) = \max_{c, a'} \left\{ u(c) + \beta E_{\varepsilon' | \varepsilon} \left[V(a', \varepsilon') \right] \right\}$$

subject to the budget constraint

$$c + a' \leq w\varepsilon + (1 + r) a$$

and lower bound constraints

$$c > 0; a' \geq \underline{a}.$$

Denote by $\Gamma(a, \varepsilon)$ the distribution of households over $\mathcal{A} \times \mathcal{E}$.

Definition 3. *A **steady-state recursive competitive equilibrium** is a set of value functions $V(a, \varepsilon)$, policy functions $g_a(a, \varepsilon)$, $g_c(a, \varepsilon)$), pricing functions, r and w, and a distribution $\Gamma(a, \varepsilon)$ such that*

1. Given prices, V, g_a and g_c solve the household's problem.
2. Firms maximize profits

   $$\omega = (1 - \alpha) \left(\frac{K}{N} \right)^{\alpha}$$

   and

   $$r = \alpha \left(\frac{K}{N} \right)^{\alpha - 1} - \delta.$$

3. Markets clear:

   $$K = \sum_{j=1}^{J} \int a d\Gamma(a, \varepsilon_j)$$

   $$N = \sum_{j=1}^{J} \int \overline{h} \varepsilon_j d\Gamma(a, \varepsilon_j).$$

4. Γ is consistent with the saving decisions of households and the process for ε.
5. The joint distribution of wealth and productivity $\Gamma(a, \varepsilon)$ is stationary.

3.5 Numerical Experiments

3.5.1 Baseline

We choose fairly standard values for our structural parameters: we let utility be logarithmic, we choose $\beta = 0.99$ and $\delta = 0.025$ as roughly consistent with quarterly aggregates for the capital/output and investment/output ratios, and we set $\alpha = 0.36$ to match capital's share of income. We also choose a zero borrowing limit.

We follow Floden and Lindé (2001) who estimate an earnings process of $\rho = 0.92$ and $\sigma_\varepsilon = 0.21$ (annual) from the PSID. The resulting 5-year wealth transition matrix is

$$\begin{bmatrix} 0.87 & 0.14 & 0.00 & 0.00 & 0.00 \\ 0.13 & 0.73 & 0.14 & 0.0 & 0.00 \\ 0.00 & 0.14 & 0.74 & 0.12 & 0.00 \\ 0.00 & 0.00 & 0.12 & 0.81 & 0.07 \\ 0.00 & 0.00 & 0.00 & 0.08 & 0.92 \end{bmatrix}$$

which features far less wealth mobility than any of the transition matrices above. Because the underlying source of both inequality and mobility in this model is the stochastic earnings process, we examine how the transition matrix above responds to different assumptions about the Markov process.

Earnings Process

The fundamental force driving the distribution of wealth in the economy is the labor productivity process. We assume the Markov process above approximates

$$\log(\varepsilon') = \rho \log(\varepsilon) + \nu', \quad \nu' \sim N\left(0, \sigma^2\right).$$

We set J, the number of individual productivity states, to 2.[14] Given this and the parameters ρ and σ, we use the Rouwenhorst method to construct the Markov chain process. Under the Rouwenhorst method, the Markov chain depends upon ρ and σ. The states are equally-space

[14] We have run our experiments with 7 productivity states as well. In general, the qualitative results do not change significantly. One issue that arises when there are more than 2 values for productivity is for very low values of ρ the transition matrix is no longer monotone (i.e., the conditional probability of moving from $\varepsilon = \varepsilon_i$ to $\varepsilon' = \varepsilon_j$, $j \neq i$, does not monotonically decrease as the distance between j and i increases). Since monotonicity of the transition matrix is important for understanding the mobility measures and this failure is simply an approximation error, we concentrate on the two-state case.

over the interval $[-\psi, \psi]$, where

$$\psi = \frac{\sqrt{(J-1)}}{\sqrt{(1-\rho^2)}}\sigma.$$

The transition matrix, Π, depends on two parameters, p and q. Following Kopecky and Suen (2010), we set

$$p = q = \frac{1+\rho}{2};$$

note that Π only depends upon the persistence parameter ρ.

A consequence of generating a Markov chain in this manner is that if one only varies ρ and keeps σ fixed, the vector of states will be different for each value of ρ. This dependence will cause the marginal distribution of effective labor to vary across experiments due solely to the approximation procedure, which could mess up our comparisons. To prevent this contamination, we make σ a function ρ. Given a baseline ρ_0 and σ_0, we define

$$\sigma(\rho) = \sigma_0\sqrt{\frac{1-\rho^2}{1-\rho_0^2}}.$$

This procedure guarantees that the ε state vector of productivity remains the same across ρ experiments and, because labor is supplied inelastically, so does N. Moreover, because Π depends solely on ρ, we can isolate changes to the transition probabilities without altering the states. In this way, ρ will increase the probability of earning the same (by construction) current labor income in the next period (it increases the weight along the diagonal of the transition matrix).

3.5.2 Understanding Mobility in the Baseline Model

We conduct a series of computational experiments to identify the fundamental ingredients governing individual wealth mobility within the model. Specifically, we vary ρ, compute the stochastic steady state, approximate the quintile wealth transition matrix via simulation, and calculate mobility. Figure 3.2 plots the relationship between ρ and several measures of mobility. Mobility is hump-shaped across persistence with mobility being low when ρ is near 0 and when ρ is near 0.9, and reaches its peak for $\rho \in (0.75, 0.80)$.

For each value of ρ the model is solved in general equilibrium, so the market clearing interest rate and the wealth distribution itself will differ in each case. Thus, our results are the combination of changes in structure, behavior, and luck. We now introduce a methodology to decompose changes in mobility as ρ changes into these three components. We discuss how each component varies with this persistence, and how they result in countervailing forces

which prevent the standard model from replicating the wealth mobility that we see in the data.

Ghost households

In order to isolate the effects of structure, behavior and luck to mobility, we introduce "ghost" households into the computed steady state wealth distributions. A ghost is single, zero-measure agent that differs from the other households in the economy in some way. Because a ghost is atomistic, its presence does not alter either equilibrium prices or the quintile boundaries of the wealth distribution. By changing the ghost's environment, policy rules, or labor productivity we can control for each of the other factors. In the first step toward constructing our decomposition, we introduce ghosts with different labor income processes into each of the steady wealth distributions found in the baseline. For exposition, we will draw a distinction between the ρ value of the process faced by normal households (that is, the value which gave rise that particular wealth distribution) and the ρ value of the ghost. Denote the first, ρ_{GE}, and the second, ρ_G. We then simulate and construct a 5×5 mobility matrix for each ghost. We will perform this exercise for two types of ghost households. The first ghosts, *informed* ghosts, understand that their process has a different autocorrelation than that of the other households around them. As a result, their saving decision rules will differ from those of the standard households in the economy, as will the realization of their productivity shocks. The second type of ghosts, *uninformed* ghosts, believe that they have the same process as the standard households but experience the productivity sequence of a household of with a different persistence ρ. Believing that their income follows the same process as their peers, these uninformed ghosts will follow the same decision rules as the remaining households in the economy, and thus they will differ only in the realization of their shocks. For a more detailed exposition on the ghost households, see Appendix 3.10.2.

Decomposing changes in mobility

We have identified three sources for the differences in mobility as the labor income process becomes more persistent. In order to disentangle the contributions of each source to the total change in steady state mobility, we will run several counterfactual experiments. Consider the steady states of two economies, one with $\rho = \rho_x$ and one with $\rho = \rho_y$; and without loss of generality, let $\rho_y > \rho_x$. Denote by $\mu_{[j,j,j]}$, the measured mobility induced by an agent acting in a distribution produced by agents with $\rho = \rho_j$, having optimal policy rules consistent with $\rho = \rho_j$, and experiencing a realized sequence of labor productivity shocks generated according to $\rho = \rho_j$. For ease of exposition, let $\mu_{[J,J,J]} = \mu_J$. Finally, let $\Delta\mu_{xy} = \mu_y - \mu_x$.

$\Delta\mu_{xy}$ is the total change in mobility between the economy with a labor income persistence of ρ_x and ρ_y.

We decompose $\Delta\mu_{xy}$ in the following manner:

$$\begin{aligned}\Delta\mu_{xy} &= \Delta structure + \Delta behavior + \Delta luck \\ &= \left(\mu_y - \mu_{[x,y,y]}\right) + \left(\mu_{[x,y,y]} - \mu_{[x,x,y]}\right) + \left(\mu_{[x,x,y]} - \mu_x\right).\end{aligned}$$

Each component removes one conflating factor in the relative mobility difference, starting with the structures of the ρ_x and ρ_y distributions, moving to differences in the decision rules (behavior) of the agents, and finishing with differences in the realized sequence of productivity shocks.

Figure 3.13 decomposes the total change in mobility as ρ rises into these three components. Across all four measures, the decomposition is qualitatively the same. Structure has a small negative effect on mobility, while behavior and luck make larger contributions, negative and positive, respectively. At low levels of ρ, mobility rises in the shock persistence because luck offsets behavior. Past a certain point, however, behavior becomes more powerful and pulls total mobility down.

Structure Figure 3.3 plots the steady state wealth distribution under different values of the persistence of the productivity process. There are two things to note about the distribution as ρ increases. First, the wealth becomes more unequally distributed as the right tail stretches out. Because there are only two productivity states, in equilibrium households with the high (low) productivity are savers (dissavers). The closer ρ is to 1, the more likely households with high ε are to draw high ε'. As a consequence, some households will receive a very long string of good productivity shocks, allowing them to amass a considerable amount of wealth. In the same way, households that draw a low productivity will be more likely to draw low productivity in the future, leading to the second feature of a larger ρ: more households are borrowing constrained. These changes in the structure of the wealth distribution affect the boundaries between quintiles. Figure 3.1 plots these boundaries for different values of ρ. The cutoffs move apart gradually as ρ approaches 0.7. As the productivity process becomes more even persistent, however, the distribution spreads out rapidly, and the boundaries become further apart. When $\rho = 0.99$, the entire first quintile is at the borrowing limit.[15]

[15]Under some measures, the narrowness of the first quintile can lead to 'spurious' mobility because households will very frequently transition between the first and second quintiles despite almost no change in wealth.

Behavior Optimal household behavior changes responds to the persistence of the shocks as well. The more sensitive is the saving policy to ε, the larger the wealth movements will be across periods, which in turn implies more rapid resorting. Here we state a proposition about the relationship between ρ and the saving policy function $g_a(a, \varepsilon)$ when the wealth distribution is fixed.

Proposition 6. *Consider two households, A and B, from the same steady state wealth distribution, and without loss of generality, let* $\rho_A > \rho_B$. *For* $a > \underline{a}$, *the distance between saving functions across productivity draws is larger for the household with a higher probability of switching productivity states, i.e.*

$$\left|g_a^B(a, \varepsilon_2) - g_a^B(a, \varepsilon_1)\right| > \left|g_a^A(a, \varepsilon_2) - g_a^A(a, \varepsilon_1)\right|$$

The proof of Proposition 6 can be found in Appendix 3.10.4. Intuitively, Proposition 6 is the permanent income hypothesis. If household A and household B have the same assets today and each draws the good shock, but A believes that its shock comes from a more persistent process than B does, then A's consumption will be more responsive and so A's saving will move less than B's will. The consequence is that, all else equal, mobility due to behavior should decrease as ρ increases.

Luck Finally, a household's mobility will be affected by the particular sequence of productivity draws. Within a given measurement window, if a household, beginning from a low wealth level, happens by chance to get higher productivity than would be expected, then that household will have high wealth mobility. The effect on mobility of more persistence in good and bad luck is not monotone. Generally, mobility will be low when persistence is either very low or very high. At very high ρ, households that start with good fortune will tend to continue having good productivity, increasing their saving and moving further away from other less fortunate households. At very low ρ, mobility is low because households switch too frequently. If the household starts in a low quintile and receives a good shock, it saves and moves up a bit in the wealth ordering, but in order to move even further up and transition through multiple quintiles over time, the household needs to get *a string* of positive shocks that is well above average. The probability of getting such a string however increases in ρ. The result is that for low ρ households tend to move around only a small region of their initial wealth position. Luck will tend to push up mobility if ρ lies in some intermediate range. In that region, households will tend to get sufficiently long strings of positive shocks to transition across quintiles, but switch between states frequently enough to support mixing.

Total mobility With these three factors in mind, the inverted U-shape of mobility over ρ can now be understood more easily. As ρ increases, agents experience longer sequences of above (below) average productivity, leading to longer strings of saving (dissaving) and a wider distribution of wealth. The expansion of the distribution should reduce mobility since it increases the distance between quintile boundaries (with the possible exception of the one between the first and second quintiles). More autocorrelated shocks should increase mobility since it allows households to experience longer strings of movement in the same direction, whether up or down; however, this effect is somewhat offset by the reduction in the sensitivity of savings to the shocks. While at higher ρ, households move in the same direction longer, they in smaller steps.

The above proposition explains the hump-shape in mobility. At low ρ, a move from state (k, ε_1) to (k, ε_2) induces a large change in k'. In itself, this would increase mobility, but because ρ is low, the probability of returning to the lower $g_k(k, \varepsilon_1)$ rule is high. Thus, it is likely that such a household will not experience a long enough string of high productivities to accumulate a lot more wealth and move up into other quintiles. By a similar logic, a household that just drew ε_1 after having been ε_2 is unlikely to move down quintiles. On average households in a low-ρ environment, are very unlikely to move far away from their initial wealth level, k, though they will move very frequently within a small neighborhood of k.

As ρ increases, the distance of between savings functions does not fall much but the likelihood of experiencing a long string of consecutive ε_2 productivities rises. This allows households to move greater distances within the wealth distribution over a fixed amount of time. At some point however, ρ becomes so large that households switch productivities very infrequently, and the distance between savings rules gets very small. A household that starts on the savings path implied by $g(k, \varepsilon_2)$ is likelihood to continue building up wealth for a long time but very slowly so that it takes many periods to transition between quintiles. In our numerical experiments, we find a ρ near 0.7 returns the highest measure of mobility over quintiles.

Figure 3.2 plots these mobility measures as functions of ρ (again where σ is normalized). While the levels of the mobility measures differ, the orderings are very similar. For instance, the correlations are nearly 1 as shown in Table 3.3.

We find an analogy to driving helpful for explaining how mobility works in this model. Think of the support of wealth as a highway that runs east and west. Take any location on that highway and call all locations to the west of it 'poorer' and all locations to the east 'richer'. 'Checkpoints' along the highway correspond to quintiles of wealth (also called 'class boundaries'). Household decision rules are lanes on a highway. Some lanes move east

Table 3.3: Correlation between mobility measures

Correlation Coefficients			
	μ_B	μ_{2E}	μ_S
μ_{MFP}	0.9991	0.9997	0.9991
μ_S	1.0000	0.9991	—
μ_{2E}	0.9991	—	—

(toward higher wealth) and others move west (toward lower wealth); and the fastest lanes are one the outside of the highway. The fastest westbound lane corresponds to the lowest labor income value, and the fastest eastbound lane to the highest value. The further the saving decision is from the 45 degree line, the faster it moves. Changing ρ alters how likely one is to switch out of their current lane and into another one. In the two ε case, there is only one westbound and one eastbound lane. If ρ is high, than a household will likely stay in its lane continuing to move up or down in the wealth ordering. As Figure 3.8 shows however, the more persistent the Markov process the closer the decision rules are to the 45 degree line and so the more slowly will be the pace of the lane in our analogy. If ρ is low the lane speeds will be faster, but the households will switch directions frequently, moving up and the moving down the ordering. Maximum wealth mobility is achieved where lanes move quickly enough to allow for distant movement, but also where they are likely not to switch too often, allowing for a sufficiently long chain of movements in the same direction.

Borrowing Limits

So far we have imposed a strict borrowing limit of zero. A large fraction of households can find this constraint binding, particularly when the labor income shocks are very persistent. As a result, the steady state wealth level separating the first and second quintiles can be very close to 0 so that even a small movement away from the borrowing limit can move a household into the second quintile. In this case, households in the first (second) quintile would appear to be very upwardly (downwardly) mobile. We have run cases with high persistence and exogenous borrowing limits near the natural borrowing limit and found that while it has little effect on our mobility measures. Therefore, we do not think that our assumption of no borrowing is restricting our findings.

3.6 Mobility and Inequality

It is well-known that a Bewley model with idiosyncratic labor income risk alone does a poor job matching the high concentration of wealth in the right tail.[16] The fundamental issue is that the sufficient amount of wealth to self-insure is low when agents are very patient and shocks are relatively small. Once a household can adequately smooth its consumption, it has no other incentive to continue saving, since interest rates are necessarily lower than time rates of preference. Several approaches have been used to generate longer right tails in the wealth distribution. We now consider two of these extensions, and study whether they improve the model's fit to mobility.

Krusell and Smith (1998) replace the scalar household discount factor with a 3-state, highly persistent Markov chain. The three values are $[0.9763, 0.9812, 0.9861]$, and the transition matrix is

$$\begin{bmatrix} 0.99654 & 0.00346 & 0 \\ 0.00043 & 0.999135 & 0.00043 \\ 0 & 0.00346 & 0.99654 \end{bmatrix};$$

these choices deliver a Gini coefficient of wealth equal to 0.78. The invariant distribution of β is $[0.1, 0.8, 0.1]$ and the average duration in either extreme-β state is 200 quarters. The 5-year wealth mobility matrix for the stochastic-β environment is

$$\begin{bmatrix} 0.84 & 0.16 & 0.00 & 0.00 & 0.00 \\ 0.16 & 0.70 & 0.15 & 0.0 & 0.00 \\ 0.00 & 0.15 & 0.73 & 0.11 & 0.00 \\ 0.00 & 0.00 & 0.11 & 0.84 & 0.05 \\ 0.00 & 0.00 & 0.00 & 0.05 & 0.95 \end{bmatrix}$$

The stochastic-β model makes the mobility match worse – the top quintiles get even more persistent, since drawing a high discount factor leads even agents with temporarily low income to save, and discount factor shocks are very persistent. It is this immobility that delivers the high wealth concentration that was the goal of Krusell and Smith (1998), but it does not come for free.[17]

In the terminology of Section 3.2.1, the economy in this model trades off increases in behavioral mobility, for decreases in structural mobility. The existence of a "saver" class with high patience implies that some agents will be upwardly mobile: their savings rate will

[16] See Quadrini and Ríos-Rull (1997) and Carroll (1998) for discussion.

[17] Carroll (2001) shows that a permanent 'two-β' model looks very much like the stochastic-β model, so the fact that the discount factors mean-revert does not seem important provided they do so slowly.

be high regardless of their income shock, increasing their behavioral mobility, and so they will tend to move upward. However, over time these agents accumulate more and more wealth, stretching out the distribution and making it harder to transition across quintiles. Our numerical experiments indicate that this structural channel dominates in the heterogeneous-β model.

Castaneda et al. (2003) add a very high productivity state with relatively low persistence and a high probability of transitioning immediately to the lowest productivity [18]. The transitory nature of this 'rockstar' state combined with the increased risk motivates households in this state to build up a substantial amount of precautionary savings. When a household draws the rockstar state, it takes advantage of its temporary good fortune by saving rapidly. This 'burst of saving' produces the matrix below which has considerably more upward mobility than the benchmark:

$$\begin{bmatrix} 0.73 & 0.17 & 0.05 & 0.04 & 0.00 \\ 0.24 & 0.52 & 0.19 & 0.05 & 0.00 \\ 0.00 & 0.34 & 0.52 & 0.14 & 0.00 \\ 0.00 & 0.00 & 0.24 & 0.59 & 0.17 \\ 0.00 & 0.00 & 0.00 & 0.17 & 0.83 \end{bmatrix}.$$

Nevertheless, the rockstar model still has too little downward mobility. The consumption-smoothing motive implies that while households save rapidly, they dissave slowly – staying away from the borrowing constraint is the reason they save, after all. And furthermore the resulting labor earnings process looks nothing like we find in the data (see Guvenen et al., 2015b).

Here we see again the tradeoff between the structural, behavioral, luck channels of mobility. Drawing the "rockstar" earnings state allows agents to move up quickly as they far out-earn their peers—this is the "luck" component of mobility. However, this is offset by the behavioral and structural components: these agents come to dominate the uppermost quintile, stretching out its bounds, and they dissave slowly once the shock has passed, rendering the highest rung of the distribution too immobile.

[18] Specifically, the state vector for labor productivity is $\mathcal{E} = [1.0, 3.15, 9.78, 250]$. The Markov process for labor earnings in Castaneda et al. (2003) has a stochastic aging component. Here, we abstract from this by isolating the submatrix associated with the worker-to-worker transition and then renormalizing the rows so that Π is stochastic. The resulting transition matrix is $\Pi = \begin{bmatrix} 0.984 & 0.012 & 0.004 & 0.000 \\ 0.031 & 0.965 & 0.004 & 0.000 \\ 0.015 & 0.004 & 0.980 & 0.000 \\ 0.109 & 0.005 & 0.062 & 0.823 \end{bmatrix}$

3.7 Factors Influencing Mobility in the Data

Comparing matrices generated by our model to those generated by the PSID wealth data shows that our models fall short of approximating the level of wealth mobility seen in the data. In this section, we use regression analysis to find suggestive evidence of the type of shocks that, if added to our model, would help us better approximate real-world mobility. In particular, we will focus on factors that affect wealth directly, such as a payoff from a risky asset, a divorce, or the receipt of a large inheritance.

