Risk-Taking, Capital Allocation and Optimal Monetary Policy*

Joel M. David[†] David Zeke[‡]
FRB Chicago USC

January 12, 2021

Abstract

We study the role of firm heterogeneity in affecting business cycle dynamics and optimal stabilization policy. Firms differ in their degree of cyclicality, and hence, exposure to aggregate risk, leading to firm-specific risk premia that influence resource allocations. The heterogeneous firm economy can be recast in a representative firm formulation, but where total factor productivity (TFP) is endogenous and depends on the resource allocation. The model uncovers a novel tradeoff between the long-run level and volatility of TFP. Inefficiencies distort this tradeoff and result in either excessive volatility or depressed output, implying a role for corrective policy. Embedding this mechanism into a workhorse New Keynesian model, we show that allocational considerations can strengthen the incentives for leaning against the wind, i.e., optimal policy is more strongly countercyclical than in an observationally equivalent economy that abstracts from heterogeneity. A quantitative exercise suggests that the losses from ignoring heterogeneity can be substantial, which stem largely from a less productive allocation of resources and so depressed TFP and output.

*We thank Jonas Fisher, Gadi Barlevy, Francois Gourio, Rohan Kekre, Moritz Lenel, Mariacristina De Nardi, Pablo Kurlat and various seminar and conference participants for helpful comments and suggestions. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System.

†joel.david@chi.frb.org.
‡zeke@usc.edu.

"The level of consumption risk in a society is, in part, subject to choice. When in an economy that is subject to larger shocks, people will live with more consumption variability and the associated loss in welfare, but they may also substitute into risk-avoiding technologies, accepting reduced average levels of production."

— Robert E. Lucas, Jr. "Macroeconomic Priorities"
2003 Presidential Address to the American Economic Association

1 Introduction

This paper studies the role of firm heterogeneity in affecting business cycle dynamics and the implications for the conduct of macroeconomic stabilization policies. We develop a general equilibrium business cycle framework featuring a cross-section of firms that differ in their degree of cyclicality, and hence, exposure to aggregate risk, which leads to firm-specific risk premia on capital investments that influence the allocation of resources across firms. The heterogenous firm economy can be recast in a representative firm formulation, but where aggregate total factor productivity (TFP), usually taken as an exogenous driving force, becomes in part endogenous and determined by the resource allocation. We theoretically explore the implications of this insight for equilibrium cyclical dynamics and optimal fiscal and monetary policy and perform a simple quantitative exercise to illustrate the potential magnitude of these allocational effects.

We augment a standard general equilibrium business cycle model featuring shocks to aggregate technology with (i) a cross-section of heterogeneous firms that differ in their cyclicality, i.e., exposure to those shocks, and (ii) cyclical "wedges" in labor supply and firm-level capital choices. These wedges are sufficiently general to capture the real effects of the policies and distortions that we consider, e.g., fiscal policy in the form of labor income taxes or wage markups (that are determined by the conduct of monetary policy) in a version with nominal rigidities in the form of sticky nominal wages. Throughout the paper, we assume that the aggregate capital stock is an exogenous fixed endowment, which is a standard approach in New Keynesian models and allows us to hone in on the new allocational considerations in our framework. Under these assumptions, we can aggregate the heterogeneous firm economy and derive a tractable log-linear representation of the equilibrium system (the solution is exact in the flexible price case), enabling us to formally prove our main results. Conditional on the properties of aggregate TFP, the flexible price version of the economy nests a simple representative firm real business cycle model with labor as the only input in production and the economy with nominal rigidities nests the standard representative firm New Keynesian model (e.g., Galí (2015)).

In the first part of the paper, we study a special case of the model – namely, where exogenous aggregate shocks are i.i.d. – that yields particularly sharp analytical characterizations of the equilibrium relationships and welfare criterion. The framework reveals a number of key theoretical results: first, the economy faces a tradeoff between the long-run (average) level of economic activity, e.g., TFP and output, and the degree of macroeconomic volatility. First, the smoothing effect: in the face of aggregate risk, agents shift resources from risky, highly procyclical firms to safer, less cyclical ones. This risk-shifting behavior endogenously reduces the economy's responsiveness to exogenous shocks and so lowers cyclical volatility. Second, the level effect: shifting capital across firms in response to aggregate risk introduces a wedge between firm-level capital investment and expected profitability, which leads to dispersion in marginal (revenue) products of capital (MRPK) – more procyclical firms with higher exposure to aggregate risk offer a higher risk premium on investment, which translates into a higher MRPK. Dispersion in the MRPK at the micro-level leads to a persistent reduction in the level of achieved aggregate productivity, i.e., TFP. Thus, the economy faces a tradeoff between cyclical volatility and the long-run level of economic activity. The result can be understood as a form of self-insurance – although there is no aggregate savings in the economy, insurance against business cycle risk is attained by shifting capital to less cyclical firms which, in aggregate, endogenously reduces the degree of cyclicality. The cost of this insurance is the foregone output caused by lower productivity due to the mis-alignment of capital and productivity. These findings provide a formalization of the mechanism posited by Robert E. Lucas, Jr. in his 2003 AEA Presidential Address quoted at the beginning of this paper.

Our second main result is that in an otherwise undistorted economy, the decentralized equilibrium is efficient – aggregate risk is driven by technological shocks alone, which are properly priced and responded to by agents. Thus, the economy attains the socially optimal tradeoff between volatility and long-run effects. Importantly, this is an interior solution – although a social planner could further reduce risk premia, incentivizing risk-taking and lowering marginal product dispersion and increasing output, doing so would raise macroeconomic volatility. In reverse, a planner could further reduce volatility by increasing risk premia and further shifting resources towards less cyclical firms, but doing so would lead to more marginal product dispersion and persistently lower TFP/output. Thus, the efficient outcome features both marginal product dispersion and macroeconomic volatility and the resource allocation reflects the optimal balance between these two considerations. However, inefficiencies in the economy distort this tradeoff and lead to suboptimal outcomes, in particular, either a depressed long-run level of TFP or excessively high aggregate volatility.

In our main contribution, we study the implications of these insights for the conduct of optimal monetary and fiscal stabilization policies. Cyclical policy affects the amount of consumption risk and firm-level risk premia and thus, the resource allocation. We focus primarily on a version with a distortion to the pricing of risk (e.g., a "risk wedge" in the stochastic discount factor) that leads to inefficient investment choices at the micro level (in addition to the standard distortion induced by nominal rigidities in a sticky price economy). Because the aggregate capital stock is in fixed supply, the distortion perturbs the distribution of capital across firms and hence the dynamics of TFP, but has no other effects on the economy. Under the assumption that the distortion is countercyclical, i.e., that agents are excessively averse to bearing risk relative to what preferences and aggregate dynamics would dictate (which we confirm in our quantitative work), it leads to an inefficiently large risk premium.[1] As a result, the capital allocation is too conservative, i.e., capital shifts more towards safer, less cyclical firms, which drives a larger wedge between firm-level profitability and capital and generates more dispersion in MRPK, and ultimately, lower TFP and output. On the flip side, the volatility of TFP is inefficiently low, since too much capital is held by firms with low degrees of cyclicality. Thus, there is scope for stabilization policy to improve equilibrium outcomes.

To make this logic transparent, we derive a representation of the welfare criterion that can be decomposed into four terms – first, volatility in inflation and the output gap, as in the standard representative firm setup. Second, the level and volatility of TFP enter directly. Intuitively, these latter two forces determine the dynamics of the natural rate of output. In the representative firm environment, monetary policy affects the output gap, but the natural rate of output is beyond its influence. Here, in contrast, monetary policy affects both the output gap and the natural rate of output through the dynamics of TFP and so both terms are relevant for understanding the welfare implications of policy.

Our third key result is that in the presence of heterogeneity, the distortion provides stronger incentives for countercyclical policy to lean against the wind – specifically, optimal policy entails a countercyclical output gap that raises output relative to the laissez-faire level in downturns and reduces it in expansions (and additionally accommodates some degree of inflation volatility). In contrast, optimal policy in a version of our model without heterogeneity – the standard representative firm New Keynesian model – entails complete stabilization of both the output gap and inflation. The usual arguments for stabilization are present in our environment, but there is an additional rationale for even more aggressive countercyclical policy – such a policy can reduce the size of the risk adjustment in firm-level input choices, which incentivizes risk-taking on the part of firms and results in a closer alignment of the capital allocation with firm-level productivity. The resulting redistribution of resources across firms leads to a higher long-run level of TFP and output (at the cost of a higher level of TFP volatility). One important corollary of this result is that stabilization policies, i.e., cyclical monetary or fiscal policy, have persistent effects – by influencing the degree of risk, the policy-maker can affect long-run resource allocations and so push the economy closer to/further from its long-run production

[1] For our main analysis, we model the distortion directly as a cyclical shock to the discount factor. However, we also provide two detailed examples showing that such a distortion arises naturally in recent models of financial frictions, e.g., due to limited household participation in financial markets or frictional financial intermediation.

possibilities frontier. A second corollary is that more aggressive policy is needed to achieve any desired level of output smoothing – when the central bank sets a countercyclical output gap to smooth cyclical fluctuations, the resource allocation responds by shifting capital to more cyclical firms, which endogenously increases the cyclicality of TFP and the natural rate of output and mitigates the direct effects of the policy action.

We show that similar results hold when fiscal policy is the tool of stabilization (i.e., through cyclical labor market taxes) in both sticky and flexible price economies. We also study the gains from monetary-fiscal coordination, i.e., when both policies are brought to bear in tandem. We further consider extensions of the framework with cost-push shocks and additional labor market distortions. Heterogeneity affects the standard tradeoff between output gap and inflation volatility in the presence of cost-push shocks, even when nominal rigidities are the only distortion in the economy. Specifically, the central bank is less responsive to the inflationary pressures from these shocks, since acting to neutralize them is more costly – such actions not only generate inefficient output gap fluctuations (in the standard way), but also distort the resource allocation and dynamics of TFP and hence, the natural rate of output. Under plausible assumptions (i.e., a countercyclical labor "wedge"), labor market distortions also strengthen the incentives for countercyclical policy. Further, these effects are increasing in the extent of heterogeneity – the labor distortion leads to inefficient cyclical fluctuations in labor supply, but also generates an additional, inefficient source of aggregate risk that influences the resource allocation. Aggressive countercyclical policy works to bring the economy closer to the efficient outcome.

In the second part of the paper, we provide a numerical illustration of these findings. The key new parameters govern the extent of firm heterogeneity and the properties of the risk distortion. We calibrate these parameters to respectively match the dispersion in cyclicality across firms and the Sharpe Ratio, which represents a direct measure of the price of risk in financial markets. Using these values, we calculate the welfare losses under a number of scenarios: (i) in a first-best case when the economy is completely undistorted. Note that this outcome is not attainable by any of the policies we consider, but serves as a useful benchmark; (ii) when monetary policy follows a Taylor rule with standard values of the reaction coefficients; (iii) under the optimal monetary policy; and (iv) under optimal monetary policy in a representative firm environment, i.e., when the central bank sets policy ignoring heterogeneity and allocational considerations.

The results of this exercise show, first, that heterogeneity can have significant effects on TFP dynamics and welfare. For example, under a standard Taylor rule, the long-run level of TFP is lower by almost 1.4% and its unconditional volatility by almost 30% (relative to the case of a representative firm facing the same shocks). In contrast, in the first-best, these values are only 0.003% and 9%, respectively. These findings imply that the equilibrium allocation is inefficiently conservative – there is an excessive shifting of capital to less cyclical firms, which reduces TFP volatility but causes excessively high marginal product dispersion, with detrimental effects on long-run TFP and output. In contrast, the first-best outcome features a more productive resource allocation with higher long-run TFP, but also higher TFP volatility. In total, including the losses from fluctuations in inflation and the output gap, equilibrium welfare is depressed by about 1.6% (in terms of lifetime consumption in the steady state) relative to the first-best. The majority of this loss (85%) stems from the reduction in long-run TFP due to the distorted resource allocation.

Second, the results suggest an important role for policy to improve on equilibrium outcomes. For example, relative to a standard Taylor rule, the optimal monetary policy increases long-run TFP by 0.44%. The welfare costs of increased TFP volatility from this policy are small, about 0.02%. The total gain from implementing the optimal policy is about 0.65%, of which roughly two-thirds is due to the effects on long-run TFP via the resource allocation. Thus, optimal policy is able to eliminate about 40% of the gap between equilibrium and first-best welfare (0.65% out of 1.6%). In contrast, if the central bank were to set policy to the optimal one ignoring heterogeneity – which, in our simple environment entails complete stabilization of inflation and the output gap – the long-run TFP loss is close to (indeed, slightly larger than) the equilibrium one, about 1.4%, as are the losses from TFP volatility. The total welfare gain relative to the equilibrium is about 0.4% – thus, accounting for heterogenity and allocational effects adds a significant contribution to the potential gains from policy, about 0.25% of steady state consumption. We further find that the results are broadly similar under fiscal policy in a flexible price economy and that the additional gains from monetary-fiscal coordination are small, i.e., once monetary policy is set optimally, the potential further improvements from fiscal policy (at least of the type we consider) are modest.

Related literature. Our paper relates, first, to a burgeoning literature exploring the implications of micro-level heterogeneity for business cycle dynamics and the transmission mechanism of monetary policy, important examples of which include Kaplan et al. (2018), Auclert (2019) and Bilbiie (2008), among others, and for the implementation of optimal policy, e.g., Challe et al. (2017), Acharya, Challe, and Dogra (2020), Bhandari et al. (2018), Bilbiie and Ragot (2020), Bilbiie (2018) and Nuño et al. (2017).[2] The focus of this work has in large part been on the role of household heterogeneity due to incomplete sharing of idiosyncratic risk in determining the consumption response to policy. Kekre and Lenel (2019) study the investment channel

[2]A large body of work studies optimal policy in a rich variety of representative agent business cycle settings. A textbook treatment of optimal monetary policy is in Galí (2015) and a recent review of the literature is in Woodford (2010). There is also a long tradition on the fiscal side, e.g., Lucas Jr and Stokey (1983), Chari et al. (1994) and Chari and Kehoe (1999).

of monetary policy through changes in risk premia when households differ in their degree of risk aversion. Ottonello and Winberry (2018) examine the investment response to policy shocks across heterogeneous firms that face financial frictions. We contribute to this line of work by exploring the implications of a different dimension of financial market heterogeneity, namely, the risk-return tradeoff highlighted in the asset pricing literature. Our findings point to an important role for this form of heterogeneity in determining the conduct and efficacy of policy – when the capital allocation can react to the policy regime, the behavior of aggregate TFP becomes endogenous to the policy. This insight influences both the conduct of optimal policy and the mapping from policy to observed outcomes such as the degree of cyclical volatility.

Angeletos and La'O (2020) and La'O and Tahbaz-Salehi (2020) study optimal policy in the presence of production-side heterogeneity in the form of dispersed private information held by firms and through input-output linkages, respectively. Our analysis contributes to this literature by introducing a new consideration into optimal policy – when firms differ in the riskiness of their technologies, policy can directly impact the dynamics of TFP – both its level and volatility. In a related point, we show that heterogeneity has indirect effects that change the slope of the standard tradeoff between output gap and inflation stabilization. An important corollary of these results is that stabilization policies, whether fiscal or monetary, can have long run effects by improving cross-sectional resource allocations and so shifting the position of the economy relative to its real production possibilities frontier.

Finally, a number of recent papers have investigated the link between business cycle volatility/policy and firm-level resource allocations. David et al. (2019) connect marginal product dispersion to firm-specific risk in a partial equilibrium setting and calculate the implications for measures of misallocation and productivity. Kurtzman and Zeke (2020) highlight the potential mis-allocative effects of central bank large-scale asset purchases of corporate securities though changes in firm-level credit spreads. Our study builds on these results by incorporating heterogeneous risk premia into a workhorse general equilibrium business cycle environment. Our model uncovers a key feedback loop between the capital allocation and macroeconomic volatility that depends crucially on its general equilibrium nature, with potentially important implications for optimal monetary and fiscal policies.

2 The Model

In this section, we introduce our model and use a number of special cases to provide sharp intuition and formally prove our main results.[3]

[3] All derivations and proofs are in Appendix A.

Preferences and technology. A continuum of households indexed by $j \in [0,1]$ seek to maximize expected lifetime utility from consumption and leisure, given by

$$\mathbb{U} = (1-\rho)\,\mathbb{E}_{-1}\left[\sum_{t=0}^{\infty}\rho^t\left(\frac{C_{jt}^{1-\gamma}}{1-\gamma} - \chi\frac{L_{jt}^{1+\varphi}}{1+\varphi}\right)\right] ,$$

where ρ denotes the time discount factor and all other notation is standard. Because we are studying the effects of risk, we assume throughout that households are sufficiently risk averse, specifically, $\gamma > 1$. Households receive labor income, income from capital, which they rent to firms, the return on holdings of risk-free bonds and any distributed profits from firms. The aggregate labor supply of households satisfies

$$T_{lt}W_t = \chi L_t^{\varphi} C_t^{\gamma} , \tag{1}$$

where W_t is the real wage (relative to the price of the final consumption good), L_t is aggregate labor supply and $T_{lt} > 0$ denotes a "wedge" in the form of an explicit or implicit tax/subsidy on labor income. This formulation is sufficiently general to capture the real effects of the policies and nominal rigidities that we study below, which show up in the wedge T_{lt}, for example, fiscal policy in the form of labor income taxes or markups (of wages over the marginal rate of substitution between consumption and labor) in an economy with sticky nominal wages. The smaller is T_{lt}, the larger the "tax" on labor income. The wedge vanishes when $T_{lt} = 1$, values below one represent a tax and values above one a subsidy. The household's intertemporal marginal rate of substitution, or stochastic discount factor (SDF), is given by

$$\Lambda_t = \rho\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma} .$$

The final consumption good is produced by a competitive representative firm, which bundles a continuum of intermediate goods, indexed by $i \in [0,1]$, using a constant elasticity of substitution (CES) aggregator:

$$Y_t = \left(\int Y_{it}^{\nu} di\right)^{\frac{1}{\nu}} , \tag{2}$$

where $\nu \in (0,1)$ and $\frac{1}{1-\nu}$ is the elasticity of substitution between intermediate goods.

Intermediate goods are produced using capital and labor according to

$$Y_{it} = A_{it}K_{it}^{\alpha_1}L_{it}^{\alpha_2}, \quad \alpha_1 + \alpha_2 \leq 1 ,$$

where A_{it} denotes the productivity of firm i in period t.

Throughout the paper, we abstract from accumulation considerations and assume the total capital stock is an exogenous and fixed endowment, i.e., $K_t = \overline{K}\ \forall\ t$.[4] This is a common assumption in the New Keynesian literature and allows us to hone in on the new allocational effects in our framework. With this assumption, the economy with nominal rigidities maps directly to the standard New Keynesian model without capital (and the flexible price version to a simple version of the RBC framework with labor as the only factor of production).

