

The Supply-Side Effects of Monetary Policy

David R. Baqaee
UCLA

Emmanuel Farhi
Harvard

Kunal Sangani
Harvard*

June 19, 2021

Abstract

We propose a supply-side channel for the transmission of monetary policy. We show that in an economy with heterogeneous firms and endogenous markups, monetary shocks have first-order effects on aggregate productivity. If high-markup firms have lower pass-throughs than low-markup firms, as is consistent with the empirical evidence, then a monetary easing generates an endogenous positive “supply shock” that amplifies the positive “demand shock” on output. The result is a flattening of the Phillips curve. This effect is distinct from another mechanism discussed at length in the real rigidities literature: a monetary easing leads to a reduction in desired markups because of strategic complementarities in pricing. We derive a tractable four-equation dynamic model, disciplined by four sufficient statistics from the distribution of firms, and use it to show that a monetary easing generates a procyclical hump-shaped response in aggregate TFP and countercyclical dispersion in firm-level TFPR. In the data, in line with the model’s predictions, we observe both procyclical aggregate productivity and reallocations to high-markup firms during expansions. A calibration using firm-level pass-throughs suggests that the supply-side channel is quantitatively as important as real rigidities, and amplifies both the impact and persistence of monetary shocks. Moreover, this channel becomes stronger, and the Phillips curve becomes flatter, with increases in industrial concentration.

*Emmanuel Farhi tragically passed away in July, 2020. He was a one-in-a-lifetime friend and collaborator and we dedicate this paper to his memory. David Baqaee and Kunal Sangani are responsible for any errors that remain. We thank Andy Atkeson, Ariel Burstein, Oleg Itskhoki, Ivan Werning, Jon Vogel, and other seminar participants for helpful comments. This paper received support from NSF grant No. 1947611.

1 Introduction

How do demand shocks affect an economy’s productivity? The standard thinking is that they do not: aggregate productivity is taken to be orthogonal to the structural shocks that move aggregate demand, such as monetary shocks.

Yet, aggregate total factor productivity (TFP), as measured by labor productivity or the Solow residual, is sensitive to demand shocks. In fact, variations in monetary and fiscal policy explain between one-quarter and one-half of the observed movements in aggregate TFP at business cycle frequencies (see, e.g. Evans, 1992). This empirical finding is robust across time and across countries.¹ One interpretation of this result is that the relationship between measured productivity and demand shocks is confounded by capacity utilization or external returns, which bias the measurement of aggregate TFP.

In this paper, we offer an alternative explanation. The aggregate TFP of an economy is not an exogenous primitive, but instead an endogenous outcome that depends on how resources are allocated across firms. We argue that in an economy with realistic firm heterogeneity, demand shocks *should* trigger changes in aggregate TFP. These changes do not arise from changes in technical efficiency—the technologies available to individual firms—but instead from shifts in the allocation of resources across firms.

The effect of monetary policy on the cross-sectional allocation of resources yields a new channel for the transmission of monetary policy, which we term *the misallocation channel*. Under conditions matching empirical patterns on firms, monetary shocks generate procyclical, hump-shaped movements in aggregate TFP, which match empirical estimates from Evans (1992), Christiano et al. (2005) and others.² The endogenous “supply shock” generated by the misallocation channel complements the traditional effects of the “demand shock” on employment and output. Incorporating the misallocation channel heightens the response of output to demand shocks and flattens the Phillips curve.

To a first-order, this supply-side effect only appears when two conditions hold: (1) the initial cross-sectional allocation of resources is inefficient, and (2) monetary policy systematically reallocates resources from low to high marginal revenue product firms. We discuss these two conditions in turn. First, if there is no initial misallocation, the

¹The failed invariance of aggregate TFP to demand shocks is also observed by Hall (1990). Cozier and Gupta (1993), Evans and dos Santos (2002), and Kim and Lim (2004) extend the analysis to Canada, the G-7 countries, and South Korea. We show that aggregate productivity in the U.S. is (unconditionally) procyclical for the period 1948-2020 and contracts after Romer and Romer (2004) monetary shocks.

²Christiano et al. (2005) estimate a positive hump-shaped response of labor productivity to monetary easing. In our one-factor model, labor productivity and aggregate TFP are the same. In Section 7, we provide our own empirical estimates of how aggregate productivity responds to identified monetary policy shocks.

marginal benefit of each input is equated across all competing uses. Therefore, starting at an efficient point, a reallocation of resources triggered by monetary policy has no first-order effects on aggregate productivity. Second, even if the initial allocation of resources is distorted, if monetary policy does not differentially affect firms with different marginal revenue products, then reallocations induced by monetary policy do not systematically raise or lower aggregate productivity.

For these reasons, the misallocation channel is absent in the workhorse log-linearized New Keynesian model. First, the benchmark model, which uses a CES demand system, assumes that the price elasticity of demand is constant across firms. This means that desired markups are the same for all firms. As a result, the flexible-price allocation of resources is efficient, and hence, reallocations starting at this point are irrelevant for aggregate productivity. Second, even starting at an equilibrium with markup dispersion, monetary policy does not differentially affect high and low marginal revenue product firms, meaning that even away from the efficient point, there is no reason to expect a monetary easing to increase productivity.

In contrast to the benchmark model, the data features substantial and persistent heterogeneity in markups across firms and systematic differences in the pass-through of marginal cost shocks into prices across the firm size distribution. Because markups are not uniform, firms with relatively high markups underproduce and those with relatively low markups overproduce compared to the efficient allocation. This dispersion in initial markups opens the door for reallocations caused by monetary policy to have first-order effects on aggregate productivity. If the pass-through of marginal cost into the price is higher for low-markup than high-markup firms, a monetary easing (or any shock that raises nominal marginal cost) systematically reallocates resources from low-markup to high-markup firms, and therefore raises aggregate productivity.

To formally analyze these reallocations, we deviate from the CES formulation of the New Keynesian model and adopt a non-parametric generalized Kimball (1995) demand system. Kimball preferences are flexible enough to generate downward-sloping residual demand curves of any desired shape while remaining tractable. We couple this flexible demand system with sticky prices. Our model is flexible enough to exactly match cross-sectional and time-series estimates of the firm-size distribution and firm-level pass-throughs, with realistic heterogeneity in firms' price elasticities of demand and desired markups. We consider how TFP and output respond to monetary shocks in such a model. Our comparative statics do not impose any additional parametric structure on preferences, and are disciplined by measurable sufficient statistics from the distribution of firms.

Our first result is that when firms' pass-throughs covary negatively with their initial

markups, then a positive demand shock, such as a monetary easing, increases aggregate TFP and moves the economy closer to the efficient frontier. This negative relationship between markups and pass-throughs has strong empirical support across countries.³ Intuitively, a monetary easing raises all firms' nominal marginal costs, but high-markup firms, which have lower pass-throughs, raise their prices by less than their low-markup counterparts. This triggers a reallocation toward high-markup firms and away from low-markup firms, which improves allocative efficiency. In principle, this heterogeneity in pass-throughs can be driven either by heterogeneity in desired pass-throughs or heterogeneity in price-stickiness.^{4,5}

Our second result shows that the response of output to a monetary shock can be decomposed into distinct demand-side and supply-side effects. The demand-side effect of an expansionary shock arises from increases in employment due to increased labor demand. Intuitively, expansionary monetary policy raises spending, but nominal rigidities prevent prices from rising by the same amount. This raises labor demand, employment, and—because of this increase in employment—output. These effects are amplified in the presence of real rigidities, which further dampen the responsiveness of prices to increases in nominal marginal costs, due to strategic complementarities in pricing.⁶ Whereas the demand-side effect raises output by raising employment, the supply-side effect boosts output by raising aggregate productivity. When markups negatively covary with pass-throughs, an expansionary shock boosts output by raising aggregate TFP and reducing the dispersion in markups across firms.

Our model suggests that the misallocation channel constitutes a quantitatively important part of monetary policy transmission mechanism. We use cross-sectional firm data from Belgium (provided by Amity et al. 2019) to calibrate a simplified version of the model.⁷ In our static model, we find that the misallocation channel reduces the slope of

³See Berman et al. (2012) in France, Chatterjee et al. (2013) in Brazil, Li et al. (2015) in China, and Amity et al. (2019) in Belgium. We use estimates from Amity et al. (2019) to calibrate the empirical results presented in this paper.

⁴By desired pass-through, we refer to the pass-through conditional on a price change.

⁵We focus on monetary shocks in this paper, but other aggregate demand shocks, such as discount factor shocks, will have similar effects on TFP.

⁶In this paper, when we refer to “real rigidities” we mean real rigidities caused by variable markups, and not real rigidities caused by other forces (e.g. decreasing returns or sticky intermediate input prices).

⁷We follow Baqaee and Farhi (2020a) and solve a series of differential equations to back out the Kimball demand system from data on firm-level sales and pass-throughs. By calibrating to pass-throughs, we avoid difficulties related to the identification of markups at high frequencies, which is notoriously difficult. This approach is also preferable to using an off-the-shelf functional form for preferences, since it does not impose the counterfactual restrictions baked in by parametric families of preferences. We provide an explicit calibration exercise in Appendix G showing that off-the-shelf functional forms are incapable of simultaneously matching all the relevant sufficient statistics in the data.

the Phillips curve by around 70%, compared to a model with demand-side effects alone. As a point of comparison, we find that real rigidities flatten the price Phillips curve by a similar amount.

Since the strength of real rigidities and the misallocation channel are governed by moments of the firm distribution, our analysis ties the strength of monetary policy to the industrial organization of the economy. In particular, we show that an increase in industrial concentration can increase the potency of both the real rigidities and misallocation channels. While the standard New Keynesian model is silent on the role of industrial concentration, in our setup increasing the Gini coefficient of firm employment from 0.80 to 0.85 flattens the Phillips curve by an additional 14%. This increase in the Gini coefficient is in line with the change in the firm employment distribution in the United States from 1978 to 2018.

While we use a static model to illustrate some of the key intuitions driving our results, we also derive a fully dynamic model. We describe the movement of aggregate TFP, output, inflation, and the interest rate using a four-equation system. This augments the classic three-equation model to account for realistic firm heterogeneity and endogenous changes in allocative efficiency. Relative to the workhorse model, the Taylor rule and the Euler equation are the same but the New Keynesian Phillips curve is different. Our model features a flattened Phillips curve and endogenous cost-push shocks due to shifts in aggregate TFP. Those movements in aggregate TFP are pinned down by the fourth equation, which closes the system. These equations are all disciplined by four sufficient statistics from the firm distribution: the average markup, the average price elasticity of demand, the average desired pass-through, and the covariance of markups and desired pass-throughs.

A calibration of the dynamic model shows that the misallocation channel deepens the loss in output following a contractionary interest rate shock by 20% on impact. The role of the misallocation channel also rises over time, increasing the half-life of the shock's effect on output by 23%. The net result is an increase in the cumulative output impact of the monetary shock by 37% compared to the workhorse model.

We find support for both macro- and micro-level predictions of our model. As mentioned above, at the macroeconomic level, our setup predicts procyclical, hump-shaped responses of aggregate TFP to monetary shocks. We confirm that aggregate productivity—as measured by labor productivity, the Solow residual, or the cost-based Solow residual—is procyclical in U.S. data. We also document that aggregate productivity contracts following Romer and Romer (2004) monetary shocks, as shown by Evans (1992).⁸

⁸We do not use capacity-utilization adjusted measures of aggregate TFP, like Basu et al. (2006) or Fernald

At the microeconomic level, our model ties the increase in aggregate productivity during demand-driven expansions to reallocations towards high-markup firms. This prediction is consistent with Compustat data on public firms: firms with above-median markups grow more in sales and costs during expansions. These results are robust to using multiple methodologies to measure markups. In particular, we document similar patterns when we use markups estimated via the user-cost approach from Gutiérrez and Philippon (2017), the production function approach from De Loecker et al. (2020), or from accounting profits. Finally, our model predicts countercyclical movements of dispersion in firm-level revenue productivity (TFPR).⁹ Countercyclical dispersion in firm-level TFPR is documented by Kehrig (2011), among others.

Other related literature. This paper contributes to the large literature on the response of firms to monetary shocks. Our analysis is rooted in models of monopolistic competition with staggered price setting originating in Taylor (1980) and Calvo (1983).

A strand of this literature is devoted to explaining the strength and persistence of monetary policy shocks, which cannot be explained by nominal rigidities alone given the frequency of price adjustment.¹⁰ Ball and Romer (1990) introduce real rigidities, which complement nominal rigidities to increase monetary nonneutrality.¹¹ A common formulation of real rigidities is incomplete pass-through, where firms are slow to reflect marginal cost shocks in their prices due to strategic complementarities in pricing. Incompleteness of pass-through is documented empirically by Gopinath et al. (2010) and Gopinath and Itskhoki (2011). Our paper complements this literature by showing that incomplete pass-through, when paired with firm-level heterogeneity, provides another mechanism through which monetary policy can affect output.

In describing changes in the allocative efficiency of the economy, we also relate to a vast literature on cross-sectional misallocation, which includes Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and Baqaee and Farhi (2020b). For the most part, the misal-

(2014), in our empirical exercises. This is because the exogeneity conditions used to identify utilization-adjusted TFP—that sectoral TFP is orthogonal to oil price shocks and monetary shocks—are invalid in our model. Indeed, our core result is that sectoral TFP is endogenous to such shocks.

⁹When firms have constant returns to scale, as in our model, firm-level TFPR is equal to the firm's markup.

¹⁰This frequency has been documented by Taylor (1999) and Nakamura and Steinsson (2008) among others.