We address this question in two ways. Following Jianakoplos and Menchik (1997), we estimate regressions using data from our samples to determine the effect of the factors mentioned above on wealth movements. First, we regress a family's change in percentile ranking on a vector X of factors that may influence changes in a family's relative ranking in the wealth distribution:

$$\Delta p_{i,t} = X_{i,t}^{\top}\beta + \alpha_t + \varepsilon_{i,t}$$

Here, $\Delta p_{i,t} = p_{i,t} - p_{i,t-1}$, that is, the change in a family's percentile ranking in the wealth distribution over a given period. The vector $X_{i,t}$ includes control variables such as a family's income and the head's age at the beginning of period $[t-1, t]$, as well as indicators for the holding of certain assets, or the head's possession of a college degree at the start of period $[t-1, t]$. We also include period fixed effects, α_t, in order to control for unobserved factors that may influence wealth mobility.[19]

Then, using the same vector of explanatory variables X_{it}, we estimate four Probit regressions. First, we define $\Delta q_{i,t} = q_{i,t} - q_{i,t-1}$, where $q_{i,t}$ represents the family's quintile in the wealth distribution in year t. We estimate two Probit models using this outcome:

$$\Pr(\Delta q_{i,t} < 0 | X) = \Phi(X^{\top}\beta) \tag{3.1}$$

$$\Pr(\Delta q_{i} > 0 | X) = \Phi(X^{\top}\beta) \tag{3.2}$$

Here, model (3.1) measures the probability that a family *fell* one or more quintiles over the period $[t-1, t]$, model (3.2) measures the probability that a family *rose* one or more quintiles, and Φ is the cumulative distribution function of a standard normal distribution.

Our final two models aim to measure factors that may influence the likelihood that a family makes a large movement through the distribution. We update models (3.1) and (3.2) to focus on families that rise or fall *two* or more quintiles over a given sample period,

[19] Such as a recession or a change in tax policy.

movements that we refer to as "jumps":

$$\Pr(\Delta q_{i,t} \leq -2|X) = \Phi(X^\top \beta) \tag{3.3}$$

$$\Pr(\Delta q_i \geq 2|X) = \Phi(X^\top \beta) \tag{3.4}$$

All of the aforementioned regressions are estimated using data on wealth mobility over the short, medium, and long time horizons.

One note should be made on these large movements. Due to the high reinterview rates in the PSID, families often appear in our samples for multiple time periods. Thus, we are able to follow some families through most or all of our 30-year sample horizon. Doing so suggests that some families have a higher tendency than others to make "jumps," movements of two or more quintiles in the wealth distribution. At any given short-horizon period in our sample horizon, the probability that a family moves two or more quintiles is between 9 and 12 percent. However, the probability that a family makes such a "jump" conditional on that family having made a "jump" in the preceding period is substantially higher: between 20 and 30 percent, depending on the period in question.

Additionally, using our panel samples taken from the PSID data, we calculate mobility matrices for subsets of our samples. Calculating these matrices for different time horizons gives us a better sense of the ways in which some of the factors in our analysis contribute to overall mobility.

3.7.1 Risk, Return, and Entrepreneurship

Evidence from the PSID suggests that a contributing factor in wealth mobility may be heterogeneity in risk preferences and returns among families. Using questions from the wealth survey, we can study the movements of families who hold at least some portion of their net worth in assets with large variance in returns, such as stocks, real estate outside of their main residence, and entrepreneurial ventures such as farms and self-owned businesses. Broadly, our Probit models do suggest that holding such assets does make a family more likely to move throughout the distribution. For example, at each time horizon, we find that ownership of stocks makes a family more likely to move up one or more quintile, and less likely to fall.

Perhaps most notably, we find that ownership of a farm or business increases both the likelihood that a family will *fall* in the distribution, and the likelihood that a family will *rise* in the distribution. The symmetric effect of these entrepreneurial activities also holds when we look at the likelihood that a family moves two or more quintiles in a given period; a

family's holding assets in this category increased the likelihood of "jumps" in both directions. This dichotomy can also be seen in Table 3.4, which shows the mobility matrices for families who respectively did (Π_B) and did not own a business (Π_{NB}) between 2003 and 2013:

Table 3.4: Mobility With and Without Business Ownership, 2003-2013

$$\Pi_B = \begin{bmatrix} 0.45 & 0.27 & 0.14 & 0.10 & 0.04 \\ 0.20 & 0.31 & 0.28 & 0.16 & 0.04 \\ 0.10 & 0.18 & 0.35 & 0.24 & 0.14 \\ 0.09 & 0.04 & 0.14 & 0.44 & 0.29 \\ 0.03 & 0.02 & 0.04 & 0.18 & 0.73 \end{bmatrix}, \Pi_{NB} = \begin{bmatrix} 0.58 & 0.29 & 0.10 & 0.02 & 0.00 \\ 0.28 & 0.43 & 0.22 & 0.05 & 0.01 \\ 0.16 & 0.22 & 0.40 & 0.19 & 0.03 \\ 0.04 & 0.09 & 0.24 & 0.45 & 0.18 \\ 0.02 & 0.02 & 0.05 & 0.27 & 0.64 \end{bmatrix}$$

Clearly, households who owned a farm or business were more likely to leave their starting quintiles, as well as more likely to make large movements in the distribution. Notice, for example, the different patterns in movements made by families who began in the first quintile: those who owned a business were about as likely as those who did not to move to the second quintile, but were far more likely to move to the third, fourth, and fifth quintiles. Similarly, families that owned a business were about twice as likely as those who did not to fall from the fourth quintile to the first.

These results suggest that a key influence in wealth movements is the opportunity for a person to invest time and resources into a project that is at least partially self-funded, and face the potential for both large gains and large losses from this project.

3.7.2 Other Shocks to Wealth

We document evidence that wealth mobility in data may be driven by other shocks directly to wealth, outside of those resulting from the realization of a return on an asset. We study two such shocks: marriages/divorce (wherein assets are combined and divided, respectively) and the receipt of inheritances. In order to capture the effect of these shocks, we include in our Probit specifications binary variables indicating whether the head went through a marriage or divorce or received an inheritance, in any of the intervening years between the start and end of the time given time horizon.

Not surprisingly, we find that the occurrence of a marriage and a divorce have symmetric effects: marriages increase the likelihood of a family rising in the distribution, while divorces make it more likely that a family will slip at least one quintile. Additionally, the occurrence of a marriage is a strong predictor of a family making an upward "jump" of two or more quintiles, and a divorce is a strong predictor of a family making a downward "jump." Importantly, the explanatory power of these events holds at the short time horizon.

We also find evidence that inheritances are strong predictors of large upward movements, particularly over short time horizons. Although this is hardly surprising, it does give us further insight into the type of features that could augment the model in order to better match wealth mobility in the data. The PSID provides us with evidence that incorporating shocks that affect wealth directly–rather than indirectly, through the labor income or savings process–may be a key component in producing realistic mobility.

In summary, the data support the conclusion that wealth mobility is driven in large part by shocks to wealth direcly, and by the holding of risky assets. In the following section, we add heterogeneity in agents' rates of return to our baseline model. The model in Section 3.8 is able to capture the degrees of both inequality and mobility in the data. In Appendix 3.10.3, we also consider a variant of our model in which we increase the space of assets, allowing for agents to purchase insurance against changes in income. While additional assets can increase mobility relative to the baseline, here again we encounter countervailing forces which prevent the model from generating sufficient wealth mobility. In brief, with persistent shocks agents are able to purchase large amounts of cheap insurance against a change in income, which results in a large swing in their wealth should this change occur. By definition, however, a high persistence imply that these changes in income happen only infrequently. As a result, the agents experiencing large swings in wealth are not numerous enough to meaningfully impact mobility.

3.8 Mobility and Returns

Benhabib et al. (2015a) use in a partial equilibrium OLG model with deterministic, heterogeneous earnings profiles and rates of return on saving to match aspects of inequality and intergenerational mobility in the US wealth distribution (see also Hubmer et al. (2017)). They argue that three factors are critical for modeling wealth inequality and wealth mobility: stochastic earnings, capital income risk, and differential saving and bequest motives. While we focus on wealth mobility over shorter time horizons than a generation, our probit estimates suggest that heterogeneous capital income risk may still be an important contributing factor.

3.8.1 Heterogeneous Return Risk

We adapt the baseline model of section 3.4 to incorporate heterogeneous return shocks: if r_t is the equilibrium interest rate in the capital market, a household with shock z_t earns return $(1 + r_t)z_t$ on their prior savings a_t. The budget constraint for an agent with state

$(a_t, \varepsilon_t, z_t)$ is given by[20]

$$c_t + a_{t+1} \leq (1 + r_t)\, z_t a_t + \omega_t \overline{h} \varepsilon_t$$

As with labor income shocks (ε_t), we assume that return shocks z_t follow an autoregressive process:

$$\begin{aligned} z_{t+1} &= \rho_z z_t + \nu_t \\ \nu_t &\sim \mathcal{N}\left(0, \sigma_z^2\right) \end{aligned}$$

Our assumption on return shocks is parsimonious, and yet has several features that aid in reconciling wealth mobility in the model with that in the data. First, note that this return risk is *uninsurable*: agents have no alternative asset, and thus all consumption smoothing is done via saving in the now-risky asset. This assumption has empirical support: data on household holdings suggest that households who hold risky assets are under-diversified, and highly exposed to these risky assets (see, for instance, Moskowitz and Vissing-Jørgensen (2002)). Additionally, these shocks affect wealth, and thus mobility, in both levels and at the margin of intertemporal substitution. The process for z_t allows for households to experience large capital income shocks, which scale with their wealth. These households will also alter their savings behavior: because returns are persistent, households with high z shocks will increase their savings rate in order to capture higher returns, and households with poor z shocks will dissave more rapidly. The combination of behavioral and mechanical changes to wealth enables our model to match the frequency of large transitions in the wealth distribution over short horizons that we observe in the data.

We are, of course, not the first to employ heterogeneous return risk in explaining trends in inequality and social mobility. Benhabib et al. (2019) use capital income risk in an estimated model to target *intergenerational* wealth mobility. By contrast, we demonstrate that return shocks are also important in matching short-run wealth mobility. Pugh (2018) argues in a calibrated model that return shocks are key drivers of transitions into and out of the top 1% of wealthiest households. We show that such shocks are critical to capturing mobility throughout the entirety of the distribution, including into and out of the bottom quintiles. As these bins are comprised of households with low consumption (and thus high marginal utilities), it is important to understand how wealth dynamics play out in these ranges of the distribution, rather than just at the top.

Having introduced return risk, we must now take a stand on when exactly in the period

[20] With heterogeneity in returns, individual wealth will evolve according to a random growth process, and as such the wealth distribution will have a Pareto tail; see e.g. Gabaix (2009). To avoid underestimating the capital stock, we employ the "Pareto extrapolation" technique of Gouin-Bonenfant and Toda (2019).

wealth is "recorded" for the purposes of measuring mobility. In the data, when households report their wealth, we assume that they include all returns earned on each of their assets. For instance, when reporting the value of retirement accounts, household include not just their contribution, but the full appreciated value of these accounts. When measuring mobility in our model, we record agents assets after the z_t return shock has been realized. The state vector for an agent is now $(w_t, \varepsilon_t, z_t)$, where $w_t = z_t a_t$. Agents now solve

$$\max_{\{c_t, w_{t+1}\}_{t=0}^{\infty}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t)$$

subject to

$$w_{t+1} = (1 + r_{t+1}) z_{t+1} (w_t - c_t)$$

The rest of the model remains the same as in Section 3.4. In order to pin down the process for z_t, we add an additional calibration target: we require that the Gini coefficient of wealth in the steady state is $G = 0.85$, as in the US data. Note that this leaves us with a degree of freedom: we have added one additional calibration target, but two extra parameters: ρ_z and σ_z. As data on household-level returns for the US is sparse, our strategy is to consider a range of calibrations. For each value of ρ_z, we calibrate σ_z (along with β and aggregate TFP Z) in order to match wealth inequality. We then measure mobility across each of the alternative parameterizations.

Figure 3.26 shows how ρ_z and σ_z relate to one another across the alternate calibrations that we consider. A clear pattern emerges here: the more persistent agents' returns, the smaller the shocks need to be in order to generate the wealth Gini in the data. This relationship is intuitive: if return shocks are persistent, then agents who enjoy good shocks will save out of them, accumulating wealth over time. Similarly, agents locked into a long stream of persistently low returns will run down their wealth. These forces stretch out the wealth distribution, allowing for a good fit to inequality with smaller shocks than are needed with less persistent returns.

3.8.2 Mobility with Uninsured Return Risk

Broadly speaking, we find that the inclusion of capital income risk–both in agents' problems, and in our measurement of wealth–improves the fit of the model to empirical mobility substantially. Figure 3.27 shows wealth mobility according to our four measures, normalized to the average of their values in the data. In contrast to the baseline model of Section 3.4, here we *overshoot* wealth mobility relative to the data. Notably, the overshoot is less pronounced at higher values of persistence (ρ_z). A sample transition matrix from this exercise

with $\rho_z = 0.9$ and $\sigma_z = 0.064$ is:

$$\begin{bmatrix} 0.32 & 0.3 & 0.25 & 0.12 & 0.01 \\ 0.25 & 0.34 & 0.27 & 0.13 & 0.01 \\ 0.04 & 0.33 & 0.4 & 0.21 & 0.02 \\ 0.0 & 0.04 & 0.35 & 0.51 & 0.1 \\ 0.0 & 0.0 & 0.01 & 0.26 & 0.73 \end{bmatrix}$$

Relative to earlier models, the most marked difference is that now, we are able to generate downward mobility similar to that in the data. As in the data, capital income risk increases the propensity with which households make large movements in both directions through the wealth distribution. In particular, note the final row in the matrix above, as compared to those in sections 3.4 (baseline model) and 3.6 (those augmented with additional features), the top quintile of wealth is very persistent; households who enter this quintile tend to dissave slowly and remain there. With uninsurable return risk, however, this quintile is far less persistent: of the households who begin there, less than three quarters remain there five years later. Over this period, some wealthy agents draw unfavorable return shocks, and fall downward in the wealth distribution as they rapidly deaccumulate wealth.

In Figure 3.27, higher values of persistence in the return shocks produce mobility more in line with the data. These shocks also generate a wealth distribution more in line with its empirical counterpart, especially at the very top. Figure 3.28 shows the share of aggregate wealth held by the wealthiest 50%, 10%, and 1% of households in the model, along with empirical counterparts, across values of persistence ρ_z. The most noted change is in the wealth share of the top 1%: although we overestimate this share across all ρ_z values, the fit at the top is better with more persistent shocks.[21] None of the moments in Figure 3.28 are targeted, and so we take the good fit here as a welcome sign that our model with return shocks, parsimonious though it is, captures important features of both inequality and social mobility throughout the wealth distribution, including at the very top.

3.8.3 Policy Implications: Risk and Returns to Wealth

Having established a way in which to augment the standard incomplete-markets model in order to generate realistic wealth mobility, we now turn to the normative implications of such an augmentation. Why is it important to match mobility? How does optimal policy change when we allow for agents to move through the wealth distribution at the rate they do in reality?

[21] Recall that we target the Gini coefficient in calibrating this version of the model, so all economies here reproduce this summary measure of the wealth distribution.

To answer these questions, we update the budget constraint of agents in our model to allow for a lump-sum transfer T, financed by a linear tax on capital income τ_k:

$$c_t + a_{t+1} \leq (1 + (1 - \tau_k)r_t)\, z_t a_t + \omega_t \overline{h} \varepsilon_t + T_t \tag{3.5}$$

For each value of ρ_z, we calculate the optimal linear capital income tax τ_k^*, shown in Figure 3.29. As the persistence ρ_z increases, bringing wealth mobility generated by the model in line with that in the data, the optimal capital income tax falls.

The intuition behind these results is as follows. Because individuals cannot save in an asset other than now-risky claims to the capital stock a, they value the insurance provided through the tax-and-transfer system. Effectively, a capital income tax here acts as a savings vehicle, and agents with low z_t shocks now prefer that capital income be taxed so that they can receive a higher transfer (T), which now allows them to smooth consumption. The capital income tax also "squeezes" returns as in Guvenen et al. (2015a), reducing the variance in post-tax returns and thereby ameliorating risk. The tradeoff, of course, is that the capital income tax discourages investment, lowering the capital stock and therefore the wage.

Recall from Figure 3.26 that when calibrating to US wealth inequality, a higher persistence of return shocks ρ_z implies that the shocks themselves need to be smaller. As such, when shocks are more persistent, risk is inherently lower: agents receive return shocks that are smaller, and today's return shock is more informative in predicting tomorrow's. Therefore, as we increase the persistence of returns, keeping inequality constant, the risk-reduction motive for capital income taxes shrinks, inducing a fall in the agents' preferred rate of capital income taxation. The relationship between mobility and capital income taxes is easy to infer: as wealth mobility rises, so too does the optimal tax rate. The predominant force seems to be an insurance motive: the more mobile agents are, the more their wealth fluctuates from period to period. All else being equal, agents value insurance against these movements, as swings in wealth make it difficult for agents to smooth consumption. The tax code provides this insurance, reducing returns but equalizing incomes via the lump-sum transfer.

3.9 Conclusion

We have studied wealth mobility in a Bewley model. In particular, we have shown how assumptions about the underlying process driving long run wealth inequality affect relative mobility. As labor income shocks become more persistent, relative mobility displays a hump-shape, starting low growing monotonically to a maximum around $\rho = 0.75$ and then declining sharply towards 0 as the process becomes closer to permanent. Using 'ghost' households, we

run several counterfactuals in order to decompose the pattern in mobility into the change in the structure of the wealth distribution, the change in optimal savings behavior in the face of different income risk, and changes in sequence of labor income itself (i.e., luck). We find that the hump-shape is generally attributable to the mixture of behavior and luck. The first contributes negatively to mobility as household's saving is less sensitive to more persistent shocks. The second contributes positively by generating longer strings of low or high income allowing wealth to accumulate or decline for longer over a fixed amount of time.

We document that the baseline Bewley model generates a stationary wealth distribution with lower short-run wealth mobility than has been found empirically. In the data, a non-trivial fraction of households experience large movements across wealth quintiles, even over fairly short horizons, while these movements do not occur in the model. We extend the baseline model in several ways commonly used in the literature to better match wealth inequality. While the inclusion of a very high income state with low persistence as in Castaneda et al. (2003) improves the model's predictions for upward mobility somewhat, it does not match the observed downward mobility. In all versions of the model studied, households move down in wealth too slowly, a natural result of the precautionary saving motive present in the incomplete markets model.

We examine the relationship between market completeness and wealth mobility. We find that replacing the non-contingent capital asset with two state-contingent claims (i.e., partial insurance) may reduce or increase mobility depending upon the underlying persistence of the income shock process. If ρ is sufficiently high, the more complete markets economy has higher mobility. Nevertheless, the model still fails to quantitatively match the observed mobility.

Finally, we study wealth mobility when agents are subject not just to uninsurable income risk, but also uninsurable risk to their rates of return. In contrast to the other models considered, the addition of return risk generates a frequency of downward movements in the distribution over short horizons that we observe in the data. Our results here suggest that highly persistent, moderately-sized shocks to rates of return perform well in replicating wealth inequality and mobility. We explore the positive implications of this feature of idiosyncratic risk, finding that the higher is wealth mobility, the higher is the optimal capital income tax, as this tax provides insurance against large swings in returns and thus wealth.

3.10 Appendix

3.10.1 Comparing Mobility Measures

Shorrocks Index This section aims to provide examples of the aspects of mobility captured by our measures. We begin with the Shorrocks Measure. As an illustrative example, the two Markov processes

$$\Pi_A = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.25 & 0.5 & 0.25 \\ 0.0 & 0.5 & 0.5 \end{bmatrix}$$

and

$$\Pi_B = \begin{bmatrix} 0.5 & 0.0 & 0.5 \\ 0.25 & 0.5 & 0.25 \\ 0.5 & 0.0 & 0.5 \end{bmatrix}$$

would be regarded as equally mobile; however, the second process moves "faster" since it admits one-period transitions between the lowest and highest wealth states, while the first process requires any movement between extreme states to first pass through the middle.

Bartholomew's Immobility Measure To see how this measure works, consider the Markov processes

$$\Pi_A = \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.25 & 0.5 & 0.25 \\ 0.0 & 0.5 & 0.5 \end{bmatrix}$$

and

$$\Pi_B = \begin{bmatrix} 0.75 & 0.0 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.0 & 0.75 \end{bmatrix}$$

According to Bartholomew's measure, these chains are equally mobile:

$$\mu_B(\Pi_A) = 0.75$$
$$\mu_B(\Pi_B) = 0.75.$$

Agents make more frequent, "small" moves in A, and less frequent "large" moves in B.

Mean First Passage Time We calculate the measure of mobility based on the Mean First Passage Time (MFP) for a Markov transition matrix M as follows. Set $A = I - M$

and partition as

$$A = I - M = \begin{bmatrix} U & c \\ d^T & \alpha \end{bmatrix}.$$

Meyer (1978) shows that

$$T = (I - K + J \operatorname{diag}(K)) (\operatorname{diag}(K))^{-1} + E$$

where

$$E = \begin{bmatrix} 0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & 0 & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix},$$

J is a matrix of all ones, and

$$K = \begin{bmatrix} U & 1^T \\ d^T & 1 \end{bmatrix}^{-1}.$$

Conlisk (1990) proposes using

$$MFP = x'Tx$$

as a measure of mobility. As mentioned in the main text, we standardize the MFP measure using the number of quintiles N:

$$\mu_{MFP}(M) = \frac{N}{MFP}$$

This normalizes the measure so that $\mu_{MFP}(M)$ lies between zero and one, as all of the remaining measures do.