The resource constraints in the economy are then given by[5]

$$C_t = Y_t, \qquad \int K_{it} di = K_t = \overline{K}, \qquad \int L_{it} di = L_t \ .$$

Demand and revenue. Profit maximization by the final good producer yields a standard demand function for intermediate good i:

$$P_{it} = \left(\frac{Y_{it}}{Y_t}\right)^{\nu-1} ,$$

where P_{it} denotes the relative price of good i in terms of the final good. Revenues for intermediate firm i at time t are

$$P_{it}Y_{it} = Y_t^{1-\nu} Y_{it}^{\nu} = Y_t^{1-\nu} A_{it}^{\nu} K_{it}^{\alpha_1 \nu} L_{it}^{\alpha_2 \nu} \ .$$

Input choices. Intermediate firms hire labor period-by-period to maximize current period profits. The optimal choice of labor satisfies

$$\alpha_2 \nu \frac{Y_t^{1-\nu} Y_{it}^{\nu}}{L_{it}} = W_t \ , \tag{3}$$

which shows that firms equalize the marginal revenue product of labor. Operating profits (revenues less labor expenses) are proportional to revenues and are equal to

$$\Pi_{it} = P_{it}Y_{it} - W_t L_{it} = (1-\alpha_2\nu)\, P_{it} Y_{it} = G Y_t^{\frac{1-\nu}{1-\alpha_2\nu}} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} W_t^{-\frac{\alpha_2\nu}{1-\alpha_2\nu}} K_{it}^{\alpha} \ ,$$

where $\alpha \equiv \frac{\alpha_1 \nu}{1-\alpha_2\nu}$ is the effective curvature of operating profits with respect to capital and $G \equiv (1-\alpha_2\nu)\,(\alpha_2\nu)^{\frac{\alpha_2\nu}{1-\alpha_2\nu}}$.

At the end of period $t-1$, firms rent capital for use in period t at rate R_t^K. The firm chooses

[4] The value of $\overline{K}$ plays no role in the analysis.

[5] In the version of the model with sticky wages, the goods market resource constraint only holds to a first-order approximation due to the costs of adjusting wages.

capital to maximize expected discounted profits, i.e., to solve

$$\max_{K_{it}} \mathbb{E}_{t-1}\left[\Lambda_t T_{\Lambda t}\Pi_{it}\right] - R_t^K K_{it} \ .$$

Firms discount the payoffs from capital using the household's discount factor, Λ_t. The term $T_{\Lambda t}$ denotes a "capital wedge" that distorts the firm's investment choice. We also refer to $T_{\Lambda t}$ as a discount rate or risk wedge since it enters the firm's problem identically to the SDF and hence captures inefficiencies in the pricing of risk (e.g., we can interpret $\Lambda_t T_{\Lambda t}$ as a distorted SDF). However, as the expression makes clear, there are a number of alternative interpretations since the wedge captures any factors that lead to an implicit/explicit tax/subsidy on firm profits.[6] For the main analysis, we do not take a stand on the precise source of the wedge, but Appendix B provides two detailed models of financial frictions that generate a wedge in the discount factor of exactly this form: the first stems from incomplete financial markets due to limited household participation as in the two agent model of Debortoli and Galí (2018) and the second from frictional financial intermediation as in Gertler and Karadi (2011). The optimal choice of capital satisfies

$$\mathbb{E}_{t-1}\left[\Lambda_t T_{\Lambda t}\alpha\frac{\Pi_{it}}{K_{it}}\right] = R_t^K \ , \tag{4}$$

which shows that firms equalize the expected discounted marginal revenue product of capital, defined as

$$MRPK_{it} = \alpha\frac{\Pi_{it}}{K_{it}} \ .$$

Substituting for the form of the profit function, the capital choice can be written as

$$K_{it} = \left(\mathbb{E}_{t-1}\left[\alpha G \Lambda_t T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} Y_t^{\frac{1-\nu}{1-\alpha_2\nu}} W_t^{-\frac{\alpha_2\nu}{1-\alpha_2\nu}}\right]\left(R_t^K\right)^{-1}\right)^{\frac{1}{1-\alpha}} \ . \tag{5}$$

The following result shows that from a macroeconomic perspective, the heterogeneous firm economy is observationally equivalent to a representative firm economy with endogenous TFP:[7]

[6] Together, the labor and capital wedges span the range of possible distortions to the two inputs in the model, i.e., although additional distortions could be added (e.g., a wedge in labor demand), they would be redundant. Similar to Chari et al. (2007), the wedges here should be interpreted as the overall distortion to the relevant equilibrium condition. Notice also that while the labor wedge is standard in the business cycle literature and the capital wedge resembles the standard "investment" wedge, the latter is actually quite different in that it distorts firm-level investment choices, but not aggregate capital. Indeed, as we detail below, this wedge only shows up in the aggregate economy through TFP (hence, would be picked up as an "efficiency" wedge in a standard business cycle accounting exercise). This result is reminiscent of the findings in Buera and Moll (2015), who show that the mapping between primitives and aggregate wedges can be complicated in economies with heterogeneity.

[7] In the sticky price economy, additional conditions are necessary to characterize the nominal side of the economy. But, given the behavior of the markup, which shows up in T_{lt}, the system laid out here is sufficient to characterize the behavior of aggregate quantities.

Proposition 1. *The aggregate variables of the heterogenous firm economy behave identically to a representative firm economy with the following equilibrium conditions:*

$$
\begin{aligned}
T_{lt} W_t &= \chi L_t^{\varphi} C_t^{\gamma} \\
W_t &= \alpha_2 \nu \frac{Y_t}{L_t} \\
C_t &= Y_t \\
Y_t &= \Psi_t K_t^{\alpha_1} L_t^{\alpha_2} \\
\Psi_t &= \left(\int A_{it}^{\frac{\nu}{1-\alpha_2 \nu}} \left(\frac{K_{it}}{K_t} \right)^{\alpha} di \right)^{\frac{1-\alpha_2 \nu}{\nu}} \\
\frac{K_{it}}{K_t} &= \frac{\left(\mathbb{E}_{t-1} \left[\Lambda_t T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2 \nu}} Y_t^{\frac{1-\nu}{1-\alpha_2 \nu}} W_t^{-\frac{\alpha_2 \nu}{1-\alpha_2 \nu}} \right] \right)^{\frac{1}{1-\alpha}}}{\int \left(\mathbb{E}_{t-1} \left[\Lambda_t T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2 \nu}} Y_t^{\frac{1-\nu}{1-\alpha_2 \nu}} W_t^{-\frac{\alpha_2 \nu}{1-\alpha_2 \nu}} \right] \right)^{\frac{1}{1-\alpha}} di} .
\end{aligned}
$$

The first four equations are identical to a simple version of standard business cycle models featuring a representative firm with TFP Ψ_t. The fifth equation shows the key result: though the economy can be written as if there was a representative firm, TFP of that firm is endogenous and depends on the efficiency of the capital allocation across the underlying set of heterogeneous firms. Specifically, TFP is equal to an average of firm-level productivities, weighted by their relative shares of aggregate capital. The last equation shows that these shares are determined by firm-level expected discounted profits relative to an appropriate average.

Stochastic processes. Firm productivity (in logs, henceforth denoted with lowercase) is given by[8]

$$a_{it} = \hat{\beta}_i a_t, \quad a_t = \delta a_{t-1} + \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim \mathcal{N}\left(0, \sigma_{\varepsilon}^2\right), \quad \hat{\beta}_i \sim \mathcal{N}\left(\bar{\hat{\beta}}, \sigma_{\hat{\beta}}^2\right) . \tag{6}$$

Here, a_t is a common aggregate shock to (log) technology that follows an AR(1) process with persistence δ and variance of the innovations σ_{ε}^2. Crucially, firms are heterogeneous in their sensitivity, or exposure, to movements in a_t, and hence their degree of cyclicality, captured by $\hat{\beta}_i$.[9] The mean beta is unity, i.e., $\bar{\hat{\beta}} = 1$. The cross-sectional variance in beta, $\sigma_{\hat{\beta}}^2$, captures the

[8] More precisely, we also include adjustments to a_{it} to offset Jensen's inequality terms when taking expectations both over time and across firms and so to maintain log-linearity of the economy. These adjustments insure that when $\kappa = 0$ (agents are risk-neutral with respect to all state prices) or firms are homogeneous, endogenous TFP always coincides with the exogenous aggregate shock. Since these terms are independent of risk and policy, they play no other role in the analysis and can safely be ignored for purposes of exposition. We provide the full expression for a_{it} in our derivations in Appendix A.1.

[9] For simplicity, we abstract from firm-level idiosyncratic shocks. Although important for matching micro-level investment moments, with complete markets, agents can perfectly diversify these shocks, implying that

extent of such heterogeneity across firms.

For simplicity, the remainder of this section considers the special case when the aggregate shock is i.i.d over time, i.e., $\delta = 0$. Under this special case, we obtain particularly sharp expressions for our main theoretical results. We relax this assumption for our quantitative exercise in Section 3, but show that the same intuition from this simpler case carries through.

We assume that the wedges are constant elasticity functions of the aggregate shock, a_t, with elasticities $\tau_{\Lambda a}$ and τ_{la}, respectively:[10]

$$\begin{aligned} \tau_{\Lambda t} &\equiv \log T_{\Lambda t} = -\tau_{\Lambda a} a_t \\ \tau_{lt} &\equiv \log T_{lt} = \tau_{la} a_t \ . \end{aligned} \tag{7}$$

We refer to the wedges as countercyclical when $\tau_{\Lambda a} > 0$, which implies a discount factor that is inefficiently countercyclical, i.e., the discount factor falls (rises) excessively in expansions (downturns) relative to what preferences and the dynamics of aggregate consumption would dictate, and when $\tau_{la} > 0$, which implies a higher "tax" on labor in downturns and thus labor supply that is inefficiently procyclical. A countercyclical risk wedge implies that agents act excessively averse to bearing aggregate risk. The risk wedge exogenous. In contrast, the labor wedge will be endogenous and a function of policy, but will take the form in expression (7). We discuss the policy maker's problem and the determination of τ_{la} in detail below.

2.1 Micro Allocations and Macro Dynamics

In this subsection, we use a special case of this framework to sharply illustrate the link between macroeconomic dynamics and micro-level resource allocations. We make the following simplifications: we set $\alpha_2 = 0$, so that capital is the only factor of production, $\nu = 1$, so that intermediate goods are perfect substitutes, and lastly, we abstract from labor and capital market distortions, i.e., $T_{lt} = T_{\Lambda t} = 1 \ \forall \ t$. Under these assumptions, there are no aggregate movements in factors of production at all – the only decision in the economy is how to allocate the fixed capital stock – yet the economy features rich dynamics arising from the resource allocation alone.

First, Proposition 1 implies that the aggregates in the economy are fully determined by the

they bear no risk premium. Additionally, they are independent of policy. Thus, they would play no role in the analysis (other than adding constant terms to a number of the equilibrium equations).

[10]We assume throughout that optimal time-invariant subsidies are in place to offset goods and labor market monopoly distortions. This ensures an efficient steady-state and that only the time-varying distortions described here play a role.

following system:

$$Y_t = \Psi_t K_t^{\alpha}, \quad \Psi_t = \int A_{it} \left(\frac{K_{it}}{K_t} \right)^{\alpha} di, \quad \frac{K_{it}}{K_t} = \frac{\left(\mathbb{E}_{t-1} \left[\Lambda_t A_{it} \right] \right)^{\frac{1}{1-\alpha}}}{\int \left(\mathbb{E}_{t-1} \left[\Lambda_t A_{it} \right] \right)^{\frac{1}{1-\alpha}} di} . \tag{8}$$

We can guess and verify that in equilibrium, (log) aggregate TFP is given by

$$\psi_t \equiv \log \Psi_t = \overline{\psi} + \psi_a a_t \ , \tag{9}$$

where $\overline{\psi}$ denotes an endogenous mean (long-run) level of TFP and ψ_a an endogenous loading on the aggregate shock, a_t, i.e., ψ_a is the elasticity of endogenous TFP to the exogenous shock. Note that $\overline{\psi}$ and ψ_a are not in general equal to zero and one, respectively, showing that endogenous TFP is not the same as exogenous technology.

Given the dynamics of TFP in expression (9), the SDF is an endogenous function of a_t. In particular, the unexpected shock to the discount factor is

$$\log \Lambda_t - \mathbb{E}_{t-1} \log \Lambda_t = -\gamma y_t = -\gamma \psi_a a_t = -\kappa a_t \ , \tag{10}$$

where $\kappa \equiv \gamma \psi_a$ is the (negative) elasticity of the discount factor, Λ_t, to the aggregate shock A_t. The term κ plays a crucial role throughout our analysis: it captures a risk adjustment in the capital allocation and indeed, is a sufficient statistic to characterize the role of risk in the cross-sectional allocation. Specifically, using (10), we can express firms' relative capital stocks from expression (8) as

$$\frac{K_{it}}{K_t} = \frac{\left(\mathbb{E}_{t-1} \left[A_{it} A_t^{-\kappa} \right] \right)^{\frac{1}{1-\alpha}}}{\int \left(\mathbb{E}_{t-1} \left[A_{it} A_t^{-\kappa} \right] \right)^{\frac{1}{1-\alpha}} di} .$$

Micro allocations. To a second order approximation, we can derive the following expression characterizing the firm's optimal capital choice (in logs):

$$k_{it} = \text{const.} - \frac{1}{1-\alpha} \beta_i \kappa \sigma_{\varepsilon}^2 \ . \tag{11}$$

The expression shows that heterogeneity in beta induces dispersion in firm-level capital choices. Specifically, the capital choice is made up of a component that is common across firms, less a term that reflects a firm-specific risk premium. The risk premium is linear and increasing in beta, scaled by the risk adjustment, κ (and by the volatility of the aggregate shock, σ_{ε}^2).

Because $\kappa > 0$, the risk adjustment causes capital to shift towards less cyclical, lower beta firms, independent of their expected productivity (in fact, in this i.i.d. case, all firms have the same expected productivity, yet do not choose the same level of capital). By introducing a wedge between the choice of capital and expected productivity, the risk adjustment induces dispersion in the (expected) marginal revenue product of capital, which (in logs) is given by:

$$Emrpk \equiv \mathbb{E}_{t-1}\left[mrpk_{it}\right] = \text{const.} + \beta_i \kappa \sigma_\varepsilon^2 \quad \Rightarrow \quad \sigma^2_{Emrpk} = \left(\kappa\sigma_\varepsilon^2\right)^2 \sigma^2_{\hat{\beta}} \,. \tag{12}$$

Without heterogeneity in $\hat{\beta}_i$, firms equalize expected *mrpk*. With heterogeneity, more procyclical firms are riskier and must offer a higher rate of return on capital as compensation. The strength of this effect is determined by the risk adjustment, κ, and the degree of aggregate volatility, σ_ε^2.

Expressions (11) and (12) illustrate the two key effects of macroeconomic risk on micro-level resource allocations: first, there is a risk-shifting effect – aggregate risk causes capital to move towards less cyclical, low beta firms. Second, there is a "misallocation" effect – because capital is allocated based on risk characteristics rather than solely on expected productivity, aggregate risk induces dispersion in the *mrpk*. These two effects are two sides of the same coin – to attract capital, riskier firms must offer a higher expected return on capital in the form of a higher *mrpk*. Due to diminishing marginal returns to capital in production, this necessarily entails a lower capital stock.

Macro dynamics. Using (11), we can aggregate firm-level capital choices and derive the following expression for aggregate TFP as a function of the risk adjustment, κ:

$$\begin{aligned} \psi_t &= \overline{\psi} + \psi_a a_t \\ &\text{where} \\ \psi_a &= 1 - \frac{\alpha}{1-\alpha}\kappa\sigma_\varepsilon^2\sigma_{\hat{\beta}}^2 \;\leq\; 1 \\ \overline{\psi} &= -\frac{1}{2}\frac{\alpha}{1-\alpha}\left(\kappa\sigma_\varepsilon^2\right)^2\sigma_{\hat{\beta}}^2 \;\leq\; 0 \,. \end{aligned} \tag{13}$$

The expression verifies our conjecture of the form of ψ_t and reveals two key effects of the resource allocation on macroeconomic dynamics that work through the risk adjustment, κ:

Risk-shifting and volatility – First, there is a smoothing effect of heterogeneity. In the absence of heterogeneity, the term ψ_a, which captures the elasticity of TFP to the exogenous shock, is equal to one – in this case, exogenous technology and endogenous TFP are the same. In contrast, with heterogeneity, ψ_a is strictly less than one – in the face of exogenous shocks, the endogenous reallocation of capital away from more procyclical firms to less cyclical ones reduces the sensitivity of TFP to those shocks. From expression (8), TFP is an average of firm-level productivities, weighted by shares of the aggregate capital stock. As capital shifts towards less cyclical, low beta firms, the weight of those firms in that average is diminished and the cyclical volatility of TFP falls. Formally, taking the derivative of ψ_a with respect to κ,

$$\frac{\partial \psi_a}{\partial \kappa} = -\frac{\alpha}{1-\alpha}\sigma_\varepsilon^2 \sigma_{\hat{\beta}}^2 < 0 \; , \tag{14}$$

which shows that the larger the risk adjustment, the lower the responsiveness of endogenous TFP to exogenous shocks.

Long-run productivity – The second effect of heterogeneity on TFP comes through the constant term in TFP, $\overline{\psi}$. By inducing dispersion in *mrpk*, the risk adjustment lowers the long-run, or average, level of aggregate productivity. Indeed, comparing expressions (12) and (13) shows that (up to a second order approximation), TFP losses due to the risk adjustment are exactly proportional to the *Emrpk* dispersion it generates, with a scaling factor that depends on the production function parameter, α (which, in this simple example, is just α_1, the capital share in production). Formally, taking the derivative of $\overline{\psi}$ with respect to κ,

$$\frac{\partial \overline{\psi}}{\partial \kappa} = -\frac{\alpha}{1-\alpha}\kappa \left(\sigma_\varepsilon^2\right)^2 \sigma_{\hat{\beta}}^2 < 0 \; , \tag{15}$$

i.e., the larger the risk adjustment, the greater the dispersion in *Emrpk* and thus, the lower is long-run productivity.

Expressions (14) and (15) reveal the key insight of this example – heterogeneity in risk implies a tradeoff between the long-run level and cyclical volatility of aggregate TFP. A smaller risk adjustment leads to a more productive allocation of resources, but more volatile TFP. A larger risk adjustment leads to the opposite. The result relies on three key ingredients: the presence of macroeconomic risk, $\sigma_\varepsilon^2 > 0$, the presence of heterogeneity with respect to that risk, $\sigma_{\hat{\beta}}^2 > 0$, and that agents are averse to bearing that risk, $\gamma > 0$. The result can also be understood as a form of self-insurance. There are no savings in the economy (bonds are in zero net supply). Yet agents can insure against cyclical fluctuations by shifting capital to less cyclical firms, which endogenously reduces the extent of aggregate risk. However, there is a cost of doing so in the form of higher marginal product dispersion, which reduces the productivity of the resource allocation and hence the long-run level of output/consumption.