¹¹Ball and Romer (1990) has also spawned a large literature of theoretical developments on real rigidities, which characterize the conditions under which real rigidities can generate observed levels of persistence in monetary shocks. Eichenbaum and Fisher (2004) and Dotsey and King (2005), for example, investigate how relaxing assumptions of constant elasticities of demand interact with other frictions to generate persistence. Klenow and Willis (2016) compare the predictions of models where real rigidities are generated by a kinked demand curve versus sticky intermediate prices.

location literature is concerned with steady-state or long-run changes in misallocation, whereas we are focused on characterizing short-run changes in misallocation following nominal shocks. Some important exceptions are Cravino (2017), Baqaee and Farhi (2017), and Meier and Reinelt (2020). In an international context, Cravino (2017) shows that heterogeneity in exporters' invoicing currency and desired markups (due to local distribution costs), coupled with nominal rigidities, implies that exchange rate changes can affect domestic productivity by changing the allocation of resources. Baqaee and Farhi (2017) provide a general framework for how allocative efficiency changes in general equilibrium and apply their results to show that if price-stickiness positively covaries with markups, then monetary policy affects TFP. Meier and Reinelt (2020) provide empirical support for this covariance, and offer a microfoundation where firms have heterogeneous Calvo parameters, so firms with more rigid prices endogenously set higher markups due to a precautionary motive. Our analysis complements, and to some extent unifies, these previous analyses by showing how heterogeneity in realized pass-throughs (driven either by variable stickiness or variable desired pass-throughs) can cause nominal shocks to have effects on productivity.¹²

The differential cross-sectional response of firms to monetary policy links the slope of the Phillips curve in our analysis to moments of the firm distribution, such as industrial concentration. Here, our study is complemented by Etro and Rossi (2015), Wang and Werning (2020), Andrés and Burriel (2018), and Corhay et al. (2020) who also discuss mechanisms by which an increase in concentration may contribute to a decline in inflation and flattening of the Phillips curve; our work is unique among these in identifying the misallocation channel of monetary policy as a potential source for this effect.

Finally, our paper is also related to a recent literature on endogenous TFP movements over the business cycle (e.g., Comin and Gertler 2006, Benigno and Fornaro 2018, Anzoategui et al. 2019, and Bianchi et al. 2019). In this literature, aggregate TFP responds to the business cycle due to frictions in technology investment, adoption, and diffusion. In contrast to this body of work, the endogenous TFP movements that arise in our model are due solely to changes in the allocation of resources across firms, rather than underlying technological primitives.

Structure of the paper. Section 2 introduces a simple static model and defines the equilibrium. Sections 3 and 4 describe the response of aggregate TFP and output (real GDP) to a monetary shock in the one-period model and elaborate on the findings using a few

¹²Productivity shocks can also affect allocative efficiency: David and Zeke (2021) show that allocative efficiency varies over the business cycle when firms have heterogeneous exposure to aggregate shocks.

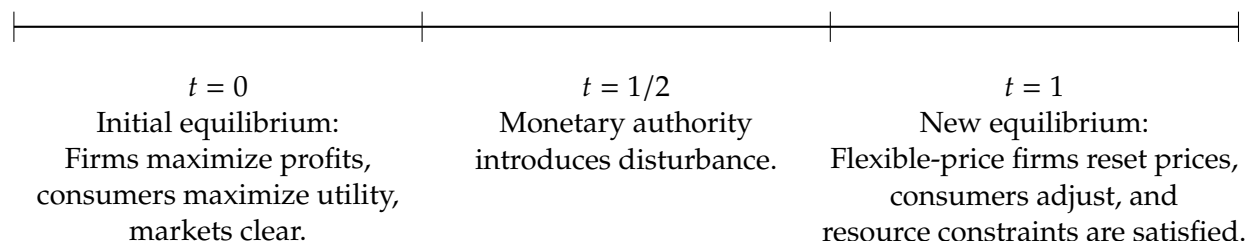
simple examples. Section 5 generalizes the static model from the previous sections to a fully dynamic setting, which yields a four-equation New Keynesian model with misallocation. Section 6 provides a quantitative illustration of the importance of the misallocation channel. Section 7 provides empirical evidence at the macro- and micro-level for the mechanisms described in the model. In Section 8, we summarize some extensions discussed in more detail in the appendices, including a model with multiple sectors, multiple factors, input-output linkages, and sticky wages, as well as a calibration in a setting with oligopolistic, rather than monopolistic, competition. Section 9 concludes.

2 Model

We start with a simple model with a single factor, labor. To build intuition, we first consider a one-period model to highlight the mechanism driving changes in allocative efficiency. In Section 5, we consider the fully dynamic model, which generalizes the findings shown here.

The timing is as follows. At time $t = 0$, the economy is in an equilibrium: households choose consumption and labor to maximize utility, firms choose prices to maximize profits, and markets clear. The monetary authority then introduces an unexpected monetary disturbance into the economy. At time $t = 1$, firms with flexible prices reset prices to maximize profits, while firms with sticky prices keep prices unchanged from the initial equilibrium. Households adjust consumption and labor to maximize utility.

Figure 1: One-period model timing.



2.1 Setup

We describe the behavior of households, firms, and the monetary authority in turn.

Households. There is a population of identical consumers. Consumers' preferences over the consumption bundle Y and labor L are given by

$$u(Y, L) = \frac{Y^{1-\gamma} - 1}{1 - \gamma} - \frac{L^{1+\frac{1}{\zeta}}}{1 + \frac{1}{\zeta}},$$

where $1/\gamma$ is the intertemporal elasticity of substitution, and ζ is the Frisch elasticity of labor supply. The consumption bundle Y consists of different varieties of goods indexed by $\theta \in [0, 1]$. Consumers have homothetic preferences over these final goods, and the utility from the consumption bundle Y is defined implicitly by¹³

$$\int_0^1 \Upsilon_\theta\left(\frac{y_\theta}{Y}\right) d\theta = 1.$$

Here, y_θ is the consumption of variety θ , and Υ_θ is an increasing and concave function. CES preferences are a special case of the general preferences above, when $\Upsilon_\theta = \Upsilon$ is a power function.

The representative consumer maximizes utility subject to the budget constraint

$$\int_0^1 p_\theta y_\theta d\theta = wL + \Pi,$$

where wL is labor income and Π is firm profit income. Maximization yields the inverse-demand curve for variety θ :

$$\frac{p_\theta}{P} = \Upsilon'_\theta\left(\frac{y_\theta}{Y}\right), \quad (1)$$

where the *price aggregator* P is defined as

$$P = \frac{P^Y}{\int_0^1 \Upsilon'_\theta\left(\frac{y_\theta}{Y}\right) \frac{y_\theta}{Y} d\theta}, \quad (2)$$

and P^Y is the ideal price index.¹⁴ As we can see in Equation (1), relative demand for a variety θ is dictated by the ratio of its price to the price aggregator P . Hence, firms compete with the rest of the market via a single price and quantity aggregator. Equation (1) also illustrates the appeal of these preferences: we can create downward-sloping demand

¹³These preferences are a generalization of Kimball (1995) preferences since the aggregator function Υ_θ is allowed to vary by variety. For more information, see Matsuyama and Ushchev (2017), who refer to these as homothetic with direct implicit additivity (HDIA) preferences.

¹⁴Recall that the ideal price index is defined as $\min_{y_\theta} \{\int p_\theta y_\theta : Y = 1\}$. Since consumer preferences are homothetic, changes in the ideal price index $d \log P^Y$ are first-order equivalent to changes in the consumer price index (CPI).

curves of any desired shape by choosing the aggregator Υ_θ .

Firms. Each variety is supplied by a single firm, and a firm of type θ has productivity A_θ . Firms produce using a constant returns to scale technology, so that the cost of producing an additional unit is constant at w/A_θ .

In the initial equilibrium, before the unexpected (zero-probability) monetary disturbance, each firm sets its price to maximize expected profits,

$$p_\theta^{\text{flex}} = \underset{p_\theta}{\text{argmax}} \mathbb{E} \left(p_\theta y_\theta - \frac{w}{A_\theta} y_\theta \right),$$

subject to its residual demand curve (1).

Unlike the CES demand system, which imposes that the price elasticity of demand is constant in both the time series and the cross-section of firms, we allow the price elasticity facing a firm to vary both with the firm's type θ and its position on the demand curve. We can use the inverse-demand function in (1) to solve for the price elasticity of demand facing a firm of type θ :

$$\sigma_\theta\left(\frac{y}{Y}\right) = -\frac{\partial \log y_\theta}{\partial \log p_\theta} = \frac{\Upsilon'_\theta\left(\frac{y}{Y}\right)}{-\frac{y}{Y}\Upsilon''_\theta\left(\frac{y}{Y}\right)}.$$

The profit-maximizing price p_θ^{flex} can be written as a markup μ_θ^{flex} times marginal cost. When the firm is able to change its price, the firm's desired price and markup are determined by

$$p_\theta^{\text{flex}} = \mu_\theta^{\text{flex}} \frac{w}{A_\theta}, \quad \text{and} \quad \mu_\theta^{\text{flex}} = \mu_\theta\left(\frac{y_\theta^{\text{flex}}}{Y}\right),$$

where the markup function is given by the Lerner formula,¹⁵

$$\mu_\theta\left(\frac{y}{Y}\right) = \frac{1}{1 - \frac{1}{\sigma_\theta\left(\frac{y}{Y}\right)}}. \quad (3)$$

Following Calvo (1983), we assume a firm of type θ has a probability δ_θ of being able to reset its price at time $t = 1$. These nominal rigidities are allowed to be heterogeneous across firm types. Flexible-price firms reset prices in $t = 1$ according to the optimal price and markup formulas above, and sticky-price firms keep their prices unchanged. As a result, the prices and markups of sticky-price firms at $t = 1$ are given by

$$p_{\theta,1}^{\text{sticky}} = p_{\theta,0} \quad \text{and} \quad \mu_{\theta,1}^{\text{sticky}} = \frac{w_1}{w_0} \mu_{\theta,0},$$

¹⁵We assume that marginal revenue curves are downward-sloping, so that the optimal choice of p_θ and y_θ is unique for each firm. In terms of primitives, this requires that $x\Upsilon_\theta'''(x) + 2\Upsilon_\theta''(x) < 0$ for every x and θ .

where the second subscript denotes the period.

A firm's desired pass-through ρ_θ is the elasticity of its optimal price with respect to its marginal cost, holding the economy-wide aggregates constant. We can express the desired pass-through of firm θ as:

$$\rho_\theta\left(\frac{y}{Y}\right) = \frac{\partial \log p_\theta^{\text{flex}}}{\partial \log mc} = \frac{1}{1 + \frac{\frac{y}{Y} \mu'_\theta\left(\frac{y}{Y}\right)}{\mu_\theta\left(\frac{y}{Y}\right)} \sigma_\theta\left(\frac{y}{Y}\right)}. \quad (4)$$

Under CES preferences, desired markups $\mu_\theta = \mu = \sigma/(\sigma - 1)$ are constant across firms. Furthermore, desired markups do not depend on the firm's location on the demand curve. When markups do not vary as a function firm size, desired pass-through is equal to one for all firms, and firms exhibit "complete desired pass-through." For brevity, we refer to ρ_θ simply as the firm's "pass-through" instead of desired pass-through. Keep in mind, however, that this pass-through is conditional on the firm's ability to change its price. For firms that are unable to change their prices, pass-through is equal to zero by assumption.

Monetary authority. At time $t = 1/2$, the monetary authority sets the nominal wage. We could easily have the monetary authority choose any other nominal variable in the economy, such as the overall price level or money supply. The nominal wage is especially convenient as it directly affects the marginal cost of every firm. We say that the monetary shock is expansionary if the nominal wage in period 1 is higher than the one in period 0, since in this case the increase in nominal marginal cost decreases markups for firms whose prices cannot adjust, and this reduction in markups boosts labor demand and hence output.

Equilibrium conditions. In equilibrium, for a given value of the nominal wage w , (1) consumers choose consumption and labor to maximize utility taking prices as given, (2) firms with flexible prices set prices to maximize profits taking other firms' prices and their residual demand curves as given, (3) firms with sticky prices produce to meet demand at fixed prices, and (4) all resource constraints are satisfied.

Notation. Throughout the rest of the paper, we use the following notation. For two variables $x_\theta > 0$ and z_θ , define the x -weighted expectation of z by

$$\mathbb{E}_x[z_\theta] = \frac{\int_0^1 z_\theta x_\theta d\theta}{\int_0^1 x_\theta d\theta}.$$

We write \mathbb{E} to denote \mathbb{E}_x when $x_\theta = 1$ for all θ . The operator \mathbb{E}_x operates a change of measure by putting more weight on types θ with higher values of x_θ . We denote the sales share density by¹⁶

$$\lambda_\theta = \frac{p_\theta y_\theta}{\int_0^1 p_\theta y_\theta d\theta},$$

and define the *aggregate markup* to be

$$\bar{\mu} = \mathbb{E}_\lambda \left[\mu_\theta^{-1} \right]^{-1}.$$

In words, $\bar{\mu}$ is the sales-weighted harmonic average of markups.

We write $d \log X$ for the differential of a variable X understood as the (infinitesimal) change in X in response to (infinitesimal) shocks. For non-infinitesimal changes in a variable, we write $\Delta \log X$ instead.

3 Productivity Response

In this section, we consider the movement of aggregate productivity from the initial allocation to the period following a monetary shock. We first introduce the concept of allocative efficiency and discuss its dependence on the distribution of markups. Then, we show that, when markups are initially dispersed, the reshuffling of resources across firms following a monetary shock changes aggregate productivity. In the next section, we show how output responds to monetary disturbances.

3.1 Allocative Efficiency, TFP, and TFPR

Recall that firms with market power set high prices and markups by restricting output. Compared to an economy with no dispersion in markups, an economy with heterogeneous markups features a distorted allocation of resources across firms: firms with higher markups are inefficiently small, and thus capture a lower share of resources than firms with lower markups. We refer to the change in *allocative efficiency* of an economy as the difference in output due to changes in the cross-sectional allocation of resources across firms, holding fixed the total supply of those resources.¹⁷

¹⁶As long as the productivity distribution is continuous, we can assume that the type distribution is uniform without loss of generality between $[0,1]$ because we can define a firm's type by the fraction of firms whose productivity is less than that firm.

¹⁷In our single-factor, single-sector model, the allocation of resources refers to the allocation of labor across firms, but this can be generalized to multiple factors and intermediate inputs. We provide an extension to multiple factors and multiple sectors in Appendix F.