Cowell-Flachaire index Cowell and Flachaire (2018) propose a superclass of mobility measures that allow for the aggregation of a broad range of mobility concepts. Members of the class take the form

$$\Omega(M) = \begin{cases} \frac{1}{\alpha|1-\alpha|} n \sum_{i=1}^{n} \left[\left[\frac{u_i}{U}\right]^{\alpha} \left[\frac{v_i}{V}\right]^{1-\alpha} - 1 \right] & \alpha \neq 0, 1 \\ -\frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{V} \log\left(\frac{u_i}{U} / \frac{v_i}{V}\right) & \alpha = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \frac{u_i}{U} \log\left(\frac{u_i}{U} / \frac{v_i}{V}\right) & \alpha = 1 \end{cases}$$

where n is the number of individuals in the population; u_i and v_i represent the "status" of individual i at the beginning and end, respectively, of the time period under consideration; and U and V are the mean status levels across individuals in each period. The

index is amenable to different definitions of status. For example, in the context of wealth mobility, status may be defined as the specific levels of wealth held by each individual, as a collection of intervals over wealth, or as subsets of a wealth distribution. The parameter α controls the weight given to downward movements relative to upward ones. For $\alpha < (>)0.5$, $\Omega(M)$ is more sensitive to upward (downward) movements. Cowell and Flachaire (2018) show that members of Ω satisfy many desirable properties in a mobility measure, including independence of population size and preservation of order under scaling.

The member of Ω that is suitable for comparing mobility between $K \times K$ quantile transition matrices is

$$\mu_{CF}(M) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left[\frac{2}{K(K+1)}\sum_{k=1}^{K}\sum_{l=1}^{K} m_{kl}k^{\alpha}l^{1-\alpha} - 1\right] & \alpha \neq 0, 1 \\ \frac{-2}{K(K+1)}\sum_{k=1}^{K}\sum_{l=1}^{K} m_{kl}l\log\left(\frac{k}{l}\right) & \alpha = 0 \\ \frac{2}{K(K+1)}\sum_{k=1}^{K}\sum_{l=1}^{K} m_{kl}k\log\left(\frac{k}{l}\right) & \alpha = 1 \end{cases}$$

3.10.2 Further Details on Ghost Households

Informed ghosts The informed ghost understands the true value of its ρ. It takes prices as given and solves the household problem. The ghost differs from the standard households in its economy in both how it responds to shocks conditional on current wealth and the shock sequence it faces. We calculate the ghost's mobility matrix under the wealth distribution generated by $\rho_{GE} \neq \rho_G$ and compare it to the mobility matrix generated by the $\rho_{GE} = \rho_G$ economy and attribute the differences to structure. Figures 3.4-3.7 plot contours of the surface generated by the (ρ_{GE}, ρ_G) pairs. The 45 degree line running the through the contour is general equilibrium mobility measures from our baseline experiments. Starting at a point on the that line, mobility declines as we move along.

On Figure 3.4, we draw an example of the structural vs. exchange mobility calculation. Comparing mobility at point A to mobility at point B, our method first picks out point C where ρ_{GE} is the same as in B but ρ_G is equal to the persistence in A. Any differences in mobility between C and A must come from facing a different distribution of wealth (i.e., structure). Movement from A to C then is 'structure' and movement from C to B is 'exchange'.

Figure 3.8 plots the savings decision rules of three households with different when the economy-wide ρ is 0.73. First notice Proposition 6 at play. Ghosts with low ρ have savings decisions that are much more distant across ε realizations, while those with ε near 1 have policy rules near the 45 degree line. Agents with $\rho = 0.05$ will experience relative large and frequent changes in wealth across one period, while those with $\rho = 0.98$ will switch infrequently but their wealth will also change very little each period. Importantly, notice

that the change in distance from $\rho = 0.05$ to $\rho = 0.73$ is much smaller than it is from $\rho = 0.73$ to $\rho = 0.98$. This is a key factor for the hump-shape in total mobility. Depending upon the measure used, the trade off between persistent shocks and smaller step sizes reaches maximum mobility value somewhere between $\rho = 0.7$ and $\rho = 0.8$. For values below 0.7, mobility is reduced because agents are switching from savers to dissavers too frequently. For values above 0.8, households are accumulating (decumulating) wealth too slowly.

Uninformed ghosts To decompose exchange mobility from between behavior and luck, we run the same type of experiment as above, but now the ghost does not realize that its labor productivity process has a different autocorrelation. This ghost uses the same decision rules as the other households in the distribution, but it realizes a different sequence of shocks. Figure 3.9-3.12 plot mobility of these agents as a function of (ρ_G, ρ_{GE}). As before with the informed ghosts, we draw path to highlight one of the three components, here being luck. We have a similar breakdown on figure 3.9. Moving from A to B is a combination of all three components, but movement between A and C is entirely due to luck because the ghosts in both cases reside in the same distribution and have the same decision rules. The only difference is that a ghost at C has a more persistent shock process (identical to the ghost at B).

The differences in the measures are also notable. The Bartholomew and Shorrocks measures show mobility increasing as the ghost's persistence parameter increases. For the mean first passage measure, the relationship has a similar hump-shaped pattern. Holding ρ_{GE} constant, mobility increases in ρ_G until it reaches a maximum somewhere between 0.70 and 0.80; then it declines rapidly. Oddly, the 2^{nd} largest eigenvalue measure actually decreases in ρ_G.

Here the we see the hump-shaped pattern in mobility. When the economy-wide ρ is low, the savings rules are far apart so non-phantom agents in a fast lane but change often. Mobility is low. The non-optimizing phantom agents with higher ρ share the same fast lanes but are much less likely to switch. They have longer chains of wealth accumulations and decumulations, and so their mobility is higher. One again, when ρ gets too high, the ghost agents remain in their lane for a very long time. They will move through the distribution but only very infrequently, and they will usually just 'pass through' one intermediate quintile. Those with low ε will spend a large number of periods in the bottom quintile before finally drawing a good shock and making a transition back through the distribution toward the top quintile where they will once again remain for a large number of periods.

3.10.3 Mobility and Market Incompleteness

We know that market incompleteness is a necessary condition for permanent mobility – mobility may be present along a transition path if agents have different preferences, but eventually it will disappear as the economy transitions to a steady state (see Caselli and Ventura, 2000 and Carroll and Young, 2011). We now take up the question of how mobility is connected to incompleteness, in the sense of the spanning of assets.

We consider two experiments. First, suppose there exist two assets, one of which pays off if $\varepsilon \geq E[\varepsilon]$ and one that pays off if $\varepsilon < E[\varepsilon]$. Second, suppose there exist three assets, which pay off if $\varepsilon > E[\varepsilon]$, $\varepsilon = E[\varepsilon]$, and $\varepsilon < E[\varepsilon]$. In each case, asset markets are 'more complete', but mobility could easily go either way. Since the price of these assets is smaller than the price of a risk-free security, portfolios that 'lever up' in certain states can lead to large changes in wealth should those states realize; the results inRampini and Viswanathan (2016) show that agents in our economy will in fact choose to endogenously hold a skewed portfolio if they are sufficiently poor.[22]

We compare the mobility results from these partial insurance cases to the baseline model. In each case, we set the number of productivity states to 7. We will mainly discuss the two-asset case; for simplicity, denote the productivity states where $\varepsilon \geq E[\varepsilon]$ 'good' states, and the other 'bad' states.

Figure 3.14 plots the portfolio decisions of several informed ghost households. In each case, the ρ value of the underlying economy is 0.73. Each subplot shows the decisions of two ghosts with the same persistence value, one with $\varepsilon = \varepsilon_{\min}$ and one with $\varepsilon = \varepsilon_{\max}$. The solid lines represent the number of claims purchased which pay off if the next period's productivity belongs to the same state as today's productivity. The dashed lines are the claims which pay off in the opposite state from today's. For example, for the $\varepsilon = \varepsilon_{\min}$ household, the solid line is the stock of claims that pay off if one of the bad states is realized next period, and the dashed line is those that pay off if the good state is realized instead.

First notice that the a household currently in the bad (good) state purchases contingent claims against the bad (good) state near the 45 degree line. In fact, the household's decision rules in this regard are similar in appearance to those in the one asset case. Just as in the baseline case, these saving rules become closer as the probability of remaining in the same state increases. Again, households consume a larger fraction of income from more persistent shocks. This feature of the portfolio induces more mobility as it allows for long strings of consistent wealth accumulation and decumulation, as we illustrated in the section above.

[22] It is straightforward conceptually to permit an arbitrary number of state-contingent claims, but the high autocorrelation of the states means that some of these assets will have essentially zero price; prices that are too low lead to instability in our solution algorithm.

The other side of the portfolio, that is the holding of claims which pay off only if the household's state switches (from good to bad or bad to good) in the next period, is quite different, and it can have a big effect on mobility, particularly in the ghost household cases. Households currently in a good productivity state purchase considerably more claims against switching to a bad productivity state. These claims compensate both for the low labor income from a bad state and provide additional precautionary savings the likely recurrence of bad state shocks. Moreover, because the probability of switching between good and bad states is low (especially for an $\varepsilon_{\max}$ or $\varepsilon_{\min}$ household), this insurance is very cheap. Naturally then, as ρ increases the good state household's claims against bad states rises, causing the balance of the portfolio to tilt more and more.

The portfolio of household's currently in a bad productivity depends on their wealth level. At sufficiently high wealth, the portfolio looks like a mirror image of the good state household's portfolio. The purchase of claims against a bad state lie close to the 45 degree line, while the purchases of claims against the good state are much lower. At lower levels of current wealth, households would like to short the claim against good states, since consumption in the bad states is very valuable. Since this shorting is not allowed, these households simply do not participate in that asset market. With the exception of the wealth region where the non-negativity constraint binds, the response of any household portfolios can be generalized in the following way: as ρ increases, the demand for claims that pay off if the current state continues become less sensitive to income shocks, while the demand for assets that payoff if the state switches becomes more sensitive.

The consequence of this portfolio behavior for mobility across ρ is that as households become less and less likely to switch states, their wealth path is characterized by small, gradual movements interspersed with infrequent large shifts. Figure 3.15 plots mobility in the partial insurance cases against the single asset baseline. Notice that mobility is lower in the partial insurance environment unless the labor income process is quite persistent. Regardless of the type of measure, mobility under partial insurance peaks at a higher ρ and may even reach a higher (absolute) level before quickly descending again as shocks approach being permanent. Figure 3.15 also shows that the pattern is strengthened by the addition of the third asset.

Although the partial insurance environment features more wealth mobility at high values of ρ, there is still less mobility than in data for our chosen value of ρ. The five-year wealth

transition matrix is

$$\begin{bmatrix} 0.75 & 0.25 & 0.00 & 0.00 & 0.00 \\ 0.24 & 0.56 & 0.19 & 0.05 & 0.00 \\ 0.01 & 0.19 & 0.66 & 0.14 & 0.00 \\ 0.00 & 0.00 & 0.15 & 0.77 & 0.08 \\ 0.00 & 0.00 & 0.00 & 0.08 & 0.92 \end{bmatrix}$$

3.10.4 Proofs

Proof of Proposition 6

Proof. Consider two households in the same wealth distribution Denote by π_{ij} the conditional probability that $\varepsilon' = \varepsilon_j$ given $\varepsilon = \varepsilon_i$. The corresponding conditional probability that $\varepsilon' = \varepsilon_{-j}$ is $1 - \pi_j$. Because $\rho^A > \rho^B$, $\pi_{11}^A > \pi_{11}^B$, and $\pi_{21}^B > \pi_{21}^A$.

We will show that $g_a^B(a, \varepsilon_1) < g_a^A(a, \varepsilon_1) < g_a^A(a, \varepsilon_2) < g_a^B(a, \varepsilon_2)$. It follows from the conditions on u and on the compactness of the budget set that $g_a^i(a, \varepsilon)$ is strictly increasing both arguments, so the inner most inequality is immediate. Next we will prove that $g_a^B(a, \varepsilon_1) < g_a^A(a, \varepsilon_1)$.

Assume not so $g_a^A(a, \varepsilon_1) \leq g_a^B(a, \varepsilon_1)$. Then by the budget constraint $c^B \leq c^A$, where c^i is consumption of household i. By the strict concavity of u,

$$u'\left(c^A\right) \leq u'\left(c^B\right)$$

which from the Euler equation implies

$$\pi_{11}^A V_1^A\left(g_a^A(a, \varepsilon_1), \varepsilon_1\right) + \left(1 - \pi_{11}^A\right) V_1^A\left(g_a^A(a, \varepsilon_1), \varepsilon_2\right) \leq \pi_{11}^B V_1^B\left(g_a^B(a, \varepsilon_1), \varepsilon_1\right) + \left(1 - \pi_{11}^B\right) V_1^B\left(g_a^B(a, \varepsilon_1), \varepsilon_2\right)$$

where V_1 is the derivative of V with respect to wealth.

We can use Theorem 6.8 from Acemoglu (2009) to establish that V is strictly concave in a.

The strict concavity of V in a leads to a contradiction since

$$\begin{aligned} V_1^A\left(g_a^A(a, \varepsilon_1), \varepsilon_1\right) &< \pi_{11}^A V_1^A\left(g_a^A(a, \varepsilon_1), \varepsilon_1\right) + \left(1 - \pi_{11}^A\right) V_1^A\left(g_a^A(a, \varepsilon_1), \varepsilon_2\right) \\ &\leq \pi_{11}^B V_1^B\left(g_a^B(a, \varepsilon_1), \varepsilon_1\right) + \left(1 - \pi_{11}^B\right) V_1^B\left(g_a^B(a, \varepsilon_1), \varepsilon_2\right) \\ &< V_1^B\left(g_a^B(a, \varepsilon_1), \varepsilon_1\right) \end{aligned}$$

which implies

$$g_a^A(a, \varepsilon_1) > g_a^B(a, \varepsilon_1).$$

Finally, we will show that $g_a^A(a,\varepsilon_2) < g_a^B(a,\varepsilon_2)$. Once again, assume not. Then

$$\begin{gathered}
g_a^B(a,\varepsilon_2) \leq g_a^A(a,\varepsilon_2) \\
u'\left(c^B\right) \leq u'\left(c^A\right) \\
\pi_{21}^B V_1^B\left(g_a^B(a,\varepsilon_2),\varepsilon_1\right) + \left(1-\pi_{21}^B\right) V_1^B\left(g_a^B(a,\varepsilon_2),\varepsilon_2\right) \leq \pi_{21}^A V_1^A\left(g_a^A(a,\varepsilon_2),\varepsilon_1\right) + \left(1-\pi_{21}^A\right) V_1^A\left(g_a^A(a,\varepsilon_2),\varepsilon_2\right) \\
V_1^B\left(g_a^B(a,\varepsilon_2),\varepsilon_1\right) < V_1^A\left(g_a^A(a,\varepsilon_2),\varepsilon_2\right)
\end{gathered}$$