Characterizing the risk adjustment. The economy features a fixed point linking the aggregate dynamics to the micro allocation: from expression (11), the risk adjustment, κ, determines micro-level allocations and by its definition in (10) is a function of ψ_a, the endogenous loading of TFP on the exogenous shock – the larger is ψ_a, the greater the extent of aggregate risk and hence the larger is the risk adjustment. However, from expression (13), through the resource allocation, ψ_a is, in turn, a function of κ – the larger is κ, the more that capital shifts to less cyclical, low beta firms and the lower is the responsiveness of TFP to shocks. Solving for κ in terms of the primitives yields:

$$\kappa = \frac{\gamma}{1 + \gamma\frac{\alpha}{1-\alpha}\sigma_\varepsilon^2\sigma_{\hat{\beta}}^2} \; .$$

The interpretation of the expression is straightforward. The numerator captures the direct (partial equilibrium) effect of risk, which is simply the coefficient of relative risk aversion. The denominator captures the general equilibrium effects that feed back from the resource allocation, which smooth TFP and hence lower the extent of aggregate risk. The equilibrium risk adjustment reflects both of these forces.

Policy. The environment reveals a tradeoff between the long-run level of TFP and its volatility. In this simple version, it is straightforward to verify that the tradeoff is efficient. The economy exhibits both TFP volatility and marginal product dispersion, but there is no scope for policy to improve on equilibrium outcomes. Policies that reduce the risk adjustment may be effective in raising long-run TFP/output, but cause a shift in the allocation of capital towards more procyclical firms, which inefficiently increases the volatility of TFP. Policies increasing the risk adjustment have the opposite effects, smoothing aggregate TFP but reducing its long-run level. In the absence of any inefficiencies, policies of either type distort the tradeoff and result in sub-optimal outcomes. In the next sections, we augment this simplified setting with a number of elements that introduce inefficiencies and study the implications for the conduct of macroeconomic stabilization policy.

2.2 TFP and Risk in the General Setting

Proposition 2 extends the results from the previous section to the more general environment featuring endogenous labor supply, imperfect substitutability across intermediate goods and labor and capital market wedges (we continue to assume aggregate shocks are i.i.d.). To keep the expressions as simple as possible, Proposition 2 works with a second order approximation, but we provide exact expressions in Appendix A.1 (all of our welfare results use the exact solution):

Proposition 2. *To a second order approximation, the optimal choice of capital is given by*[11]

$$k_{it} = const. - \frac{1}{1-\alpha}\beta_i\kappa\sigma_\varepsilon^2 \; . \tag{16}$$

[11] More precisely, there are also terms that reflect differences in the variance of shocks. Because these terms are small and are independent of risk/policy, we suppress them here. The full approximate expression and the exact analogs are in Appendix A.1.

Aggregate TFP is equal to

$$\psi_t = \overline{\psi} + \psi_a a_t \tag{17}$$

where

$$\psi_a = 1 - \omega\kappa\sigma_\varepsilon^2\sigma_\beta^2 \tag{18}$$

$$\overline{\psi} = -\frac{1}{2}\omega\left(\kappa\sigma_\varepsilon^2\right)^2\sigma_\beta^2 \,. \tag{19}$$

The equilibrium risk adjustment is

$$\kappa \equiv \kappa_\psi\psi_a + \tau_{\Lambda a} + \kappa_l\tau_{la} \tag{20}$$

where

$$\kappa_\psi = (\gamma - 1)\,\phi_\psi + \frac{\nu}{1-\alpha_2\nu} > 0 \qquad \text{with} \qquad \phi_\psi = \frac{1}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}$$

$$\kappa_l = (\gamma - 1)\,\phi_l > 0 \qquad \phi_l = \frac{\alpha_2}{1+\varphi}\frac{1}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}$$

and $\omega \equiv \frac{\alpha_1}{1-\alpha} = \frac{\alpha_1(1-\alpha_2\nu)}{1-(\alpha_1+\alpha_2)\nu}$, $\beta_i \equiv \frac{\nu}{1-\alpha_2\nu}\hat{\beta}_i$, $\sigma_\beta^2 \equiv \left(\frac{\nu}{1-\alpha_2\nu}\right)^2\sigma_{\hat{\beta}}^2$.

Conditional on the risk adjustment, κ, expression (16) is the clear analog to (11) with modified versions of α and β, which now reflect the presence of labor in production ($\alpha_2 > 0$) and imperfect substitutability across intermediate goods ($\nu < 1$). Similarly, equations (17) to (19) are analogous to (13), with ω representing a composite parameter that captures both capital and labor elasticities in production and curvature in demand.

The key difference between this more general version and the simpler one in the last subsection is in the definition of the risk adjustment itself. Here, the adjustment term, κ, is a composite that measures the (negative of the) elasticity of the discounted profitability of capital to movements in a_t operating through discount factor effects (as was the case in the simpler example) but now additionally through equilibrium effects on the wage, W_t, and aggregate demand, Y_t. Thus, κ captures the aggregate risk facing the firm through the joint movement of all of these variables.

First, κ_ψ captures the effects of a_t through changes in endogenous aggregate TFP, ψ_t. To see this more clearly, we can write

$$\kappa_\psi = \left(\gamma - \frac{1-\nu}{1-\alpha_2\nu} + \frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1 - \frac{1-\gamma}{1+\varphi}\right)\right)\phi_\psi \,. \tag{21}$$

The term in parentheses measures the (negative) elasticity of discounted profitability to movements in aggregate output, Y_t. These come through movements in the (undistorted) SDF (γ), changes in aggregate demand and so firm-level prices (the term $\frac{1-\nu}{1-\alpha_2\nu}$), and lastly through wages, $\left(\frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1-\frac{1-\gamma}{1+\varphi}\right)\right)$. Multiplying by ϕ_ψ translates this elasticity with respect to output into the elasticity with respect to TFP, ψ_t, and multiplying through by ψ_a translates the entire term into the elasticity with respect to the exogenous shock, a_t.

Second, movements in a_t affect the discount rate wedge with elasticity $\tau_{\Lambda a}$, which directly enters the risk adjustment. Finally, a_t moves the labor wedge with elasticity τ_{la} and the coefficient κ_l measures the (negative) elasticity of discounted profits with respect to this wedge.[12] Analogous to expression (21), Appendix A.1 provides a decomposition of κ_l into a component coming via the direct effect of wages on profitability and a component coming indirectly via changes in labor supply. In that appendix we also provide full solutions for κ, ψ_a and $\overline{\psi}$ as functions of model primitives.

Optimal risk adjustment. Before turning to our analysis of monetary and fiscal policies, we study the allocation chosen by a planner who faces – but takes as given – the distortions τ_{lt} and $\tau_{\Lambda t}$. From the definitions of $\overline{\psi}$ and ψ_a in Proposition 2, we can see that the allocation of capital is summarized by κ. This implies that the optimal allocation can be characterized by allowing the planner to directly choose κ to maximize household welfare, subject to the definitions of $\overline{\psi}$ and ψ_a in (18) and (19). The planner chooses the risk adjustment subject to the responses of the private sector to its choice. The optimal κ satisfies[13]

$$\frac{\partial\overline{\psi}}{\partial\kappa} = (\gamma-1)\,\phi_\psi\psi_a\sigma_\varepsilon^2\frac{\partial\psi_a}{\partial\kappa}\ . \tag{22}$$

The optimal risk adjustment weighs the productivity gain from reducing κ against the losses from additional TFP volatility. Substituting for the derivatives (the expressions are in Appendix A.2) and rearranging yields the following result:

Proposition 3. *The optimal risk adjustment satisfies $\kappa^* = \kappa_\psi\psi_a^*$ where κ_ψ is defined in Proposition 2 and ψ_a^* is the value of ψ_a without distortions.*

The proposition has two key implications. First, in the absence of distortions, i.e., $\tau_{la} = \tau_{\Lambda a} = 0$, the equilibrium risk adjustment corresponds to the optimal one. Thus, the equilibrium of the undistorted economy achieves the first-best outcome. The result replicates our finding from the simpler version in the previous subsection, and follows from the fact that in the undistorted economy there is no scope for policy to improve allocations – macroeconomic volatility

[12] Expressions (30) and (31) in Appendix A.1 show that ϕ_ψ and ϕ_l are the elasticities of output with respect to TFP and the labor wedge, respectively, and make explicit all the relevant elasticities.

[13] The exact form of the objective function depends on whether prices are sticky or flexible, but the result holds in both cases.

and marginal product dispersion co-exist and both are symptoms of the economy's efficient response to fundamental shocks.[14]

Second, the proposition shows that even in the distorted economy, the constrained optimal risk adjustment is independent of those distortions. In other words, even though the economy may exhibit inefficiencies on other other margins (e.g., labor supply), the optimal κ does not further distort the allocation of capital across firms. Thus, the optimal and constrained optimal risk adjustments are the same. In contrast, the distorted equilibrium does not achieve either the optimal or constrained optimal allocation – from the definition of κ in Proposition 2, if the labor and risk wedges are countercyclical, they increase the risk adjustment, causing a reallocation of capital to less cyclical, lower beta firms. In this case, the allocation of capital is too conservative, leading to inefficiently high marginal product dispersion and low macroeconomic volatility (the effects are reversed if the wedges are procyclical). In the next sections, we show that in the presence of distortions, there is a role for monetary and/or labor market fiscal policies to improve on equilibrium outcomes, but neither is able to attain the optimal κ – intuitively, neither policy can fully correct the allocation without further distorting the economy on other margins.

2.3 Monetary Policy

We now study the positive and normative implications of heterogeneity for macroeconomic policy. First, we turn to our main results on monetary policy, perhaps the most standard tool of stabilization policy. To do so, we first flesh out some necessary details regarding the nature of nominal rigidities and the behavior of aggregate prices. In Section 2.4, we show that similar insights go through when fiscal policy is the instrument of stabilization, even when prices are perfectly flexible. We also study a version with monetary-fiscal coordination, i.e., where policy makers optimally employ both tools simultaneously.

New Keynesian system. We assume nominal rigidities in the form of sticky wages. The setup is standard and we provide only a broad overview. Households monopolistically supply differentiated labor services, which are then bundled into the final labor input using a CES aggregator with elasticity of substitution ν_w. Wage changes are subject to quadratic adjustment costs a la Rotemberg (1982), given by $\frac{\theta_w}{2}\left(\Pi_t^w - 1\right)^2 Y_t$, where Π_t^w denotes gross nominal wage inflation.[15] In deviations from the steady state, the log-linearized equilibrium system is

[14] Taking κ directly as the policy instrument implies that we are restricting our attention to linear policies. However, we can show that the optimal linear policy corresponds exactly to the optimal global policy (including non-linear ones) in the undistorted economy. With distortions, the optimal policy is not exactly linear, but the linear policy can be shown to approximate the global one.

[15] In the standard way, the framework also accommodates Calvo pricing frictions with a modified definition of the slope of the Phillips Curve.

characterized by:

$$\begin{aligned} w_t &= y_t - l_t \\ w_t + \mu_t &= \gamma y_t + \varphi l_t \\ y_t &= \psi_t + \alpha_2 l_t \\ \pi_t^w &= \rho \mathbb{E}_t\left[\pi_{t+1}^w\right] + \lambda_w \mu_t \\ y_t &= \mathbb{E}_t\left[y_{t+1}\right] - \frac{1}{\gamma}\left(i_t - \mathbb{E}_t\left[\pi_{t+1}^p\right]\right) \ , \end{aligned} \tag{23}$$

where ψ_t is as defined in Proposition 2. The first equation is the labor demand condition (3), which relates the real wage, w_t, to the marginal product of labor. The second defines (the inverse of) the wage markup, μ_t, which comes from the wage-setting problem of households, as the (negative of the) difference between the real wage and the marginal rate of substitution between consumption and labor. A comparison to (1) shows that the labor wedge defined earlier, τ_{lt}, now stems explicitly from the nominal wage rigidity: specifically, τ_{lt} defined above is exactly equal to the (inverse) wage markup μ_t. The third equation defines the aggregate production function, which, crucially, depends on endogenous aggregate TFP, ψ_t, which is a function of the micro-level resource allocation, rather than only on the exogenous aggregate shock. The fourth equation is the New Keynesian Wage Phillips Curve, which relates wage inflation, π_t^w, to expected wage inflation and the markup. The slope of the Phillips Curve is determined by the composite parameter $\lambda_w \equiv \frac{\alpha_2 \nu_w}{\theta_w}$. The last equation is the standard consumption Euler equation, relating expected output/consumption growth to the nominal interest rate, i_t, and expected price inflation, π_{t+1}^p. Note that the distortion $\tau_{\Lambda t}$ does not show up in any of the equilibrium equations in (23) – the only effect of the distortion comes via the resource allocation and dynamics of TFP as shown in Proposition 2. In other words, conditional on TFP, the macroeconomy appears efficient (though in reality, it may not be).

The conduct of policy. The wage markup is the only nominal variable that enters the real side of the equilibrium system in (23) and so determines the impact of monetary policy on quantity variables. Further, as we show next, the (inverse) markup, μ_t, is proportional to the output gap (the difference between realized output and the "natural" level of output in the absence of nominal rigidities, appropriately defined). Thus, it proves convenient to abstract from an explicit representation of the transmission mechanism of monetary policy and characterize policy in terms of the output gap/markup, μ_t, directly, i.e., to assume the central bank directly controls the output gap. This is the main path that we take.[16] In particular,

[16] This is a common approach to characterizing optimal policy, for example, Galí (2015) .

we assume the central bank chooses a value μ_a that determines the cyclicality of the output gap/markup, defined by $\mu_t = \mu_a a_t$.[17] More aggressive countercyclical policy entails a more negative μ_a, i.e., a more countercyclical output gap.

The Euler equation implies a relationship between the output gap and the nominal interest rate. In particular, given the behavior of the output gap, μ_a, we can derive an associated value, i_a, that determines the cyclicality of the nominal rate, defined by $i_t = i_a a_t$.[18] Put another way, any desired behavior of the output gap can be implemented through the central bank's choice of the nominal rate. We use this fact to also characterize optimal policy in terms of the cyclicality of the nominal rate, in particular, the optimal response of the nominal rate to the aggregate shock. More aggressive countercyclical policy entails a more positive i_a, i.e., a more procyclical nominal rate. We can also relate μ_a to other standard representations of monetary policy, e.g., the coefficients in a Taylor Rule. We discuss this mapping in more detail below.

Monetary policy and TFP. The system in (23) is identical to the textbook representative firm New Keynesian model (e.g., Galí (2015)), the only difference being the endogeneity of TFP. However, this difference is crucial for our study and can have dramatic implications for the conduct and effects of monetary policy. To see this, recall the results from Proposition 2, in particular, expression (20): the output gap, which, as just discussed, is directly determined by monetary policy and hence a choice variable for the central bank, influences the risk adjustment, κ, with a more procyclical output gap (higher μ_a) increasing the risk adjustment (simply replace τ_{la} in that expression by μ_a, since we have $\tau_{la} = \mu_a$). Intuitively, the cyclicality of the economy – the source of aggregate risk – depends on both the cyclicality of the natural rate of output and the cyclicality of the output gap. A procyclical output gap increases the cyclicality of realized output and thus the amount of aggregate risk; a countercyclical output gap has the opposite effects. Expressions (16), (18) and (19) show that these effects feed through to the resource allocation and via this channel, to the behavior of TFP, determined by ψ_a and $\overline{\psi}$.

Thus, through the risk adjustment κ, monetary policy determines the allocation of capital and the dynamics of TFP. As an example, consider the case of more aggressive stabilization, i.e., a lower value of μ_a and hence, κ. As shown in (16), there is a reallocation of capital towards more cyclical, higher beta firms. From (18) and (19), this leads to (i) a higher value of ψ_a – TFP becomes more volatile – and (ii) a higher value of $\overline{\psi}$ – the long-run level of TFP increases. Intuitively, more aggressive stabilization by the central bank mitigates the extent of aggregate risk by reducing the cyclicality of output. This incentivizes further risk-taking on the part of the private sector, i.e., capital shifts toward more cyclical firms, leading to more volatile TFP,

[17] We abstract from issues of commitment and assume the central bank can commit to this simple rule.

[18] Specifically, $i_a = -i_0 - i_\mu \mu_a$ where i_0 and i_μ are positive constants.

but also a more productive resource allocation, i.e., one that more closely aligns capital with expected productivity at the firm-level and thus features less marginal product dispersion.

There are two key implications: first, monetary policy has permanent effects – the level of long-run TFP (and hence output/consumption) is in part determined by the extent of stabilization, which, by influencing the capital allocation and dispersion in $mrpk$, moves the economy closer to/further from its real production possibilities frontier. This results holds despite our use of a standard form of nominal rigidities.

Second, for any desired degree of output smoothing, more aggressive policy is needed than implied by a representative firm model. In particular, the deviations of output from its long-run mean (which is endogenous and depends on $\overline{\psi}$) can be written as

$$y_t = \phi_\psi \psi_a a_t + \phi_l \mu_t = (\phi_\psi \psi_a + \phi_l \mu_a)\, a_t \ .$$

Define the output gap as $\tilde{y}_t \equiv y_t - y_t^n$, where y_t^n denotes the natural level of output in a flexible price economy (e.g., a notion of potential output). Holding the capital allocation fixed across actual and potential output, since upon realization of the shock, capital is quasi-fixed, we can decompose output into its natural rate and the output gap, i.e.,

$$\begin{aligned} y_t &= y_t^n + \tilde{y}_t \\ &\text{where} \\ y_t^n &= \phi_\psi \psi_a a_t \\ \tilde{y}_t &= \phi_l \mu_a a_t \ , \end{aligned}$$

where we have used the fact that ψ_a is fully determined by the capital allocation and thus does not affect the gap between y_t and y_t^n. In a representative firm environment, ψ_a is simply equal to one and outside the influence of policy. Thus, the central bank can act to smooth the output gap through its choice of μ_a leaving the dynamics of the natural rate of output unchanged. This stems directly from the fact that TFP in the representative firm setup is exogenous. Here, in contrast, TFP, and so the natural rate of output, also depend on the actions of the central bank. In particular, as the central bank acts to smooth the output gap though a lower μ_a, it reduces the risk adjustment in the capital allocation, κ, which leads to a shifting of capital towards more cyclical firms, increasing ψ_a and so the volatility of TFP and the natural rate of output. In other words, increased risk-taking on the part of the private sector in response to the attempted stabilization partially offsets the effects of the policy – as the central bank smooths the output gap, the natural rate of output itself becomes more volatile. Thus, if the central bank is targeting the overall volatility of output, a more aggressive smoothing policy is required.