Changes in aggregate output can then be decomposed, to a first-order, into changes in the total supply of resources, in this case labor, and changes in allocative efficiency

$$d \log Y = d \log L + [d \log Y - d \log L] = d \log L + d \log A.$$

Here, $d \log A$ is the change in the distortion-adjusted (or cost-based) Solow residual, or equivalently, the change in aggregate labor productivity.¹⁸ We refer to $d \log A$ as the change in aggregate TFP. The following lemma, which is an application of the main result in Baqaee and Farhi (2020b), shows how aggregate TFP is related to changes in markups.

Lemma 1. *Following a monetary shock, the response of aggregate TFP at $t = 1$ is*

$$d \log A = d \log \bar{\mu} - \mathbb{E}_\lambda [d \log \mu_\theta]. \quad (5)$$

Recall that $\bar{\mu}$ is the (harmonic) average markup. Therefore, Proposition 1 shows that allocative efficiency increases when the average of markups rises more than markups on average. This can only happen if there is a composition effect whereby high-markup firms expand their use of inputs relative to low-markup firms. Since these reallocations occur due to changes in markups, (5) can also be viewed as a measure of the change in the dispersion in markups. In particular, if high-markup firms expand relative to low-markup firms, this must be because their markups are falling relative to those of low-markup firms — that is, $d \log A$ is positive if the distribution of markups is being compressed, and the allocation is moving closer to the efficient frontier.

Therefore, Equation (5) links TFP at the economy level to the dispersion in markups at the firm level. It is useful to note a connection here with a literature that has documented cyclical variation in the dispersion of plant- and firm-level revenue productivity, or TFPR.¹⁹ When production has constant returns to scale, variation in markups is exactly equal to variation in TFPR. Hence, an improvement in aggregate TFP driven by a compression of the markup distribution will also imply that TFPR dispersion should be countercyclical. Specifically, TFPR is usually measured by subtracting input growth from revenue growth. In our model, this is just

$$\Delta \log \text{TFPR} = \Delta \log \mu_\theta + \Delta \log w,$$

¹⁸The distortion-adjusted Solow residual weighs the change in each factor by its share of total factor costs, rather than its share in aggregate income. See Hall (1990) and Baqaee and Farhi (2020b) for more information on why the Solow residual must be corrected in the presence of markups.

¹⁹See Foster et al. (2008) for more on the relationship between TFPR and “physical productivity” (TFPQ, or A_θ in our setting).

where changes in the wage $\Delta \log w$ do not vary by firm. As we will see, our model predicts that a monetary easing can simultaneously increase aggregate TFP and reduce dispersion in TFPR.²⁰

A corollary of Lemma 1 is that when markups are not dispersed in the initial equilibrium, first-order changes in aggregate productivity are necessarily zero.²¹

Corollary 1. *If $\mu_\theta = \mu$ in the initial equilibrium, then following a monetary shock, the response of aggregate TFP at $t = 1$ is*

$$d \log A = 0,$$

regardless of changes in markups $d \log \mu_\theta$.

Corollary 1 can also be derived as a consequence of the envelope theorem: when markups are identical across firms, the cross-sectional allocation of resources is efficient. Hence, changes in aggregate TFP due to reallocations are zero to a first order. This confirms that in models without initial markup dispersion, the response of aggregate TFP to monetary shocks is zero to a first-order, regardless of how markups respond.

3.2 Productivity Response: Two Mechanisms

Proposition 1 uses Lemma 1 to show how aggregate TFP responds to a monetary shock in general.

Proposition 1. *Following a monetary shock, the response of aggregate TFP at $t = 1$ is*

$$\frac{d \log A}{d \log w} = \underbrace{\kappa_\rho \text{Cov}_\lambda [\rho_\theta, \sigma_\theta | \text{flex}]}_{\substack{\text{Reallocation due to} \\ \text{heterogeneous} \\ \text{pass-through}}} + \underbrace{\kappa_\delta \text{Cov}_\lambda [\sigma_\theta, \delta_\theta]}_{\substack{\text{Reallocation due to} \\ \text{heterogeneous} \\ \text{price-stickiness}},} \quad (6)$$

where $\text{Cov}_\lambda [\rho_\theta, \sigma_\theta | \text{flex}]$ is the covariance of pass-throughs and elasticities for the subset of flexible-price firms,²² and κ_ρ and κ_δ are positive constants

$$\kappa_\rho = \frac{\bar{\mu} \mathbb{E}_\lambda [\delta_\theta] \mathbb{E}_\lambda [1 - \delta_\theta]}{\mathbb{E}_\lambda [[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \sigma_\theta]}, \quad \kappa_\delta = \frac{\bar{\mu} \mathbb{E}_{\lambda\delta} [\rho_\theta]}{\mathbb{E}_\lambda [[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \sigma_\theta]}.$$

²⁰In this sense, our model will be consistent with the empirical findings of Kehrig (2011), who finds that dispersion in plant-level TFPR is countercyclical, increasing about 10% during recessions compared to booms.

²¹To show Corollary 1 write $d \log \bar{\mu} = \mathbb{E}_{\lambda\mu^{-1}} [d \log \mu_\theta] - \mathbb{E}_{\lambda\mu^{-1}} [d \log \lambda_\theta]$ and impose uniformity of markups.

²²The reader may note that this is equivalent to $\text{Cov}_{\lambda\delta} [\rho_\theta, \sigma_\theta]$.

A first glance reveals that the response of aggregate TFP is nonzero only when markups are dispersed. If markups μ_θ and thus elasticities $\sigma_\theta = \mu_\theta/(\mu_\theta - 1)$ are equal across firms, both covariance terms in Equation (6) are zero.

However, dispersion in markups is not sufficient for monetary policy to affect aggregate productivity. It must also be the case that markups (or equivalently, σ_θ) covary systematically with either desired pass-throughs or price-stickiness. Either of these two mechanisms cause *realized* pass-throughs to covary with the level of the markups, and as long as realized pass-through covaries negatively with the level of markups, an increase in nominal marginal costs will result in productivity-increasing reallocations.²³ To build more intuition, we now consider each of the two mechanisms mentioned above in isolation.

Mechanism I: heterogeneous pass-through. If price-stickiness is homogeneous across firms ($\delta_\theta = \delta$), then the second covariance term in Proposition 1 is zero, and the productivity response depends on the covariance between desired pass-throughs ρ_θ and elasticities σ_θ alone:

Corollary 2. *If price stickiness is homogeneous across firms ($\delta_\theta = \delta$), then*

$$\frac{d \log A}{d \log w} = \kappa_\rho \text{Cov}_\lambda [\rho_\theta, \sigma_\theta].$$

If markups negatively covary with pass-throughs, then $\frac{d \log A}{d \log w} > 0$.

One of the salient reasons why markups may covary negatively with pass-throughs is related to firm size.

Definition 1. *Marshall's strong second law of demand* requires that desired markups are increasing in quantity and desired pass-throughs are decreasing in quantity. That is,²⁴

$$\mu'(\frac{y}{Y}) > 0 \quad \text{and} \quad \rho'(\frac{y}{Y}) < 0.$$

²³For concreteness, in this paper, we interpret increases in nominal marginal cost $d \log w > 0$ to be the consequence of monetary easing. However, the basic intuition will apply to other kinds of demand shocks as well, since other shocks to aggregate demand will also raise nominal marginal costs, and hence lead to productivity-increasing reallocations.

²⁴Marshall's strong second law of demand is equivalent to requiring that the individual marginal revenue curve be log-concave. There is also a weaker version of Marshall's second law, which requires $\mu'(\frac{y}{Y}) \geq 0$ (and hence $\rho(\frac{y}{Y}) \leq 1$) alone. This is equivalent to requiring that the residual demand curve be log-concave in log price. The strong version implies the weak version. See Melitz (2018) for more information.

If Marshall’s strong second law holds, then markups are increasing and pass-throughs are decreasing in firm size. This guarantees that monetary expansions will raise aggregate productivity. Marshall’s second law of demand has strong empirical support (see, for example, empirical estimates of pass-throughs by firm size from Amiti et al. 2019).²⁵

While Marshall’s strong second law is sufficient for a supply-side channel of monetary policy to operate, it is not necessary. Markups and pass-throughs may be correlated for reasons unrelated to firm size, such as quality or nicheness (e.g. as shown empirically by Chen and Juvenal, 2016 or Auer et al., 2018).

To understand the intuition for Corollary 2, consider an expansionary shock ($d \log w > 0$). The higher nominal wage increases marginal costs, leading flexible-price firms to increase their prices. The optimal price satisfies

$$d \log p_{\theta}^{\text{flex}} = (1 - \rho_{\theta})d \log P + \rho_{\theta}d \log w,$$

where $d \log P$ is the change in the price aggregator defined in Equation (2). The optimal price of a firm with a high pass-through moves closely with shocks to marginal cost from the nominal wage. Firms with low pass-through, on the other hand, exhibit “pricing-to-market” behavior: they place less weight on their own marginal cost, and more weight on the expected aggregate price level in the economy. Sticky-price firms, of course, cannot adjust their prices after observing the nominal wage shock.

Following an increase in the nominal wage, flexible-price firms shrink and sticky-price firms, whose prices are kept artificially low, expand²⁶:

$$d \log\left(\frac{y_{\theta}^{\text{flex}}}{Y}\right) = -\sigma_{\theta}\rho_{\theta}(d \log w - d \log P) < 0, \quad \text{and} \quad d \log\left(\frac{y_{\theta}^{\text{sticky}}}{Y}\right) = \sigma_{\theta}d \log P > 0.$$

Among flexible-price firms, the ones with low pass-throughs and high markups expand relative to the ones with high pass-throughs and low markups. Among sticky-price firms, the opposite effect prevails, where firms with low markups expand relative to firms with high markups.

The former effect dominates when $Cov_{\lambda}[\rho_{\theta}, \sigma_{\theta}] > 0$. When this is the case, firms with high markups also have low pass-throughs, allowing them to cut prices and stay competitive. As a result, the allocation of output shifts toward high-markup firms in

²⁵Oligopolistic competition models, such as Atkeson and Burstein (2008), also satisfy Marshall’s strong second law of demand. In Appendix H, we show that our results can also be derived under such a framework.

²⁶Due to nominal and real rigidities, the price aggregator P will move more slowly than the nominal wage, so generically $\frac{d \log P}{d \log w} \in [0, 1]$.

aggregate. Since high-markup firms are initially too small relative to the efficient cross-sectional allocation, the expansion of high-markup firms boosts allocative efficiency and hence aggregate TFP.

Mechanism II: heterogeneous price-stickiness. Now consider the case where pass-through is homogeneous, but price-stickiness is not.

Corollary 3. *If desired pass-through is homogeneous across firms ($\rho_\theta = \rho$),²⁷ then*

$$\frac{d \log A}{d \log w} = \kappa_\delta \text{Cov}_\lambda [\sigma_\theta, \delta_\theta]. \quad (7)$$

If high-markup firms have higher price-stickiness, then $\frac{d \log A}{d \log w} > 0$.

Consider an expansionary shock ($d \log w > 0$). If high markups firms are less likely to adjust prices, low-markup firms will tend to increase their prices more on average than high-markup firms. This causes high-markup firms to expand relative to low-markup firms, compressing the markup distribution, and increasing aggregate TFP.

The mechanism by which heterogeneous price-stickiness can result in endogenous aggregate TFP changes was previously pointed out in Baqaee and Farhi (2017) and has been recently analyzed by Meier and Reinelt (2020). Meier and Reinelt (2020) show that in a CES model with heterogeneous Calvo parameters, firms with greater price rigidity endogenously set higher markups due to a precautionary motive; this generates an increase in markup dispersion following contractionary shocks. Although we allow for the possibility that price-stickiness may vary as a function of firm size, we will abstract from this mechanism in our quantitative applications. When we quantify the model, we assume there is no variation in price-stickiness, and instead focus on heterogeneity in desired pass-through, where there is more robust empirical support for Marshall’s second law of demand.

3.3 Discussion

To recap, changes in aggregate productivity are caused by changes in the cross-sectional distribution of markups. There is evidence supporting both the macro- and micro-level predictions of the model in the data. We document that aggregate productivity responds procyclically to monetary shocks in Section 7, corroborating previous evidence from Evans

²⁷Homogeneous desired pass-throughs are generated when the Kimball aggregator takes the form, $\Upsilon(x) = -\text{Ei}(-Ax^{\rho-1})$ where $E_i(x) = \int_{-x}^{\infty} \frac{e^{-t}}{t} dt$ is the exponential integral function.

(1992), among others. At the plant-level, Kehrig (2011) documents countercyclical dispersion in TFPR. In addition, we show in Section 7 that expansions are accompanied by reallocations toward high-markup firms using markups estimated on public firms in Compustat. In our model, these empirical patterns—procyclical aggregate TFP, countercyclical TFPR dispersion, and the differential growth of high-markup firms during expansions—are all linked to the reallocation of resources toward high-markup firms triggered by the demand shock.

4 Output Response and the Phillips Curve

In the previous section, we showed that aggregate TFP responds to monetary shocks. In this section, we show how monetary shocks are transmitted to output, taking into account the movements in aggregate productivity, and characterize the slope of the Phillips curve. We show that the change in output can be decomposed into three channels: (1) nominal rigidities (as in a CES economy with sticky prices), (2) real rigidities due to imperfect pass-through (which arise from strategic complementarities in pricing), and (3) the endogenous response of aggregate TFP, which we term the misallocation channel.

This section is organized as follows. We first characterize the response of output to a monetary shock. Then, we characterize the slope of the Phillips curve and formalize two channels—real rigidities and the misallocation channel—which flatten the slope of the Phillips curve relative to the benchmark model. Finally, to gain intuition, we compute the slope of the Phillips curve in a few simple example economies.

4.1 Output Response

Proposition 2 describes the response of output to a monetary shock.