Again by strict concavity of V in a,

$$g_a^B(a,\varepsilon_2) > g_a^A(a,\varepsilon_2)$$

which is a contradiction. □

3.10.5 Additional Figures

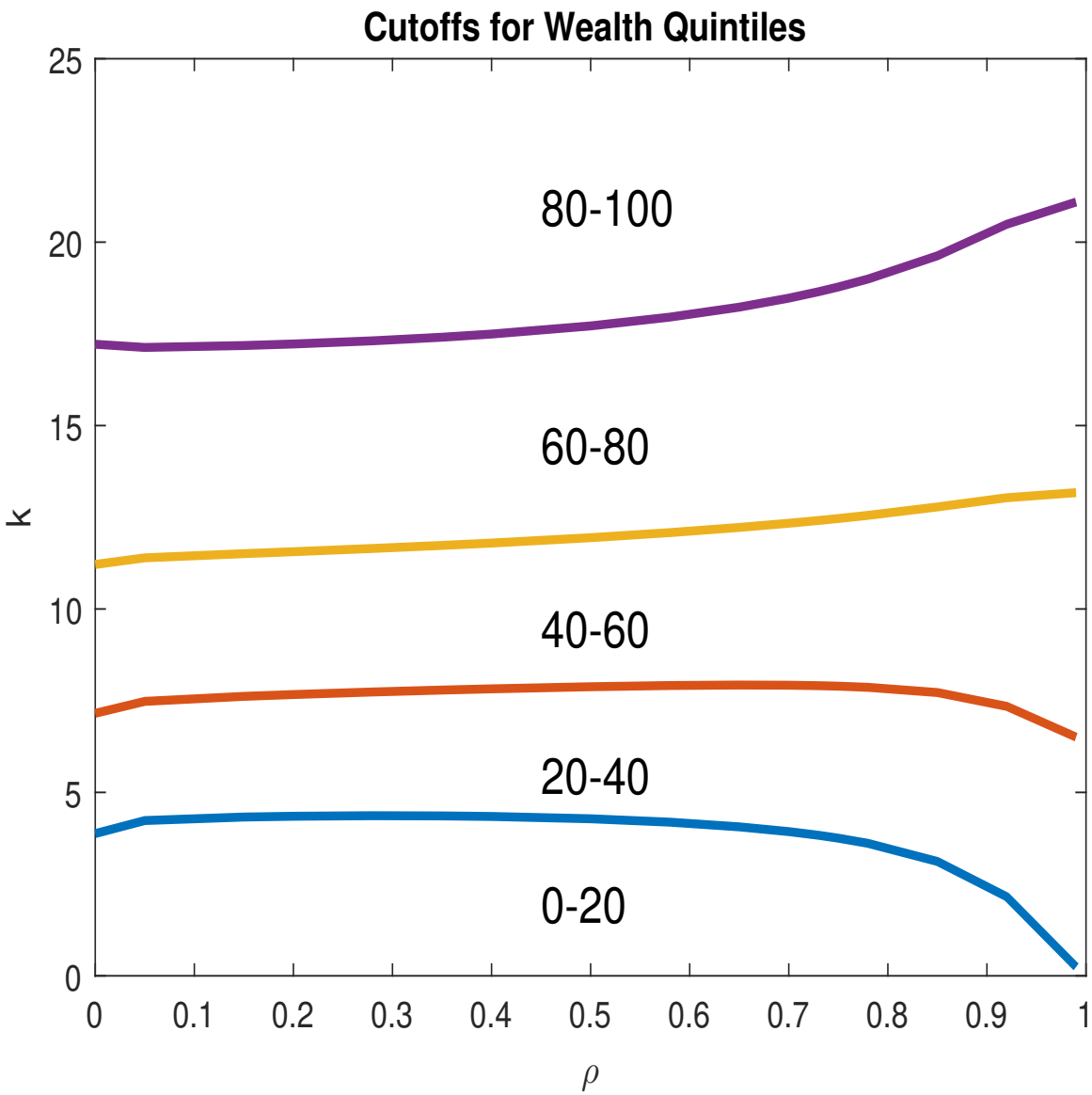


Figure 3.1: Cutoffs for wealth quintiles across persistence

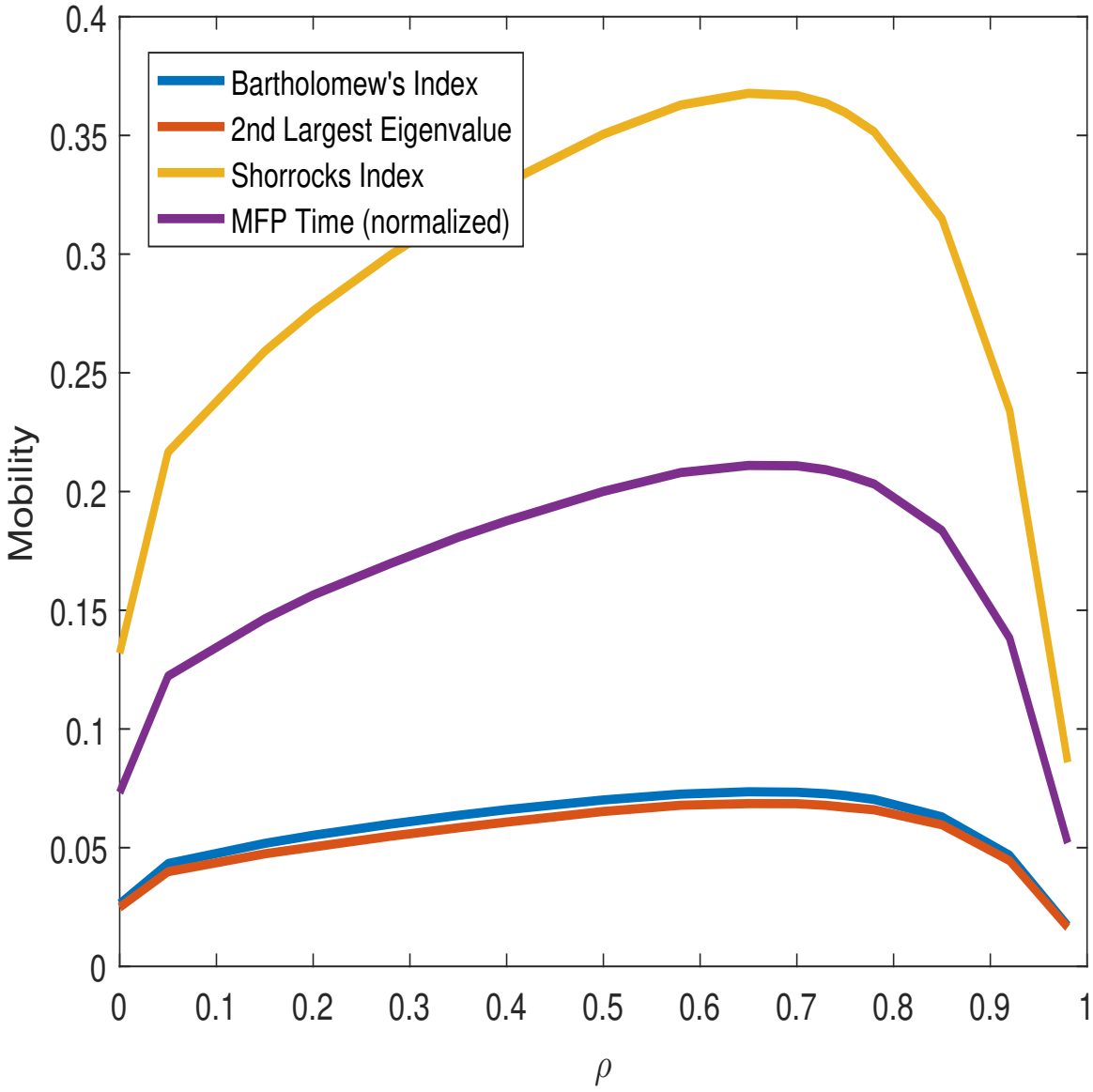


Figure 3.2: Mobility across persistence

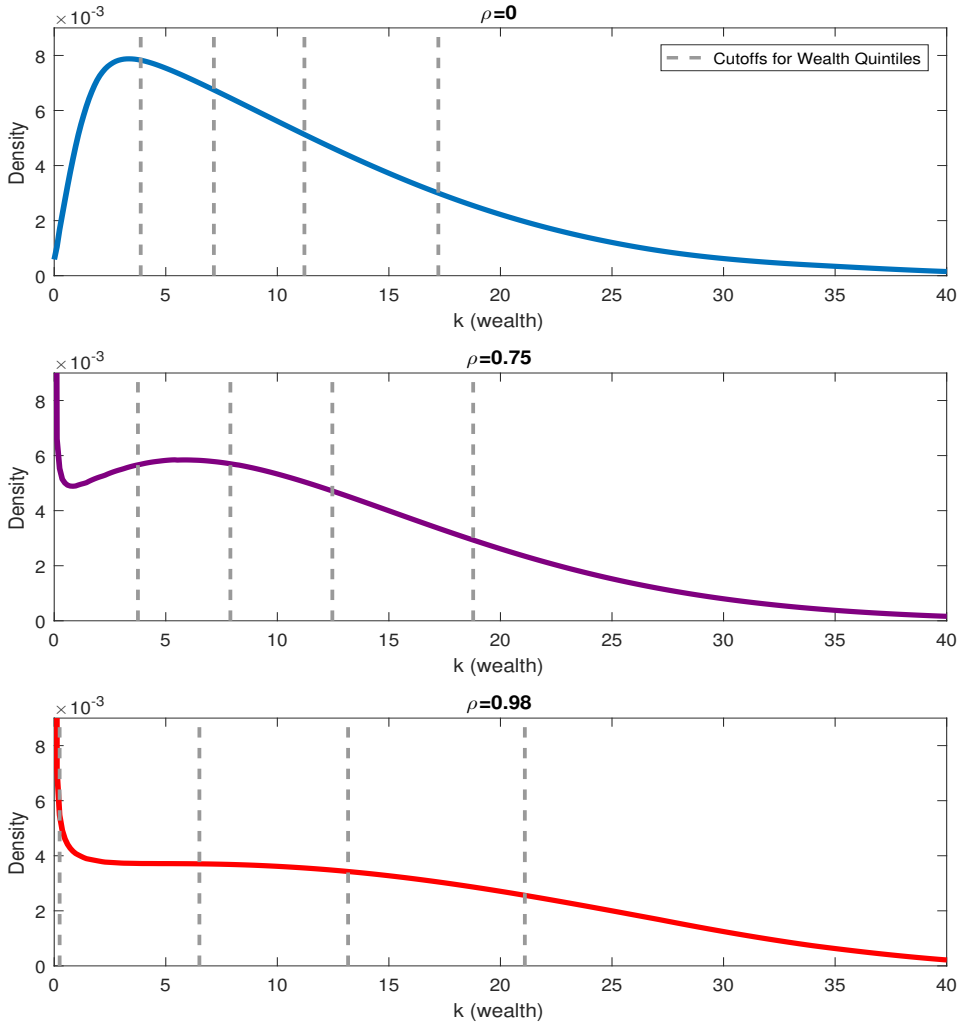


Figure 3.3: Boundaries between quintiles for different values of ρ

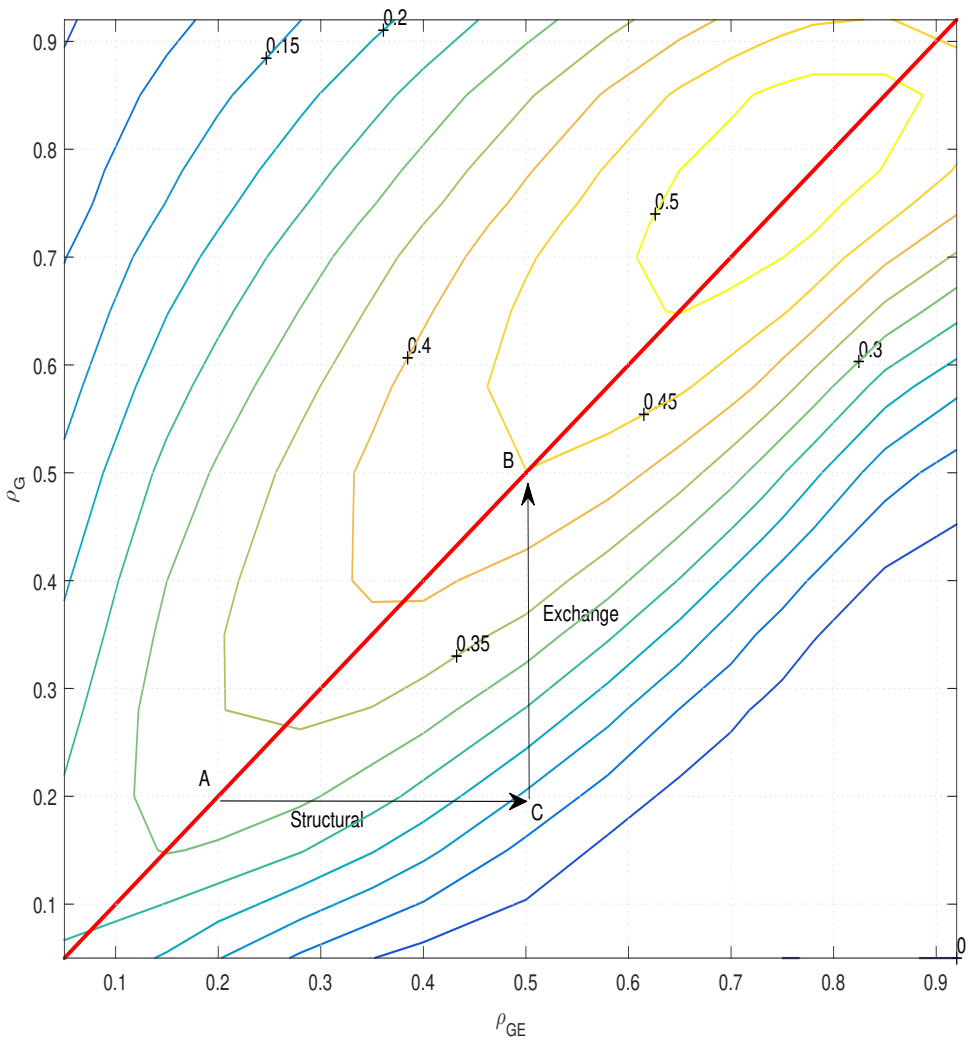


Figure 3.4: Mobility of optimizing ghost: μ_{MFP}

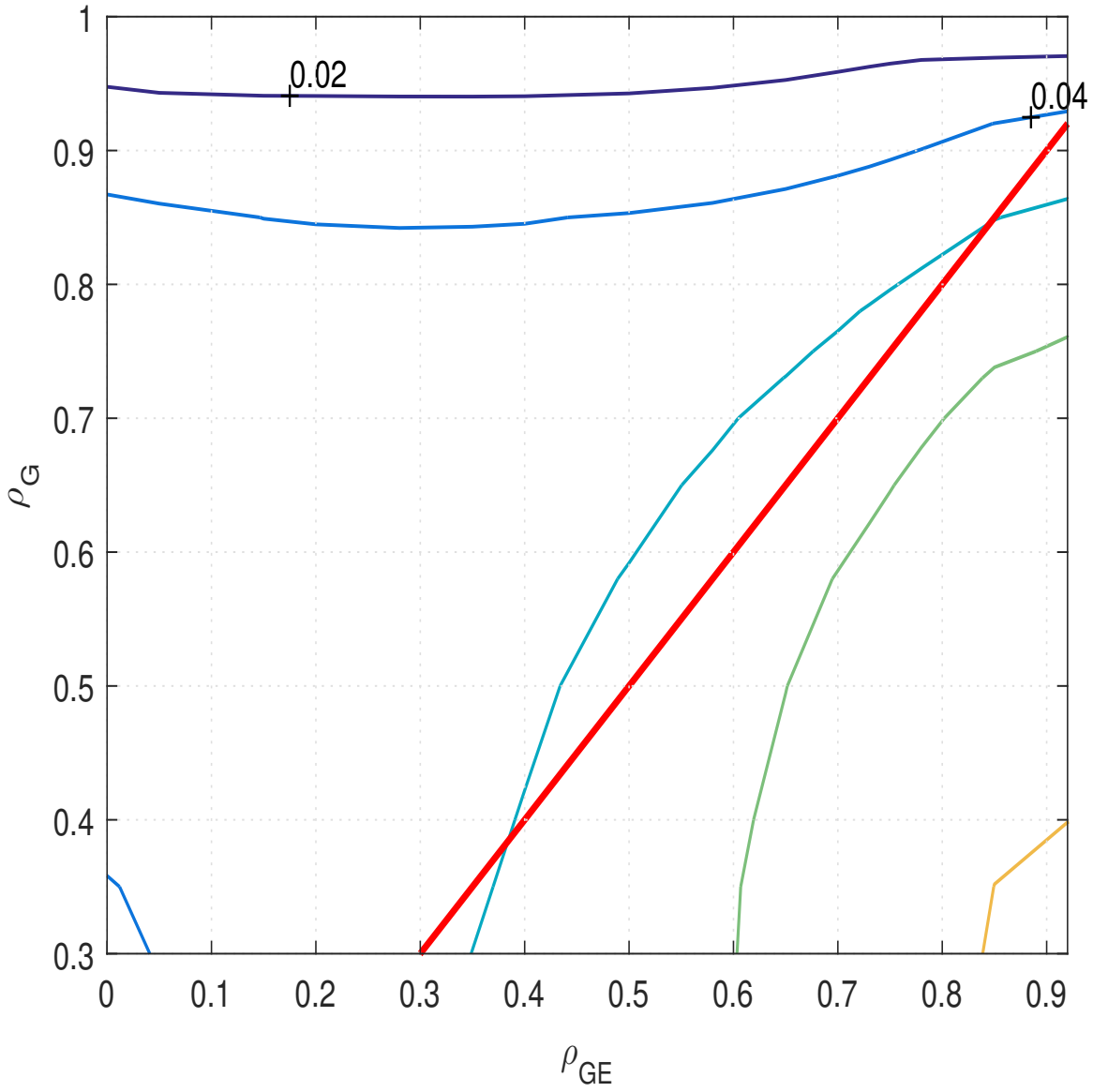


Figure 3.5: Mobility of optimizing ghost: μ_{2E}

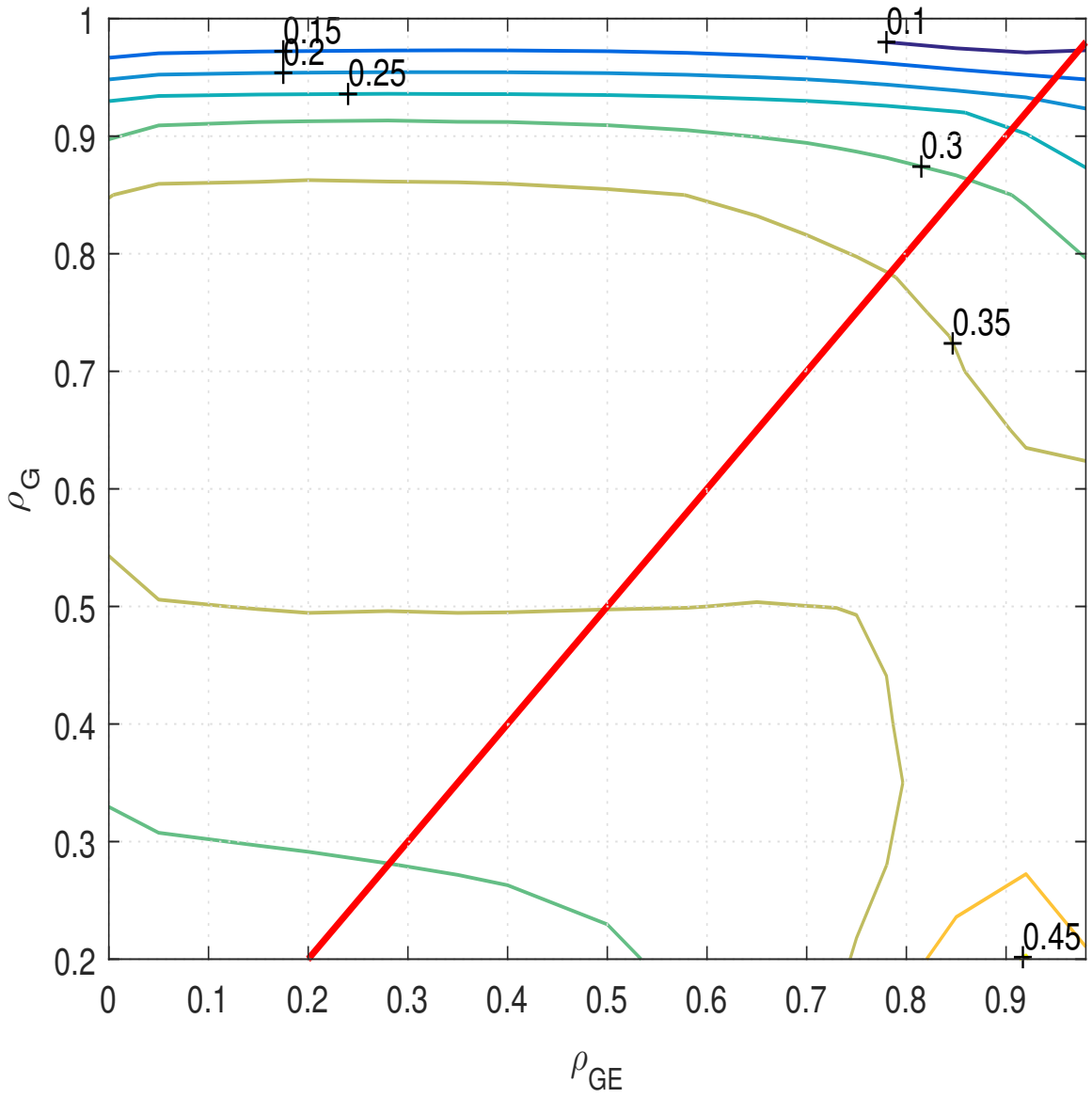


Figure 3.6: Mobility of optimizing ghost: μ_S

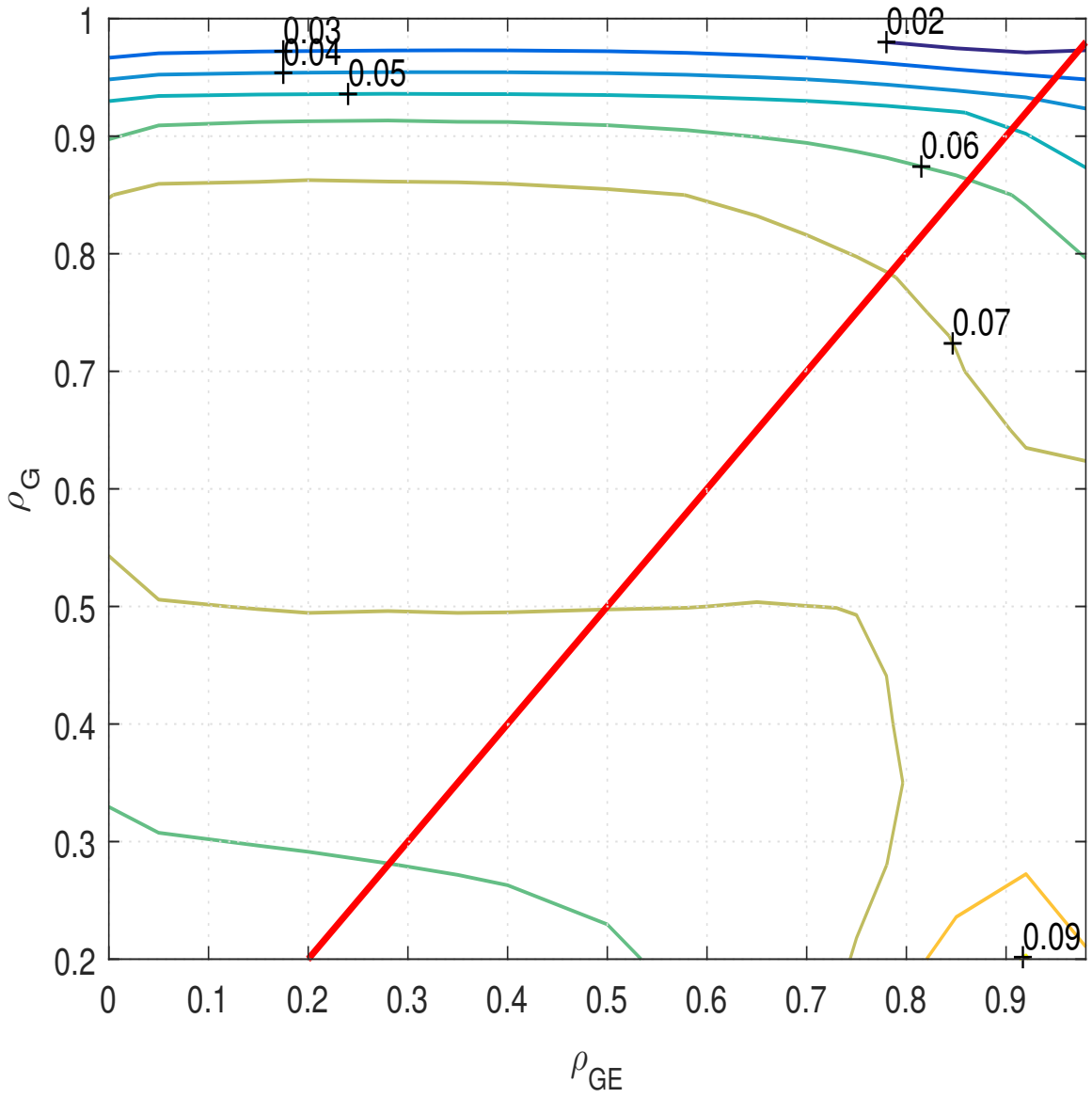


Figure 3.7: Mobility of optimizing ghost: μ_B