This is a form of the Lucas Critique at work – in the representative firm setup, the dynamics of TFP are assumed to be exogenous and hence invariant to policy; in the heterogenous firm economy, this is no longer the case and the central bank must take into account the effects of its actions on TFP and the natural rate of output.

Optimal monetary policy. Appendix A.2 derives the following second order approximation to the welfare loss function, expressed in terms of the equivalent consumption decline measured as a fraction of consumption in the non-stochastic steady state:

$$\mathbb{W} = -\overline{\psi} + \frac{\text{var}(\psi_t)}{2}(\gamma - 1)\phi_\psi + \frac{\text{var}(\tilde{y}_t)}{2}\frac{1}{\phi_l} + \frac{\text{var}(\pi_t^w)}{2}\frac{\alpha_2 \nu_w}{\lambda_w},$$

which expresses the loss as a function of (i) the level of TFP, (ii) the volatility of TFP, (iii) the volatility of the output gap, and (iv) the volatility of wage inflation, all with appropriate weights. It is straightforward to verify that the last two terms correspond exactly to the welfare function in the textbook New Keynesian model and capture the standard losses from output gap and inflation fluctuations, respectively, conditional on the dynamics of TFP.[19] The effects of heterogeneity enter the loss function through the first two terms, which are new to our setting – welfare is increasing in the long-run level of TFP (the first term), and decreasing in its volatility (the second term).[20] The first term reflects the long-run productivity losses due to marginal product dispersion. The second term reflects the negative effects of TFP volatility, both through fluctuations in consumption/output and labor supply. Due to the presence of heterogeneity, these two effects are endogenous, dependent on the conduct of monetary policy and are not independent – indeed, they are both determined by the risk adjustment, κ.

Substituting for the variances,

$$\mathbb{W} = -\overline{\psi} + (\gamma - 1)\phi_\psi \psi_a^2 \frac{\sigma_\varepsilon^2}{2} + \mu_a^2 \phi_l \frac{\sigma_\varepsilon^2}{2} + \frac{\alpha_2 \nu_w}{\lambda_w}(\lambda_w \mu_a)^2 \frac{\sigma_\varepsilon^2}{2}, \tag{24}$$

which makes explicit the new role of monetary policy in the presence of heterogeneity: through $\overline{\psi}$ and ψ_a, both the level and volatility of TFP are functions of policy (they both depend on κ, which, in turn, depends on μ_a). In contrast, in the absence of heterogeneity, $\overline{\psi} = 0$ and $\psi_a = 1$, i.e., both terms are exogenous and invariant to policy and hence can be ignored for the purposes of determining optimal policy. As above, the last two terms capture the standard losses from output gap and inflation volatility in terms of monetary policy and model primitives.

[19] For example, compare these terms to the welfare function in Galí (2015), equation (26) of Chapter 6, and note that $\frac{1}{\phi_l} = \gamma + \frac{\varphi + 1 - \alpha_2}{\alpha_2}$.

[20] We can also interpret the first two terms as capturing the level and volatility of the natural rate of output, y_t^n: multiply and divide to obtain $-\frac{1}{\phi_\psi}\overline{y^n} + \frac{\text{var}(y_t^n)}{2}\frac{\gamma - 1}{\phi_\psi}$, where $\overline{y^n}$ is the long-run mean of the natural rate.

We first characterize optimal policy in terms of the μ_a that minimizes (24). Next, we show that this policy can be implemented with a nominal interest rate rule. The optimal choice of μ_a satisfies

$$\frac{\partial \overline{\psi}}{\partial \mu_a} = (\gamma - 1)\, \phi_\psi \sigma_\varepsilon^2 \psi_a \frac{\partial \psi_a}{\partial \mu_a} + \phi_l \sigma_\varepsilon^2 \mu_a + \alpha_2 \nu_w \sigma_\varepsilon^2 \lambda_w \mu_a \ . \tag{25}$$

The central bank chooses μ_a so that the productivity gain from a marginal increase in stabilization (decrease in μ_a, left-hand side) equals the marginal welfare costs from changes in (i) the volatility of TFP, (ii) the volatility of the output gap and (iii) the volatility of inflation. Optimal policy balances all of these considerations. Substituting for the derivatives in (25) yields the following result:

Proposition 4. *Optimal monetary policy satisfies*

$$\begin{aligned} \mu_a^* &= -\tau_{\Lambda a} \Phi_\Lambda \\ &\textit{where} \\ \Phi_\Lambda &= \frac{\kappa_l \Omega}{\phi_l + \kappa_l^2 \Omega + \alpha_2 \nu_w \lambda_w} > 0, \qquad \frac{\partial \Phi_\Lambda}{\partial \sigma_\beta^2} > 0 \end{aligned}$$

and $\Omega = \frac{\omega \Upsilon\left(\sigma_\beta^2 \sigma_\varepsilon^2\right)}{1+\kappa_\psi \omega \Upsilon\left(\sigma_\beta^2 \sigma_\varepsilon^2\right)}$, $\Upsilon\left(\sigma_\beta^2 \sigma_\varepsilon^2\right) = \frac{\sigma_\beta^2 \sigma_\varepsilon^2}{1+\sigma_\beta^2 \sigma_\varepsilon^2}$.

To build intuition, consider first the representative firm New Keynesian model, where $\sigma_\beta^2 = 0$. Clearly, we have $\mu_a^* = 0$: TFP is exogenous and thus outside the control of policy, optimal policy completely stabilizes both inflation and the output gap, and the economy is at first-best. Thus, in the standard way, the "divine coincidence" holds and the central bank faces no tradeoff in setting policy. Note that this result holds independently of whether the risk wedge is active or not – as noted above, this distortion only affects aggregate variables through the resource allocation. In the absence of allocational considerations in the representative firm version, the wedge does not change optimal policy. The result also holds in the presence of heterogeneity, but when the risk wedge is absent (i.e., $\tau_{\Lambda a} = 0$), but for a different reason – in this case, TFP is in fact endogenous, but it is efficient, and again the central bank is able to achieve the first-best by completely stabilizing inflation and the output gap. In other words, when firms are heterogeneous but the resource allocation is undistorted, TFP is endogenous, but there is no role for policy in correcting the allocation or dynamics of TFP and optimal policy coincides with the representative firm economy.

In contrast, this logic breaks down when firms are heterogeneous ($\sigma_\beta^2 > 0$) and the allocation is distorted ($\tau_{\Lambda a} \neq 0$). If the distortion is countercyclical, i.e., $\tau_{\Lambda a} > 0$ (the empirically relevant case), it strengthens the incentives for countercyclical policy. From expression (20), the wedge generates an inefficiently high risk adjustment, which, from (16)-(19), leads to an overly conservative capital allocation, an inefficiently low cyclicality of TFP and excessively high marginal product dispersion, which ultimately, depresses the level of productivity. Thus, the optimal response entails more aggressive countercyclical policy – specifically, a negative value for μ_a^* (countercyclical output gap) – that leans against cyclical fluctuations in labor supply to partially offset these effects.

As we have seen, the optimal value for μ_a^* is zero if the wedge is zero and there is complete stabilization of inflation and the output gap. Thus, correcting the wedge requires the central bank to allow some fluctuations in these latter two variables. However, the policy response to the wedge is tempered by the presence of wage-setting frictions, which show up through ν_w and λ_w in the denominator – the central bank accounts for the fact that using countercyclical policy to counteract the risk distortion generates volatility in inflation and the output gap and so does not act to fully offset the distortion (doing so would increase the losses from inflation/output gap fluctuations).

Finally, the positive derivative of Φ_Λ with respect to σ_β^2 shows that the optimal policy responds to the real friction more aggressively when there is more heterogeneity – with larger differences across firms, the distortionary effects of the wedge on the capital allocation and the resulting impacts on the level and volatility of TFP are more costly relative to output gap and inflation volatility. In contrast, as just discussed, in the extreme case of no heterogeneity, the risk wedge has no effects at all and optimal policy can safely ignore it. Thus, a coherent message emerges from these results – the presence of firm-level risk in conjunction with a countercyclical distortion in the pricing of risk leads monetary policy to be more countercyclical and strengthens incentives for leaning against the wind, specifically, countercyclical movements in the output gap. The opposite holds if the distortion is procyclical.

Optimal nominal interest rate. We can use the Euler equation to derive the nominal interest rate that implements the desired behavior of the output gap/markup, which takes the form $i_t = i_a^* a_t$, where

$$\begin{aligned} i_a^* &= \Phi^i + \tau_{\Lambda a}\Phi_\Lambda^i \\ &\text{where} \\ \Phi^i &< 0, \quad \frac{\partial \Phi^i}{\partial \sigma_\beta^2} > 0 \\ \Phi_\Lambda^i &> 0, \quad \frac{\partial \Phi_\Lambda^i}{\partial \sigma_\beta^2} > 0 , \end{aligned}$$

and Φ^i and Φ_Λ^i are constants defined in Appendix A.2. The result aligns closely with Proposition 4: (i) the cyclicality of the optimal nominal rate is a linear and increasing function of the risk wedge, so that a countercyclical wedge ($\tau_{\Lambda a} > 0$) leads to a more aggressive rise in the nominal rate in response to positive technology shocks and (ii) the strength of this effect is increasing in heterogeneity, i.e., for a given value of the wedge, more heterogeneity leads the optimal nominal rate to be more procyclical. Indeed, from the definition of Φ^i_Λ in Appendix A.2, Φ^i_Λ is zero in the absence of heterogeneity, in which case $i^*_a = \Phi^i$, which sets the nominal rate such that the real interest rate always equals the natural interest rate. The same result clearly holds in an economy with nominal rigidities but no additional distortion.

In Appendix A.2, we also show that μ^*_a can also be implemented using standard formulations of a Taylor Rule. In particular, we specify the rule in terms of the output gap and expected inflation, which, in deviations takes the form[21]

$$i_t = \phi_y \tilde{y}_t + \phi_\pi \mathbb{E}_t \left[\pi^p_{t+1}\right] \ ,$$

and derive a mapping between the policy coefficients, ϕ_y and ϕ_π, and the markup, μ_a.

Additional distortions. We have also studied the effects of two additional distortions: first, a cost-push shock, which generates further cyclical inflation dynamics and a trade-off for the central bank even in the absence of heterogeneity. We specify this shock as a constant elasticity function of a_t, i.e., $\eta_t = \eta_a a_t$, where $\eta_a > 0$ represents a procyclical shock to (wage) inflation and $\eta_a < 0$ the opposite. Second, we add a non-policy based labor market distortion that generates inefficient cyclicality in labor supply. In some abuse of notation, following expression (7), we specify this distortion as $\tau_{lt} = \tau_{la} a_t$, where $\tau_{la} > 0$ represents a countercyclical distortion to labor supply (i.e., the distortion leads labor supply to be inefficiently procyclical). We continue to denote with μ_a the cyclicality of the (inverse) markup induced by the nominal rigidity.[22] With these distortions, the welfare loss function is given by

$$\mathbb{W} = -\overline{\psi} + (\gamma - 1)\,\phi_\psi \psi_a^2 \frac{\sigma_\varepsilon^2}{2} + (\tau_{la} + \mu_a)^2\,\phi_{\tau_l} \frac{\sigma_\varepsilon^2}{2} + \frac{\alpha_2 \nu_w}{\lambda_w} (\lambda_w \mu_a + \eta_a)^2 \frac{\sigma_\varepsilon^2}{2} \ ,$$

[21] See, e.g., Clarida et al. (2000) for empirical support for this specification. Similar results hold for other common formulations of the rule, i.e., in terms of wage inflation, etc.

[22] The cost-push shock enters as an additional term in the Phillips curve. The labor distortion enters the aggregate labor supply condition. The derivations of the welfare criterion and optimal policy follow closely those in the baseline model in Appendix A.2.

and optimal policy takes the form

$$\mu_a^* = -\tau_{\Lambda a}\Phi_\Lambda - \eta_a \Phi_\eta - \tau_{la}\Phi_l \tag{26}$$

where

$$\Phi_l = \frac{\phi_{\tau_l} + \kappa_{\tau_l}^2 \Omega}{\phi_{\tau_l} + \kappa_{\tau_l}^2 \Omega + \alpha_2 \nu_w \lambda_w} > 0, \qquad \frac{\partial \Phi_l}{\partial \sigma_\beta^2} > 0$$

$$\Phi_\eta = \frac{\alpha_2 \nu_w}{\phi_{\tau_l} + \kappa_{\tau_l}^2 \Omega + \alpha_2 \nu_w \lambda_w} > 0, \qquad \frac{\partial \Phi_\eta}{\partial \sigma_\beta^2} < 0 \, ,$$

where Φ_Λ and Ω are the same as in Proposition 4. Since Φ_Λ is unchanged from that proposition, we do not discuss it further here.

Because of the cost-push shock, in the standard way, the central bank faces a tradeoff and cannot simultaneously stabilize inflation and the output gap. Stabilizing the output gap requires setting $\mu_a^* = 0$, but this implies complete accommodation of the cost-push shock. In reverse, stabilizing inflation requires setting $\mu_a^* = -\frac{\eta_a}{\lambda_w} \neq 0$, so cyclical policy that induces a negative output gap in the face of inflationary pressure (e.g., when productivity is high if $\eta_a > 0$ or when productivity is low if $\eta_a < 0$), but clearly allows for fluctuations in the output gap. How do heterogeneity and allocational considerations change this tradeoff? Expression (26) shows that Φ_η is strictly decreasing in σ_β^2 – in the presence of greater heterogeneity, the central bank responds less aggressively to the inflationary pressure from the cost-push shock. The typical cost of offsetting this shock is output gap volatility; here, there is an additional cost – allowing μ_a to depart from zero to lean against the shock affects the risk adjustment, κ, and distorts the capital allocation. For example, consider the case where $\tau_{\Lambda t} = 0$. Then, the capital allocation and dynamics of TFP are efficient without the effects of policy and the impact of the policy response to the cost-push shock on these margins is unambiguously distortionary. Thus, heterogeneity reduces the magnitude of the optimal response to the cost-push shock and affects the tradeoff between inflation and output gap volatility (even with no additional distortions). The central bank must account for the fact that responding to this shock induces an inefficient reallocation of capital across firms.

The effect of the labor distortion, captured by τ_{la}, is similar to the risk wedge: if the distortion is countercyclical ($\tau_{la} > 0$), it further strengthens the incentives for countercyclical policy.[23] For example, consider the extreme case with no other distortions and flexible prices, i.e., $\tau_{\Lambda a} = \eta_a = \lambda_w = 0$. Optimal policy entails completely neutralizing the labor distortion, i.e., $\mu_a^* = -\tau_{la}$.[24] The distortion generates a procyclical output gap; with no other inefficiencies,

[23] A large body of work dating back at least to Chari et al. (2007) documents a countercyclical labor wedge.

[24] Of course, with flexible prices, monetary policy cannot achieve this outcome. However, cyclical fiscal policy can.

optimal policy works to exactly undo it. With nominal rigidities, the optimal response is less than one-for-one: countercyclical policy aimed to stabilize the labor distortion generates costly inflation volatility. How does heterogeneity change this tradeoff? The positive derivative of Φ_l with respect to σ_β^2 shows that greater heterogeneity strengthens the response of policy to the labor distortion and allows for more inflation volatility. Intuitively, the cost of the labor distortion is increasing in heterogeneity: with heterogeneity, the distortion not only has a direct effect on the cyclicality of labor supply, but also leads the risk adjustment, κ, to be inefficiently large (when the distortion is countercyclical), which distorts the allocation of capital and the dynamics of TFP. This latter effect is larger when there are more opportunities for reallocation, i.e., σ_β^2 is large. Thus, optimal policy responds more aggressively to the distortion than in the case of a representative firm.

2.4 Fiscal Policy

Before turning to the numerical exercise, we show that the main insights go through when cyclical fiscal policy is the tool of the policy maker, both in sticky and flexible price economies. Although we view our main contribution as characterizing optimal monetary policy in the presence of heterogeneity, it is important to note that the main results do not hinge on the presence of nominal rigidities or the use of monetary policy as the instrument of stabilization.

Flexible prices. With flexible prices, the nominal side of the economy is completely disentangled from real quantities and the first three equations in (23) along with the definition of TFP in Proposition 2 fully characterize the aggregate dynamics. Here, we revert to the notation τ_{lt} for the labor wedge, which now represents a cyclical labor income tax that is chosen by the fiscal authority (rather than μ_t in (23) which denoted the labor wedge from sticky wages). The welfare loss function takes the form[25]

$$\mathbb{W} = -\overline{\psi} + (\gamma - 1)\,\phi_\psi \psi_a^2 \frac{\sigma_\varepsilon^2}{2} + \tau_{la}^2 \phi_l \frac{\sigma_\varepsilon^2}{2}\ , \tag{27}$$

which is clearly a special case of (24) when there are no nominal rigidities.

Optimal fiscal policy sets the cyclicality of the tax, τ_{la}, to minimize (27), which yields the first order condition:

$$\frac{\partial \overline{\psi}}{\partial \tau_{la}} = (\gamma - 1)\,\phi_\psi \sigma_\varepsilon^2 \psi_a \frac{\partial \psi_a}{\partial \tau_{la}} + \phi_l \sigma_\varepsilon^2 \tau_{la}\ .$$

Comparing to expression (25) shows that, again, the expression is a special case when there are no nominal rigidities. The policy-maker chooses τ_{la} to optimally weigh the level of TFP

[25] In this flexible price, i.i.d. case, expression (27) is exact.

(left-hand side) against its volatility and inefficient volatility in labor supply (the output gap) induced by the tax. Similar to monetary policy, fiscal policy of this form cannot fully correct inefficiencies in the risk adjustment – and hence, capital allocation – without distorting the labor supply margin, and the optimal τ_{la} takes this into account. We can prove the following result:

Proposition 5. *The optimal labor market fiscal policy satisfies*[26]

$$\tau_{la}^{*} = -\tau_{\Lambda a}\Phi_{\Lambda}$$

where

$$\Phi_{\Lambda} = (\gamma-1)\frac{\omega\Upsilon\left(\sigma_{\beta}^{2}\sigma_{\varepsilon}^{2}\right)}{1+\left((\gamma-1)+\frac{\nu}{1-\alpha_{2}\nu}\right)\omega\Upsilon\left(\sigma_{\beta}^{2}\sigma_{\varepsilon}^{2}\right)}>0, \quad \frac{\partial\Phi_{\Lambda}}{\partial\sigma_{\beta}^{2}}>0\,.$$

In the absence of distortions, the optimal policy is a laissez-faire one – the policy-maker sets the tax to zero. This replicates the result above that the undistorted economy is efficient and optimally weighs the level of productivity against its volatility. The capital market imperfection distorts this tradeoff and the tax is set to partially offset this distortion. If the risk wedge is countercyclical, i.e., $\tau_{\Lambda a}>0$, it provides incentives for countercyclical tax policy (a procyclical labor tax) – because such a wedge generates inefficient cyclical volatility, optimality entails a procyclical labor tax that induces lower (higher) labor supply in expansions (downturns). Notice, however, that even with flexible prices, the policy-maker cannot replicate the first-best capital allocation – although within the set of attainable allocations, doing so would necessarily further distort the labor supply margin.[27]

As was the case for monetary policy, Proposition 5 also reveals a key role for heterogeneity in changing the incentives for stabilization – the optimal fiscal policy response to the risk wedge is zero if there is no heterogeneity and is strictly increasing in the extent of heterogeneity. As above, if firms are homogeneous, the distortion has no effect on the economy. As heterogeneity increases, the distortionary impact of the wedge on the allocation grows, necessitating a more aggressive policy response. Thus, the more heterogeneous are firms, the more policy should lean against the wind through a procyclical tax regime.