Proposition 2. *Following a shock to the nominal wage $d \log w$, the response of output at $t = 1$ is*

$$d \log Y = \underbrace{\frac{1}{1 + \gamma \zeta} d \log A}_{\text{Supply-side effect}} - \underbrace{\frac{\zeta}{1 + \gamma \zeta} \mathbb{E}_\lambda [d \log \mu_\theta]}_{\text{Demand-side effect}}, \quad (8)$$

where

$$\frac{\mathbb{E}_\lambda [d \log \mu_\theta]}{d \log w} = - \underbrace{\mathbb{E}_\lambda [1 - \delta_\theta]}_{\text{Nominal rigidities}} - \underbrace{\frac{\mathbb{E}_\lambda [\delta_\theta (1 - \rho_\theta)] \mathbb{E}_\lambda [\sigma_\theta (1 - \delta_\theta)]}{\mathbb{E}_\lambda [[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \sigma_\theta]}}_{\text{Real rigidities}}. \quad (9)$$

Equation (8) breaks down the response of output into a supply-side and demand-side effect. The demand-side effect of an expansionary shock arises from the average reduction in markups, which increases labor demand (and employment). The supply-side effect is due to changes in aggregate TFP and arises from changes in the economy's allocative efficiency.

Equation (9) further decomposes the demand-side effect into the effect of sticky prices and the effect of real rigidities. The first is the standard New Keynesian channel: nominal rigidities prevent sticky-price firms from responding to the marginal cost shock. As a result, markups fall for a fraction $\mathbb{E}_\lambda [\delta_\theta]$ of firms. This reduction in the markups of sticky-price firms boosts labor demand, and hence output.

This sticky price effect in (9) is exacerbated by real rigidities, which arise from imperfect pass-through. When pass-through is incomplete, flexible-price firms increase prices less than one-for-one with the marginal cost shock. As a result, the markups of flexible-price firms also fall. Together, the reduction in the markups of both sticky-price and flexible-price firms increase labor demand, which spurs employment and output.

The supply-side effect is concerned with the productivity of these resources. Returning to (8), we find that when aggregate TFP increases following an expansionary shock, $d \log A / d \log w > 0$, the endogenous positive "supply shock" complements the effects of the positive "demand shock" on output.

Interestingly, whereas the demand-side effect is increasing in the size of the elasticity of labor supply ζ , the supply-side effect is decreasing in ζ . In fact, the supply-side effect is strongest when labor is inelastically supplied $\zeta = 0$. On the other hand, when the Frisch elasticity of labor supply goes to infinity, the supply side effect becomes irrelevant for output. This is because when the Frisch is infinite, the reallocations that boost productivity, and raise the market share of high-markup firms, are exactly cancelled out by reductions in labor supply, which contracts due to the expansion of high-markup firms.

4.2 Flattening of the Phillips Curve

We can rearrange the output response given in Proposition 2 to get the slope of the wage Phillips curve. To get the price Phillips curve, we use the relationship between the consumer price index P^Y , the nominal wage, and average markups,

$$d \log P^Y = d \log w + \mathbb{E}_\lambda [d \log \mu_\theta].$$

Both are presented in Proposition 3.

Proposition 3. *The wage Phillips curve is given by*

$$d \log w = (1 + \gamma \zeta) \frac{1}{\left[\frac{d \log A}{d \log w} - \zeta \mathbb{E}_\lambda \left[\frac{d \log \mu_\theta}{d \log w} \right] \right]} d \log Y.$$

The price Phillips curve is given by

$$d \log P^Y = (1 + \gamma \zeta) \frac{1 + \mathbb{E}_\lambda \left[\frac{d \log \mu_\theta}{d \log w} \right]}{\left[\frac{d \log A}{d \log w} - \zeta \mathbb{E}_\lambda \left[\frac{d \log \mu_\theta}{d \log w} \right] \right]} d \log Y.$$

Note that the supply-side effect flattens both the price and wage Phillips curves. Proposition 2 allows us to quantify the amount of flattening caused by real rigidities and the misallocation channel relative to the CES baseline. To do this, we calculate the flattening of the Phillips curve due to real rigidities by dividing the slope of the Phillips curve due to sticky-prices alone by the slope of the Phillips curve due to sticky prices and real rigidities. Since real rigidities flatten the Phillips curve, this ratio is greater than one. If this quantity is, say, 1.5, this means that incorporating real rigidities decreases the responsiveness of output to the price level by 50%. Similarly, we calculate the flattening of the Phillips curve due to misallocation channel by comparing the slope of the Phillips curve due to sticky prices and real rigidities by the slope of the Phillips curve where we account for changes in allocative efficiency. As long as $\frac{d \log A}{d \log w} > 0$, this quantity is also greater than one.

Proposition 4 presents the flattening of the price Phillips curve due to each channel. For simplicity, we present the case where pass-throughs are heterogeneous and price-stickiness is constant across firms (the general version is Proposition 6 in Appendix A).

Proposition 4. *Suppose $\delta_\theta = \delta$ for all firms. The flattening of the price Phillips curve due to real rigidities, compared to nominal rigidities alone, is*

$$\text{Flattening due to real rigidities} = 1 + \frac{\mathbb{E}_\lambda [\sigma_\theta] \mathbb{E}_\lambda [1 - \rho_\theta]}{\delta \text{Cov}_\lambda [\rho_\theta, \sigma_\theta] + \mathbb{E}_\lambda [\rho_\theta] \mathbb{E}_\lambda [\sigma_\theta]}. \quad (10)$$

The flattening of the price Phillips curve due to the misallocation channel is

$$\text{Flattening due to the misallocation channel} = 1 + \frac{\bar{\mu}}{\zeta} \frac{\delta \text{Cov}_\lambda [\rho_\theta, \sigma_\theta]}{\delta \text{Cov}_\lambda [\rho_\theta, \sigma_\theta] + \mathbb{E}_\lambda [\sigma_\theta]}. \quad (11)$$

In Equation (10), we see that the flattening of the price Phillips curve due to real rigidities increases as average pass-throughs fall. The flattening due to real rigidities in (10)

is also decreasing in price flexibility δ . As price flexibility increases, the price aggregator moves more closely with shocks to marginal cost; hence the “pricing-to-market” effect from incomplete pass-throughs is less powerful.

The flattening of the price Phillips curve due to the misallocation channel depends positively on covariance of the pass-throughs and elasticities. When $Cov_\lambda[\rho_\theta, \sigma_\theta] = 0$, there is no allocative efficiency effect on the slope of the Phillips curve. Equation (11) also shows that the flattening due to misallocation is decreasing in the Frisch elasticity ζ . A higher aggregate markup, $\bar{\mu}$, increases the strength of the misallocation channel, since the productivity response is increasing in $\bar{\mu}$. Finally, since the expansion of high-markup firms relative to low-markup firms occurs only for flexible-price firms, the misallocation channel is stronger relative to real rigidities when price flexibility is higher (δ closer to one).

To cement intuition, we now calculate the change in allocative efficiency and the slope of the Phillips curve in three simple benchmark economies: an economy with CES preferences, an economy with real rigidities but a representative firm, and an economy with two firm types.

CES Example. We obtain the CES benchmark by setting $\Upsilon_\theta(x) = x^{\frac{\sigma-1}{\sigma}}$, where $\sigma > 1$. Under CES, desired markups for all firms are fixed at $\mu = \frac{\sigma}{\sigma-1}$, and all firms exhibit complete desired pass-through of cost shocks to price ($\rho = 1$).

Since desired markups are uniform, the initial allocation of the economy is efficient. By Corollary 1, $d \log A = 0$. Applying Proposition 3, the slope of the price Phillips curve is

$$d \log P^\Upsilon = \frac{1 + \gamma \zeta}{\zeta} \frac{\delta}{1 - \delta} d \log \Upsilon.$$

This is the traditional New Keynesian Phillips Curve.²⁸ Nominal rigidities, captured by the Calvo parameter $\delta < 1$, flatten the Phillips curve. As δ approaches one, prices become perfectly flexible and the Phillips curve becomes vertical.

Representative Firm Example. We now relax the assumption of CES preferences, but consider an economy with a representative firm: all firms have the same price-stickiness ($\delta_\theta = \delta$), the same residual demand curve $\Upsilon'_\theta = \Upsilon'$, and productivity level ($A_\theta = 1$).

The homogeneous firms in this economy have identical markups, $\mu_\theta = \mu$, and pass-throughs, $\rho_\theta = \rho$. By deviating from CES, however, we allow firms’ desired pass-throughs

²⁸See, for example, Galí (2015). Section 4.2 can be replicated exactly from Galí (2015) pg. 63 by setting $\beta = 0$ and assuming constant returns to scale.

to be incomplete, i.e., $\rho < 1$.

Since markups are uniform, the cross-sectional allocation of resources across firms in the initial equilibrium is efficient. Applying Corollary 1, we have $d \log A = 0$. Unlike the CES case, incomplete pass-throughs imply that flexible-price firms will not fully adjust prices to reflect increases in marginal cost from a monetary shock. Compared to the CES economy, prices in this economy are slower to respond, and hence, the slope of the price Phillips curve is flatter:

$$d \log P^Y = \frac{1 + \gamma\zeta}{\zeta} \frac{\delta}{1 - \delta} \rho d \log Y.$$

In particular, Proposition 4 implies that the amount of flattening due to the real rigidities channel is ρ^{-1} . That is, the amount of flattening is decreasing in the average desired pass-through ρ .

Two Type Example. We now allow for heterogeneous firms of two types: high- and low-markup firms. High- and low-markup firms differ in their markups and pass-throughs, and we denote them with subscripts H and L .

Following Lemma 1, the change in aggregate TFP following a nominal shock is

$$d \log A = d \log \bar{\mu} - \mathbb{E}_\lambda [d \log \mu_\theta] = \lambda_H \left(1 - \frac{\bar{\mu}}{\mu_H} \right) (d \log l_H - d \log l_L),$$

where l_H and l_L are employment by H and L firms. Reallocations in inputs (labor in our single factor model) across high- and low-markup firms, paired with the initial distribution of sales and markups, determines the change in aggregate TFP. In particular, aggregate TFP increases if the growth in employment at high-markup firms outpaces the growth of employment at low-markup firms.²⁹

For simplicity, we again impose homogeneous price-stickiness ($\delta_H = \delta_L = \delta$). Proposition 3 implies that the price Phillips curve is

$$d \log P^Y = \frac{1 + \gamma\zeta}{\zeta} \frac{\delta}{1 - \delta} \frac{\delta(\sigma_L - \sigma_H)(\rho_L - \rho_H) + (\lambda_L^{-1}\sigma_H + \lambda_H^{-1}\sigma_L)(\lambda_H\rho_H + \lambda_L\rho_L)}{\delta\left(1 + \frac{\bar{\mu}}{\zeta}\right)(\sigma_L - \sigma_H)(\rho_L - \rho_H) + (\lambda_L^{-1}\sigma_H + \lambda_H^{-1}\sigma_L)} d \log Y.$$

This price Phillips curve is flatter than the CES economy if $\rho_L > \rho_H$, i.e., if low-markup firms have higher pass-throughs than high-markup firms. An increase in the covariance of elasticities and pass-throughs, $(\sigma_L - \sigma_H)(\rho_L - \rho_H)$, further flattens the Phillips curve.

²⁹See Section 7 for supporting empirical evidence.

4.3 Discussion

Before moving onto the dynamic version of the model, we discuss some of implications and extensions of the results in this section.

First, unlike the standard model, our model links the slope of the Phillips curve to the industrial organization of the economy, via statistics like the covariance of pass-throughs and price-elasticities. This means that industrial concentration plays a role in shaping the Phillips curve. We consider this effect quantitatively in Section 6, where we illustrate the effect of increasing industrial concentration on the Phillips curve slope.

Second, the results in Sections 3 and 4 can also be derived in models of oligopolistic competition that are populated by a discrete number of firms instead of a continuum of infinitesimal firms in monopolistic competition. As discussed above, the nested CES model of Atkeson and Burstein (2008) generates markups and pass-throughs that conform with Marshall’s second law of demand, and hence yields similar implications (we show this in Appendix H). In the body of the paper we focus on the monopolistic competition model because monopolistic competition is much more tractable in a fully dynamic environment.

5 Four-Equation Dynamic Model

We now present a general dynamic model, which generalizes the findings from the static model above. We present a four-equation system that generalizes the workhorse three-equation model presented in Galí (2015) to account for imperfect pass-through and endogenous aggregate productivity. We provide a high-level walk-through of the derivation to highlight the key intuitions; the detailed derivation is in Appendix B.

The setup is as follows: each firm sets price to maximize discounted future profits, subject to a Calvo friction. For expositional simplicity, we present a version with homogeneous price-stickiness across firms. Households consume according to a standard Euler equation. As in Galí (2015), we log-linearize around the no-inflation steady state. The model is closed by the actions of the monetary authority, which we assume follow a Taylor rule.

Firm i sets its price today to maximize the expected value of discounted future profits, given by

$$\max_{p_{i,t}} \mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1 + r_{t+j})} (1 - \delta_i)^k y_{i,t+k} \left(p_{i,t} - \frac{w_{t+k}}{A_i} \right) \right], \quad (12)$$

where r_{t+j} is the interest rate, δ_i is the Calvo parameter, $y_{i,t+k}$ is the quantity firm i sells in period $t + k$ if it last set its price in period t . The solution to the firms’ maximization

problem describes how prices move in the economy. We use this to describe the movement of inflation, output, and aggregate TFP by aggregating across firms.

5.1 The New Keynesian Model with Misallocation

Proposition 5 is the main result in this section, characterizing the equilibrium allocation up to a first-order approximation.

Proposition 5. *Changes in aggregates are described by the following four-equation system.*

Taylor rule.

$$d \log i_t = \phi_\pi d \log \pi_t + \phi_y d \log Y_t + v_t,$$

where $d \log i_t$ is the nominal interest rate, $d \log \pi_t = d \log P_t^Y - d \log P_{t-1}^Y$ is inflation, ϕ_π and ϕ_y are parameters, and v_t is a monetary policy shock.

Dynamic IS equation.

$$d \log Y_t = d \log Y_{t+1} - \frac{1}{\gamma} (d \log i_t - d \log \pi_{t+1}),$$

where $1/\gamma$ is the intertemporal elasticity of substitution.

New Keynesian Phillips curve with misallocation.

$$d \log \pi_t = \beta d \log \pi_{t+1} + \varphi \mathbb{E}_\lambda [\rho_\theta] \frac{1 + \gamma \zeta}{\zeta} d \log Y_t - \alpha d \log A_t, \quad (13)$$

where $\varphi = \frac{\delta}{1-\delta}(1 - \beta(1 - \delta))$ and $\alpha = \frac{\varphi}{\bar{\mu}} \left(\mathbb{E}_\lambda [\rho_\theta] \left(1 + \frac{\bar{\mu}}{\zeta} \right) - 1 \right)$.