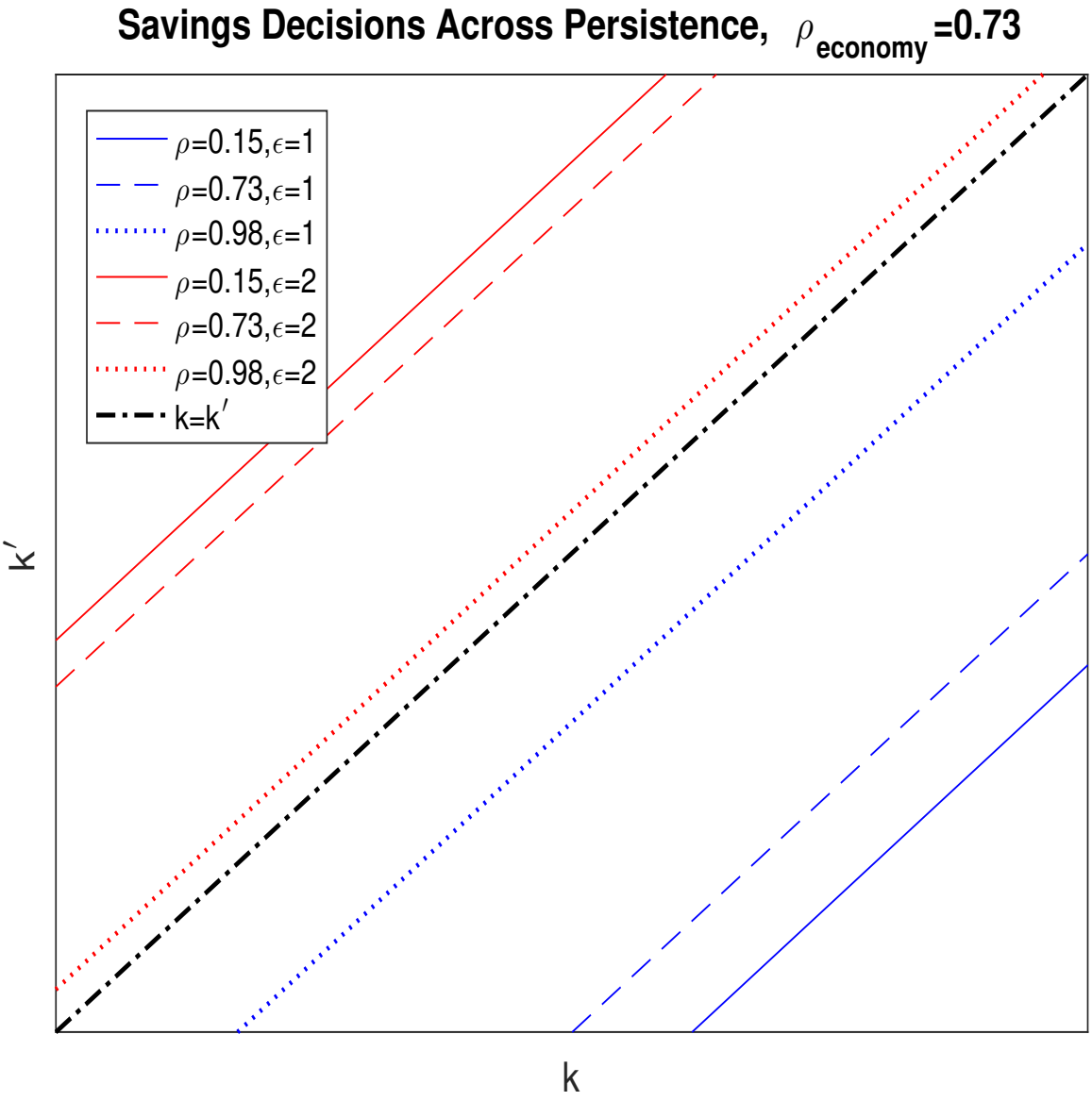


Figure 3.8: Saving Decisions across persistence

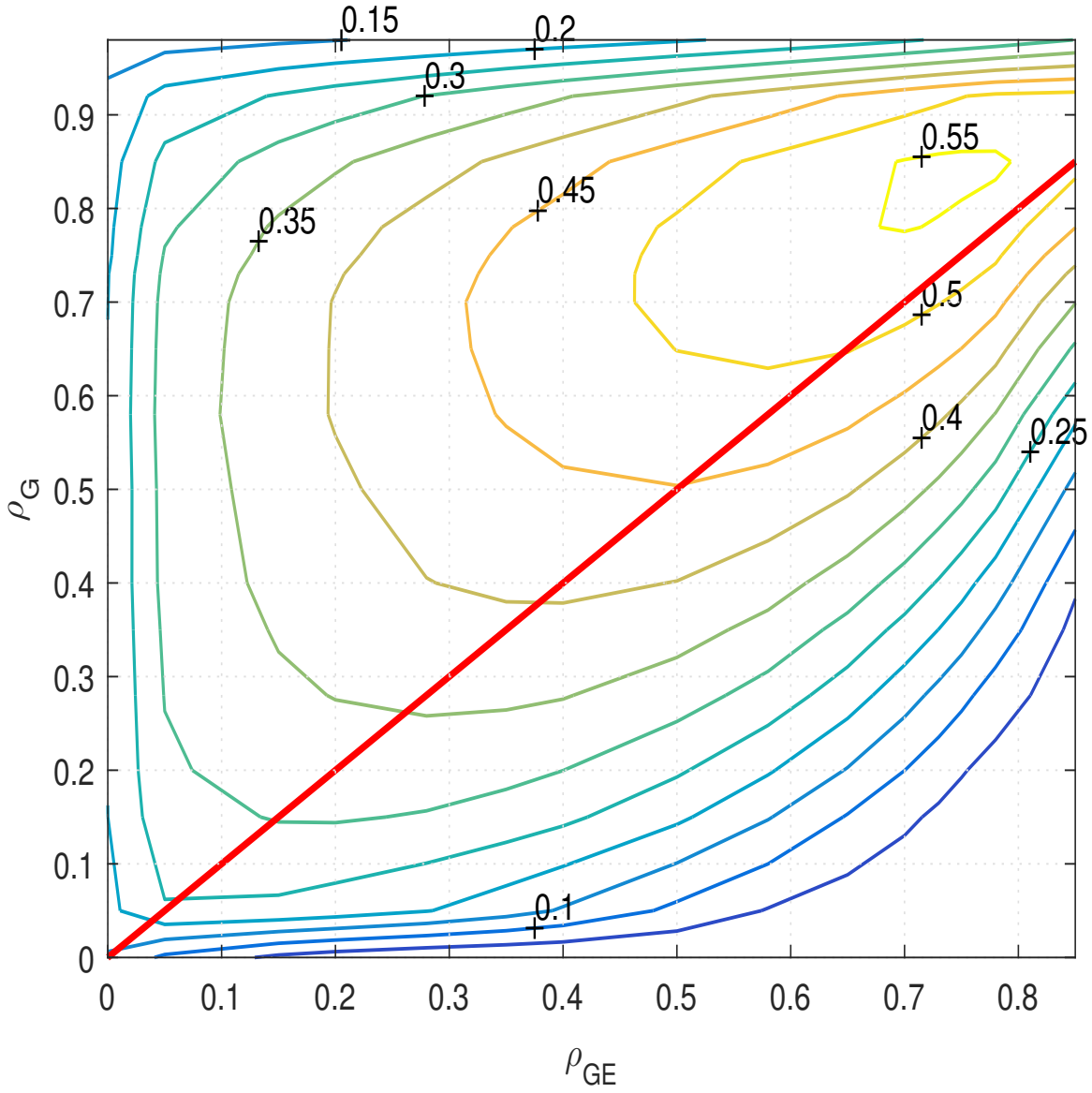


Figure 3.9: Mobility of non-optimizing ghost: μ_{MFP}

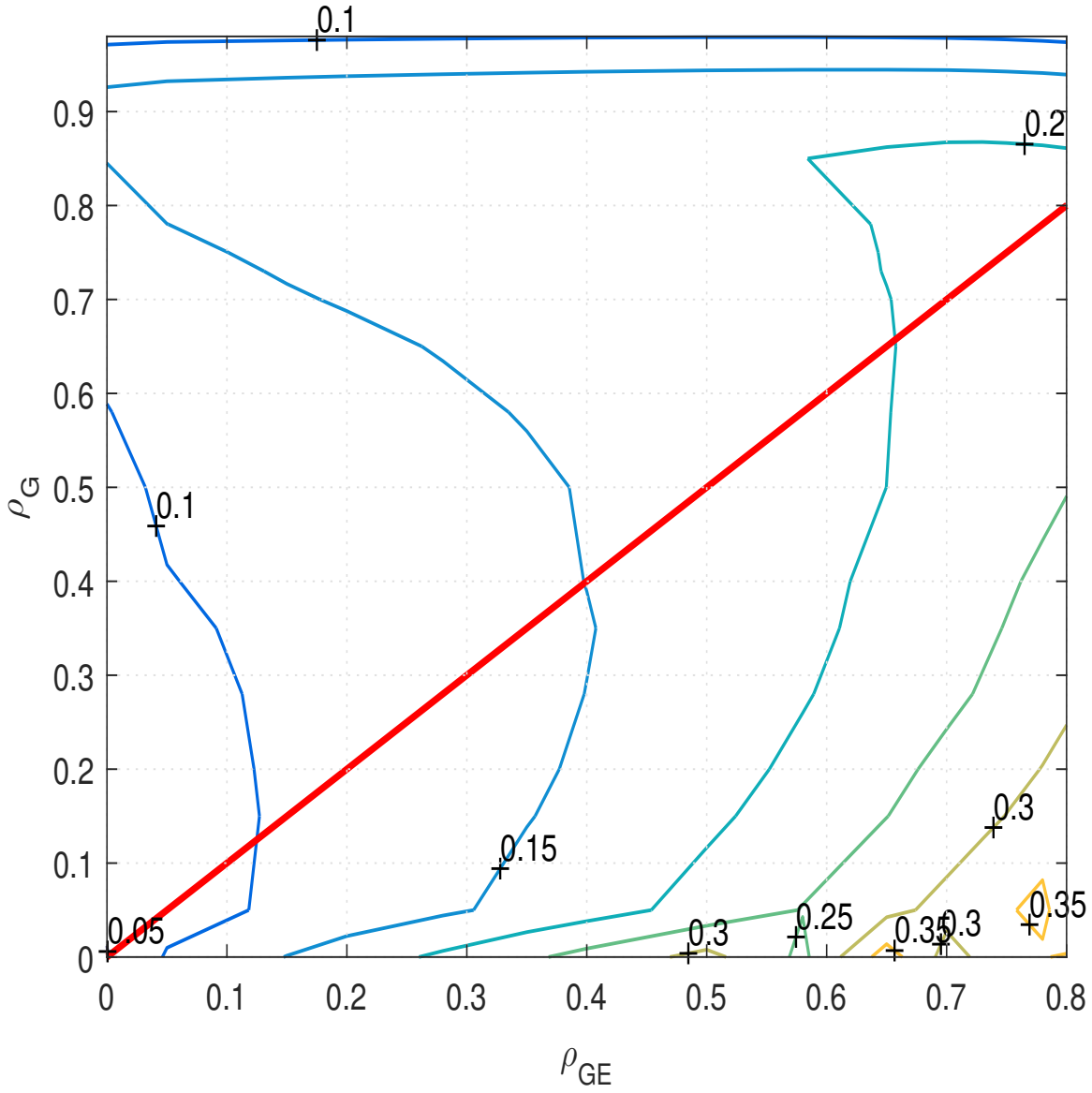


Figure 3.10: Mobility of non-optimizing ghost: μ_{2E}

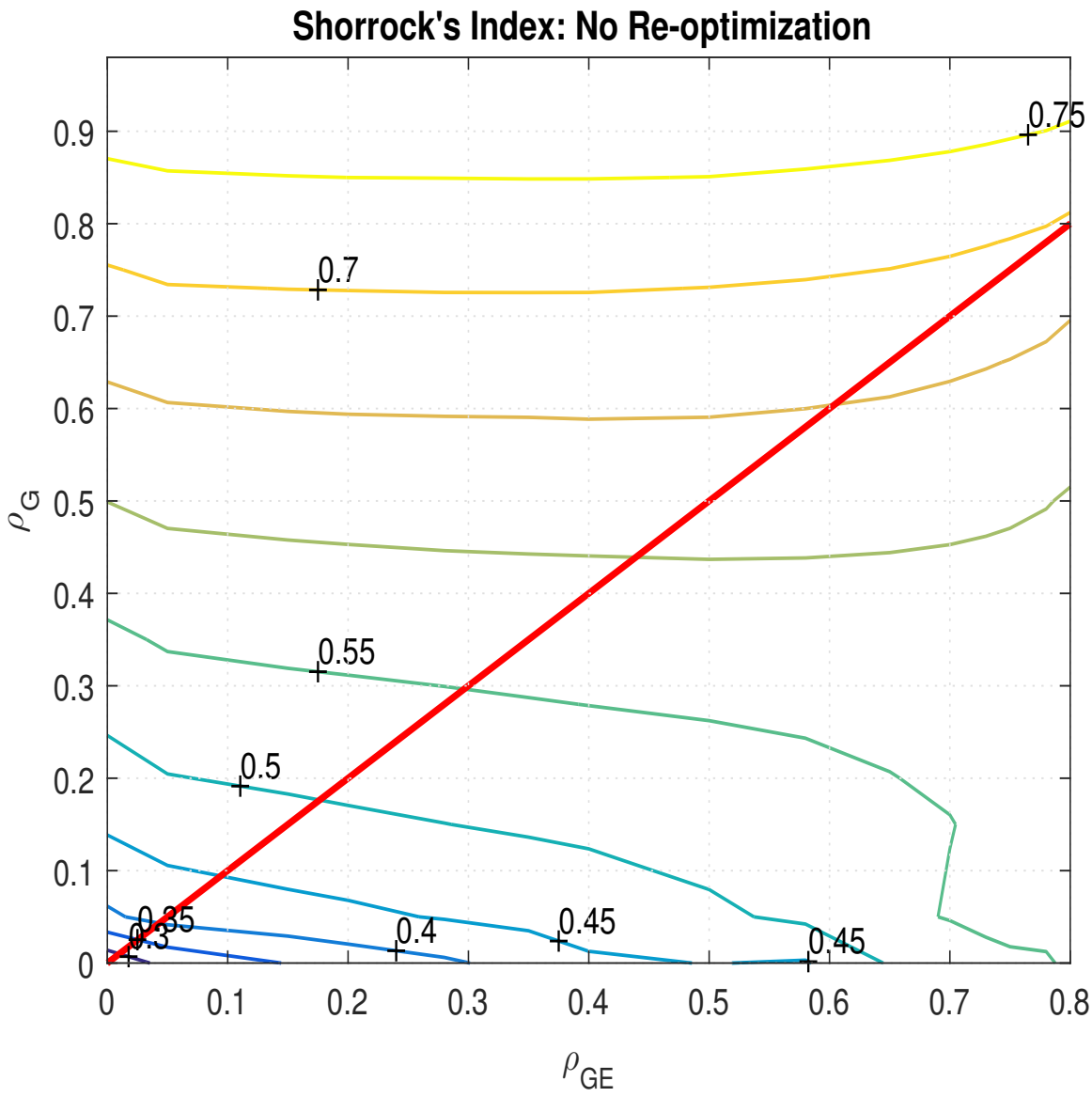


Figure 3.11: Mobility of non-optimizing ghost: μ_Sm

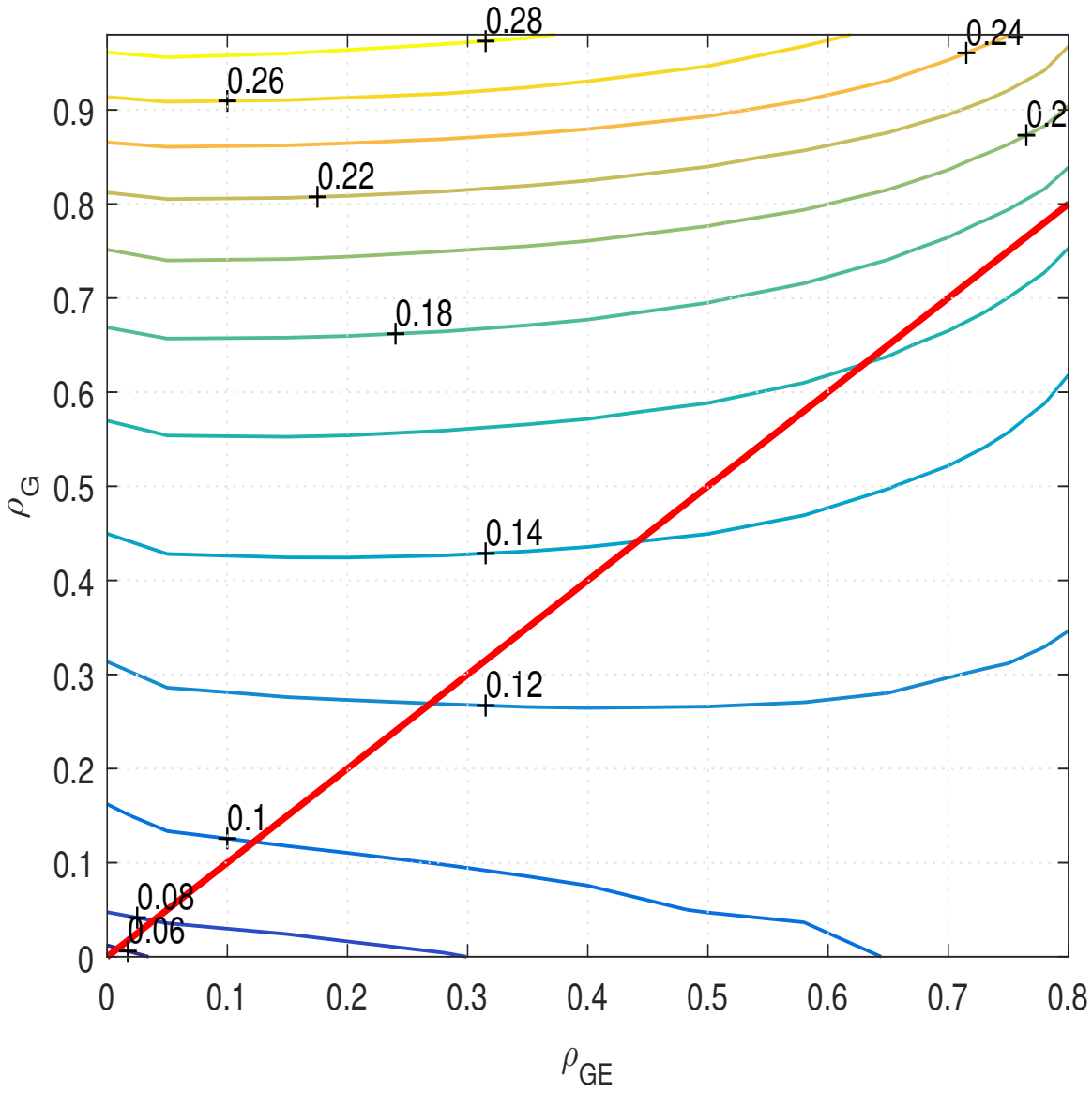


Figure 3.12: Mobility of non-optimizing ghost: μ_B

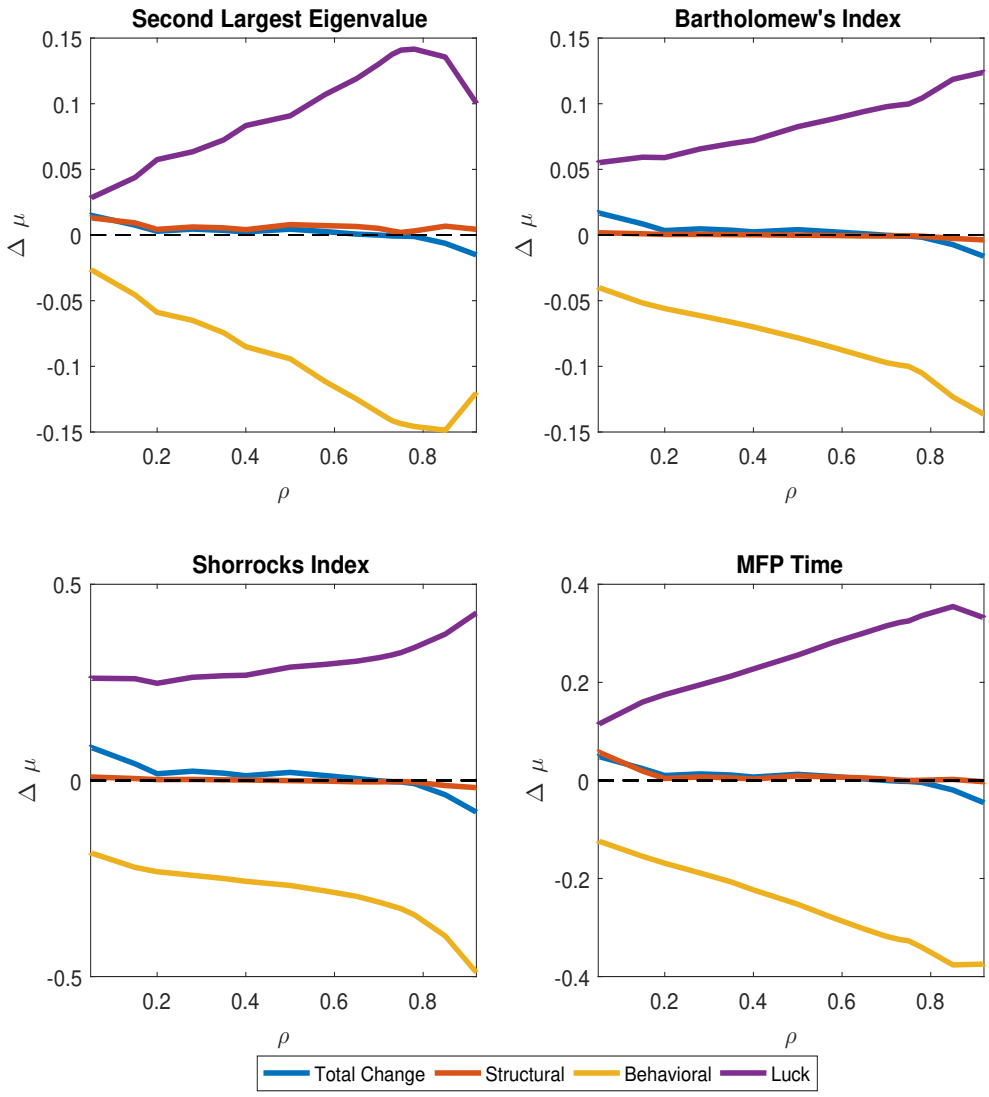


Figure 3.13: Decomposition of change in mobility as ρ increases

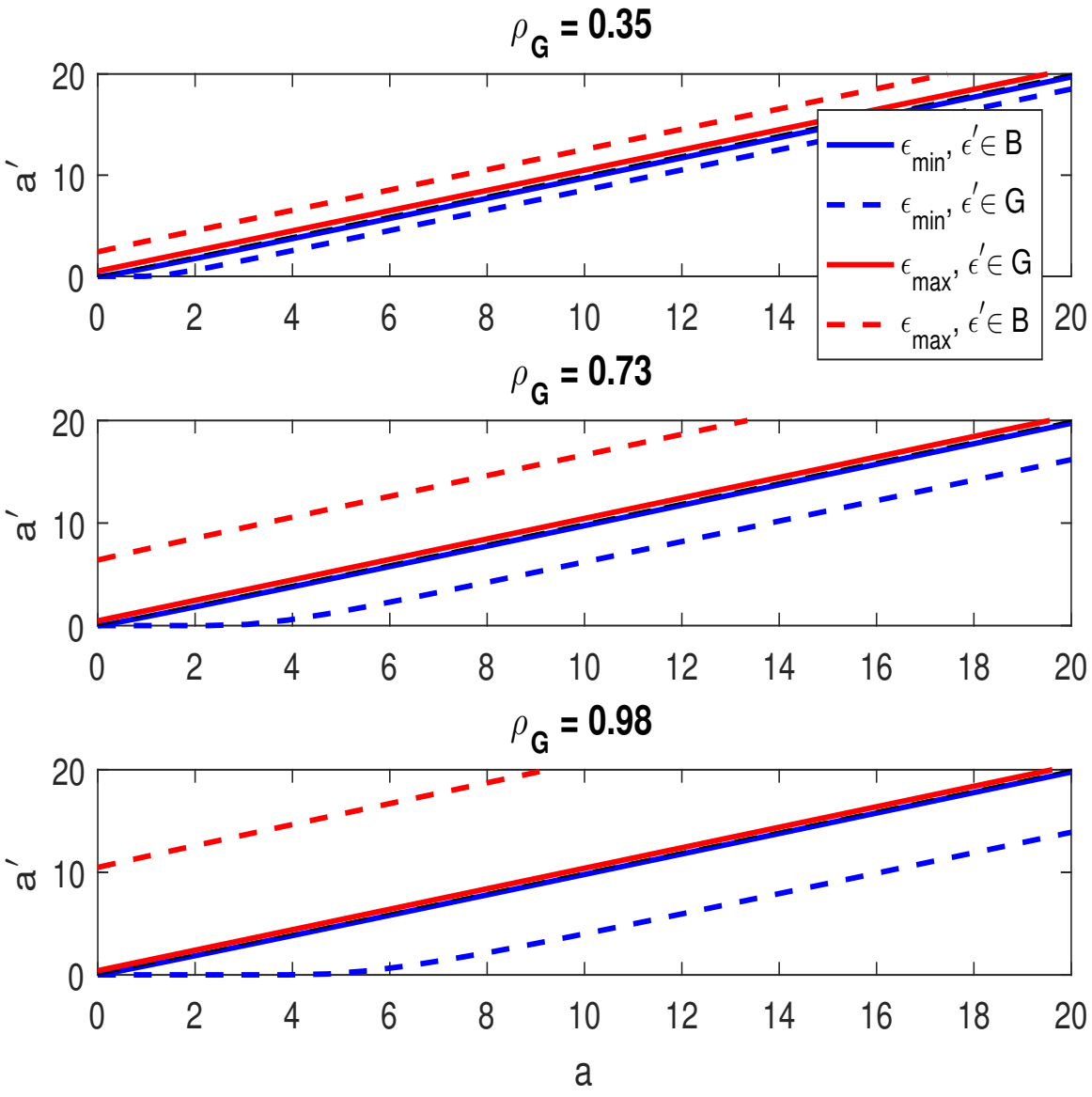
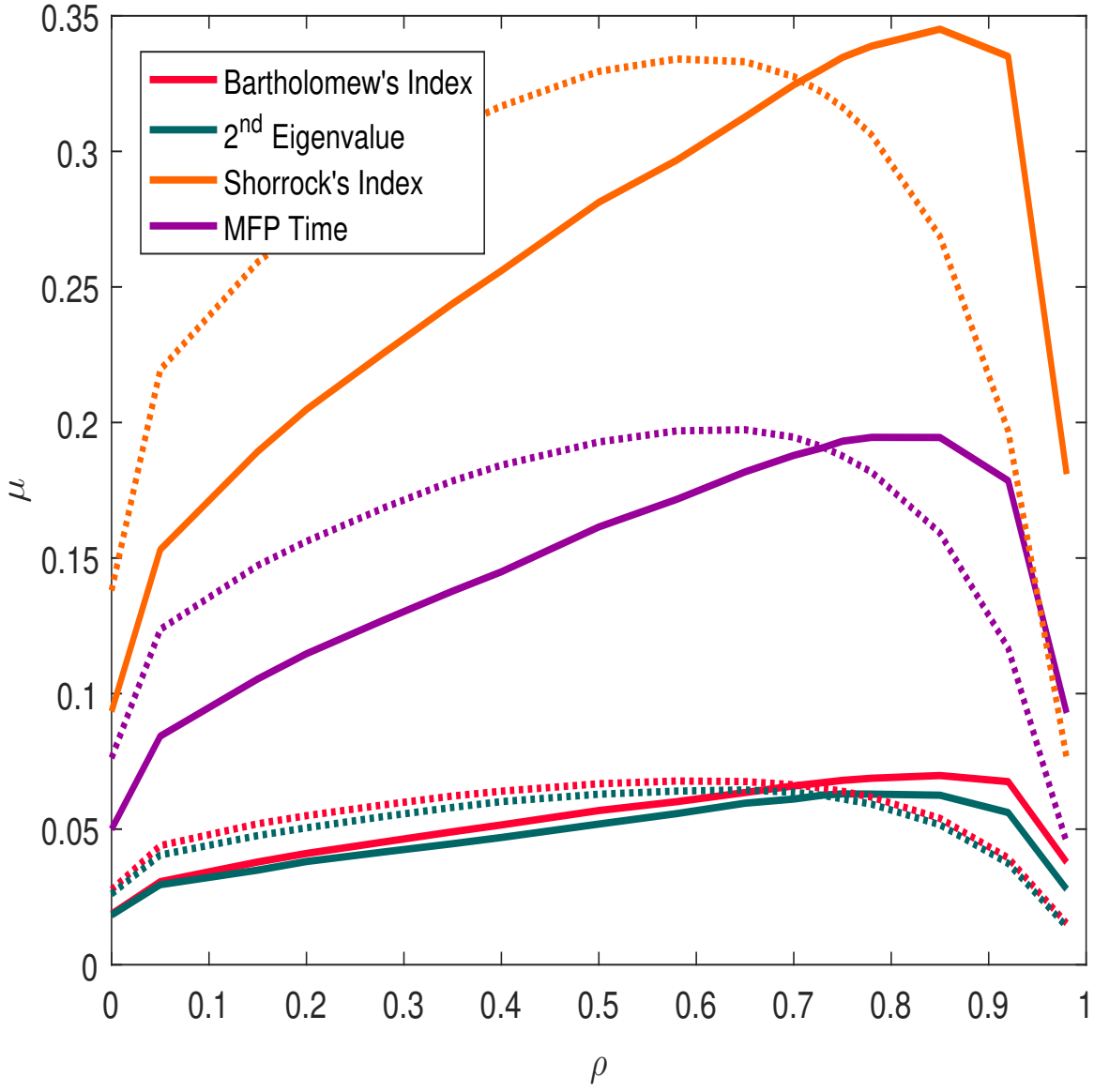


Figure 3.14: Portfolio decisions

Figure 3.15: Mobility across ρ; Incomplete markets vs. partial insurance