Fiscal-monetary coordination. A last case we consider is when fiscal and monetary policy are set optimally in tandem, i.e., there is coordination between the fiscal and monetary authorities. In this case, we can prove that (i) optimal monetary policy sets $\mu_a=0$, i.e., monetary

[26] The expression for Φ_{Λ} here can be shown to be a special case of the one in Proposition 4 when prices are perfectly flexible.

[27] Replicating the first-best capital allocation would entail setting τ_{la} such that $\tau_{\Lambda a}+\kappa_{l}\tau_{la}=0$ or $\tau_{la}=-\frac{1}{\kappa_{l}}\tau_{\Lambda a}$.

policy is the same as in the representative firm economy and completely stabilizes the markup (and hence, the portion of the output gap that stems from wage stickiness) and inflation, and (ii) optimal fiscal policy satisfies Proposition 5. Thus, when both fiscal and monetary policy are jointly put to work, a natural ordering emerges: first, monetary policy is set so that the economy replicates the flexible price outcome. Then, fiscal policy is set as it would be if prices were indeed truly flexible. We also study this case and calculate the potential gains from such coordination in our quantitative exercise in the next section.

3 Quantitative Exercise

In this section, we provide a numerical evaluation of the policy impacts studied in the last section.

First, we return to the more general case of persistent aggregate shocks, i.e., we allow δ in equation (6) to be non-zero. With this modification, the system in (23) exactly nests the textbook New Keynesian model in, e.g., Galí (2015). All of the equilibrium equations remain unchanged from above, with the exception of aggregate TFP. With persistence, TFP takes the form

$$\psi_t = \overline{\psi} + \delta a_{t-1} + \psi_a \varepsilon_t \; , \tag{28}$$

which is easily shown to be an ARMA(1,1). In other words, although the exogenous shock follows an AR(1), due to heterogeneity and risk-shifting, endogenous TFP follows an altered dynamic process. The other macroeconomic variables, such as output and labor, follow a similar process. Because of this, optimal policy takes the form of a rule for the markup/output gap that responds to the lagged value of aggregate technology and the current realization of the shock to technology with different coefficients, i.e.,

$$\mu_t = \mu_{a_{-1}} a_{t-1} + \mu_a \varepsilon_t$$

The welfare loss function is given by

$$\begin{aligned} \mathbb{W} &= -\overline{\psi} + (\gamma - 1)\,\phi_\psi \psi_a^2 \frac{\sigma_\varepsilon^2}{2} + \left(\mu_a^2 + \frac{\rho}{1-\rho\delta^2}\mu_{a_{-1}}^2\right)\phi_l \frac{\sigma_\varepsilon^2}{2} \\ &+ \frac{\alpha_2 \nu_w}{\lambda_w}\left(\left(\frac{\rho\lambda_w \mu_{a_{-1}}}{1-\rho\delta} + \lambda_w \mu_a\right)^2 + \frac{\rho}{1-\rho\delta^2}\left(\frac{\lambda_w \mu_{a_{-1}}}{1-\rho\delta}\right)^2\right)\frac{\sigma_\varepsilon^2}{2} \; , \end{aligned}$$

which is an extended version of (24) that reflects the effects of persistence. The interpretation of each of the terms is the same as in that equation: the expression captures the losses from (i) the level of TFP, (ii) the volatility of TFP, (iii) the volatility of the output gap and (iv) the volatility of inflation. Minimizing this expression with respect to μ_a and $\mu_{a_{-1}}$ yields a pair of first order conditions that characterize optimal policy.

3.1 Calibration

We begin by assigning values to the more standard preference and production function parameters of our model. We assume a period length of one year and set the annual discount factor, ρ, to 0.96 and the inverse Frisch elasticity of labor supply, φ, to 1. We set risk aversion, γ, to 10. Although this is somewhat high for the macro literature (and at the upper bound of what is typically deemed the "reasonable" range, e.g., Mehra and Prescott (1985)), it is a standard value (indeed, at the lower end) in the finance literature studying issues of premia, e.g., Bansal and Yaron (2004), as do we. We assume constant returns to scale in production and set α_1 and α_2 to standard values of one-third and two-thirds, respectively. We set the substitution parameter across intermediate goods (corresponding to firm revenue returns to scale) to $\nu = 0.8$, which is a common value in the literature, e.g., Atkeson and Burstein (2010).

The parameters governing the aggregate shock process are chosen to match moments in measured aggregate TFP. Estimates of the TFP process in expression (28) imply a persistence parameter of $\delta = 0.7$. We calibrate the standard deviation of the shocks, σ_ε, so that the model replicates the standard deviation of annual TFP growth rates (approximately 0.03), which results in a value of $\sigma_\varepsilon = 0.05$.[28] The cross-sectional standard deviation of firm exposures to the aggregate shock, $\sigma_{\hat{\beta}}$, is chosen to match the observed dispersion in cyclicality among Compustat firms. Specifically, we estimate time-series regressions of firm-level productivity growth on aggregate TFP growth. We can then use the coefficient estimates and the equilibrium processes on firm and aggregate productivity to recover measures of firm-level betas (up to an additive constant). The cross-sectional dispersion in the estimates yields a value of $\sigma_{\hat{\beta}} = 3.2$.[29]

Following, e.g., Broer et al. (2020) and Galí (2015), we set the elasticity of substitution across labor types, ν_w, to 6. We then pin down the adjustment cost parameter, θ_w, using an indirect inference procedure on the slope of the wage Phillips curve. Specifically, we regress wage inflation on measures of the output gap both in the model and data, and set θ_w so that

[28] The persistence estimate is calculated using HP-filtered annual TFP constructed from data on real GDP and aggregate capital and labor from the Bureau of Economic Analysis. Estimates of the standard deviation differ across TFP series, e.g., the annual BEA data give about 0.025, where the series calculated by John Fernald gives about 0.036. 0.03 is roughly the midpoint.

[29] Firm-level data are obtained from Compustat. We include firms with at least 40 observations (the data are at the annual frequency) and trim the 0.5% tails of the estimates and additionally adjust for sampling error. The estimated dispersion is somewhat smaller than that in David et al. (2019) (4.8), which is estimated using stock market returns in conjunction with a structural model, and so is likely conservative (there are other differences in the estimates as well, e.g., assumptions on curvature, the sample of firms studied and the frequency of the data, among others).

Table 1: Calibration – Summary

Parameter	Description	Value
Preferences		
ρ	Discount factor	0.96
φ	Inverse Frisch elasticity	1
γ	Risk aversion	10
Production		
α_1	Capital share	1/3
α_2	Labor share	2/3
ν	Intermediate good substitutability	0.8
δ	Persistence of agg. shock	0.7
σ_ε	Std. dev. of agg. shock	0.05
$\sigma_{\hat{\beta}}$	Std. dev. of risk exposures	3.2
Wage-setting & distortions		
ν_w	Labor elasticity of substitution	6
θ_w	Wage adjustment cost	866
$\tau_{\Lambda a}$	Risk wedge	-14.7

the estimated slope coefficients are identical.[30] The procedure yields a value of $\theta_w = 866$. This implies a relatively high level of these costs, which is needed to match the flat slope of the empirical wage Phillips curve. We assume that monetary policy in the baseline equilibrium follows a standard Taylor rule in expected (price) inflation and the output gap (see, e.g., Clarida et al. (2000)), given by $i_t = 1.5\mathbb{E}_t\left[\pi^p_{t+1}\right] + 0.5\tilde{y}_t$. Lastly, we calibrate the risk wedge, $\tau_{\Lambda a}$, so that the model generates a maximum annual Sharpe Ratio equal to 0.75, which is roughly the annualized value calculated for the S&P 500 Index in, e.g., Lo (2002) (and indeed, may be a conservative value, since achieved Sharpe ratios on other investment strategies have been shown to be even higher). This approach yields a significantly countercyclical wedge, $\tau_{\Lambda a} = -14.7$, implying that the wedge amplifies the degree of aggregate risk. Table 1 summarizes the parameter values.

[30] We use the difference between potential real GDP as computed by the BEA and realized real GDP as a measure of the output gap. Using various (annual, HP-filtered) measures of wages from the Bureau of Labor Statistics yields slope coefficients ranging from about 0.1 (using, e.g., average hourly earnings of production and nonsupervisory employees) to 0.3 (using, e.g., business sector compensation per hour). We target a slope of 0.2, approximately the midpoint of this range. Note that the coefficient from this regression does not directly map into a structural parameter, since from equation (23), inflation expectations, which are correlated with the output gap, are in the error term. Rather, the identification is indirect and follows from matching a salient moment from the model and data.

Table 2: Heterogeneity and Monetary Policy

	Baseline (1)	First-Best (2)	Optimal Policy (3)	Ignoring Hetero. (4)
Welfare loss (%)				
Total	1.796	0.193	1.155	1.403
TFP level	1.370	0.003	0.932	1.396
TFP volatility	0.008	0.190	0.025	0.008
Output gap volatility	0.100	0.000	0.172	0.000
Inflation volatility	0.318	0.000	0.026	0.000
Equilibrium statistics				
$\Delta\sigma(\psi_t)$ (%)	-28.92	-8.59	-26.77	-29.00
$\sigma(\tilde{y}_t)$	1.34	0.00	1.69	0.00
$\sigma(\pi_t^w)$	0.28	0.00	0.08	0.00
ϵ_{i_t,ψ_t}	0.24	-	0.83	-0.22

3.2 Optimal Monetary Policy

Table 2 presents the equilibrium and counterfactual policy exercises. Each column displays welfare losses (top panel) and a number of equilibrium statistics (bottom panel) under alternative policy regimes. We report the total welfare loss as well as a decomposition of the loss into its four components: the level of TFP and the volatilities of TFP, the output gap and (wage) inflation. We report four salient statistics of the equilibrium under each policy: the reduction in TFP volatility relative to the case of a representative firm facing the same shocks (denoted $\Delta\sigma(\psi_t)$), the volatilities of the output gap and inflation, and lastly, a measure of the cyclicality of monetary policy, namely, the elasticity of the nominal interest rate to the realization of TFP (denoted ϵ_{i_t,ψ_t}). Column (1) ("Baseline") corresponds to the equilibrium in the baseline calibrated model; column (2) ("First-Best") reports results from the first-best allocation; column (3) ("Optimal Policy") does so under the optimal monetary policy and column (4) ("Ignoring Hetero.") when monetary policy is set to the optimal one in the representative firm case, i.e., ignoring micro-level heterogeneity and allocational concerns.

The results suggest (i) firm-level heterogeneity can have sizable effects on TFP dynamics and welfare, and (ii) accounting for this heterogeneity can have important implications for the conduct of monetary policy through its effects on the resource allocation.

First, column (1) implies that in the baseline equilibrium – which assumes a standard formulation of the Taylor rule – long-run TFP is lower by almost 1.4% (relative to a representative firm economy facing the same exogenous shocks). At the same time, the volatility of TFP is also lower, by about 29%. The welfare costs of depressed TFP are directly equal to the TFP loss itself. The welfare costs of TFP volatility turn out to be relatively small. Column (2), the first-best allocation, provides a natural benchmark for these values. In this case, TFP losses are extremely small, only about 0.003%. TFP is more volatile, only about 9% less so than with a representative firm facing the same shocks. These findings imply that due to the risk distortion, the equilibrium allocation is inefficiently conservative – there is an excessive shifting of capital towards less cyclical firms, which reduces TFP volatility relative to the first-best, but also causes excessively high marginal product dispersion, which depresses long-run TFP. In contrast, the first-best allocation features a more productive allocation of capital with higher long-run TFP, but also higher TFP volatility. In total, welfare in the baseline equilibrium is about 1.6% lower than in the first-best. Decomposing this loss shows that the large majority stems from the reduction in long-run TFP due to the distorted resource allocation. Of the total 1.6 percentage point loss relative to first-best($1.80\% - 0.19\%$), about 1.37 percentage points (1.37%-0.00%), or roughly 85%, is due to the losses in long-run TFP. Indeed, since the first-best allocation features higher TFP volatility than the baseline equilibrium, equilibrium welfare losses from this source are smaller than in the first-best. The remaining welfare difference is due to changes in inflation and output gap volatility – these are modest, but non-negligible, in the equilibrium outcome and are zero in the first-best, which by construction, is completely undistorted.

Second, the results suggest an important role for policy to improve on equilibrium outcomes and further, highlight the importance of accounting for heterogeneity when determining optimal policy. Column (3) shows that relative to the Taylor rule in column (1), the optimal monetary policy increases long-run TFP by 0.44 percentage points ($1.37\% - 0.93\%$). The central bank achieves these gains by setting a more countercyclical policy, which reduces the risk extent of aggregate risk and induces the private sector to take on a riskier allocation that more closely aligns firm-level capital and productivity – for example, the elasticity of the nominal interest rate to the realization of TFP is roughly four times larger under optimal policy than under the Taylor rule, i.e., the optimal nominal rate is significantly more procyclical. At the same time, there is a costly increase in TFP volatility, but this offsetting effect turns out to be small (the standard deviation is of TFP is about 2 percentage points higher than under the Taylor rule), as are the resulting welfare losses, which are about 0.02% higher. The total gain from implementing the optimal policy relative to the Taylor rule is about 0.65%, of which about two-thirds is due to the effects on long-run TFP via the resource allocation. Thus, optimal policy eliminates about 40% of the gap between equilibrium and first-best welfare ($\frac{1.80\%-1.15\%}{1.80\%-0.19\%}$). However, even the optimal policy cannot achieve the first-best, since using (counter)cyclical policy to influence the resource allocation also affects the output gap and inflation (under this calibration, optimal policy reduces inflation volatility but increases output gap volatility relative to the Taylor rule).

Finally, how does the presence of heterogeneity change the conduct of optimal policy? Comparing columns (3) and (4) shows that if the central bank were to set policy to the optimal one ignoring heterogeneity – which, in this simple environment entails complete stabilization of inflation and the output gap – the gain relative to the baseline is about 0.4%, essentially entirely due to the elimination of fluctuations in these variables. The properties of TFP change only minimally, so that the central bank is almost entirely missing out on the gains from improving the resource allocation (indeed the level of TFP under this policy is even lower than the baseline). Thus, accounting for micro-level heterogeneity makes a significant contribution to the potential gains from policy, about 0.25 percentage points $(1.40\% - 1.15\%)$ of steady state consumption. These gains come wholly from an improved allocation and higher long-run TFP – TFP under optimal policy is almost 0.5% higher than under the policy ignoring heterogeneity. In reverse, the cost of this gain is somewhat higher volatility on all dimensions, with inflation volatility being the costliest form. Again, the central bank achieves this gains through a more aggressively countercyclical policy – the interest rate elasticity to TFP is not procyclical enough when not accounting for heterogeneity (indeed, in this framework, since the the output gap and inflation respond inversely to TFP, optimal policy ignoring heterogeneity entails reducing the interest rate when TFP is high in order to stabilize these variables).

3.3 Optimal Fiscal Policy

Table 3 displays analogous results under optimal fiscal policy, in the form of a cyclical labor market tax/subsidy captured by τ_{la} and perfectly flexible wages. Column (1) ("Baseline") reports results from a baseline equilibrium where we assume that there are no cyclical labor income taxes, i.e.,. $\tau_{la} = 0$. In column (2), we show outcomes under the optimal fiscal policy and in column (3) under the optimal policy ignoring heterogeneity.[31] In this case, since the risk wedge is the only distortion (in contrast to the environment above with the pricing friction), the optimal fiscal policy when not accounting for heterogeneity is a laissez-faire one, i.e., $\tau_{la} = 0$. Thus, columns (1) and (3) coincide. The first-best allocation is the same as in Table 2 so we do not repeat it here.

The results are qualitatively similar to those in Table 2. In the baseline equilibrium, long-run TFP is 1.4% lower than in an equivalent representative firm economy and TFP is 29% less volatile. Notice that these outcomes are identical to column (4) in Table 2 – the flexible price economy with laissez-faire fiscal policy is the same as the sticky price economy with complete stabilization of inflation and the output gap (and inactive fiscal policy). Optimal fiscal policy works to reduce both of these effects: long run TFP increases by over 0.5% and TFP volatility also rises, though quite modestly (to 26% less volatile than the representative firm economy).

[31]For purposes of comparison, we do not recalibrate the value of $\tau_{\Lambda a}$ in Table 3 versus Table 2. However, doing so leads to only small changes in the results.

Table 3: Heterogeneity and Fiscal Policy

	Baseline (1)	Optimal Policy (2)	Ignoring Hetero. (3)
Welfare loss (%)			
Total	1.403	1.123	1.403
TFP level	1.396	0.879	1.396
TFP volatility	0.008	0.028	0.008
Output gap volatility	0.000	0.215	0.000
Equilibrium statistics			
$\Delta\sigma(\psi_t)$ (%)	-29.00	-26.40	-29.00
$\sigma(\tilde{y}_t)$	0.00	1.90	0.00
τ_{la}	0.00	-4.54	0.00

In total, welfare under the optimal policy is 0.28% higher than in the baseline equilibrium. The value of τ_{la} shows that these gains are achieved through aggressive countercyclical tax policy (i.e., a procyclical tax): the elasticity of the tax with respect to TFP shocks is large and positive (the negative of τ_{la}). Further, the flexible price economy gives a particularly sharp illustration of the importance of accounting for heterogeneity when setting policy – optimal policy when ignoring heterogeneity corresponds to the policy in the baseline equilibrium and thus, the entirety of the welfare gains from the true optimal policy stem from addressing allocational considerations.

Finally, we can use the results in Table 3 to gauge the benefits from monetary-fiscal coordination. Recall that in this case, monetary policy completely stabilizes the markup and thus the portion of the output gap that stems from price stickiness and optimal fiscal policy is then set as it would be in a flexible price economy. This is exactly the scenario in column (2) in Table 3. Thus, comparing the results in that column to column (3) in Table 2 gives the incremental gains of monetary-fiscal coordination over monetary policy alone. It turns out that these gains are modest, about 0.03% of steady state consumption. In other words, once monetary policy is optimally determined accounting for heterogeneity, the scope for additional improvements from labor market fiscal policies is small.[32]

[32] In the simple environment here, the result largely follows from our focus on labor income taxes. If the fiscal authority had access to two distinct cyclical taxes that did not have exactly proportional effects on labor supply and the capital allocation (e.g., a cyclical tax on firm profits), the first-best allocation could be achieved. However, this would not be the case in a richer environment with additional distortions, such as the labor market distortions and cost-push shocks we study in Section 2.3.