Endogenous TFP equation.

$$d \log A_t = \frac{1}{\kappa_A} d \log A_{t-1} + \frac{\beta}{\kappa_A} d \log A_{t+1} + \frac{\varphi}{\kappa_A} \frac{1 + \gamma \zeta}{\zeta} \bar{\mu} \frac{\text{Cov}_\lambda[\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda[\rho_\theta]} d \log Y_t, \quad (14)$$

where $\kappa_A = 1 + \beta + \varphi \left[1 + \frac{\text{Cov}_\lambda(\rho_\theta, \sigma_\theta)}{\mathbb{E}_\lambda[\sigma_\theta]} \left(1 + \frac{\bar{\mu}}{\zeta} \right) \right]$.

The Taylor rule and Dynamic IS equations are the same as the workhorse three-equation model. Differences arise in the last two equations. We discuss the last two equations in turn.

Consider the New Keynesian Phillips Curve (NKPC) with misallocation. We note two key differences: first, in the standard NKPC, the coefficient on $d \log Y_t$ is $\varphi \frac{1+\gamma\zeta}{\zeta}$.³⁰ In the NKPC with misallocation, this coefficient is multiplied by $\mathbb{E}_\lambda [\rho_\theta]$. Imperfect pass-through moderates the response of prices to nominal shocks, and hence flattens the NKPC, as in the static version of the model. Second, changes in aggregate TFP enter the Phillips curve as endogenous, negative cost-push shocks, given by $\alpha d \log A_t$.³¹ This means that procyclical movements in aggregate TFP dampen the response of inflation to an expansionary shock, similarly flattening the inflationary effects of a monetary expansion.

The path of aggregate TFP is pinned down by (14). When markups covary negatively with pass-throughs, changes in output $d \log Y$ are positively associated with aggregate TFP. Note that, unlike the standard New Keynesian model's equations, which are first-order difference equations, aggregate TFP follows a second-order difference equation. As a result, the augmented four-equation model can generate hump-shaped impulse responses to monetary shocks.

Proposition 5 also generalizes the static model presented in Sections 2–4 as shown by the following corollary.

Corollary 4. *Suppose output, aggregate TFP, and the price level are in steady state at $t = 0$ (i.e., $d \log Y_0 = d \log A_0 = d \log P_0^Y = 0$). When the discount factor $\beta = 0$, the effect of shocks on impact are the same as the static results from Proposition 1 and Proposition 2.*

5.2 Solution Strategy

We present a high-level walk-through of Equations (13) and (14). Start with the firm maximization problem described in Equation (12). The optimal reset price $p_{i,t}^*$ for profit maximization satisfies

$$\mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} (1-\delta_i)^k y_{i,t+k} \left[\frac{dy_{i,t+k}}{dp_{i,t}} \frac{p_{i,t}^*}{y_{i,t+k}} \frac{p_{i,t}^* - \frac{w_{t+k}}{A_i}}{p_{i,t}^*} + 1 \right] \right] = 0. \quad (15)$$

We log-linearize this equation around the perfect foresight zero inflation steady state. Note that the steady state is characterized by a constant discount factor such that $\frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} = \beta^k$.

³⁰See, e.g., Galí (2015) with constant returns.

³¹We find that $\alpha > 0$ when $\mathbb{E}_\lambda [\rho_\theta] > \frac{\bar{\mu}^{-1}\zeta}{1+\bar{\mu}^{-1}\zeta}$. The reciprocal of the average markup $\bar{\mu}^{-1}$ is bounded above by 1, and estimates of the Frisch elasticity place ζ between 0.1 and 0.4. Average pass-through is greater than 0.5, which suggests that $\alpha > 0$ holds nearly always.

With some manipulation, the log-linearization of Equation (15) yields,

$$d \log p_{i,t}^* = [1 - \beta(1 - \delta_i)] \sum_{k=0}^{\infty} \beta^k (1 - \delta_i)^k [\rho_i d \log w_{t+k} + (1 - \rho_i) d \log P_{t+k}]. \quad (16)$$

When prices are fully flexible, this simplifies to just

$$d \log p_{i,t}^* = (1 - \rho_i) d \log P_t + \rho_i d \log w_t.$$

Compared to the case without nominal rigidities, a firm with sticky prices is forward looking and incorporates expected future prices and marginal costs into its reset price today. Just as in the completely flexible benchmark, firms with high pass-throughs are more responsive to expected changes in their own marginal costs, while firms with low pass-throughs are more responsive to expected changes in the economy's price aggregator.

Rewrite Equation (16) recursively, and for each firm type θ , as

$$d \log p_{\theta,t}^* = [1 - \beta(1 - \delta_{\theta})] [\rho_{\theta} d \log w_t + (1 - \rho_{\theta}) d \log P_t] + \beta(1 - \delta_{\theta}) d \log p_{\theta,t+1}^*.$$

The price level of a firm of type θ at time t is equal to the firm's reset price with probability δ_{θ} , or else pinned at the last period price with probability $(1 - \delta_{\theta})$. In expectation,

$$\mathbb{E}[d \log p_{\theta,t}] = \delta_{\theta} \mathbb{E}[d \log p_{\theta,t}^*] + (1 - \delta_{\theta}) \mathbb{E}[d \log p_{\theta,t-1}]$$

Combining the above two equations and rearranging yields a second-order difference equation for the expected price of firm type θ ,

$$\begin{aligned} \mathbb{E}[d \log p_{\theta,t} - d \log p_{\theta,t-1}] - \beta \mathbb{E}[d \log p_{\theta,t+1} - d \log p_{\theta,t}] \\ = \varphi [-\mathbb{E}[d \log p_{\theta,t}] + \rho_{\theta} d \log w_t + (1 - \rho_{\theta}) d \log P_t], \quad (17) \end{aligned}$$

where

$$\varphi = \frac{\delta_{\theta}}{1 - \delta_{\theta}} (1 - \beta(1 - \delta_{\theta})).$$

We are almost finished. Equation (17) pins down the movement of the price of a firm with type θ over time and, by extension, the movement in the firm's markup over time. Our aggregate variables, such as the consumer price index, aggregate TFP, labor supply, and output, can all be recovered by manipulating this expression and averaging over firm types.

For instance, by taking the sales-weighted expectation of both sides in Equation (17),

we can recover the movement of the consumer price index.³² We get,

$$d \log \pi_t - \beta d \log \pi_{t+1} = \varphi \left[\mathbb{E}_\lambda [\rho_\theta] (d \log w_t - d \log P_t) + (d \log P_t - d \log P_t^Y) \right]. \quad (18)$$

The objects that remain—the difference between the price aggregator $d \log P_t$ and the nominal wage $d \log w_t$, and the difference between the aggregator $d \log P_t$ and the consumer price index $d \log P_t^Y$ —depend solely on the average markup and the distribution of markups. In particular, the following identities allow us to express these endogenous objects in terms of output and aggregate TFP:

$$d \log P_t - d \log P_t^Y = \bar{\mu}^{-1} d \log A_t, \quad (19)$$

$$d \log P_t^Y - d \log w_t = \frac{1}{\zeta} [d \log A_t - (1 + \gamma \zeta) d \log Y_t]. \quad (20)$$

Equation (19) can be derived by log-linearizing and rearranging the expression for the price aggregator in (2),³³ and (20) comes from rearranging (2) for the average change in markups. Substituting these identities into (18) yields the NKPC with misallocation in Proposition 5.

The Endogenous TFP equation can also be derived by rearranging (17). In particular, from Lemma 1, we have

$$d \log A_t = d \log \bar{\mu} - \mathbb{E}_\lambda [d \log \mu_\theta] = \bar{\mu} (\mathbb{E}_{\sigma\lambda} [d \log \mu_{\theta,t}] - \mathbb{E}_\lambda [d \log \mu_{\theta,t}]). \quad (21)$$

The changes in markups can in turn be derived from (17) by subtracting changes in marginal cost (the nominal wage) from changes in prices. This yields a second-order difference equation for the change in markups for each firm-type. Taking sales-weighted averages over these markup changes and rearranging yields expressions for the two terms on the right-hand side of (21).

5.3 Discussion

The model presented in Proposition 5 provides a tractable, four-equation system that can be used to calibrate economies with realistic heterogeneity in markups and pass-throughs. This model incorporates real rigidities and the misallocation channel while being parsimoniously governed by four objects from the firm distribution: the average

³²The CPI price index, log linearized around the steady state, is $\mathbb{E}_\lambda [\mathbb{E} [d \log p_\theta]] = d \log P^Y$.

³³In particular, $d \log P_t = d \log P_t^Y - \mathbb{E}_\lambda \left[\left(1 - \frac{1}{\sigma_\theta}\right) d \log \left(\frac{y_{\theta,t}}{Y}\right) \right] = d \log P_t^Y - \mathbb{E}_\lambda [\mu_\theta^{-1}] \mathbb{E}_{\lambda\mu^{-1}} \left[d \log \left(\frac{y_{\theta,t}}{Y}\right) \right] = d \log P_t^Y + \bar{\mu}^{-1} d \log A_t$.

sales-weighted elasticity $\mathbb{E}_\lambda[\sigma_\theta]$, the average sales-weighted pass-through $\mathbb{E}_\lambda[\rho_\theta]$, the covariance of elasticities and pass-throughs $Cov_\lambda[\sigma_\theta, \rho_\theta]$, and the aggregate markup $\bar{\mu}$. We present one such calibration in the section below.

The second-order difference equation for aggregate TFP generates hump-shaped patterns for $d \log A$ and potentially other aggregate variables. This is a virtue since empirical estimates of the impulse response of labor productivity to monetary shocks (see, e.g., Christiano et al. 2005) exhibit this shape. Equation (14) is a second-order equation because changes in prices, and hence changes in markups, are second-order difference equation. The backward looking part comes from the Calvo friction, and the forward looking part comes from expectations.

For the purposes of our discussion, we have focused on monetary policy shocks. However, the four-equation model introduced here can also accommodate other demand shocks, such as discount rate shocks or expansionary fiscal policy. These demand shocks would trigger similar reallocations across firms and hence also raise aggregate TFP, like the monetary shocks highlighted here.

6 Calibration

We now calibrate both the static and dynamic versions of the model to assess the quantitative importance of the misallocation channel. This section is organized as follows. First, we describe how to calibrate the model without relying on an off-the-shelf functional form for the Kimball aggregator. Second, we calibrate the model using empirical pass-through estimates from Amiti et al. (2019) with Belgian firm-level data. We then report results from this calibration exercise: we compute the flattening of the Phillips curve due to real rigidities and misallocation in the static model and discuss comparative statics with respect to the Frisch elasticity and the degree of industrial concentration. At the end of the section, we turn to the dynamic model, where we present impulse response functions following a monetary policy shock.

6.1 Non-parametric Estimation Procedure

It may be tempting to use an off-the-shelf functional form for the Kimball aggregator and tune parameters to match moments from the data. However, there is no guarantee that parametric specifications of preferences are able to match the relevant features of the data

required for generating correct aggregate properties.³⁴ Instead, we follow Baqaee and Farhi (2020a) and back out the shape of the Kimball aggregator non-parametrically from the data. We summarize this approach below.

We start by assuming that all firms face the same residual demand curve: $\Upsilon'_\theta = \Upsilon'$. This means that in our calibration, time series and cross-sectional changes in markups are related. Firms charge different markups, and have different pass-throughs, because they are on different parts of the same demand curve. Order firms by their size and let $\theta \in [0, 1]$ be firm θ 's centile in the size distribution. Baqaee and Farhi (2020a) show that, in the cross-section, markups and sales must satisfy the following differential equation³⁵

$$\frac{d \log \mu_\theta}{d\theta} = (\mu_\theta - 1) \frac{1 - \rho_\theta}{\rho_\theta} \frac{d \log \lambda_\theta}{d\theta}. \quad (22)$$

Given data on sales shares λ_θ and pass-throughs ρ_θ , we can use this differential equation to solve for markups μ_θ up to a boundary condition. We choose the boundary condition to target a given value of the (harmonic) sales-weighted average markup, $\bar{\mu}$. We can then use $\sigma_\theta = 1/(1 - 1/\mu_\theta)$ to recover price-elasticities. The distributions of pass-throughs, markups, price elasticities, and sales shares are the sufficient statistics we need to calibrate the model.

6.2 Data and Parameter Values

We follow Baqaee and Farhi (2020a) to implement this procedure and provide additional details in Appendix C. We order firms by size and present the distributions of pass-through ρ_θ and sales share density $\log \lambda_\theta$ in Figure 2. Pass-throughs are strictly decreasing with firm size, which means that Marshall's strong second law holds.

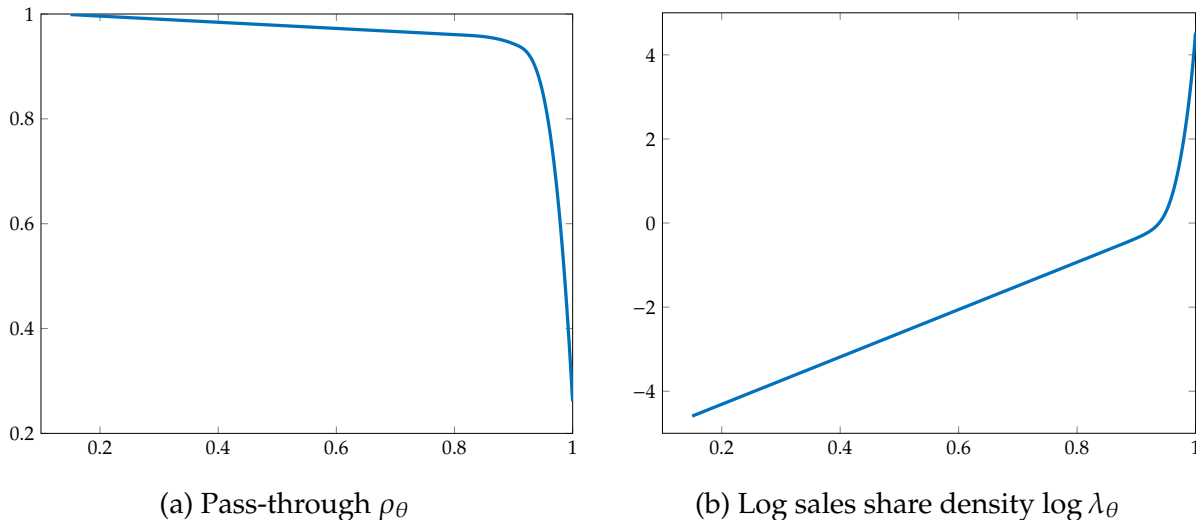
To impute the distribution of markups and elasticities from this data, our estimation procedure requires that we take a stance on the average markup. We assume that the average markup $\bar{\mu} = \mathbb{E}_\lambda [\mu_\theta^{-1}]^{-1} = 1.15$, in line with estimates from micro-data.³⁶

³⁴As an example, see Section 8 for a discussion of the unsuitability of the popular parametric family of preferences considered by Klenow and Willis (2016) for our application.