Solid - 2 assets; Dashed - 1 asset

Figure 3.16: *

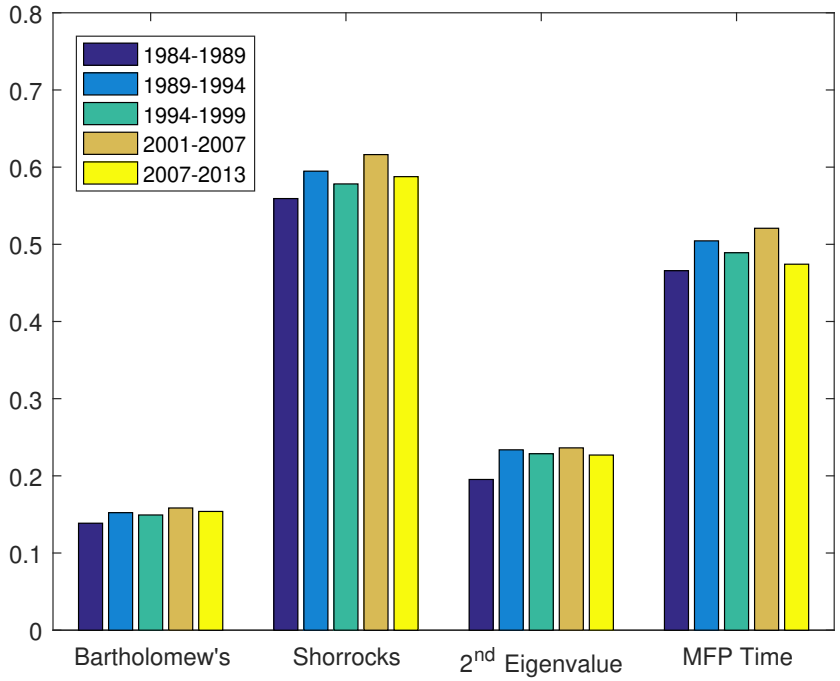


Figure 3.17: Short-Horizon Mobility

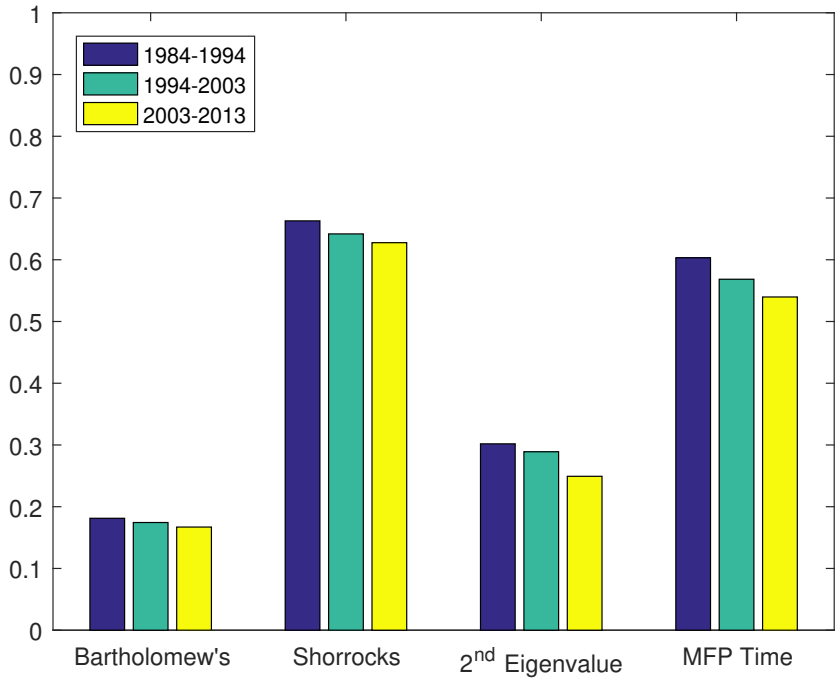


Figure 3.18: Medium-Horizon Mobility

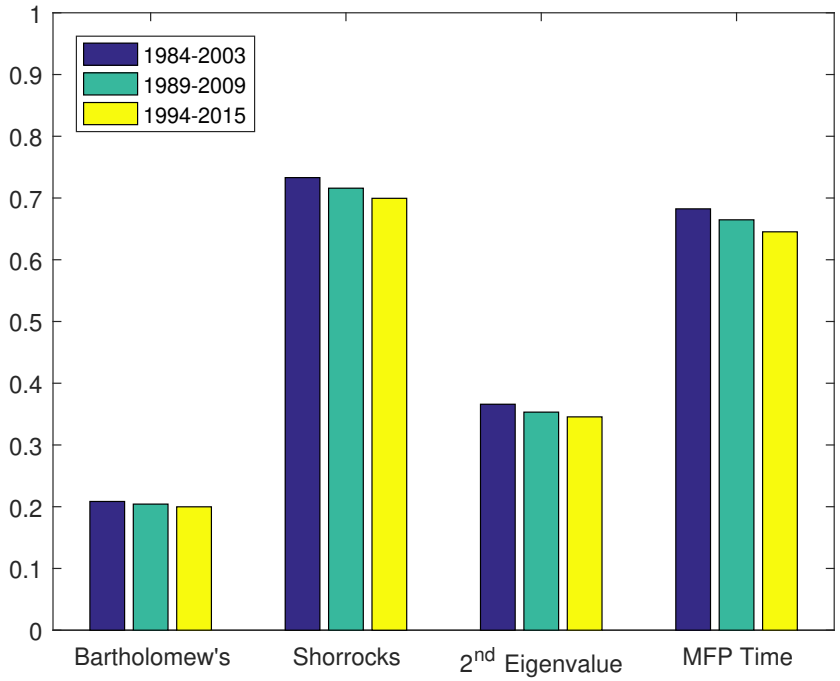


Figure 3.19: Long-Horizon Mobility

Figure 3.20: Bootstrapped Mobility Measures, Short Horizon

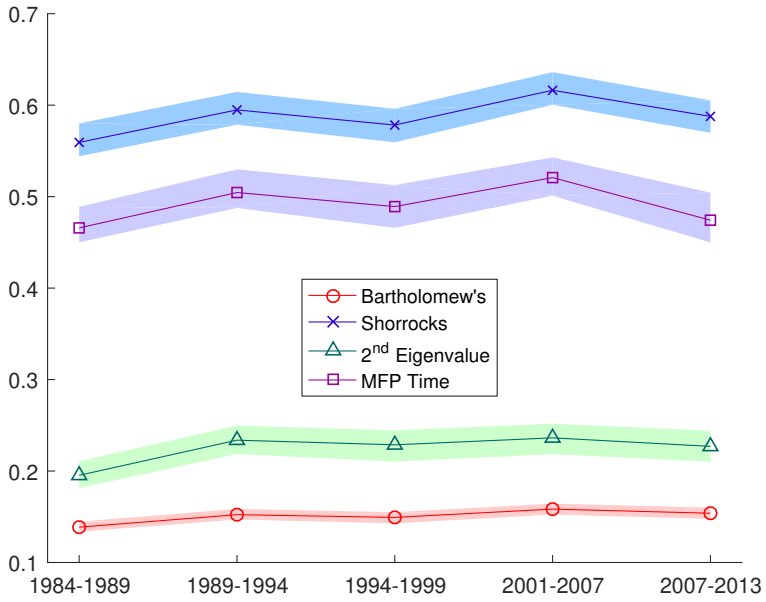


Figure 3.21: Bootstrapped Mobility Measures, Medium Horizon

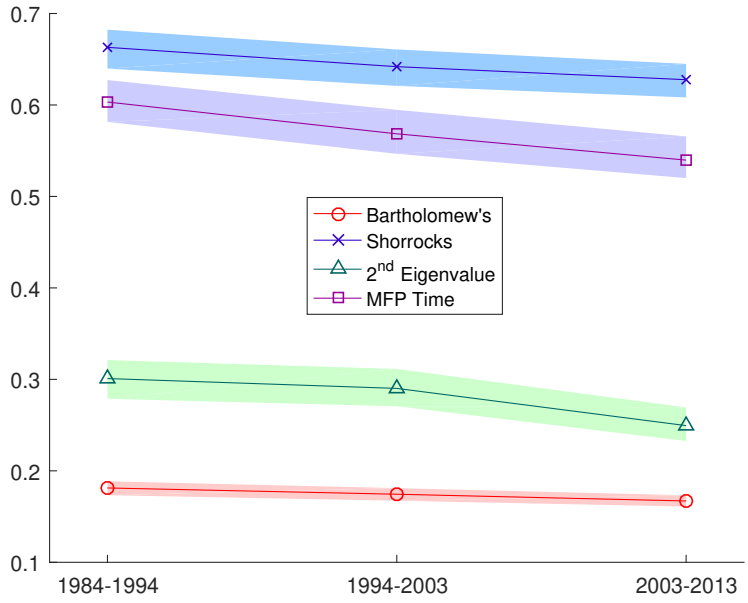


Figure 3.22: Bootstrapped Mobility Measures, Long Horizon

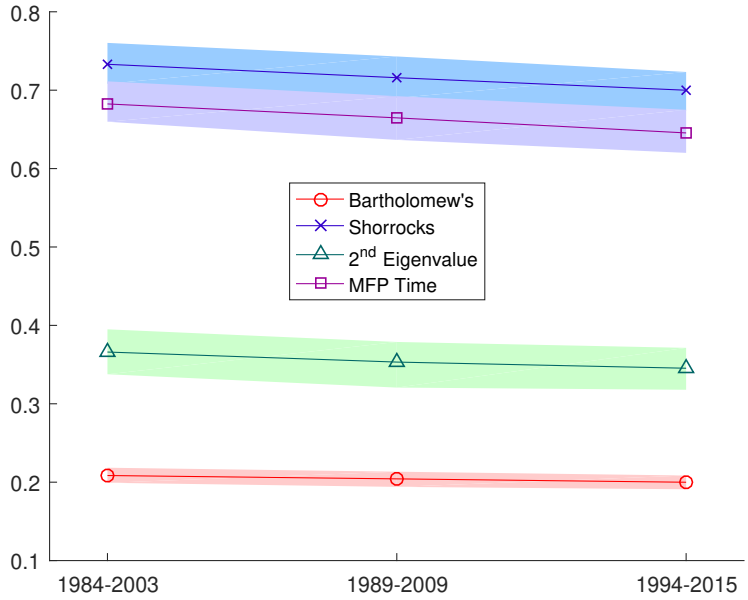


Figure 3.23: Decomposition: Short Horizon

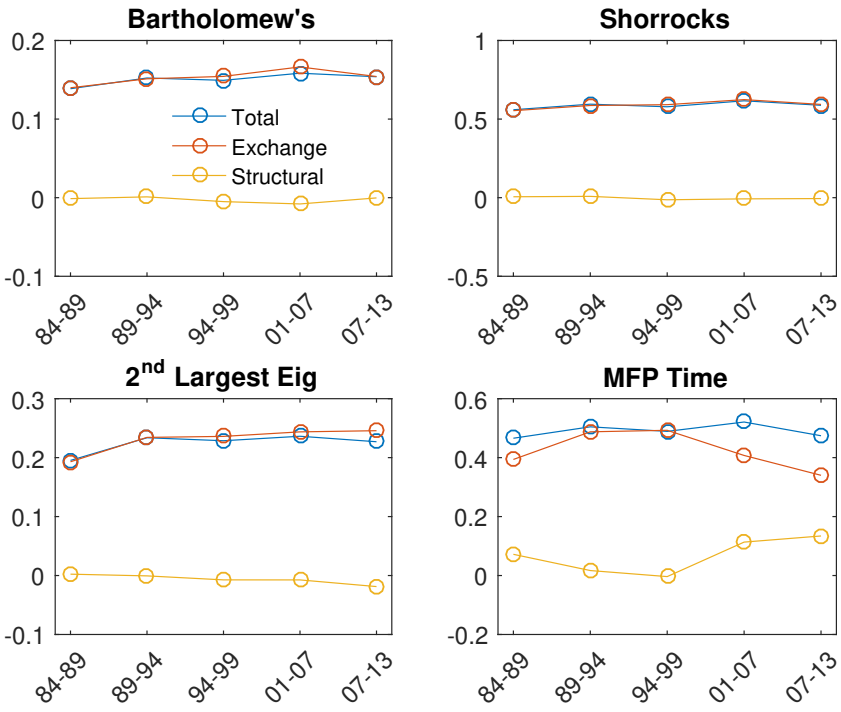


Figure 3.24: Decomposition: Medium Horizon

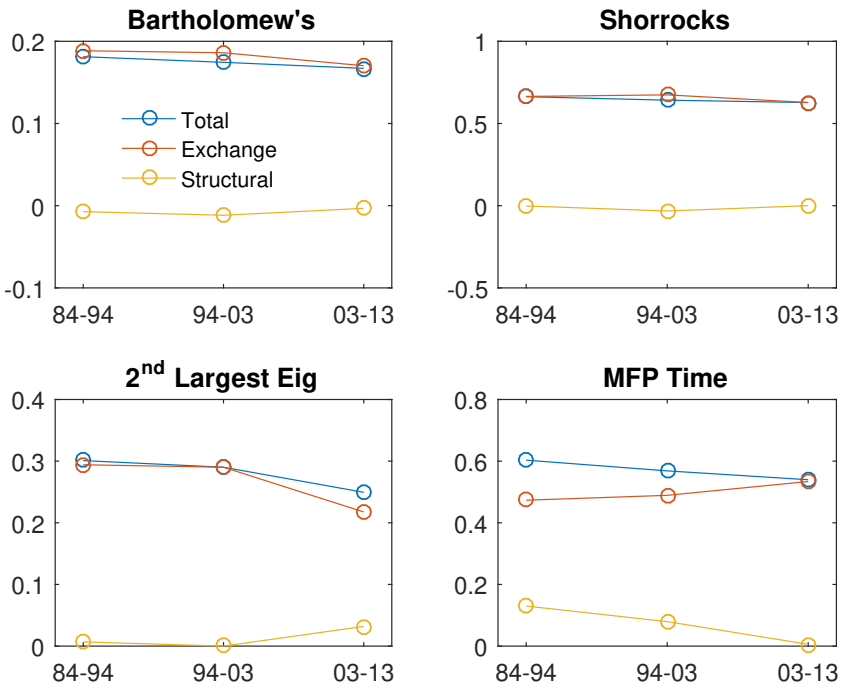
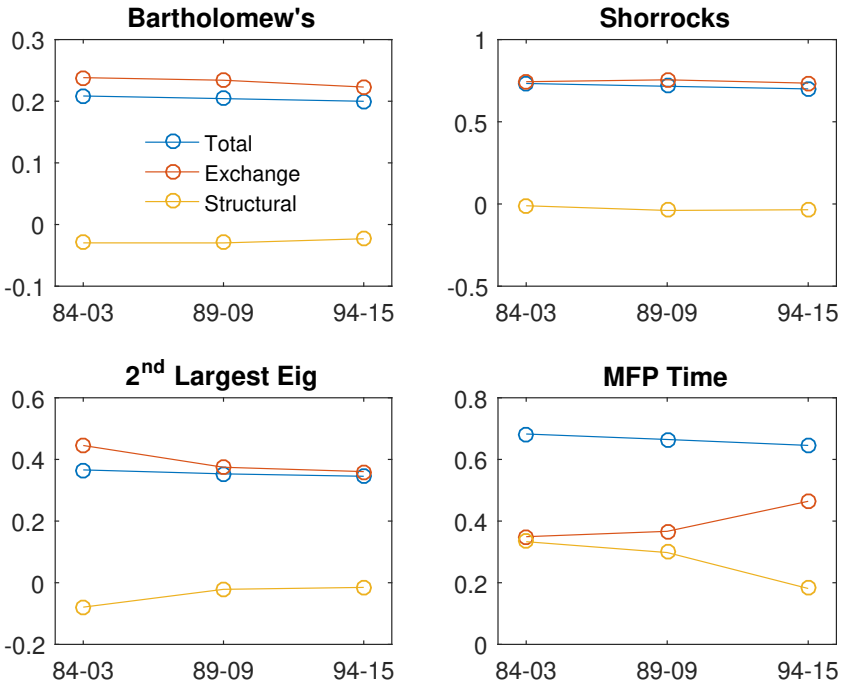