4 Conclusion

In this paper, we have studied the implications of firm heterogeneity – specifically, differences in cyclicality – for business cycle dynamics and optimal stabilization policy. The heterogeneous firm economy can be recast in a representative firm formulation but where the resource allocation and hence aggregate TFP are endogenous and depend on the conduct of policy. We show that (i) the economy faces a tradeoff between the long-run level of TFP and TFP volatility, (ii) the tradeoff is efficient in an otherwise undistorted economy and (iii) empirically founded assumptions on distortions in capital or labor markets lead optimal policy to be more aggressively countercyclical than in an observationally equivalent representative firm model. Thus, firm heterogeneity tends to strengthen the incentives for lean against the wind policies. A quantitative exercise suggests that the welfare gains from implementing policies that account for these allocational considerations can be significant.

We have deliberately kept our framework simple in order to highlight the new insights while taking only small departures from textbook business cycle models. A fruitful, though challenging next step would be to add additional ingredients that enable the model to match a wider set of business cycle and micro-level moments, e.g., adjustment costs, financial frictions, more complicated preferences, etc., and evaluate the effects of heterogeneity in a state-of-the-art quantitative DSGE model. Of particular interest would be the implications for capital accumulation and the dynamics of aggregate investment, as well as the inclusion of additional distortions/shocks that have been highlighted in the literature (for example, such as those studied qualitatively in Section 2.3). One broader lesson of our paper is that understanding the properties of inefficiencies – and their heterogeneous effects – is crucial to reaching accurate conclusions regarding effective macroeconomic policies.

References

ACHARYA, S., E. CHALLE, AND K. DOGRA (2020): "Optimal Monetary Policy According to HANK," CEPR Discussion Paper No. DP14429.

ANGELETOS, G.-M. AND J. LA'O (2020): "Optimal monetary policy with informational frictions," *Journal of Political Economy*, 128, 1027–1064.

ATKESON, A. AND A. BURSTEIN (2010): "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, 118, 433–484.

AUCLERT, A. (2019): "Monetary Policy and the Redistribution Channel," *American Economic Review*, 109, 2333–67.

Bansal, R. and A. Yaron (2004): "Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles," *Journal of Finance*, 59.

Bhandari, A., D. Evans, M. Golosov, and T. J. Sargent (2018): "Inequality, business cycles, and monetary-fiscal policy," Tech. rep., National Bureau of Economic Research.

Bilbiie, F. O. (2008): "Limited asset markets participation, monetary policy and (inverted) aggregate demand logic," *Journal of economic theory*, 140, 162–196.

——— (2018): "Monetary policy and heterogeneity: An analytical framework," CEPR Discussion Paper No. DP12601.

Bilbiie, F. O. and X. Ragot (2020): "Optimal monetary policy and liquidity with heterogeneous households," *Review of Economic Dynamics*.

Broer, T., N.-J. H. Hansen, P. Krusell, and E. Oberg (2020): "The New Keynesian Transmission Mechanism: A Heterogeneous-Agent Perspective," *The Review of Economic Studies*, 87, 77–101.

Buera, F. J. and B. Moll (2015): "Aggregate Implications of a Credit Crunch: The Importance of Heterogeneity," *American Economic Journal: Macroeconomics*, 7, 1–42.

Challe, E. et al. (2017): "Uninsured unemployment risk and optimal monetary policy," *American Economic Journal: Macroeconomics*.

Chari, V. V., L. J. Christiano, and P. J. Kehoe (1994): "Optimal fiscal policy in a business cycle model," *Journal of Political Economy*, 102, 617–652.

Chari, V. V. and P. J. Kehoe (1999): "Optimal fiscal and monetary policy," *Handbook of macroeconomics*, 1, 1671–1745.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007): "Business Cycle Accounting," *Econometrica*, 75, 781–836.

Clarida, R., J. Gali, and M. Gertler (2000): "Monetary policy rules and macroeconomic stability: evidence and some theory," *The Quarterly journal of economics*, 115, 147–180.

David, J. M., L. Schmid, and D. Zeke (2019): "Risk-Adjusted Capital Allocation and Misallocation," Working paper, Duke University.

David, J. M. and V. Venkateswaran (2019): "The sources of capital misallocation," *American Economic Review*, 109, 2531–67.

Debortoli, D. and J. Galí (2018): "Monetary policy with heterogeneous agents: Insights from TANK models," Universitat Pompeu Fabra Economics Working Paper No. 1686.

Galí, J. (2015): *Monetary Policy, Inflation, and the Business Cycle: an Introduction to the New Keynesian Framework and its Applications*, Princeton University Press.

Gertler, M. and P. Karadi (2011): "A model of unconventional monetary policy," *Journal of monetary Economics*, 58, 17–34.

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

Kekre, R. and M. Lenel (2019): "Monetary Policy, Redistribution, and Risk Premia," Tech. rep., Princeton University.

Kurtzman, R. J. and D. Zeke (2020): "Misallocation Costs of Digging Deeper Into the Central Bank Toolkit," *Review of Economic Dynamics.*

La'O, J. and A. Tahbaz-Salehi (2020): "Optimal Monetary Policy in Production Networks," Tech. rep., National Bureau of Economic Research.

Lo, A. W. (2002): "The statistics of Sharpe ratios," *Financial analysts journal*, 58, 36–52.

Lucas Jr, R. E. and N. L. Stokey (1983): "Optimal fiscal and monetary policy in an economy without capital," *Journal of monetary Economics*, 12, 55–93.

Mehra, R. and E. C. Prescott (1985): "The equity premium: A puzzle," *Journal of monetary Economics*, 15, 145–161.

Nuño, G., C. Thomas, et al. (2017): "Optimal monetary policy with heterogeneous agents," Tech. rep., Mimeo New York.

Ottonello, P. and T. Winberry (2018): "Financial Heterogeneity and the Investment Channel of Monetary Policy," Tech. rep., National Bureau of Economic Research.

Rotemberg, J. J. (1982): "Sticky prices in the United States," *Journal of Political Economy*, 90, 1187–1211.

Woodford, M. (2010): "Optimal Monetary Stabilization Policy," in *Handbook of Monetary Economics*, Elsevier, vol. 3, 723–828.

Appendix

A Derivations and Proofs

This appendix provides detailed derivations and proofs for the results in the body of the paper.

A.1 Equilibrium Dynamics

Proof of Proposition 1. The labor supply condition and resource constraint are directly stated in the text. Aggregating the labor demand condition is straightforward and simply involves rearranging and using the fact that $\int Y_{it}^{\nu} di = Y_t^{\nu}$.

To derive the aggregate production function and TFP, note that the labor demand condition implies $\frac{L_{it}}{L_t} = \left(\frac{Y_{it}}{Y_t}\right)^{\nu}$ and substituting into the firm-level production function,

$$Y_{it} = A_{it} K_{it}^{\alpha_1} \left(\left(\frac{Y_{it}}{Y_t} \right)^{\nu} L_t \right)^{\alpha_2}$$

so that

$$Y_{it}^{\nu} = \left(A_{it} K_{it}^{\alpha_1} \left(Y_t^{-\nu} L_t \right)^{\alpha_2} \right)^{\frac{\nu}{1-\alpha_2 \nu}}$$

Substituting into the final good production function (2) and rearranging yields

$$Y_t = \Psi_t K_t^{\alpha_1} L_t^{\alpha_2}$$

where

$$\Psi_t = \left(\int \left(A_{it}^{\frac{\nu}{1-\alpha_2 \nu}} \left(\frac{K_{it}}{K_t} \right)^{\alpha} \right) di \right)^{\frac{1}{1-\alpha}}$$

To solve for relative capital, integrate firm-level capital choices in (5), impose capital market clearing and rearrange.

□

Proof of Proposition 2. First, we derive expressions for L_t and W_t as a function of Y_t and distortions. We can combine the labor supply and labor demand conditions from Proposition 1 along with the final good resource constraint to obtain

$$\begin{aligned} L_t &= \left(\frac{\alpha_2 \nu}{\chi} \right)^{\frac{1}{1+\varphi}} T_{lt}^{\frac{1}{1+\varphi}} Y_t^{\frac{1-\gamma}{1+\varphi}} \\ W_t &= (\alpha_2 \nu)^{1-\frac{1}{1+\varphi}} \chi^{\frac{1}{1+\varphi}} T_{lt}^{-\frac{1}{1+\varphi}} Y_t^{1-\frac{1-\gamma}{1+\varphi}} \end{aligned} \tag{29}$$

and substituting into the production function from Proposition 1 and rearranging yields an expression for Y_t as a function of distortions and TFP only:

$$Y_t = \left(\left(\frac{\alpha_2\nu}{\chi}\right)^{\frac{\alpha_2}{1+\varphi}} K_t^{\alpha_1}\right)^{\frac{1}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}} \Psi_t^{\phi_\psi} T_{lt}^{\phi_l} \tag{30}$$

where ϕ_ψ and ϕ_l are as defined in the text.

Using this result along with the expressions for W_t in (29), and for Y_t in (30) and rearranging, we can write the expectation of the stochastic terms in the capital choice in (5) as

$$\begin{aligned}
& \mathbb{E}_{t-1}\left[\Lambda_t T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} Y_t^{\frac{1-\nu}{1-\alpha_2\nu}} W_t^{-\frac{\alpha_2\nu}{1-\alpha_2\nu}}\right] \\
= \;& \left((\alpha_2\nu)^{\varphi}\chi\right)^{-\frac{1}{1+\varphi}\frac{\alpha_2\nu}{1-\alpha_2\nu}} \rho Y_{t-1}^{\gamma} \mathbb{E}_{t-1}\left[T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} T_{lt}^{\frac{1}{1+\varphi}\frac{\alpha_2\nu}{1-\alpha_2\nu}} Y_t^{\frac{1-\nu}{1-\alpha_2\nu}-\gamma-\frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1-\frac{1-\gamma}{1+\varphi}\right)}\right] \qquad (31) \\
= \;& \left((\alpha_2\nu)^{\varphi}\chi\right)^{-\frac{1}{1+\varphi}\frac{\alpha_2\nu}{1-\alpha_2\nu}} \left(\left(\frac{\alpha_2\nu}{\chi}\right)^{\frac{\alpha_2}{1+\varphi}} K_t^{\alpha_1}\right)^{\frac{\frac{1-\nu}{1-\alpha_2\nu}-\gamma-\frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1-\frac{1-\gamma}{1+\varphi}\right)}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}} \rho Y_{t-1}^{\gamma} \\
\times \;& \mathbb{E}_{t-1}\left[T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} T_{lt}^{\frac{1}{1+\varphi}\frac{\alpha_2\nu}{1-\alpha_2\nu}+\frac{\alpha_2}{1+\varphi}\frac{\frac{1-\nu}{1-\alpha_2\nu}-\gamma-\frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1-\frac{1-\gamma}{1+\varphi}\right)}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}} \Psi_t^{\frac{\frac{1-\nu}{1-\alpha_2\nu}-\gamma-\frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1-\frac{1-\gamma}{1+\varphi}\right)}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}}\right] \\
= \;& \left((\alpha_2\nu)^{\varphi}\chi\right)^{-\frac{1}{1+\varphi}\frac{\alpha_2\nu}{1-\alpha_2\nu}} \left(\left(\frac{\alpha_2\nu}{\chi}\right)^{\frac{\alpha_2}{1+\varphi}} K_t^{\alpha_1}\right)^{\frac{1-\gamma}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}-\frac{\nu}{1-\alpha_2\nu}} \rho Y_{t-1}^{\gamma} \\
\times \;& \mathbb{E}_{t-1}\left[T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} T_{lt}^{\frac{\alpha_2}{1+\varphi}\frac{1-\gamma}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}} \Psi_t^{\frac{1-\gamma}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}-\frac{\nu}{1-\alpha_2\nu}}\right] \\
= \;& \text{Const.} \times \mathbb{E}_{t-1}\left[T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} T_{lt}^{-\kappa_l} \Psi_t^{-\kappa_\psi}\right]
\end{aligned}$$

where κ_ψ and κ_l are as defined in the text and the constant term varies through time, but not across firms.

We conjecture (and later verify) a log-linear form for aggregate productivity as a function of the exogenous shock, i.e.,

$$\psi_t \equiv \log \Psi_t = \overline{\psi} + \psi_a a_t$$

and use this, along with the functional forms of the distortions, to write the capital choice in logs as

$$
\begin{aligned}
k_{it} &= \frac{1}{1-\alpha}\log\left(\mathbb{E}_{t-1}\left[\Lambda_t T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} Y_t^{\frac{1-\nu}{1-\alpha_2\nu}} W_t^{-\frac{\alpha_2\nu}{1-\alpha_2\nu}}\right]\right) + \text{const.} \qquad (32)\\
&= \frac{1}{1-\alpha}\log\left(\mathbb{E}_{t-1}\left[T_{\Lambda t} A_{it}^{\frac{\nu}{1-\alpha_2\nu}} T_{lt}^{-\kappa_l} \Psi_t^{-\kappa_\psi}\right]\right) + \text{const.}\\
&= \frac{1}{1-\alpha}\log\left(\mathbb{E}_{t-1}\left[e^{\frac{\nu}{1-\alpha_2\nu}a_{it} - \left(\kappa_\psi\psi_a + \tau_{\Lambda a} + \kappa_l\tau_{la}\right)a_t}\right]\right) + \text{const.}\\
&= \frac{1}{1-\alpha}\log\left(\mathbb{E}_{t-1}\left[e^{\frac{\nu}{1-\alpha_2\nu}a_{it} - \kappa a_t}\right]\right) + \text{const.}
\end{aligned}
$$

where κ is as defined in equation (20) and, again, the constant picks up terms that may vary through time (it is constant in the i.i.d. case), but not across firms.

The text discusses the intuition behind the first two terms in κ. To gain intuition for the third, $\kappa_l\tau_{la}$, write

$$
-\kappa_l = \frac{1}{1+\varphi}\frac{\alpha_2\nu}{1-\alpha_2\nu} + \frac{\alpha_2}{1+\varphi}\frac{1}{1-\alpha_2\frac{1-\gamma}{1+\varphi}}\left(\frac{1-\nu}{1-\alpha_2\nu} - \gamma - \frac{\alpha_2\nu}{1-\alpha_2\nu}\left(1 - \frac{1-\gamma}{1+\varphi}\right)\right)
$$

The first term in the expression captures the direct effect of τ_{lt} on profitability through changes in the wage, which is the product of the elasticity of profits with respect to the wage, $-\frac{\alpha_2\nu}{1-\alpha_2\nu}$ and the elasticity of the wage with respect to the wedge, $-\frac{1}{1+\varphi}$. The second term captures the effect of the wedge on profitability through changes in output due to labor supply, which equals the elasticity of profits to output, which is the term in parentheses, multiplied by the elasticity of output with respect to the wedge. Multiplying through by τ_{la} translates the term into the elasticity with respect to a_t.

Next, we derive the approximate solution outlined in the Proposition and then the exact solution. First, we provide an exact expression for firm productivity.

Firm productivity. The exact expression for firm productivity is

$$
a_{it} = \hat{\beta}_i a_t - \xi_0 - \xi_\beta\hat{\beta}_i^2 - \xi_{a_{-1}}\delta a_{t-1} - \xi_{a_{-1}^2}\left(\delta a_{t-1}\right)^2 - \xi_a\varepsilon_t^2 - \xi_{a_{-1},a}\delta a_{t-1}\varepsilon_t \qquad (33)
$$

We define the adjustment terms below to ensure that when $\kappa = 0$ or $\sigma_{\hat{\beta}}^2 = 0$, TFP satisfies $\psi_t = a_t$. All of the adjustment terms are independent of risk and policy and so do not affect the main results.

Approximate solution. Following a standard aggregation approach (see, e.g. David and Venkateswaran (2019)), we can derive aggregate output as

$$
\begin{aligned}
Y_t &= \frac{\left(\int A_{it}^{\frac{\nu}{1-(\alpha_1+\alpha_2)\nu}} MRPK_{it}^{-\frac{\alpha_1\nu}{1-(\alpha_1+\alpha_2)\nu}} di\right)^{\frac{1-\alpha_2\nu}{\nu}}}{\left(\int A_{it}^{\frac{\nu}{1-(\alpha_1+\alpha_2)\nu}} MRPK_{it}^{-\frac{1-\alpha_2\nu}{1-(\alpha_1+\alpha_2)\nu}} di\right)^{\alpha_1}} K_t^{\alpha_1} L_t^{\alpha_2} \\
&= \Psi_t K_t^{\alpha_1} L_t^{\alpha_2}
\end{aligned}
$$

and taking logs,

$$
\begin{aligned}
\psi_t &= \frac{1-\alpha_2\nu}{\nu} \log\left(\int A_{it}^{\frac{\nu}{1-(\alpha_1+\alpha_2)\nu}} MRPK_{it}^{-\frac{\alpha_1\nu}{1-(\alpha_1+\alpha_2)\nu}} di\right) \\
&- \alpha_1 \log\left(\int A_{it}^{\frac{\nu}{1-(\alpha_1+\alpha_2)\nu}} MRPK_{it}^{-\frac{1-\alpha_2\nu}{1-(\alpha_1+\alpha_2)\nu}} di\right)
\end{aligned}
$$

A second order approximation yields

$$
\psi_t = \overline{a}_t + \frac{\nu}{1-(\alpha_1+\alpha_2)\nu}\frac{\sigma_{a_t}^2}{2} - \frac{\alpha_1(1-\alpha_2\nu)}{1-(\alpha_1+\alpha_2)\nu}\frac{\sigma_{mrpk_t}^2}{2} \tag{34}
$$

where $\overline{a}_t$ and $\sigma_{a_t}^2$ are the mean and cross-sectional variance of of a_{it}, respectively, and $\sigma_{mrpk_t}^2$ is the cross-sectional variance of $mrpk_{it}$.