³⁵This follows from combining the following two differential equations: $\frac{d \log \lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1} \frac{d \log A_\theta}{d\theta}$, and $\frac{d \log \mu_\theta}{d\theta} = (1 - \rho_\theta) \frac{d \log A_\theta}{d\theta}$. Intuitively, compared to a firm of type θ , a firm of type $\theta + d\theta$ has higher productivity $\log A_{\theta+d\theta} - \log A_\theta = d \log A_\theta / d\theta$. The first differential equation uses the fact that the firm of type $\theta + d\theta$ will have lower price due to the pass-through of marginal cost, $\log p_{\theta+d\theta} - \log p_\theta = \rho_\theta d \log A_\theta / d\theta$, and higher sales $d \log \lambda_{\theta+d\theta} - \log \lambda_\theta = (\sigma_\theta - 1) \rho_\theta d \log A_\theta / d\theta$, with $\sigma_\theta - 1 = 1/(\mu_\theta - 1)$. The second differential equation uses the fact that $d \log \mu_\theta / d \log mc_\theta = \rho_\theta - 1$.

³⁶Konings et al. (2005) use micro-evidence to estimate price-cost margins in Bulgaria and Romania, and find that average price-cost margins range between 5-20% for nearly all sectors. In an earlier version of their paper, Amity et al. (2019) report that small firms in their calibration have a markup of around 14%, and large

Figure 2: Pass-through ρ_θ and sales share density $\log \lambda_\theta$ for Belgian manufacturing firms ordered by type θ .



This choice of the average markup, as well as the remaining parameter values, are listed in Table 1: We set the intertemporal elasticity of substitution $\gamma = 1$, which is a focal point for empirical estimates, and set the Frisch elasticity $\zeta = 0.2$ in line with recent estimates (see, for example, Chetty et al. 2011, Martinez et al. 2018, Sigurdsson 2019). For calibrating the static model, we consider a time period of approximately six months. Given an average price duration of one year (see Taylor 1999, Nakamura and Steinsson 2008), this means $\delta_\theta = \delta = 0.5$.

For the calibration of the dynamic model, we choose the coefficients on the Taylor rule, ϕ_π and ϕ_y , to match the calibration of the standard New Keynesian model given in Galí (2015). We also match Galí (2015) by setting the discount factor $\beta = 0.99$, corresponding to a 4% annual interest rate. We assume that monetary disturbances follow an AR(1) process $v_t = \rho_v v_{t-1} + \epsilon_t$, and set $\rho_v = 0.7$, indicating strong persistence to the interest rate shock, and set the size of the initial interest rate shock to 25 basis points. Finally, we set the period length in our dynamic calibration to one quarter, and hence set $\delta_\theta = \delta = 0.25$.

With our data for pass-throughs ρ_θ , sales shares λ_θ , and our choices for parameter values, we are now ready to present the estimates from our calibrated model. We first present estimates from the static model, and discuss comparative statics with respect to the Frisch elasticity and the degree of industrial concentration. Then, we present impulse

firms have markups of around 30%. These micro-estimated average markups are also broadly in line with macro estimates from Gutiérrez and Philippon (2017) and Barkai (2020), who estimate average markups on the order of 10-20%, but lower than estimates by De Loecker et al. (2020), who estimate the average markup for public firms at 61%.

Table 1: Parameters for empirical calibration.

Parameter	Description	Estimate
<i>Static model:</i>		
$\bar{\mu}$	(Harmonic) average markup	1.15
$1/\gamma$	Intertemporal elasticity of substitution	1
ζ	Frisch elasticity	0.2
δ	Calvo friction	0.5
<i>Dynamic model:</i>		
δ	Calvo friction (quarterly)	0.25
ϕ_y	Taylor rule coefficient on output gap	0.5 / 4
ϕ_π	Taylor rule coefficient on inflation gap	1.5
β	Discount factor	0.99
ϵ_0	Initial interest rate shock	25bp
ρ_v	Shock persistence	0.7

response functions from the dynamic model.

6.3 Results from Static Model

Table 2 reports the estimated flattening of the Phillips curve due to real rigidities and the misallocation channel (as given by Proposition 4). We find that the misallocation channel is quantitatively important: compared to the real rigidities channel, which flattens the wage Phillips curve by 27% and the price Phillips curve by 73%, the misallocation channel flattens both Phillips curves by 71%. This means that including supply-side effects increases the responsiveness of output to a monetary shock by 71%.

Table 2: Non-parametric estimates of Phillips curve flattening due to real rigidities and the misallocation channel.

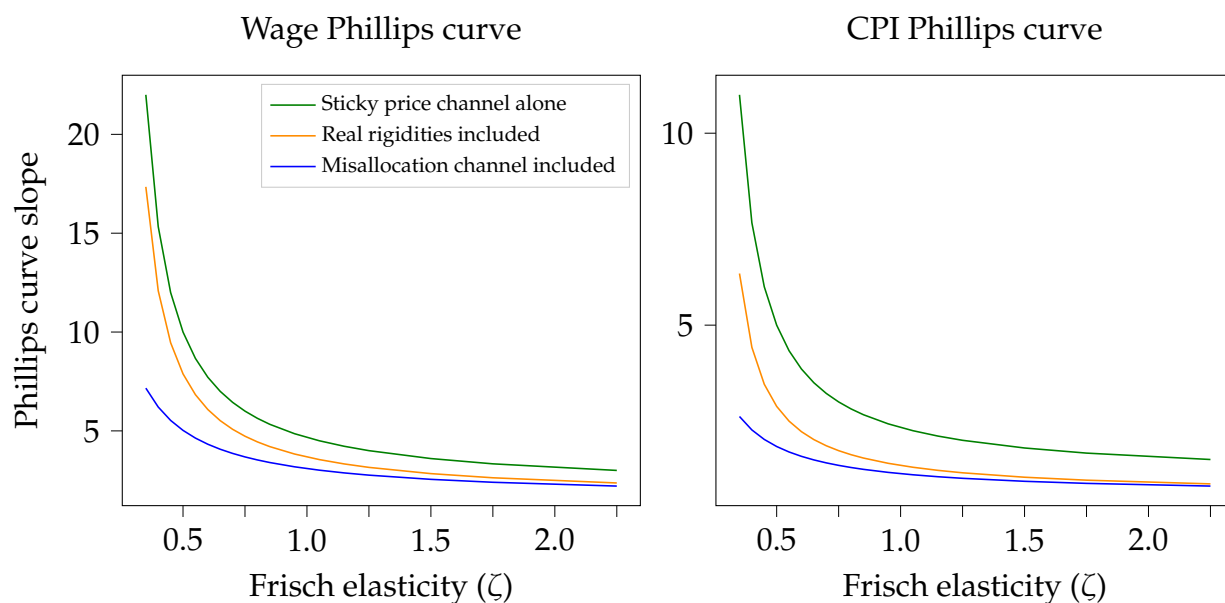
Flattening	Wage Phillips curve	CPI Phillips curve
Real rigidities	1.27	1.73
Misallocation channel	1.71	1.71

To highlight the key forces at play in this calibration, we consider how these estimates change as we vary the Frisch elasticity, the degree of industrial concentration, the average markup, and the level of price-stickiness.

The Frisch elasticity. The discussion following Propositions 1 and 2 shows that the misallocation channel should be more important for lower values of the Frisch elasticity of labor supply. This intuition is confirmed in Figure 3, where we plot the slope of the Phillips curve as a function of the Frisch elasticity. The flattening of the Phillips curve due to real rigidities does not depend on the Frisch elasticity. However, the flattening due to the misallocation channel increases dramatically as the Frisch elasticity approaches zero.

The introduction of the misallocation channel—and its increased strength at low Frisch elasticities—may help explain the discrepancy between micro-evidence on the Frisch elasticity and those required to explain the slope of the Phillips curve in traditional models. Evidence accumulated from quasi-experimental studies suggests that the labor supply elasticity is on the order of 0.1-0.4. In order to match the slope of the Phillips curve that the model with real rigidities and misallocation predicts at $\zeta = 0.2$, the model with nominal rigidities alone would require $\zeta \approx 1$. Incorporating the misallocation channel allows us to generate flatter Phillips curves at lower levels of the Frisch elasticity.

Figure 3: Decomposition of Phillips curve slope, varying the Frisch elasticity ζ .



Industrial concentration. Our analysis explicitly links the slope of the Phillips curve to characteristics of the firm distribution. A natural question, then, is how varying that firm distribution will affect the strength of the real rigidities and misallocation channels.

In order to illustrate the role of industrial concentration, we consider counterfactual firm distributions. To do so, we use a beta distribution for firm productivities, A_θ .³⁷ We

³⁷We choose the beta distribution since, as a bounded distribution, it allows us to remain within the

choose the shape parameters of the beta distribution, $a = 0.14$ and $b = 15.7$, to match the Gini coefficient of firm employment in the Belgian data and the slope of the price Phillips curve in our baseline calibration.

We then perturb the distribution by scaling a and b by a constant. Scaling the parameters of the beta distribution preserves the mean of the distribution while decreasing the variance, hence decreasing the concentration of firm employment. In Figure 4, we plot the slope of the Phillips curve against the Gini coefficient as we scale the parameters of the beta distribution. As the distribution in productivity becomes less concentrated, the employment distribution becomes more equal, and the Gini coefficient falls. As expected, the slope of the Phillips curve under nominal rigidities alone (as in the CES demand system) is unchanged as we vary the concentration of employment over this range. However, the strength of real rigidities and the misallocation channel do depend on the firm size distribution: the strength of both channels increases as we increase concentration.

To put these numbers into context, our model predicts that increasing the Gini coefficient from 0.80 to 0.85 flattens the price Phillips curve by an additional 14%. This experiment is in line with the increase in the Gini coefficient in firm employment from 1978 to 2018 in the United States (measured using the Census Business Dynamics Statistics, see Appendix J). Increasing the Gini coefficient from 0.72 to 0.86 (the increase in the Gini coefficient in the retail sector over the same period) flattens the price Phillips curve by 41%.

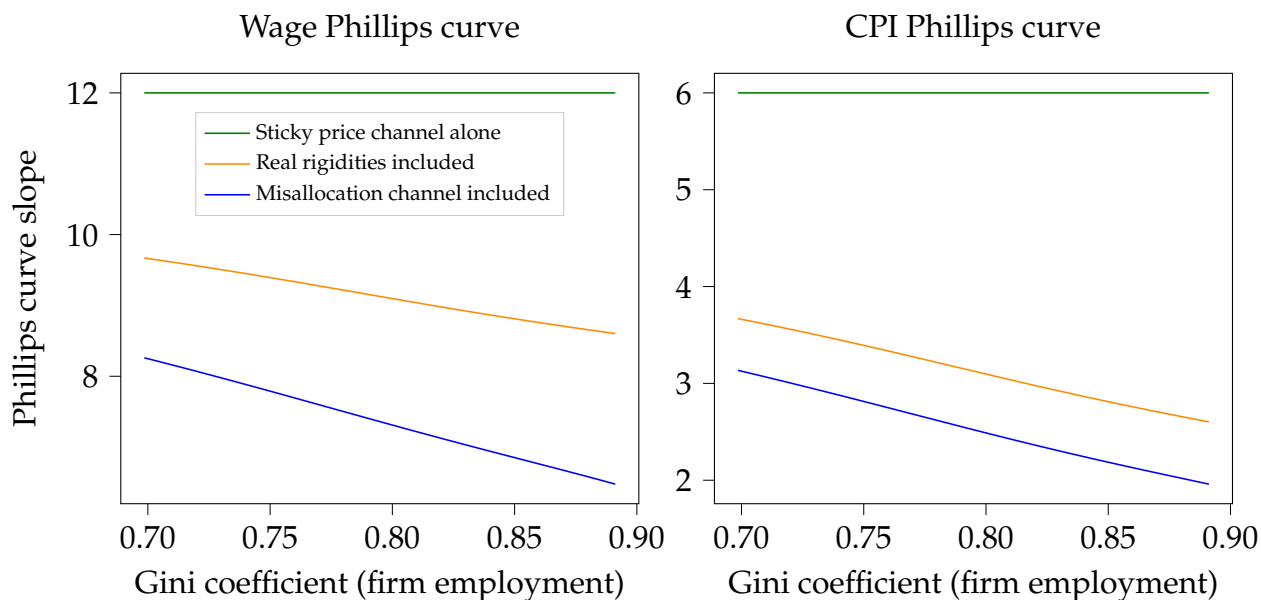
Other parameters. We show how the estimated slope of the Phillips curve changes as we vary the average markup $\bar{\mu}$ and the price-stickiness δ in Appendix C. We briefly summarize the results: Increasing the average markup $\bar{\mu}$ has no effect on the flattening due to real rigidities, but increases the flattening due to misallocation channel linearly. This means that our estimates of the misallocation channel are conservative relative to those implied by the average markups estimated by De Loecker and Eeckhout (2018). Increasing the price flexibility parameter δ increases the flattening of the price Phillips curve due to the misallocation channel and decreases the flattening due to real rigidities, for reasons explained after the statement of Proposition 4.