Figure 3.25: Decomposition: Long Horizon



3.10.6 Regression Results

Table 3.5: Movement Through Distribution (OLS)

	Dependent variable:		
	$\Delta p_{i,t}$		
	Short	Medium	Long
Married During Interval	0.018^{***}	0.061^{***}	0.080^{***}
	(0.004)	(0.007)	(0.011)
Divorced During Interval	-0.049^{***}	-0.058^{***}	-0.075^{***}
	(0.006)	(0.007)	(0.013)
Non-white (Head)	0.006^{*}	0.009^{*}	0.016
	(0.003)	(0.005)	(0.012)
Mean Income (Thousands)	0.0001^{***}	0.0001^{***}	0.0002^{***}
	(0.00001)	(0.00002)	(0.0001)
College Degree (Head)	0.022^{***}	0.034^{***}	0.058^{***}
	(0.003)	(0.004)	(0.009)
Owned Real Estate	-0.002	-0.006	0.001
	(0.003)	(0.004)	(0.009)
Owned a Farm or Business	-0.004	0.0003	-0.014
	(0.003)	(0.004)	(0.009)
Owned Stocks	0.012^{***}	0.014^{***}	0.027^{***}
	(0.003)	(0.004)	(0.010)
Received Inheritance	0.061^{***}	0.044^{***}	0.047^{***}
	(0.004)	(0.005)	(0.010)
Observations	27,334	13,798	3,219
R^2	0.049	0.088	0.201
Adjusted R^2	0.048	0.087	0.198
Residual Std. Error	0.836	0.943	1.087
F Statistic	93.426^{***}	101.387^{***}	67.040^{***}

Note: *p<0.1; **p<0.05; ***p<0.01
All dollar values in constant 2016 USD

Table 3.6: Upward Movements (Probit)

	Dependent variable:		
	$\Pr(\Delta q_{i,t} > 0 \mid X)$		
	Short	Medium	Long
Married During Interval	0.168^{***}	0.398^{***}	0.456^{***}
	(0.007)	(0.009)	(0.013)
Divorced During Interval	-0.165^{***}	-0.253^{***}	-0.323^{***}
	(0.010)	(0.010)	(0.016)
Non-white (Head)	0.042^{***}	0.047^{***}	-0.140^{***}
	(0.005)	(0.007)	(0.016)
Mean Income (Thousands)	-0.0002^{***}	-0.0001^{***}	-0.0002^{**}
	(0.00002)	(0.00003)	(0.0001)
College Degree (Head)	0.067^{***}	0.146^{***}	0.249^{***}
	(0.004)	(0.006)	(0.012)
Owned Real Estate	0.012^{***}	0.019^{***}	0.011
	(0.004)	(0.006)	(0.012)
Owned a Farm or Business	0.0003	0.018^{***}	-0.048^{***}
	(0.005)	(0.006)	(0.012)
Owned Stocks	0.091^{***}	0.110^{***}	0.195^{***}
	(0.004)	(0.006)	(0.013)
Received Inheritance	0.279^{***}	0.175^{***}	0.254^{***}
	(0.007)	(0.008)	(0.012)
Observations	27,334	13,798	3,219

Note: *p<0.1; **p<0.05; ***p<0.01
All dollar values in constant 2016 USD

Table 3.7: Downward Movements (Probit)

	Dependent variable:		
	$\Pr(\Delta q_{i,t} < 0 \mid X)$		
	Short	Medium	Long
Married During Interval	-0.014^{**}	-0.112^{***}	-0.347^{***}
	(0.007)	(0.010)	(0.015)
Divorced During Interval	0.334^{***}	0.330^{***}	0.408^{***}
	(0.009)	(0.010)	(0.016)
Non-white (Head)	-0.089^{***}	-0.090^{***}	-0.180^{***}
	(0.005)	(0.007)	(0.015)
Mean Income (Thousands)	-0.002^{***}	-0.002^{***}	-0.003^{***}
	(0.00004)	(0.0001)	(0.0001)
College Degree (Head)	-0.122^{***}	-0.128^{***}	-0.237^{***}
	(0.005)	(0.006)	(0.012)
Owned Real Estate	0.022^{***}	0.061^{***}	0.092^{***}
	(0.004)	(0.006)	(0.011)
Owned a Farm or Business	0.039^{***}	0.042^{***}	0.157^{***}
	(0.005)	(0.006)	(0.012)
Owned Stocks	-0.094^{***}	-0.072^{***}	-0.016
	(0.004)	(0.006)	(0.012)
Received Inheritance	-0.387^{***}	-0.287^{***}	-0.228^{***}
	(0.008)	(0.008)	(0.012)
Observations	27,334	13,798	3,219

Note: *p<0.1; **p<0.05; ***p<0.01
All dollar values in constant 2016 USD

Table 3.8: Upward Jumps (Probit)

	Dependent variable:		
	$\Pr(\Delta q_{i,t} \geq 2 \mid X)$		
	Short	Medium	Long
Married During Interval	0.195^{***}	0.435^{***}	0.280^{***}
	(0.010)	(0.011)	(0.016)
Divorced During Interval	-0.053^{***}	-0.121^{***}	-0.137^{***}
	(0.014)	(0.014)	(0.019)
Non-white (Head)	0.129^{***}	0.083^{***}	-0.161^{***}
	(0.007)	(0.010)	(0.022)
Mean Income (Thousands)	-0.00005	0.0001	-0.0001
	(0.00003)	(0.00004)	(0.0001)
College Degree (Head)	0.139^{***}	0.189^{***}	0.283^{***}
	(0.007)	(0.008)	(0.015)
Owned Real Estate	0.107^{***}	0.056^{***}	0.088^{***}
	(0.007)	(0.008)	(0.015)
Owned a Farm or Business	0.182^{***}	0.235^{***}	0.084^{***}
	(0.007)	(0.008)	(0.015)
Owned Stocks	0.005	0.075^{***}	0.268^{***}
	(0.007)	(0.008)	(0.017)
Received Inheritance	0.354^{***}	0.229^{***}	0.137^{***}
	(0.010)	(0.010)	(0.015)
Observations	27,334	13,798	3,219

Note: *p<0.1; **p<0.05; ***p<0.01
All dollar values in constant 2016 USD

Table 3.9: Downward Jumps (Probit)

	Dependent variable:		
	$\Pr(\Delta q_{i,t} \leq -2 \mid X)$		
	Short	Medium	Long
Married During Interval	0.011	-0.085^{***}	-0.302^{***}
	(0.010)	(0.015)	(0.021)
Divorced During Interval	0.299^{***}	0.412^{***}	0.470^{***}
	(0.013)	(0.013)	(0.021)
Non-white (Head)	0.074^{***}	-0.041^{***}	-0.153^{***}
	(0.007)	(0.010)	(0.021)
Mean Income (Thousands)	-0.002^{***}	-0.002^{***}	-0.002^{***}
	(0.0001)	(0.0001)	(0.0002)
College Degree (Head)	-0.090^{***}	-0.117^{***}	-0.226^{***}
	(0.007)	(0.009)	(0.017)
Owned Real Estate	0.033^{***}	0.141^{***}	-0.010
	(0.007)	(0.008)	(0.015)
Owned a Farm or Business	0.214^{***}	0.178^{***}	0.309^{***}
	(0.007)	(0.009)	(0.015)
Owned Stocks	-0.019^{***}	-0.089^{***}	-0.015
	(0.007)	(0.008)	(0.016)
Received Inheritance	-0.358^{***}	-0.252^{***}	-0.140^{***}
	(0.014)	(0.013)	(0.017)
Observations	27,334	13,798	3,219

Note: *p<0.1; **p<0.05; ***p<0.01
All dollar values in constant 2016 USD

3.10.7 Directional Mobility and Demographics

3.10.8 A Measure of Directional Mobility

We find that splitting our panels along demographic groups–particularly, along educational attainment and race–reveals notable differences in mobility. All of our measures treat upward and downward mobility identically, so for the purposes of this section, we consider the following measure of directional mobility: for an $n \times n$ matrix Π, we define the measure $\mathbf{x} = (x_d, x_u)$, where

$$x_d = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{i-1} \Pi_{i,j} \, |i-j|$$

and

$$x_u = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Pi_{i,j} \, |i-j|$$

This measure is similar to the Bartholomew measure, but it is restricted to the entries below (above) the diagonal for downward (upward) mobility.

For the purposes of exposition, we will refer to the non-college and nonwhite groups as the "disadvantaged" group. We also note that, although our measure of directional mobility places greater weight on larger movements, the difference in upward and downward mobility across groups is not driven by a few large movements. Irrespective of the horizon or specific time period, for the disadvantaged group nearly every below (above) diagonal element of the transition matrix is greater (less) than the corresponding element in the transition matrix for the advantaged group.

3.10.9 Differences in Mobility Across Demographics

Education

First, we divide our sample into families wherein the head had a college degree at the start of the sample period, and those who did not. We then use the same method as before to construct mobility matrices that capture the wealth transitions of families in each subsample.

Evidence from the PSID data suggests that families whose heads have a college degree experience higher upward mobility. As an example, consider the following two matrices from 2003-2013:

These matrices show that families with a college-educated head experience higher upward mobility, lower persistence in the lower wealth quintiles, and higher persistence in the upper wealth quintiles. Families in this subsample have a good chance of getting to high levels

$$\Pi_C^{03-13} = \begin{bmatrix} 0.435 & 0.261 & 0.179 & 0.103 & 0.022 \\ 0.245 & 0.328 & 0.250 & 0.130 & 0.047 \\ 0.114 & 0.140 & 0.327 & 0.301 & 0.118 \\ 0.041 & 0.031 & 0.168 & 0.430 & 0.330 \\ 0.008 & 0.003 & 0.036 & 0.180 & 0.774 \end{bmatrix} \Pi_{NC}^{03-13} = \begin{bmatrix} 0.592 & 0.294 & 0.092 & 0.019 & 0.003 \\ 0.273 & 0.428 & 0.227 & 0.059 & 0.014 \\ 0.154 & 0.240 & 0.408 & 0.168 & 0.030 \\ 0.064 & 0.101 & 0.229 & 0.453 & 0.153 \\ 0.042 & 0.039 & 0.064 & 0.293 & 0.563 \end{bmatrix}$$

of wealth, and when they do so, they tend to stay there. By contrast, families wherein the head did not have a college degree experienced higher downward mobility, higher persistence in lower quintiles, and lower persistence in upper quintiles. These patterns are even more pronounced at long horizon (see Table 3.10). These results suggest that the mobility matrices over longer time periods may be built by two distinct groups: college-educated families making larger contributions to the above-diagonal elements, and the families without a college-educated making larger contributions to the below-diagonal elements.

Race

Splitting our sample by race yields similar results. Below, we report the ten-year wealth transition matrices over the period 2003-2013 for households with white (Π_W) and nonwhite (Π_{NW}) families:

$$\Pi_W^{03-13} = \begin{bmatrix} 0.480 & 0.314 & 0.137 & 0.055 & 0.014 \\ 0.232 & 0.397 & 0.245 & 0.093 & 0.033 \\ 0.118 & 0.167 & 0.390 & 0.254 & 0.071 \\ 0.033 & 0.062 & 0.188 & 0.473 & 0.244 \\ 0.014 & 0.010 & 0.045 & 0.223 & 0.709 \end{bmatrix}$$

$$\Pi_{NW}^{03-13} = \begin{bmatrix} 0.624 & 0.274 & 0.084 & 0.016 & 0.001 \\ 0.299 & 0.424 & 0.218 & 0.051 & 0.008 \\ 0.183 & 0.284 & 0.382 & 0.126 & 0.025 \\ 0.136 & 0.124 & 0.271 & 0.345 & 0.124 \\ 0.090 & 0.090 & 0.077 & 0.282 & 0.462 \end{bmatrix}$$

Here we see that families with a white head experience high levels of upward mobility, low persistence in lower quintiles, and high persistence in upper quintiles. Nonwhite families, by contrast, are more likely to make downward movements, and experience relatively higher persistence in low quintiles, and lower persistence in higher quintiles. We can see, for example, that a nonwhite family who started off in the first quintile had about a one in ten

chance of reaching one of the top three quintiles, compared to the roughly one in five chance faced by a white family at making the same transition from the same starting point. Once again, this is a pattern that holds true at a long horizon (Table 3.11).

Table 3.10: Twenty-Year College Breakdown

$$\Pi_{C}^{89-09}: \begin{bmatrix} 0.289 & 0.211 & 0.244 & 0.122 & 0.133 \\ 0.167 & 0.188 & 0.333 & 0.210 & 0.101 \\ 0.059 & 0.129 & 0.294 & 0.329 & 0.188 \\ 0.049 & 0.079 & 0.238 & 0.311 & 0.323 \\ 0.019 & 0.023 & 0.047 & 0.211 & 0.700 \end{bmatrix}$$

$$\Pi_{NC}^{89-09}: \begin{bmatrix} 0.600 & 0.286 & 0.081 & 0.022 & 0.010 \\ 0.298 & 0.422 & 0.155 & 0.088 & 0.038 \\ 0.144 & 0.312 & 0.293 & 0.166 & 0.086 \\ 0.099 & 0.131 & 0.314 & 0.325 & 0.131 \\ 0.011 & 0.082 & 0.142 & 0.290 & 0.475 \end{bmatrix}$$

Table 3.11: Twenty-Year Race Breakdown

$$\Pi_{W}^{89-09}: \begin{bmatrix} 0.402 & 0.303 & 0.176 & 0.077 & 0.042 \\ 0.204 & 0.327 & 0.229 & 0.163 & 0.076 \\ 0.096 & 0.192 & 0.295 & 0.268 & 0.149 \\ 0.065 & 0.093 & 0.290 & 0.334 & 0.218 \\ 0.013 & 0.045 & 0.092 & 0.247 & 0.603 \end{bmatrix}$$

$$\Pi_{NW}^{89-09}: \begin{bmatrix} 0.657 & 0.258 & 0.058 & 0.010 & 0.017 \\ 0.371 & 0.433 & 0.143 & 0.040 & 0.013 \\ 0.174 & 0.428 & 0.290 & 0.072 & 0.036 \\ 0.192 & 0.250 & 0.250 & 0.212 & 0.096 \\ 0.059 & 0.176 & 0.059 & 0.235 & 0.471 \end{bmatrix}$$