Next, to derive expressions for k_{it} and $\sigma_{mrpk_t}^2$, we take a second order approximation to the expectational term in (32) around the means of a_{it} and a_t, which yields

$$
\begin{aligned}
\mathbb{E}_{t-1}\left[A_{it}^{\frac{\nu}{1-\alpha_2\nu}} A_t^{-\kappa}\right] &\approx e^{\frac{\nu}{1-\alpha_2\nu}\overline{a}_i - \kappa\overline{a} + \left(\frac{\nu}{1-\alpha_2\nu}\right)^2 \frac{\text{var}_{t-1}(a_{it})}{2} + \kappa^2\frac{\text{var}_{t-1}(a_t)}{2} - \kappa\frac{\nu}{1-\alpha_2\nu}\text{cov}(a_{it},a_t)} \\
&= e^{\left(\frac{\nu}{1-\alpha_2\nu}\right)^2 \frac{\hat{\beta}_i^2\sigma_\varepsilon^2 + \xi_a^2 \text{var}\left(\varepsilon_t^2\right)}{2} + \kappa^2\frac{\sigma_\varepsilon^2}{2} - \kappa\frac{\nu}{1-\alpha_2\nu}\hat{\beta}_i\sigma_\varepsilon^2}
\end{aligned}
$$

where the second line uses the functional form in (33) (here, we set $\xi_0 = \xi_\beta = 0$) and the special case in Proposition 2 of i.i.d. shocks. Substituting,

$$
\begin{aligned}
k_{it} &= \frac{1-\alpha_2\nu}{1-(\alpha_1+\alpha_2)\nu}\left(\hat{\beta}_i^2\left(\frac{\nu}{1-\alpha_2\nu}\right)^2\frac{\sigma_\varepsilon^2}{2} - \kappa\frac{\nu}{1-\alpha_2\nu}\hat{\beta}_i\sigma_\varepsilon^2\right) + \text{const.} \\
&\approx \text{const.} - \frac{1-\alpha_2\nu}{1-(\alpha_1+\alpha_2)\nu}\kappa\frac{\nu}{1-\alpha_2\nu}\hat{\beta}_i\sigma_\varepsilon^2 \\
&= \text{const.} - \frac{1}{1-\alpha}\beta_i\kappa\sigma_\varepsilon^2
\end{aligned}
$$

where the second line suppresses the $\hat{\beta}_i^2$ terms that stem from differences in the variance of shocks. This is expression (16).

Next,

$$\begin{aligned} mrpk_{it} &\approx \frac{\nu}{1-\alpha_2\nu}a_{it} - \frac{1-(\alpha_1+\alpha_2)\nu}{1-\alpha_2\nu}k_{it} + \text{const.} \\ &= \hat{\beta}_i\left(\frac{\nu}{1-\alpha_2\nu}a_t + \kappa\frac{\nu}{1-\alpha_2\nu}\sigma_\varepsilon^2\right) + \text{const.} \end{aligned}$$

and the cross-sectional variance is

$$\begin{aligned} \sigma_{mrpk_t}^2 &= \left(\frac{\nu}{1-\alpha_2\nu}a_t + \kappa\frac{\nu}{1-\alpha_2\nu}\sigma_\varepsilon^2\right)^2\sigma_{\hat{\beta}}^2 \\ &= \left(\frac{\nu}{1-\alpha_2\nu}a_t\right)^2\sigma_{\hat{\beta}}^2 + 2\left(\frac{\nu}{1-\alpha_2\nu}\right)^2\kappa\sigma_\varepsilon^2\sigma_{\hat{\beta}}^2 a_t + \left(\frac{\nu}{1-\alpha_2\nu}\kappa\sigma_\varepsilon^2\right)^2\sigma_{\hat{\beta}}^2 \end{aligned}$$

and letting $\xi_a = \frac{\nu}{1-\alpha_2\nu}\frac{\sigma_{\hat{\beta}}^2}{2}$ and substituting into (34), we obtain

$$\psi_t = a_t\left(1 - \frac{\alpha_1(1-\alpha_2\nu)}{1-(\alpha_1+\alpha_2)\nu}\left(\frac{\nu}{1-\alpha_2\nu}\right)^2\kappa\sigma_\varepsilon^2\sigma_{\hat{\beta}}^2\right) - \frac{\alpha_1(1-\alpha_2\nu)}{1-(\alpha_1+\alpha_2)\nu}\left(\frac{\nu}{1-\alpha_2\nu}\right)^2\left(\kappa\sigma_\varepsilon^2\right)^2\frac{\sigma_{\hat{\beta}}^2}{2}$$

or

$$\psi_t = \overline{\psi} + \psi_a a_t$$

where, using the definitions of β, σ_β^2, α, ω and κ,

$$\begin{aligned} \psi_a &= 1 - \omega\kappa\sigma_\varepsilon^2\sigma_\beta^2 \\ &= 1 - \omega\left(\kappa_\psi\psi_a + \tau_{\Lambda a} + \kappa_l\tau_{la}\right)\sigma_\varepsilon^2\sigma_\beta^2 \\ &= \frac{1-\left(\tau_{\Lambda a}+\kappa_l\tau_{la}\right)\omega\sigma_\varepsilon^2\sigma_\beta^2}{1+\kappa_\psi\omega\sigma_\varepsilon^2\sigma_\beta^2} \end{aligned}$$

and

$$\begin{aligned} \overline{\psi} &= -\omega\left(\kappa\sigma_\varepsilon^2\right)^2\frac{\sigma_\beta^2}{2} \\ &= -\omega\left(\left(\kappa_\psi\psi_a + \tau_{\Lambda a} + \kappa_l\tau_{la}\right)\sigma_\varepsilon^2\right)^2\frac{\sigma_\beta^2}{2} \\ &= -\omega\left(\left(\frac{\kappa_\psi + \tau_{\Lambda a} + \kappa_l\tau_{la}}{1+\kappa_\psi\omega\sigma_\varepsilon^2\sigma_\beta^2}\right)\sigma_\varepsilon^2\right)^2\frac{\sigma_\beta^2}{2} \end{aligned}$$

and

$$\kappa = \frac{\kappa_\psi + \tau_{\Lambda a} + \kappa_l \tau_{la}}{1 + \kappa_\psi \omega \sigma_\varepsilon^2 \sigma_\beta^2}$$

which are equations (17)-(19) in Proposition 2 and the full solutions.

Exact solution. For our welfare results, we work with an exact solution rather than the more transparent approximation. We also generalize the result to the non-i.i.d. case, which applies to the quantitative analysis in Section 3. The i.i.d. case is nested when $\delta = 0$. With appropriate assumptions on the correction terms in (33) we can obtain just such a solution.

Substituting (32) into the expression for relative capital in Proposition 1 and then into the expression for TFP, we obtain:

$$\Psi_t = \frac{\left(\int \left(A_{it}\left(\mathbb{E}_{t-1}\left[A_{it}^{\frac{\nu}{1-\alpha_2\nu}} A_t^{-\kappa}\right]\right)^{\alpha_1 \frac{1-\alpha_2\nu}{1-(\alpha_1+\alpha_2)\nu}}\right)^{\frac{\nu}{1-\alpha_2\nu}} di\right)^{\frac{1-\alpha_2\nu}{\nu}}}{\left(\int \left(\mathbb{E}_{t-1}\left[A_{it}^{\frac{\nu}{1-\alpha_2\nu}} A_t^{-\kappa}\right]\right)^{\frac{1-\alpha_2\nu}{1-(\alpha_1+\alpha_2)\nu}} di\right)^{\alpha_1}}$$

To explicitly evaluate the time-series expectation and the cross-sectional integrals, we make use of the fact that for a normal random variable $x \sim \mathcal{N}(\bar{x}, \sigma_x^2)$ where $1 - 2b\sigma_x^2 > 0$, the properties of Gaussian integrals imply

$$\mathbb{E}\left[e^{ax+bx^2}\right] = \frac{1}{\sqrt{1-2b\sigma_x^2}} e^{\frac{a\bar{x} + \frac{a^2}{2}\sigma_x^2 + b\bar{x}^2}{1-2b\sigma_x^2}} \tag{35}$$

Defining the terms in equation (33) as

$$\begin{aligned}
\xi_0 &= -\frac{\alpha_1}{2}\log\left(1+\sigma_\beta^2\sigma_\varepsilon^2\right) - \alpha_1\left(\frac{\nu}{1-\alpha_2\nu}\right)^2\frac{\sigma_\varepsilon^2}{2} \\
\xi_\beta &= \alpha_1\frac{\left(\frac{\nu}{1-\alpha_2\nu}\right)^2\frac{\sigma_\varepsilon^2}{2}}{1+\sigma_\beta^2\sigma_\varepsilon^2} \\
\xi_{a_{-1}} &= -\frac{\alpha\sigma_\beta^2\sigma_\varepsilon^2}{1-\alpha+\sigma_\beta^2\sigma_\varepsilon^2} \\
\xi_{a_{-1}^2} &= \frac{\nu}{1-\alpha_2\nu}(1-\alpha)\left(\frac{1+\sigma_\beta^2\sigma_\varepsilon^2}{1-\alpha+\sigma_\beta^2\sigma_\varepsilon^2}\right)^2\frac{\sigma_{\hat\beta}^2}{2} \\
\xi_a &= \frac{\nu}{1-\alpha_2\nu}\frac{\sigma_{\hat\beta}^2}{2} \\
\xi_{a_{-1},a} &= \frac{\nu}{1-\alpha_2\nu}\frac{1+\sigma_\beta^2\sigma_\varepsilon^2}{1-\alpha+\sigma_\beta^2\sigma_\varepsilon^2}\sigma_{\hat\beta}^2
\end{aligned}$$

and repeatedly applying result (35), some lengthy but relatively straightforward algebra yields

$$\psi_t = \overline{\psi} + \delta a_{t-1} + \psi_a\varepsilon_t$$

where

$$\psi_a = 1-\omega\kappa\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right) \tag{36}$$

$$\overline{\psi} = \omega\frac{\nu}{1-\alpha_2\nu}\kappa\sigma_\varepsilon^2\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right) - \alpha_1\left(\frac{\frac{1}{1-\alpha}\kappa\sigma_\varepsilon^2}{1+\sigma_\beta^2\sigma_\varepsilon^2}\right)^2\left(1-\alpha+\sigma_\beta^2\sigma_\varepsilon^2\right)\frac{\sigma_\beta^2}{2} \tag{37}$$

and $\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right) = \frac{\sigma_\beta^2\sigma_\varepsilon^2}{1+\sigma_\beta^2\sigma_\varepsilon^2}$. Finally, substitute for κ into the expression for ψ_a to obtain:

$$\psi_a = \frac{1-\left(\tau_{\Lambda a}+\kappa_l\tau_{la}\right)\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}$$

and

$$\kappa = \frac{\kappa_\psi+\tau_{\Lambda a}+\kappa_l\tau_{la}}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}$$

which are the exact analogs to the approximate solution.

□

A.2 Welfare and Optimal Policy

Welfare criterion. A second order approximation to the utility function yields

$$\mathbb{W} \equiv \frac{\mathbb{U} - \overline{\mathbb{U}}}{U_{\overline{C}}\overline{C}} = (1-\rho)\,\mathbb{E}_{-1}\sum_{t=0}^{\infty}\left(c_t - \alpha_2 l_t + (1-\gamma)\frac{c_t^2}{2} - \alpha_2(1+\varphi)\frac{l_t^2}{2}\right)$$

where we have also assumed the optimal time-invariant production subsidy, $\frac{1}{\nu}$, is in place.

To evaluate the square terms, we use the following first order expressions:

$$\begin{aligned}
c_t &= \phi_\psi \psi_t + \phi_l \mu_t \\
&= \left(\phi_\psi \delta + \phi_l \mu_{a_{-1}}\right) a_{t-1} + \left(\phi_\psi \psi_a + \phi_l \mu_a\right)\varepsilon_t \\
l_t &= \frac{1-\gamma}{1+\varphi} c_t + \frac{1}{1+\varphi}\mu_t \\
&= \left(\frac{1-\gamma}{1+\varphi}\phi_\psi \delta + \frac{1}{1+\varphi}\phi_\psi \mu_{a_{-1}}\right) a_{t-1} + \left(\frac{1-\gamma}{1+\varphi}\phi_\psi \psi_a + \frac{1}{1+\varphi}\phi_\psi \mu_a\right)\varepsilon_t
\end{aligned}$$

from which we can solve

$$\begin{aligned}
\mathbb{E}_{-1}\left[\sum \rho^t c_t^2\right] &= \frac{\left(\phi_\psi \psi_a + \phi_l \mu_a\right)^2 \sigma_\varepsilon^2}{1-\rho} + \frac{\rho\left(\phi_\psi \delta + \phi_l \mu_{a_{-1}}\right)^2 \sigma_\varepsilon^2}{(1-\rho)(1-\rho\delta^2)} \\
\mathbb{E}_{-1}\left[\sum \rho^t l_t^2\right] &= \frac{\left(\frac{1-\gamma}{1+\varphi}\phi_\psi \psi_a + \frac{1}{1+\varphi}\phi_\psi \mu_a\right)^2 \sigma_\varepsilon^2}{1-\rho} + \frac{\rho\left(\frac{1-\gamma}{1+\varphi}\phi_\psi \delta + \frac{1}{1+\varphi}\phi_\psi \mu_{a_{-1}}\right)^2 \sigma_\varepsilon^2}{(1-\rho)(1-\rho\delta^2)}
\end{aligned}$$

To evaluate the linear terms, we use the following second order approximation to the resource constraint

$$c_t = y_t - \frac{\theta_w}{2}\left(\pi_t^w\right)^2$$

where the second term captures the losses from wage adjustment costs. Substituting,

$$c_t = \psi_t + \alpha_2 l_t - \frac{\theta_w}{2}\left(\pi_t^w\right)^2$$

so that

$$\begin{aligned}
c_t - \alpha_2 l_t &= \psi_t - \frac{\theta_w}{2}\left(\pi_t^w\right)^2 \\
&= \psi_t - \frac{1}{2}\frac{\alpha_2 \nu_w}{\lambda_w}\left(\pi_t^w\right)^2
\end{aligned}$$

and

$$\mathbb{E}_{-1}\left[c_t - \alpha_2 l_t\right] = \overline{\psi} - \frac{\alpha_2 \nu_w}{\lambda_w}\frac{\mathbb{E}_{-1}\left[\left(\pi_t^w\right)^2\right]}{2}$$

To solve for π_t^w, we use the Phillips Curve. Conjecture that

$$\pi_t^w = \zeta_{a_{-1}}^w a_{t-1} + \zeta_a^w \varepsilon_t$$

Then,

$$\zeta_{a_{-1}}^w a_{t-1} + \zeta_a^w \varepsilon_t = \rho \zeta_{a_{-1}}^w \left(\delta a_{t-1} + \varepsilon_t\right) + \lambda_w \left(\mu_{a_{-1}} a_{t-1} + \mu_a \varepsilon_t\right)$$

and the method of undetermined coefficients yields

$$\begin{aligned}
\zeta_{a_{-1}}^w &= \frac{\lambda_w \mu_{a_{-1}}}{1-\rho\delta} \\
\zeta_a^w &= \frac{\rho\lambda_w \mu_{a_{-1}}}{1-\rho\delta} + \lambda_w \mu_a
\end{aligned}$$

Thus,

$$\pi_t^w = \frac{\lambda_w \mu_{a_{-1}}}{1-\rho\delta} a_{t-1} + \left(\frac{\rho\lambda_w \mu_{a_{-1}}}{1-\rho\delta} + \lambda_w \mu_a\right) \varepsilon_t$$

and

$$\mathbb{E}_{-1}\left[\sum \rho^t \left(\pi_t^w\right)^2\right] = \frac{\left(\frac{\rho\lambda_w \mu_{a_{-1}}}{1-\rho\delta} + \lambda_w \mu_a\right)^2 \sigma_\varepsilon^2}{1-\rho} + \frac{\rho \left(\frac{\lambda_w \mu_{a_{-1}}}{1-\rho\delta}\right)^2 \sigma_\varepsilon^2}{(1-\rho)\left(1-\rho\delta^2\right)}$$

Finally, substituting back into welfare function and suppressing terms that are independent of policy:

$$\begin{aligned}
\mathbb{W} &= \overline{\psi} - (\gamma-1)\,\phi_\psi \psi_a^2 \frac{\sigma_\varepsilon^2}{2} - \left(\mu_a^2 + \frac{\rho}{1-\rho\delta^2}\mu_{a_{-1}}^2\right)\phi_l \frac{\sigma_\varepsilon^2}{2} \\
&- \frac{\alpha_2 \nu_w}{\lambda_w}\left(\left(\frac{\rho\lambda_w \mu_{a_{-1}}}{1-\rho\delta} + \lambda_w \mu_a\right)^2 - \frac{\rho}{1-\rho\delta^2}\left(\frac{\lambda_w \mu_{a_{-1}}}{1-\rho\delta}\right)^2\right)\frac{\sigma_\varepsilon^2}{2}
\end{aligned} \tag{38}$$

The negative of the expression is the welfare loss function in Section 3 and in the i.i.d. case where $\delta = \mu_{a_{-1}} = 0$ in expression (24).

Proof of Proposition 3. To derive (22) the planner chooses κ to maximize (38), taking all else as given. To derive the optimal κ, note that from (36) and (37), the derivatives of ψ_t with respect to κ are

$$\begin{aligned}
\frac{\partial \psi_a}{\partial \kappa} &= -\omega \Upsilon \left(\sigma_\beta^2 \sigma_\varepsilon^2\right) \\
\frac{\partial \overline{\psi}}{\partial \kappa} &= \omega \frac{\nu}{1-\alpha_2 \nu} \sigma_\varepsilon^2 \Upsilon \left(\sigma_\beta^2 \sigma_\varepsilon^2\right) - \alpha_1 \kappa \left(\frac{\frac{1}{1-\alpha}\sigma_\varepsilon^2}{1+\sigma_\beta^2 \sigma_\varepsilon^2}\right)^2 \left(1-\alpha+\sigma_\beta^2 \sigma_\varepsilon^2\right)\sigma_\beta^2
\end{aligned}$$

Substituting these expressions and equation (36) into the first order condition (22) and simplifying yields the optimal κ:

$$\kappa^* = \frac{\kappa_\psi}{1 + \kappa_\psi \omega \Upsilon \left(\sigma_\beta^2 \sigma_\varepsilon^2\right)}$$

which satisfies

$$\kappa^* = \psi_a^* \kappa_\psi$$

where ψ_a^* denotes the value of ψ_a at κ^*.