6.4 Results from Dynamic Model

Figure 5 shows the impulse response functions of aggregate variables following a persistent, 25 basis point (100bp annualized) shock to the interest rate in the dynamic model. We

range of productivities for which we have estimated the Kimball aggregator.

Figure 4: The slope of the Phillips curve, and its decomposition, as a function of the Gini coefficient of the employment distribution.



compare the benchmark heterogeneous firm model to a homogeneous firm model, which has real rigidities but no misallocation channel, and a CES model, which has neither real rigidities nor the misallocation channel.

In the CES and homogeneous firms case, aggregate TFP does not react to the monetary shock, as implied by Corollary 1. In contrast, when firms have heterogeneous markups, the dispersion in TFPR across firm types increases by around 30 basis points following the contractionary shock, and the response of aggregate TFP is procyclical and hump-shaped.³⁸ The fall in aggregate TFP dampens the extent of disinflation following the contractionary monetary shock and deepens the immediate response of output to the shock.

Figure 6 zooms in on the output response following the shock. The panel on the right computes the ratio of the output response across models: the orange line is the ratio of the output response in the homogeneous firm Kimball model relative to the CES model, and the blue line is the ratio of the output response in the full model relative to the homogeneous firm Kimball model. We find that the contraction in output in the full model is about 20% deeper on impact than in the homogeneous firm model. The difference widens over time, because the shock's effect on output is more persistent in the

³⁸For comparison, Kehrig (2011) finds that TFPR dispersion increases about 10% during recessions and increased over 20% from 2007 to the trough of the recession in 2009.

Figure 5: Impulse response functions (IRFs) following a 25bp monetary shock. Green, orange, and blue IRFs indicate the CES, homogeneous firms, and heterogeneous firms models respectively.

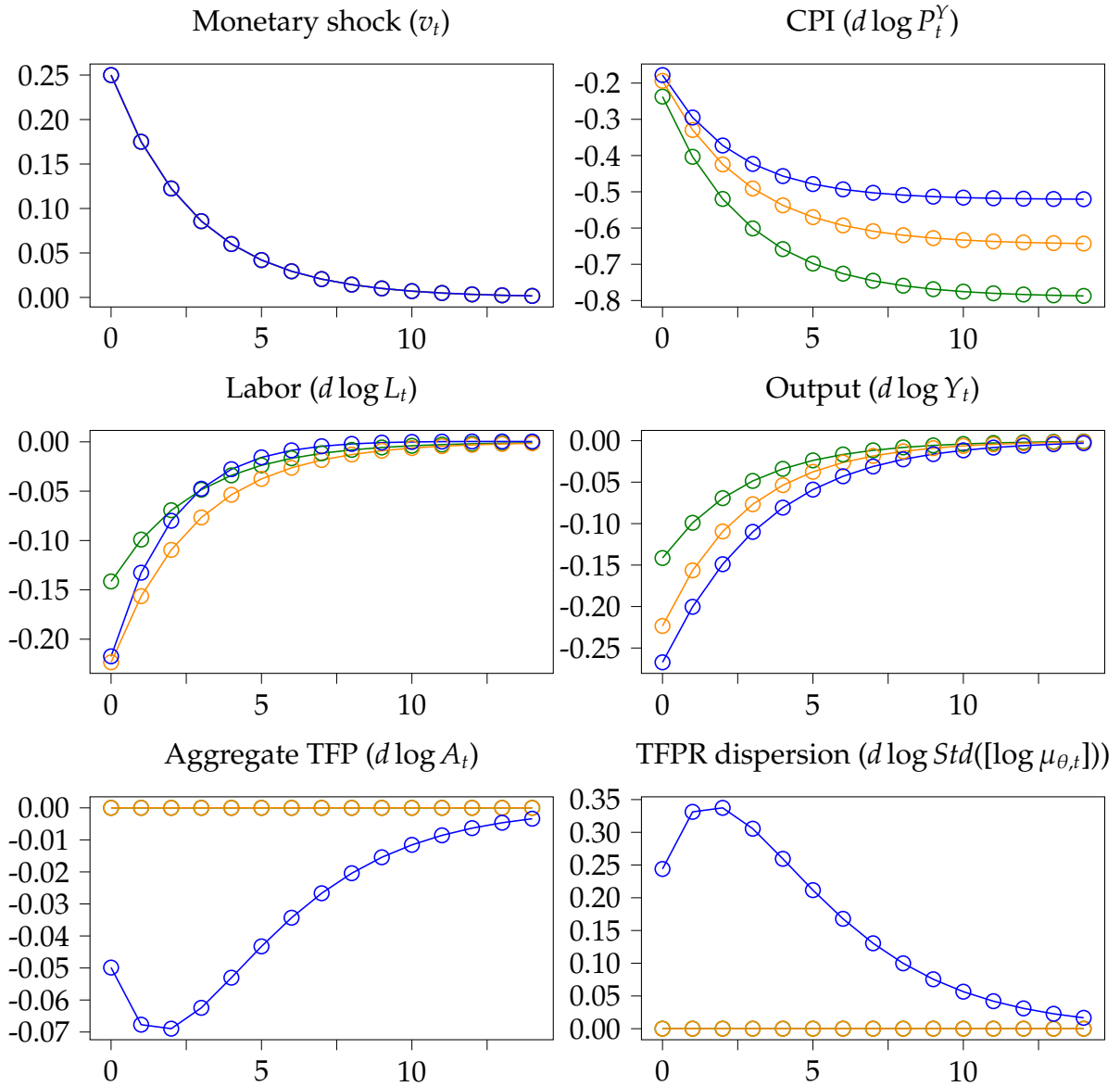
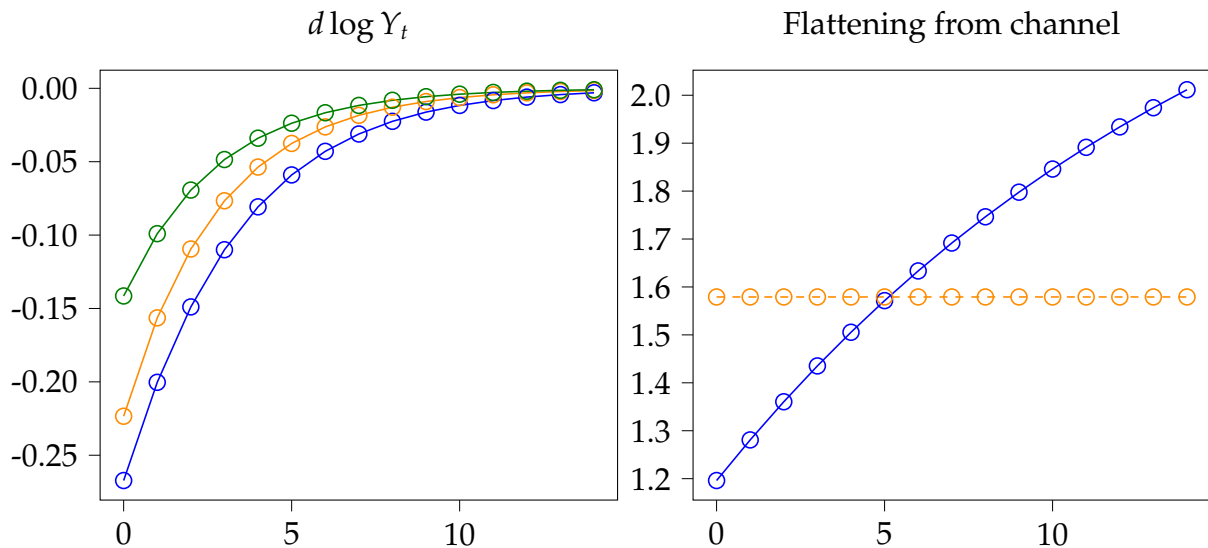


Figure 6: Impulse response function of output following a 25bp monetary policy shock. In the left panel green, orange, and blue IRFs indicate the CES, homogeneous firms, and heterogeneous firms models respectively. In the right panel, the orange and blue lines are the ratios of the output response of the homogeneous firm model relative to the CES model and the heterogenous firm model relative to the homogeneous firm model, respectively.



full model.

We quantify this effect on persistence by calculating the half-life of the shock on output in Table 3. The CES and homogeneous firm models feature a constant half-life of just under two quarters; the misallocation channel increases the half-life of the shock by 23% to about 2.4 quarters.³⁹ Table 3 also reports the cumulative impact on output of the monetary shock in the different models. The misallocation channel increases the cumulative output impact of monetary shocks by around 25 log points, which is about the same as the real rigidities channel.

Table 3: Effect of monetary policy shock on output.

Model	Output effect at $t = 0$	Half life	Cumulative output impact
CES	-0.14	1.95	-0.47
Homogeneous Firms	-0.22	1.95	-0.75
Heterogeneous Firms	-0.27	2.39	-1.02

We provide additional calibration results in Appendix C. In particular, we report the

³⁹Due to the second-order difference equation in aggregate TFP, the full model no longer features a constant half-life.

Table 4: Procyclical aggregate productivity.

% ΔTFP	Labor productivity			Solow residual			Cost-based Solow residual		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
<i>Unemp.</i>	-0.355** (0.126)			-0.465** (0.141)			-0.477** (0.142)		
<i>Recession</i>		-0.878** (0.394)			-2.114** (0.414)			-2.082** (0.500)	
% ΔGDP			0.221** (0.087)			0.209* (0.106)			0.354** (0.097)
Period	1948-2020			1948-2020			1961-2014		

Notes: *Unemp.* is the average unemployment rate in year $t + 1$, % ΔGDP is real GDP growth from year $t - 1$ to t , and *Recession* = 1 if any quarter in the year is marked an NBER recession. Robust standard errors in parentheses. * indicates significance at 10%, ** at 5%.

change in sales shares for firms at different percentiles of the size distribution. The sales shares of small firms are about one order of magnitude more volatile than output, whereas the sales shares of the largest firms are less volatile than output. In Appendix E, we show that results are quantitatively similar when monetary policy is implemented via changes in money supply (with a cash-in-advance constraint) rather than an interest rate rule.

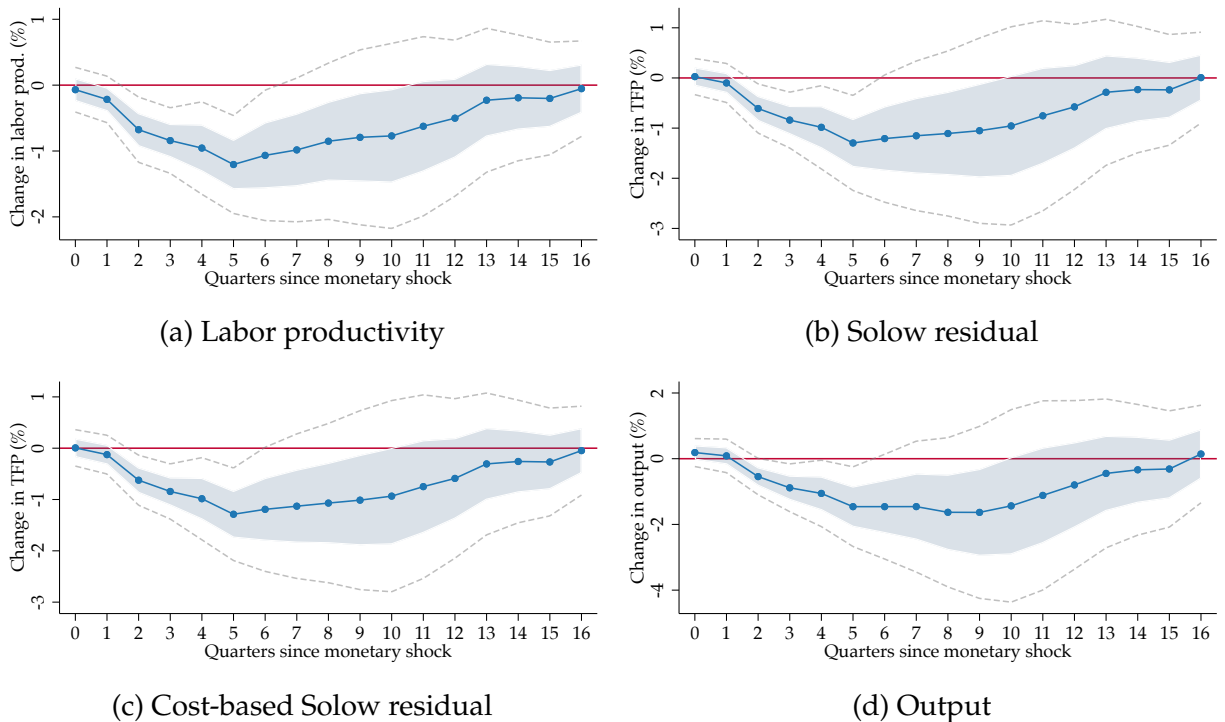
All in all, our results suggest that the misallocation channel is potentially as powerful as the real rigidities channel in affecting the transmission of monetary policy.

7 Empirical Evidence

In this section, we provide some suggestive empirical evidence in support of the reallocation mechanism described in this paper. We first present macro-level evidence on the cyclicity of aggregate TFP and its response to exogenous monetary shocks. We then show, at the micro-level, that expansions are accompanied by procyclical reallocations to high-markup firms, as predicted by our model.

Macro-level evidence. Table 4 shows the unconditional association between various measures of aggregate productivity and the business cycle. We use three different measures of aggregate productivity—labor productivity, the Solow residual, and the cost-based Solow residual (see Hall, 1990)—and three different measures of the business cycle—the unemployment rate, NBER recession dates, and real GDP growth. For all mea-

Figure 7: Local projection of a 25bp contractionary Romer and Romer (2004) shock on aggregate productivity. The shaded region indicates robust standard errors, and dashed lines are 95% confidence intervals.



tures, we find that productivity covaries significantly with business cycle indicators.⁴⁰

We do not use utilization-adjusted TFP (e.g. Basu et al., 2006; Fernald, 2014) in Table 4. This is because these series are identified using the assumption that sectoral productivity is orthogonal to monetary shocks, and this exogeneity condition fails in our model.