Figure 3.26: Calibration: σ_z across ρ_z for fixed Gini

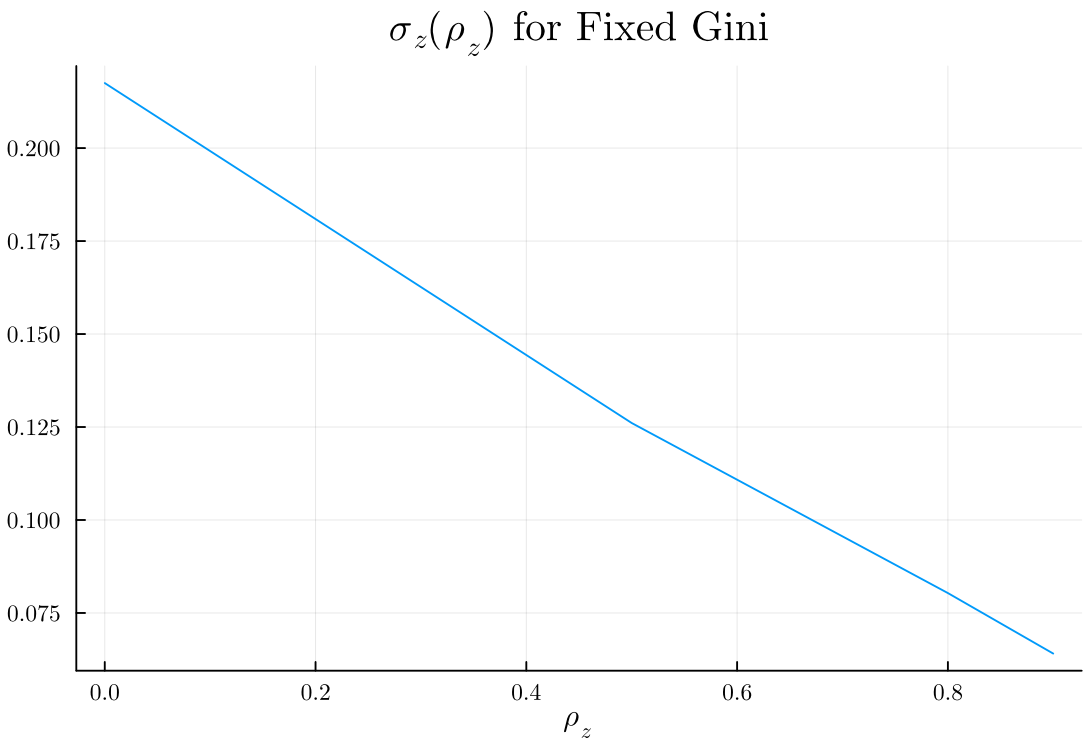


Figure 3.27: Mobility Measures Across Return Persistence ρ_z

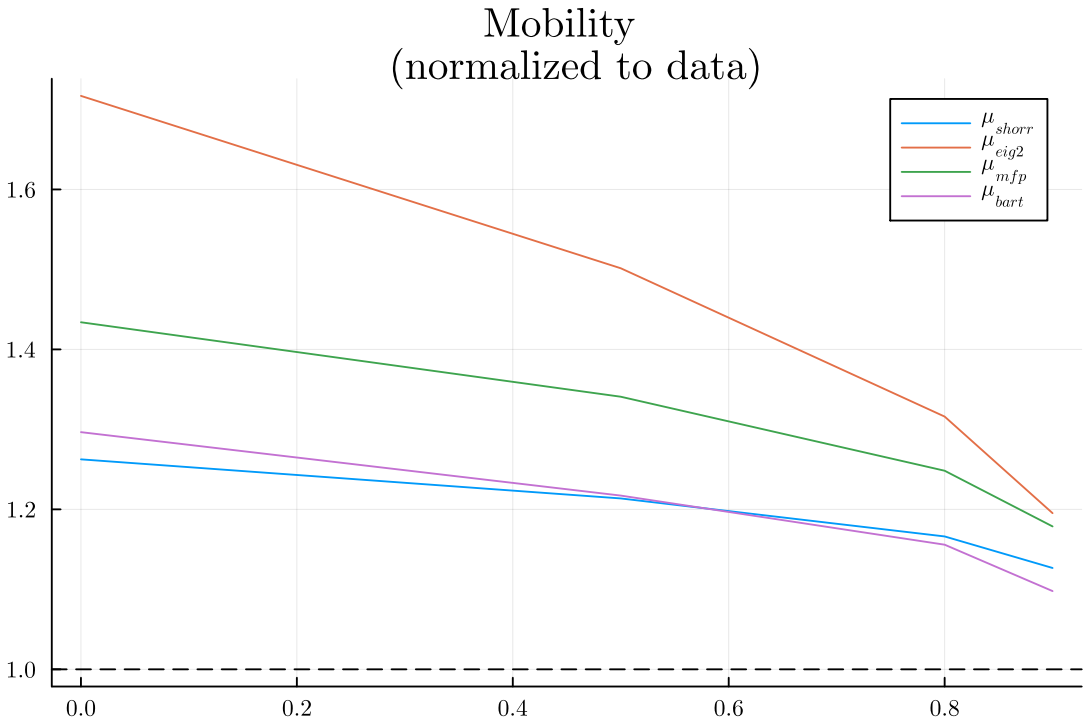


Figure 3.28: Top Wealth Shares Across Return Persistence ρ_z

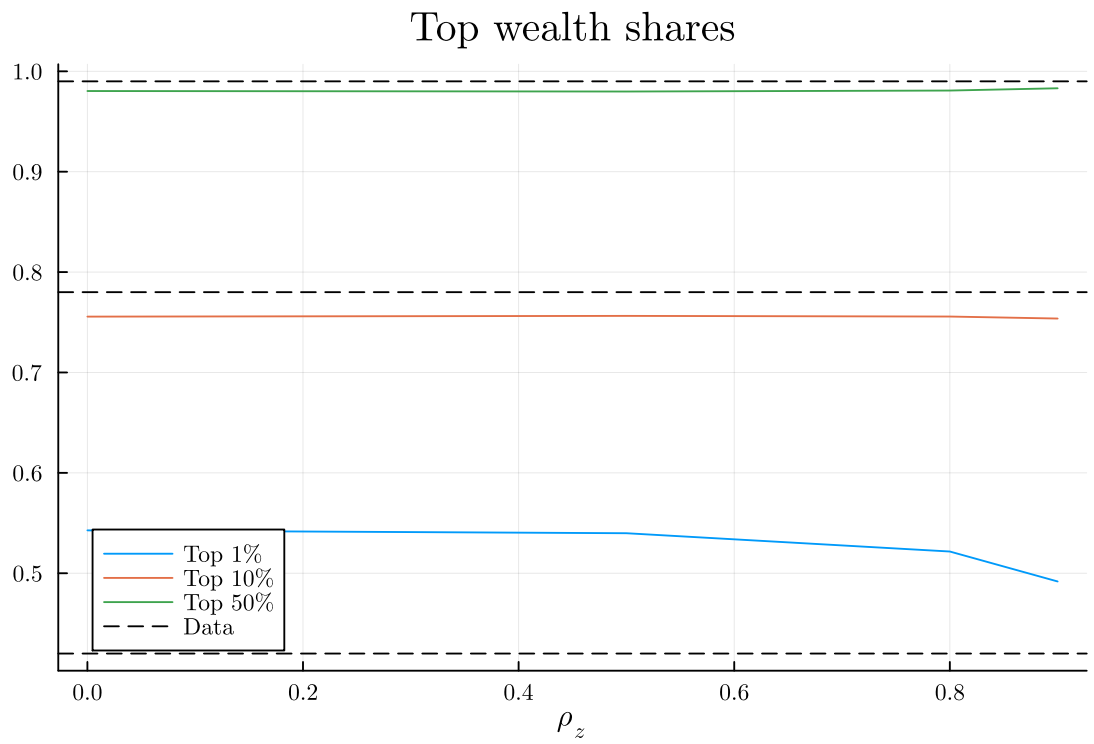
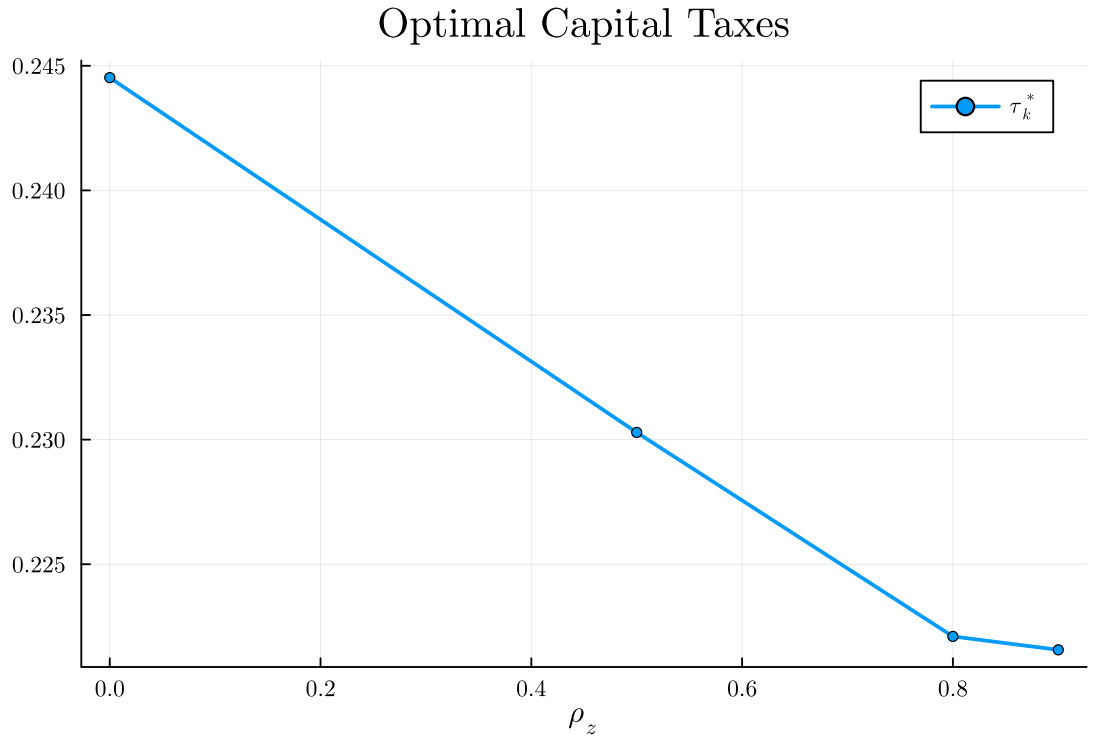


Figure 3.29: Optimal Taxes Across Return Persistence ρ_z



Bibliography

Daron Acemoglu. *Introduction to modern economic growth*. Princeton University Press, 2009.

Sushant Acharya, Edouard Challe, and Keshav Dogra. Optimal monetary policy according to hank. 2020.

S Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics*, 109(3):659–684, 1994a. doi: 10.2307/2118417. URL http://qje.oxfordjournals.org/content/109/3/659.abstract.

S Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. *The Quarterly Journal of Economics*, 109(3):659–684, 1994b.

Stefania Albanesi. Optimal taxation of entrepreneurial capital with private information, 2006.

Stefania Albanesi and Christopher Sleet. Dynamic optimal taxation with private information. *The Review of Economic Studies*, 73(1):1–30, 2006.

George-Marios Angeletos. Uninsured idiosyncratic investment risk and aggregate saving. *Review of Economic dynamics*, 10(1):1–30, 2007.

Anthony Barnes Atkinson and Joseph E Stiglitz. The design of tax structure: direct versus indirect taxation. *Journal of public Economics*, 6(1-2):55–75, 1976.

Adrien Auclert. Monetary policy and the redistribution channel. *American Economic Review*, 109(6):2333–67, 2019.

David Baqaee, Emmanuel Farhi, and Kunal Sangani. The supply-side effects of monetary policy. Technical report, National Bureau of Economic Research, 2021.

David J Bartholomew and David J Bartholomew. *Stochastic models for social processes*. Wiley London, 1967.

Jess Benhabib, Alberto Bisin, and Shenghao Zhu. The distribution of wealth and fiscal policy in economies with finitely lived agents. *Econometrica*, 79(1):123–157, 2011.

Jess Benhabib, Alberto Bisin, and Mi Luo. Wealth distribution and social mobility in the us: a quantitative approach. Technical report, National Bureau of Economic Research, 2015a.

Jess Benhabib, Alberto Bisin, and Shenghao Zhu. The wealth distribution in bewley economies with capital income risk. *Journal of Economic Theory*, 159:489–515, 2015b.

Jess Benhabib, Alberto Bisin, and Mi Luo. Wealth distribution and social mobility in the us: A quantitative approach. *American Economic Review*, 109(5):1623–47, 2019.

Ben S Bernanke, Mark Gertler, and Simon Gilchrist. The financial accelerator in a quantitative business cycle framework. *Handbook of macroeconomics*, 1:1341–1393, 1999.

Truman F Bewley. Stationary monetary equilibrium with a continuum of independently fluctuating consumers. In Werner Hildenbrand and Andreu Mas-Collel, editors, *Contributions to mathematical economics in honor of Gerard Debreu*, pages 79–102. North Holland, 1986.

Florin Bilbiie. Monetary Policy and Heterogeneity. 2021.

Florin O. Bilbiie. The New Keynesian cross. *Journal of Monetary Economics*, 114:90–108, October 2020. ISSN 03043932. doi: 10.1016/j.jmoneco.2019.03.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0304393219300492.

Florin O Bilbiie, Diego R Känzig, and Paolo Surico. Capital and Income Inequality: an Aggregate-Demand Complementarity. 2021.

Jean Boivin and Marc P Giannoni. Assessing changes in the monetary transmission mechanism: A var approach. *Economic Policy Review*, 8(1), 2002.

Jean Boivin and Marc P Giannoni. Has monetary policy become more effective? *The Review of Economics and Statistics*, 88(3):445–462, 2006.

Jean Boivin, Michael T Kiley, and Frederic S Mishkin. How has the monetary transmission mechanism evolved over time? In *Handbook of monetary economics*, volume 3, pages 369–422. Elsevier, 2010.

Francisco J Buera and Benjamin Moll. Aggregate implications of a credit crunch: The importance of heterogeneity. *American Economic Journal: Macroeconomics*, 7(3):1–42, 2015.

Francisco J Buera, Joseph P Kaboski, and Yongseok Shin. Finance and development: A tale of two sectors. *American economic review*, 101(5):1964–2002, 2011.

Marco Cagetti and Mariacristina De Nardi. Entrepreneurship, frictions, and wealth. *Journal of political Economy*, 114(5):835–870, 2006.

Fabio Canova and Luca Gambetti. Structural changes in the us economy: Is there a role for monetary policy? *Journal of Economic dynamics and control*, 33(2):477–490, 2009.

Charles T Carlstrom and Timothy S Fuerst. Agency costs, net worth, and business fluctuations: A computable general equilibrium analysis. *The American Economic Review*, pages 893–910, 1997.

Charles T Carlstrom and Timothy S Fuerst. Monetary policy in a world without perfect capital markets. Technical report, 2001.

Christopher D Carroll. Why do the rich save so much? Technical report, National Bureau of Economic Research, 1998.

Christopher D Carroll. A theory of the consumption function, with and without liquidity constraints. *The Journal of Economic Perspectives*, 15(3):23–45, 2001.

Daniel R Carroll and Eric R Young. The long run effects of changes in tax progressivity. *Journal of Economic Dynamics and Control*, 35(9):1451–1473, 2011.

Daniel R Carroll, Jim Dolmas, and Eric R Young. The politics of flat taxes. Technical report, Federal Reserve Bank of Cleveland, 2016.

Francesco Caselli and Jaume Ventura. A representative consumer theory of distribution. *American Economic Review*, 90(4):909–926, 2000.

Ana Castaneda, Javier Díaz-Giménez, and José-Víctor Ríos-Rull. Accounting for the us earnings and wealth inequality. *Journal of Political Economy*, 111(4):818–857, 2003.

Christophe Chamley. Optimal taxation of capital income in general equilibrium with infinite lives. *Econometrica: Journal of the Econometric Society*, pages 607–622, 1986.

Lawrence J Christiano, Martin Eichenbaum, and Charles L Evans. Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of political Economy*, 113(1): 1–45, 2005.

James Cloyne, Clodomiro Ferreira, and Paolo Surico. Monetary policy when households have debt: new evidence on the transmission mechanism. *The Review of Economic Studies*, 87 (1):102–129, 2020.

Andrea Colciago, Anna Samarina, and Jakob de Haan. Central bank policies and income and wealth inequality: A survey. *Journal of Economic Surveys*, 33(4):1199–1231, 2019.

John Conlisk. Monotone mobility matrices. *Journal of Mathematical Sociology*, 15(3-4): 173–191, 1990.

Frank A Cowell and Emmanuel Flachaire. Measuring mobility. *Quantitative Economics*, 9 (2):865–901, 2018.

Valentino Dardanoni. Measuring social mobility. *Journal of Economic Theory*, 61(2):372–394, 1993.

Jason DeBacker, Vasia Panousi, and Shanthi Ramnath. A risky venture: Income dynamics among pass-through business owners. *American Economic Journal: Macroeconomics*, 15 (1):444–474, 2023.

Marco Del Negro, Michele Lenza, Giorgio E Primiceri, and Andrea Tambalotti. What's up with the phillips curve? Technical report, National Bureau of Economic Research, 2020.

Javier Díaz-Giménez, Andy Glover, and José-Víctor Ríos-Rull. Facts on the distributions of earnings, income, and wealth in the united states: 2007 update. *Federal Reserve Bank of Minneapolis Quarterly Review*, 34(1):2–31, 2011.

Charles L Evans. Productivity shocks and real business cycles. *Journal of Monetary Economics*, 29(2):191–208, 1992.

Andreas Fagereng, Luigi Guiso, Davide Malacrino, and Luigi Pistaferri. Heterogeneity and persistence in returns to wealth. *Econometrica*, 88(1):115–170, 2020.

Emmanuel Farhi and Iván Werning. Insurance and taxation over the life cycle. *Review of Economic Studies*, 80(2):596–635, 2013.

Ethan Feilich. Monetary policy and the dynamics of wealth inequality. 2021.

Gary S Fields and Efe A Ok. The measurement of income mobility: an introduction to the literature. In *Handbook of Income Inequality Measurement*, pages 557–598. Springer, 1999.

Martin Floden and Jesper Lindé. Idiosyncratic risk in the united states and sweden: is there a role for government insurance? *Review of Economic Dynamics*, 4(2):406–437, 2001.

Martin Flodén, Matilda Kilström, Jósef Sigurdsson, and Roine Vestman. Household debt and monetary policy: Revealing the cash-flow channel. *The Economic Journal*, 131(636): 1742–1771, 2021.

Xavier Gabaix. Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294, 2009.

Jordi Galí. *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press, 2015.

Mikhail Golosov, Narayana Kocherlakota, and Aleh Tsyvinski. Optimal indirect and capital taxation. *The Review of Economic Studies*, 70(3):569–587, 2003.

Beatriz González, Galo Nuño, Dominik Thaler, and Silvia Albrizio. Firm heterogeneity, capital misallocation and optimal monetary policy. 2024.

Emilien Gouin-Bonenfant and Alexis Akira Toda. Pareto extrapolation: An analytical framework for studying tail inequality. *Available at SSRN 3260899*, 2019.

Jeremy Greenwood, Zvi Hercowitz, and Gregory W Huffman. Investment, capacity utilization, and the real business cycle. *The American Economic Review*, pages 402–417, 1988.

F Guvenen, G Kambourov, B Kuruscu, and D Chen. Use it or lose it: Efficiency gains from wealth taxation. Technical report, Working paper, 2015a.

Fatih Guvenen, Fatih Karahan, Serdar Ozkan, and Jae Song. What do data on millions of us workers reveal about life-cycle earnings risk? Technical report, National Bureau of Economic Research, 2015b.

Fatih Guvenen, Gueorgui Kambourov, Burhanettin Kuruscu, Sergio Ocampo-Diaz, and Daphne Chen. Use it or lose it: Efficiency gains from wealth taxation. Technical report, National Bureau of Economic Research, 2019.

Jonathon Hazell, Juan Herreno, Emi Nakamura, and Jón Steinsson. The slope of the phillips curve: evidence from us states. *The Quarterly Journal of Economics*, 137(3):1299–1344, 2022.

Bengt Holmström. Moral hazard and observability. *The Bell journal of economics*, pages 74–91, 1979.

Hugo A Hopenhayn. Firms, misallocation, and aggregate productivity: A review. *Annu. Rev. Econ.*, 6(1):735–770, 2014.

Hugo A Hopenhayn and Edward C Prescott. Stochastic monotonicity and stationary distributions for dynamic economies. *Econometrica*, pages 1387–1406, 1992.

Joachim Hubmer, Per Krusell, Anthony A Smith, et al. The historical evolution of the wealth distribution: A quantitative-theoretic investigation. Technical report, National Bureau of Economic Research, 2017.

Mark Huggett. The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control*, 17(5):953–969, 1993.

Erik Hurst, Ming Ching Luoh, Frank P Stafford, and William G Gale. The wealth dynamics of american families, 1984-94. *Brookings Papers on Economic Activity*, 1998(1):267–337, 1998.

Priit Jeenas. *Firm balance sheet liquidity, monetary policy shocks, and investment dynamics*. Universitat Pompeu Fabra, Department of Economics and Business, 2023.

Ian Jewitt. Justifying the first-order approach to principal-agent problems. *Econometrica: Journal of the Econometric Society*, pages 1177–1190, 1988.

Nancy A Jianakoplos and Paul L Menchik. Wealth mobility. *Review of Economics and Statistics*, 79(1):18–31, 1997.

Kenneth L Judd. Redistributive taxation in a simple perfect foresight model. Technical report, Discussion Paper, 1982.

Greg Kaplan, Benjamin Moll, and Giovanni L Violante. Monetary policy according to hank. *American Economic Review*, 108(3):697–743, 2018.

Rohan Kekre and Moritz Lenel. Monetary policy, redistribution, and risk premia. *Econometrica*, 90(5):2249–2282, 2022.

John G Kemeny and James Laurie Snell. Finite markov chains, undergraduate texts in mathematics. 1976.

Youngju Kim and Hyunjoon Lim. Transmission of monetary policy in times of high household debt. *Journal of Macroeconomics*, 63, 2020.

Nobuhiro Kiyotaki. Credit and business cycles. *The Japanese Economic Review*, 49(1): 18–35, 1998.

Narayana R Kocherlakota. Zero expected wealth taxes: A mirrlees approach to dynamic optimal taxation. *Econometrica*, 73(5):1587–1621, 2005.

Karen A Kopecky and Richard MH Suen. Finite state markov-chain approximations to highly persistent processes. *Review of Economic Dynamics*, 13(3):701–714, 2010.

Per Krusell and Anthony A Smith, Jr. Income and wealth heterogeneity in the macroeconomy. *Journal of Political Economy*, 106(5):867–896, 1998.

Robert E Lucas. On the size distribution of business firms. *The Bell Journal of Economics*, pages 508–523, 1978.

Alexander Matusche and Johannes Wacks. Monetary policy and wealth inequality: The role of entrepreneurs. 2021.

Robert McClelland and Shannon Mok. A review of recent research on labor supply elasticities. 2012.

Cara McDaniel. Average tax rates on consumption, investment, labor and capital in the oecd 1950-2003. *Manuscript, Arizona State University*, 19602004, 2007.

Alisdair McKay, Emi Nakamura, and Jón Steinsson. The power of forward guidance revisited. *American Economic Review*, 106(10):3133–3158, 2016.

Aaron Medlin. Federal reserve monetary policy and wealth inequality: An instrumental-variable local projections approach. *Available at SSRN 4650805*, 2023.

Davide Melcangi and Vincent Sterk. Stock market participation, inequality, and monetary policy. *Review of Economic Studies*, page rdae068, 2024.

Carl D Meyer. An alternative expression for the mean first passage matrix. *Linear algebra and its applications*, 22:41–47, 1978.

James A Mirrlees. An exploration in the theory of optimum income taxation. *The review of economic studies*, 38(2):175–208, 1971.

James A Mirrlees. The theory of moral hazard and unobservable behaviour: Part i. *The Review of Economic Studies*, 66(1):3–21, 1999.

Benjamin Moll. Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review*, 104(10):3186–3221, 2014.

Tobias J Moskowitz and Annette Vissing-Jørgensen. The returns to entrepreneurial investment: A private equity premium puzzle? *American Economic Review*, 92(4):745–778, 2002.

Pablo Ottonello and Thomas Winberry. Financial heterogeneity and the investment channel of monetary policy. *Econometrica*, 88(6):2473–2502, 2020.

Thomas Phelan. The optimal taxation of business owners. *Federal Reserve Bank of Cleveland Working Paper Series*, 2021.

SJj Prais. Measuring social mobility. *Journal of the Royal Statistical Society. Series A (General)*, 118(1):56–66, 1955.

Thomas Pugh. Wealth and mobility: superstars, returns heterogeneity and discount factors. Technical report, Tech. rep., Mimeo, September. 1, 2018.

Vincenzo Quadrini. The importance of entrepreneurship for wealth concentration and mobility. *Review of Income and Wealth*, 45(1):1–19, 1999.

Vincenzo Quadrini. Entrepreneurship, saving, and social mobility. *Review of Economic Dynamics*, 3(1):1–40, 2000a.

Vincenzo Quadrini. Entrepreneurship, saving, and social mobility. *Review of economic dynamics*, 3(1):1–40, 2000b.

Vincenzo Quadrini and José-Víctor Ríos-Rull. Understanding the us distribution of wealth. *Quarterly Review Federal Reserve Bank of Minneapolis*, 21(2):22, 1997.

Adriano A Rampini and S Viswanathan. Household risk management. Technical report, National Bureau of Economic Research, 2016.

José-Vıctor Ríos-Rull. Computing equilibria in models with heterogenous agents. *Computational Methods for the Study of Dynamic Economics*, pages 238–264, 1998.

Julio J Rotemberg. Monopolistic price adjustment and aggregate output. *The Review of Economic Studies*, 49(4):517–531, 1982.

Florian Scheuer. Entrepreneurial taxation with endogenous entry. *American Economic Journal: Economic Policy*, 6(2):126–63, 2014.

Frank Schorfheide. Dsge model-based estimation of the new keynesian phillips curve. *FRB Richmond Economic Quarterly*, 94(4):397–433, 2008.

Anthony F Shorrocks. The measurement of mobility. *Econometrica*, pages 1013–1024, 1978.

Paul M Sommers and John Conlisk. Eigenvalue immobility measures for markov chains 1. *Journal of Mathematical Sociology*, 6(2):253–276, 1979.

John Stachurski and Alexis Akira Toda. An impossibility theorem for wealth in heterogeneous-agent models with limited heterogeneity. *Journal of Economic Theory*, 182:1–24, 2019.

James Tobin. *Asset accumulation and economic activity: Reflections on contemporary macroeconomic theory*. University of Chicago Press, 1982.

Philip Vermeulen. How fat is the top tail of the wealth distribution? *Review of Income and Wealth*, 64(2):357–387, 2018.