$\square$

Proof of Proposition 4. The optimal policy in the general model is a pair μ_{a-1}, μ_a that maximizes (38), accounting for the effects on $\overline{\psi}$ and ψ_a. The first order conditions give

$$\begin{aligned}
0 &= \frac{\partial \overline{\psi}}{\partial \mu_a} - \alpha_2 \nu_w \left(\frac{\rho \lambda_w \mu_{a-1}}{1 - \rho\delta} + \lambda_w \mu_a \right) \sigma_\varepsilon^2 + (1 - \gamma) \phi_\psi \sigma_\varepsilon^2 \psi_a \frac{\partial \psi_a}{\partial \mu_a} - \phi_l \mu_a \sigma_\varepsilon^2 \\
0 &= -\alpha_2 \nu_w \left(\left(\frac{\rho \lambda_w \mu_{a-1}}{1 - \rho\delta} + \lambda_w \mu_a \right) \frac{\rho}{1 - \rho\delta} + \frac{\rho}{1 - \rho\delta^2} \frac{\lambda_w \mu_{a-1}}{1 - \rho\delta} \frac{1}{1 - \rho\delta} \right) \sigma_\varepsilon^2 - \frac{\rho}{1 - \rho\delta^2} \phi_l \mu_{a-1} \sigma_\varepsilon^2
\end{aligned}$$

Substituting for the derivatives and rearranging, we can obtain the following matrix equation representation:

$$A\mu = B$$

where

$$\mu = \begin{bmatrix} \mu_{a-1} & \mu_a \end{bmatrix}'$$

and the elements of A and B are:

$$
\begin{aligned}
A_{11} &= \frac{\rho}{1-\rho\delta^2}\phi_l + \frac{\alpha_2\nu_w\rho\lambda_w}{(1-\rho\delta)^2}\left(\rho + \frac{1}{1-\rho\delta^2}\right) \\
A_{12} &= \frac{\alpha_2\nu_w\rho\lambda_w}{1-\rho\delta} \\
B_{11} &= 0 \\
A_{21} &= \frac{\alpha_2\nu_w\rho\lambda_w\sigma_\varepsilon^2}{1-\rho\delta} \\
A_{22} &= \alpha_1\left(\frac{\kappa_l}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}\right)^2\left(\frac{\frac{1}{1-\alpha}\sigma_\varepsilon^2}{1+\sigma_\beta^2\sigma_\varepsilon^2}\right)^2\left(1-\alpha+\sigma_\beta^2\sigma_\varepsilon^2\right)\sigma_\beta^2 \\
&+ \alpha_2\nu_w\lambda_w\sigma_\varepsilon^2 + \phi_l\sigma_\varepsilon^2 - (1-\gamma)\phi_\psi\sigma_\varepsilon^2\left(\frac{\kappa_l\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}\right)^2 \\
B_{21} &= \kappa_l\frac{\omega\frac{\nu}{1-\alpha_2\nu}\sigma_\varepsilon^2\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right) - \alpha_1\frac{\kappa_\psi+\tau_{\Lambda a}}{1-\kappa_\psi\omega\Upsilon}\left(\frac{\frac{1}{1-\alpha}\sigma_\varepsilon^2}{1+\sigma_\beta^2\sigma_\varepsilon^2}\right)^2\left(1-\alpha+\sigma_\beta^2\sigma_\varepsilon^2\right)\sigma_\beta^2}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)} \\
&- (1-\gamma)\phi_\psi\sigma_\varepsilon^2\left(1-\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)\frac{\kappa_\psi+\tau_{\Lambda a}}{1+\kappa_\psi\omega\Upsilon}\right)\frac{\kappa_l\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon}
\end{aligned}
$$

The first order condition in the i.i.d. case where $\delta = \mu_{a_{-1}} = 0$ yields expression (25) and rearranging yields the result in the proposition.

□

Optimal nominal interest rate. The Euler equation is

$$
y_t = -\frac{1}{\gamma}\left(i_t - \mathbb{E}_t\left[\pi_{t+1}^p\right]\right) + \mathbb{E}_t\left[y_{t+1}\right]
$$

In the i.i.d. case, we have

$$
\begin{aligned}
\pi_t^w &= \lambda_w\mu_a\varepsilon_t \\
&= \zeta_w^w\varepsilon_t \\
y_t &= \left(\phi_\psi\psi_a + \phi_l\mu_a\right)\varepsilon_t \\
&= y_a\varepsilon_t \\
w_t &= \left(1-\frac{1-\gamma}{1+\varphi}\right)y_t - \frac{1}{1+\varphi}\mu_t
\end{aligned}
$$

and by definition,

$$\begin{aligned}\pi_{t+1}^{p} &= w_t - w_{t+1} + \pi_{t+1}^{w} \\ &= \left(\left(1-\frac{1-\gamma}{1+\varphi}\right) y_a - \frac{1}{1+\varphi}\mu_a\right)\varepsilon_t + \left(-\left(1-\frac{1-\gamma}{1+\varphi}\right) + \frac{1}{1+\varphi}\mu_a + \zeta_a^w\right)\varepsilon_{t+1}\end{aligned}$$

so

$$\mathbb{E}_t\left[\pi_{t+1}^{p}\right] = \left(\left(1-\frac{1-\gamma}{1+\varphi}\right) y_a - \frac{1}{1+\varphi}\mu_a\right)\varepsilon_t$$

and substituting into the Euler equation:

$$\begin{aligned}i_t &= \left(\left(1-\frac{1-\gamma}{1+\varphi}\right) y_a - \frac{1}{1+\varphi}\mu_a - \gamma y_a\right)\varepsilon_t \\ &= \left(\frac{(1-\gamma)\varphi}{1+\varphi}\phi_\psi\psi_a - \frac{1-\alpha_2(1-\gamma)}{1+\varphi}\phi_\psi\mu_a\right)\varepsilon_t \\ &= i_a\varepsilon_t\end{aligned}$$

where

$$i_a = \frac{(1-\gamma)\varphi}{1+\varphi}\phi_\psi\frac{1-\tau_{\Lambda a}\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)} + \left(\frac{(1-\gamma)\varphi}{1+\varphi}\frac{\kappa_l\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)} - \frac{1-\alpha_2(1-\gamma)}{1+\varphi}\right)\phi_\psi\mu_a$$

Substituting for the optimal μ_a, we obtain an expression

$$i_a = \Phi^i + \tau_{\Lambda a}\Phi_\Lambda^i$$

as in the main text, where the constants satisfy the properties discussed there.

Taylor Rule. The optimal policy can also be implemented with standard formulations of a Taylor Rule. We specify the rule as

$$i_t = \phi_y\tilde{y}_t + \phi_\pi\mathbb{E}_t\left[\pi_{t+1}^{p}\right]$$

In the general (non-i.i.d. case) we have

$$\begin{aligned}y_t &= \left(\delta\phi_\psi + \phi_l\mu_{a_{-1}}\right)a_{t-1} + \left(\phi_\psi\psi_a + \phi_l\mu_a\right)\varepsilon_t \\ &= y_{a_{-1}}a_{t-1} + y_a\varepsilon_t \\ \tilde{y}_t &= \phi_l\mu_t \\ &= \phi_l\left(\mu_{a_{-1}}a_{t-1} + \mu_a\varepsilon_t\right)\end{aligned}$$

Following similar steps as above, we can solve for expected price inflation as

$$\mathbb{E}_t\left[\pi_{t+1}^p\right] = \zeta_{a_{-1}}^p a_{t-1} + \zeta_a^p \varepsilon_t$$

where

$$\begin{aligned}
\zeta_{a_{-1}}^p &= \left(1-\frac{1-\gamma}{1+\varphi}\right)(1-\delta)\, y_{a_{-1}} - \frac{1-\delta}{1+\varphi}\mu_{a_{-1}} + \delta\zeta_{a_{-1}}^w \\
\zeta_a^p &= \left(1-\frac{1-\gamma}{1+\varphi}\right)\left(y_a - y_{a_{-1}}\right) - \frac{1}{1+\varphi}\left(\mu_a - \mu_{a_{-1}}\right) + \zeta_{a_{-1}}^w
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{E}_t\left[\pi_{t+1}^p\right] &= \left\{\left(1-\frac{1-\gamma}{1+\varphi}\right)(1-\delta)\left(\phi_\psi\delta + \phi_l\mu_{a_{-1}}\right) - \frac{1-\delta}{1+\varphi}\mu_{a_{-1}} + \delta\frac{\lambda_w\mu_{a_{-1}}}{1-\rho\delta}\right\} a_{t-1} \\
&+ \left\{\left(1-\frac{1-\gamma}{1+\varphi}\right)\left(\phi_\psi\left(\psi_a-\delta\right) + \phi_l\left(\mu_a-\mu_{a_{-1}}\right)\right) - \frac{1}{1+\varphi}\left(\mu_a-\mu_{a_{-1}}\right) + \frac{\lambda_w\mu_{a_{-1}}}{1-\rho\delta}\right\}\varepsilon_t
\end{aligned}$$

Substituting into the Taylor Rule, along with the definition of ψ_a:

$$\begin{aligned}
i_t &= \left\{\phi_\pi\left(1-\frac{1-\gamma}{1+\varphi}\right)(1-\delta)\,\delta\phi_\psi\right. \\
&+ \left.\mu_a\left(\left(\left(1-\frac{1-\gamma}{1+\varphi}\right)\phi_l - \frac{1}{1+\varphi}\right)(1-\delta)\,\phi_\pi + \frac{\delta\lambda_w}{1-\rho\delta}\phi_\pi + \phi_l\phi_y\right)\right\} a_{t-1} \\
&+ \left\{\phi_\pi\left(1-\frac{1-\gamma}{1+\varphi}\right)\phi_\psi\frac{1-\tau_{\Lambda a}\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)} - \phi_\pi\left(1-\frac{1-\gamma}{1+\varphi}\right)\phi_\psi\delta\right. \\
&- \mu_{a_{-1}}\left(\left(\left(1-\frac{1-\gamma}{1+\varphi}\right)\phi_l - \frac{1}{1+\varphi}\right)\phi_\pi - \frac{\lambda_w}{1-\rho\delta}\phi_\pi\right) \\
&+ \left.\mu_a\left(\left(\left(1-\frac{1-\gamma}{1+\varphi}\right)\phi_l - \frac{1}{1+\varphi}\right)\phi_\pi + \left(1-\frac{1-\gamma}{1+\varphi}\right)\phi_\psi\frac{\kappa_l\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}\phi_\pi + \phi_l\phi_y\right)\right\}\varepsilon_t
\end{aligned}$$

Following a similar approach as above, the Euler equation gives a second representation of the nominal rate:

$$\begin{aligned} i_t &= \left\{ \frac{(1-\gamma)\varphi}{1+\varphi}(1-\delta)\delta\phi_\psi + \left(\frac{\delta\lambda_w}{1-\rho\delta} - (1-\alpha_2(1-\gamma))\phi_\psi \frac{1-\delta}{1+\varphi} \right) \mu_{a_{-1}} \right\} a_{t-1} \\ &+ \left\{ \frac{(1-\gamma)\varphi}{1+\varphi}\phi_\psi \left(\frac{1-\tau_{\Lambda a}\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)} - \delta \right) \right. \\ &+ \left(\frac{(1-\gamma)\varphi}{1+\varphi} \frac{\kappa_l\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)}{1+\kappa_\psi\omega\Upsilon\left(\sigma_\beta^2\sigma_\varepsilon^2\right)} - \frac{1-\alpha_2(1-\gamma)}{1+\varphi} \right) \phi_\psi\mu_a \\ &+ \left. \left(\frac{1-\alpha_2(1-\gamma)}{1+\varphi}\phi_\psi + \frac{\lambda_w}{1-\rho\delta} \right) \mu_{a_{-1}} \right\} \varepsilon_t \end{aligned}$$

Equating coefficients, we obtain an equation of the form

$$A \begin{bmatrix} \mu_{a_{-1}} \\ \mu_a \end{bmatrix} = B$$

where the elements of the matrices A and B are function of the Taylor Rule coefficients. Thus, there is a one-to-one mapping between the pairs ϕ_y, ϕ_π and $\mu_{a_{-1}}, \mu_a$ so that optimal policy can be implemented with an appropriate Taylor Rule of this form. To see this another way, note that the mapping can be inverted to derive an equation of the form

$$C \begin{bmatrix} \phi_\pi \\ \phi_y \end{bmatrix} = D$$

which shows explicitly that any pair $\mu_{a_{-1}}, \mu_a$ can be implemented with the appropriate choice of ϕ_y, ϕ_π. Note that in the i.i.d. case, $\mu_{a_{-1}} = 0$ and so there are two Taylor Rule coefficients to set one policy parameter. Thus, the mapping from a particular policy to these coefficients is not unique, i.e., there are many combinations of the coefficients that implement the same policy.

B Discount Rate Wedges and Financial Frictions

This appendix provides two detailed examples of financial frictions that lead to a discount rate, or risk wedge as described in the main analysis. The first is based on a simplified version of the framework in Debortoli and Galí (2018), which features limited asset market participation in conjunction with shocks to goods price markups. The second shows that such a wedge arises naturally in recent models of frictional financial intermediation a la Gertler and Karadi (2011).

B.1 Limited Asset Market Participation

The setup follows closely the two agent New Keynesian (TANK) model developed in Debortoli and Galí (2018). There are two types of households of time invariant measures θ and $1-\theta$, respectively. The first type – "constrained" households – do not participate in financial markets and simply consume their labor income in each period. The second type – "unconstrained" – hold all capital and equity shares in firms. Under Rotemberg wage setting frictions, wages are common across households. We assume that employment is uniformly distributed across households.

Because only unconstrained households own capital, the relevant SDF for pricing capital returns is given by $\Lambda_t^U = \rho \left(\frac{C_t^U}{C_{t-1}^U} \right)^{-\gamma}$, where C_t^U denotes the average consumption of an unconstrained household. By definition, aggregate consumption is the sum of consumption across constrained and unconstrained households:

$$C_t = \theta C_T^K + (1-\theta) C_T^U$$

Define $G_t = \frac{C_t^U - C_t^K}{C_t^U}$ as the gap between the average consumption of unconstrained and constrained households. We can rewrite this as

$$C_t = C_t^U (1 - \theta G_t)$$

and log-linearizing and rearranging yields an expression for unconstrained consumption as a function of aggregate consumption and the consumption gap

$$c_t^U = c_t + \frac{\theta}{1-\theta\overline{G}} g_t \tag{39}$$

where $\overline{G}$ denotes the consumption gap in the non-stochastic steady state.

Next, we can also write the gap as a function of the ratio of profits to labor income:

$$G_t = 1 - \frac{C_t^K}{C_t^U} = \frac{\frac{\Pi_t}{W_t L_t}}{1-\theta + \frac{\Pi_t}{W_t L_t}}$$

Firms face a common, but time-varying shock to their price markup, T_t^p, when making labor decisions.[33] Though not necessary, it simplifies the algebra slightly to assume that the markup follows a random walk (in logs), where the innovation is a linear function of the exogenous

[33]More generally, the results go through when T_t^p captures any wedge that leads to a time-varying labor share. For concreteness, we use the example of a time-varying price markup.

aggregate shock, i.e.,

$$\tau_t^p \equiv \log T_t^p = \tau_{t-1}^p + \tau_a^p \varepsilon_t$$

Firm choose labor to satisfy

$$\max_{L_{it}} \; P_{it} Y_{it} - T_t^p W_t L_{it}$$

and first-order conditions, aggregating and rearranging yields

$$\frac{\Pi_t}{W_t L_t} = \frac{Y_t}{W_t L_t} - 1 = \frac{1}{\alpha_2 \nu} T_t^p - 1$$

Use this expression to rewrite the consumption gap as

$$G_t = \frac{\frac{1}{\alpha_2 \nu} T_t^p - 1}{\frac{1}{\alpha_2 \nu} T_t^p - \theta}$$

and log-linearizing,

$$\begin{aligned} g_t &= \Theta \tau_t^p \\ &\text{where} \\ \Theta &= \frac{\frac{1}{\alpha_2 \nu}(1-\theta)}{\left(\frac{1}{\alpha_2 \nu} - \theta\right)\left(\frac{1}{\alpha_2 \nu} - 1\right)} > 0 \end{aligned}$$

Substituting into (39) gives an expression for unconstrained consumption as a function of aggregate consumption and the markup shock:

$$c_t^U = c_t + \frac{\theta}{1 - \theta \overline{G}} \Theta \tau_t^p$$

From here, we obtain an expression for the relevant SDF

$$\begin{aligned} \lambda_t^U &\equiv \log \Lambda_t^U = \lambda_t + \tau_{\Lambda t} \\ &\text{where} \\ \lambda_t &= -\gamma (y_t - y_{t-1}) \\ \tau_{\Lambda t} &= -\tau_{\Lambda a} \varepsilon_t \\ \tau_{\Lambda a} &= \frac{\theta}{1 - \theta \overline{G}} \gamma \Theta \tau_a^p \end{aligned}$$

It is straightforward to see that the "undistorted" SDF, λ_t, takes exactly the same form as in the main text, and $\tau_{\Lambda t}$ takes exactly the form of the risk wedge. Thus, incomplete markets in the form of limited participation in conjunction with cyclical price markup shocks lead to a risk wedge and distorted SDF precisely of the forms assumed in the text. Note that $\tau_{\Lambda a} > 0$ so long as $\tau_a^p > 0$, which is the case when the price markup shock is procyclical, i.e., when the profit share increases in expansions.

B.2 Frictional Financial Intermediation

The setup is a simplified version of that in, e.g., Gertler and Karadi (2011). A continuum of identical financial intermediaries own the equity shares of firms. Intermediaries borrow from households at the risk-free rate in order to provide this funding to firms. The balance sheet of the representative intermediary is given by

$$\int Q_{it} S_{it} di = N_t + D_t \tag{40}$$

Intermediary assets consist of the total market value of its ownership claims on firms, where Q_{it} and S_{it} denote the price and quantity of claims on firm i. Intermediary liabilities consist of deposits, D_t, and equity, or net worth, N_t. Net worth is equal to the gross return on assets less costs of borrowing:

$$N_t = \int R^s_{it} Q_{it-1} S_{it-1} di - R_t D_{t-1} \tag{41}$$

The intermediary has an exogenous net dividend payout rate of $1 - \sigma$. It acts to maximize the expected discounted stream of dividends payed out to households,

$$V_t = \mathbb{E}_t \left[\Lambda_{t+1} \left((1-\sigma) N_{t+1} + \sigma V_{t+1} \right) \right] \tag{42}$$

Due to a moral hazard/costly enforcement problem, intermediaries face collateral constraints that limit their ability to obtain deposits. Specifically, the intermediary can divert a fraction θ of its assets, which leads to the the following incentive constraint that limits the intermediary's collateral:

$$V_t \geq \theta \int Q_{it} S_{it} di \tag{43}$$

The intermediary chooses its holdings of firms' securities, S_{it}, to maximize (42) subject to (40), (41) and (43).

Assume that the collateral constraint binds. We can use the first order conditions of the intermediary's problem and the form of the value function to obtain

$$\mathbb{E}_t \left[\Lambda_{t+1} T_{\Lambda t+1} \left(R^s_{it+1} - R_{t+1} \right) \right] = \theta \frac{\lambda_t}{1 + \lambda_t}, \quad \forall\ i, t$$

where λ_t is the Lagrange multiplier on the collateral constraint and

$$T_{\Lambda t} = 1 - \sigma + \sigma \frac{\partial V_t}{\partial N_t} = 1 - \sigma + \sigma \frac{V_t}{N_t}$$

Thus, the relevant SDF for pricing assets takes exactly the form as in the main text and is equal to the household SDF, Λ_t, multiplied by a term, $T_{\Lambda t}$, that captures the shadow marginal value of net worth to financial intermediaries, which, in this simple setup, is the intermediary leverage ratio. Log-linearizing this latter term gives a functional form for the wedge analogous to the one assumed in the text.[34]

[34]More precisely, there will also be terms that depend on lagged state variables in addition to ε_t. Because these terms are known at time $t-1$ and are common across firms, they do not affect the capital allocation and hence play no role in the main analysis.