In Figure 7, we show that aggregate productivity falls after contractionary monetary policy shocks, using Romer and Romer (2004) monetary shocks from 1969-1996. Following a 25bp contractionary shock, there is a significant negative contraction in aggregate productivity for two to six quarters. The magnitude of the decline in aggregate productivity is roughly half of the effect on output. These movements in aggregate productivity are larger than those predicted by our model, which suggests that allocative effects explain part (but perhaps not all) of the procyclical movements of aggregate productivity.

⁴⁰We use measures of labor productivity and the Solow residual for the U.S. business sector provided by the Federal Reserve Bank of San Francisco for the period 1948-2020. To calculate cost-based Solow residual, we use the aggregate markup, estimated using sales and accounting profits of Compustat firms from 1961-2014, to estimate input cost shares.

Micro-level evidence. In our model, the procyclical movements of aggregate TFP arise from endogenous, procyclical reallocations toward high-markup firms. We now turn to micro-level evidence on these reallocations. To do so, we use estimates of markups for publicly traded firms in Compustat and explore how the growth of high- and low-markup firms changes over the business cycle. Of course, this exercise must be interpreted with caution since measuring markups accurately at high frequency is notoriously challenging and Compustat is not a representative sample of all US producers. Nevertheless, our empirical results, which are only suggestive, are supportive of the basic mechanism underlying the misallocation channel.

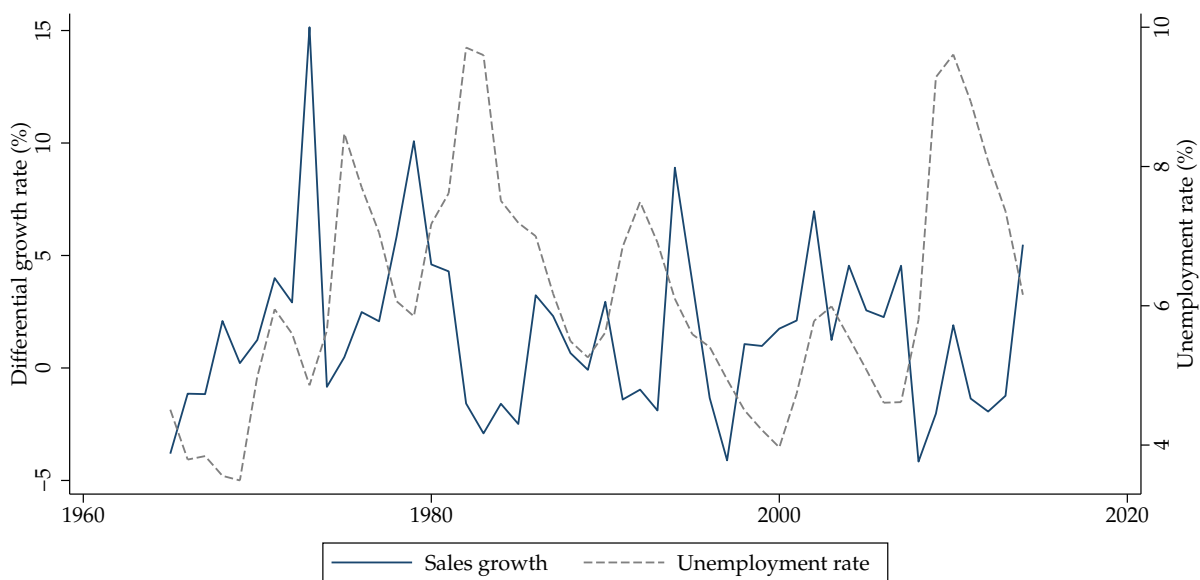


Figure 8: Procyclical reallocations to high-markup firms. A firm is categorized as high-markup (low-markup) if its markup is above (below) median in year t . The solid line shows the difference in sales growth of high- and low-markup firms from t to $t + 1$.

Figure 8 shows the difference in the sales growth of high- and low-markup firms from 1965-2015. We use accounting profits to estimate firm markups in year t and split public firms into high-markup (above median) and low-markup (below median) groups.⁴¹ We then calculate the difference in the sales growth of both groups from year t to $t + 1$.⁴² As shown in Figure 8, the differential growth rate shows substantial variance over the sample and appears to be correlated with the business cycle (as shown by the unemployment rate).

⁴¹Specifically, we assume operating income is profit and infer the markup by assuming firms have constant returns to scale. We use accounting profits in Figure 8 since that allows us to plot the series for the longest sample.

⁴²For each year t , we limit our analysis to firms in the sample in both years t and $t + 1$.

We formalize this analysis by considering several markup series used in the literature and several measures of the business cycle. To estimate markups, we use (1) the user-cost approach (UC), (2) the production function estimation approach (PF), and the accounting profits approach (AP). The user cost approach relies on measuring each firm’s capital stock and user-cost of capital, for which we follow Gutiérrez and Philippon (2017) and Gutiérrez (2017), both excluding and including intangibles in a firm’s capital stock (UC1 and UC2).⁴³ The production function approach, estimated following De Loecker and Warzynski (2012) and De Loecker et al. (2020), infers the markup by comparing the expenditure share of a variable input to its output elasticity (in our case, the variable input is “cost of goods sold”). Finally, the accounting profits approach assumes that firms have constant returns to scale and that operating income is equal to profits (though simplistic, this approach requires very little manipulation of the raw data and produces similar results to the more sophisticated approaches.) Appendix D describes the data sources and assumptions underlying each series in more detail. While each markup series has advantages and disadvantages, our hope is that, by comparing results across all four series, our results are generalizable beyond the idiosyncrasies of any specific estimation procedure.

We estimate two specifications,

$$SalesGrowth_{i,t \rightarrow t+1} = b_1 \cdot (HighMarkup_{i,t} \times Cycle_t) + c_1 Cycle_t + d_1 HighMarkup_{i,t} + \chi_t + \epsilon_{i,t},$$

$$CostGrowth_{i,t \rightarrow t+1} = b_2 \cdot (HighMarkup_{i,t} \times Cycle_t) + c_2 Cycle_t + d_2 HighMarkup_{i,t} + \chi_t + \epsilon_{i,t},$$

where $HighMarkup_{i,t}$ is an indicator for whether the firm’s markup is above or below the median among Compustat firms in year t , $SalesGrowth_{i,t \rightarrow t+1}$ is the growth in firm sales from year t to $t + 1$, and $CostGrowth_{i,t \rightarrow t+1}$ is the growth in a firm’s input usage, measured as sales divided by the firm’s markup, from t to $t + 1$.⁴⁴ As in our analysis of procyclical aggregate productivity, we use three measures of the business cycle: (1) lagged unemployment, (2) NBER recessions, and (3) GDP growth.

Table 5 shows estimates of \hat{b}_1 and \hat{b}_2 for different markup series and measures of the business cycle. In nearly all specifications, expansions are associated with reallocations of both sales and input costs toward high-markup firms, in line with the predictions of the model.⁴⁵

⁴³The user-cost approach is also used by Foster et al. (2008) and by De Loecker et al. (2020) as a robustness check.

⁴⁴Under constant returns to scale, λ/μ is equal to a firm’s total input costs. More generally, λ/μ is the cost inclusive of payments to fixed factors (or, equivalently, total costs plus Ricardian rents).

⁴⁵In Appendix D, we replicate the analysis grouping firms by above or below median within industry (2-digit NAICS). The results are similar, suggesting that this effect is not driven exclusively by reallocations across industries.

Table 5: Procyclical reallocations to high-markup firms.

Cycle measure	Markup series	Sales ($\% \Delta \lambda_{t \rightarrow t+1}$)		Costs ($\% \Delta (\lambda/\mu)_{t \rightarrow t+1}$)	
		Estimate	Standard error	Estimate	Standard error
<i>Unemployment</i>	UC1	-1.117*	(0.560)	-0.929**	(0.453)
<i>Unemployment</i>	UC2	-1.487**	(0.527)	-1.103**	(0.420)
<i>Unemployment</i>	PF	-0.006	(0.645)	-0.005	(0.646)
<i>Unemployment</i>	AP	-0.776**	(0.285)	-0.630**	(0.229)
<i>Recession</i>	UC1	-5.873**	(2.120)	-4.884**	(1.736)
<i>Recession</i>	UC2	-4.261**	(1.974)	-2.903**	(1.375)
<i>Recession</i>	PF	-3.797**	(1.735)	-4.248**	(1.812)
<i>Recession</i>	AP	-0.683	(1.189)	-0.395	(0.993)
$\% \Delta GDP$	UC1	1.172**	(0.387)	0.931**	(0.245)
$\% \Delta GDP$	UC2	1.448**	(0.351)	1.034**	(0.347)
$\% \Delta GDP$	PF	0.510	(0.414)	0.580	(0.403)
$\% \Delta GDP$	AP	0.496**	(0.161)	0.296*	(0.149)

Notes: The estimate columns report \hat{b}_1 and \hat{b}_2 from the specifications:

$$SalesGrowth_{i,t \rightarrow t+1} = b_1 \cdot (HighMarkup_{i,t} \times Cycle_t) + c_1 Cycle_t + d_1 HighMarkup_{i,t} + \chi_t + \epsilon_{i,t},$$

$$CostGrowth_{i,t \rightarrow t+1} = b_2 \cdot (HighMarkup_{i,t} \times Cycle_t) + c_2 Cycle_t + d_2 HighMarkup_{i,t} + \chi_t + \epsilon_{i,t}.$$

$HighMarkup_{i,t}$ is an indicator for whether a firm's markup in year t is above the median markup of all public firms in year t . *Unemployment* is the average unemployment rate from Q4 of year t to Q4 of year $t + 1$, $\% \Delta GDP$ is real GDP growth to Q4 of year t , and *Recession* = 1 if any quarter in the previous year is marked an NBER recession. Markups are estimated using the user-cost approach (UC1 and UC2, where UC2 includes intangibles in measured capital), the production function estimation approach (PF), and the accounting profits approach (AP). Regressions are weighted by sales, and standard errors are clustered by year and firm. * indicates significance at 10%, ** at 5%.

8 Extensions

Before concluding, we summarize some extensions that are developed in detail in the appendices.

Multiple sectors, multiple factors, input-output linkages, and sticky wages. The model we use in the main text of the paper is deliberately stylized for clarity. In particular, it is missing some ingredients that are quantitatively important for how output responds to monetary shocks, but that are unrelated to the mechanism this paper studies. In particular, our model has only one sector and only one factor of production, labor. In Appendix F, we show how to extend the model to have a general production network structure, with multiple sectors and multiple factors. As an example, in Appendix F.1 we consider an economy with two factors (labor and capital), a firm sector, and a “labor union” sector that generates sticky wages. The intuition underlying the supply-side effects of a monetary shock are unchanged in this extension compared to the model presented in the main text, and we find that the misallocation channel remains similar in magnitude.

Variation in markups and pass-throughs unrelated to size. In our calibrations, we assume that there are no firm-specific taste shifters and all firms face the same residual demand curve. Furthermore, we also calibrate the model starting at a zero-inflation steady-state where all firms charge their desired markups. This means that, at the initial point, markups are a monotone function of firm-size in our model. Whilst markups and pass-throughs do vary as a function of firm size (e.g. see Burstein et al., 2020 or Amiti et al., 2019), in practice, firm markups and pass-throughs also vary for reasons unrelated to size, such as firm-specific shifters in demand curves, quality differences, or markup dispersion inherited from previous periods. In Appendix I, we show how our baseline results change if there is variation in markups and pass-throughs unrelated to size. We show that the supply-side effects of monetary policy are strengthened if the excess variation in markups is negatively correlated with the excess variation in pass-throughs, and weakened if this correlation is positive. When excess variation in markups and pass-throughs are orthogonal, then the presence of the noise does not affect the strength of supply-side effects of monetary policy relative to our benchmark calibration.

Klenow and Willis (2016) calibration. In the main text of the paper, we caution against using off-the-shelf functional forms for preferences. We illustrate this by calibrating our model with the commonly used Klenow and Willis (2016) specification in Appendix G. We

show that to match the observed relationship between pass-through and firm-size, with near complete pass-through for small firms and very incomplete pass-through for large firms, large firms must have markups that are on the order of 10,000%. Under standard calibrations, which do not produce astronomically large markups for large firms, the implied pass-through function does not vary much as a function of firm-size. Therefore, under standard calibrations, these preferences fail to capture the cross-sectional covariance between pass-throughs and markups, and hence imply counterfactually small supply-side effects.

Oligopoly calibration. In the main text of the paper, we model a continuum of firms in monopolistic competition. An alternative is to consider an economy composed of oligopolistic markets. We develop our static model under the nested CES structure used by Atkeson and Burstein (2008) and compute the flattening of the Phillips curve due to real rigidities and the misallocation channel in this setting. The results, reported in Appendix H, are qualitatively and quantitatively similar to the monopolistic competition calibration in Section 6.

Additional calibration results. In the calibration section of the paper, we provide comparative statics of our calibration results with respect to the Frisch elasticity and the degree of industrial concentration. In Appendix C, we also provide comparative statics with respect to the aggregate markup and the level of price rigidity. We also report impulse response functions for an exogenous money supply shock in Appendix E.

9 Conclusion

We analyze the transmission of monetary policy in an economy with heterogeneous firms, variable desired markups and pass-throughs, and sticky prices. In contrast to the benchmark New Keynesian model, where the envelope theorem renders reallocations irrelevant to output, we find that in the enriched model monetary shocks have quantitatively significant effects on aggregate output and productivity via reallocations.

These results accord with evidence at both the micro level, where previous studies document that dispersion in plant- and firm-level revenue productivity is countercyclical, and at the macro level, where previous studies document procyclical movements in aggregate TFP. We link these pieces of evidence and show how monetary shocks can generate both effects.

In this paper, we focus on monetary policy shocks, but the same intuition applies to other kinds of demand shocks, such as discount factor or fiscal policy shocks. In general, demand shocks that raise nominal marginal costs will tend to increase TFP and reduce firm-level TFPR dispersion as long as realized pass-throughs covary negatively with markups.

Furthermore, in this paper, we focus on heterogeneous markups in product markets, but it is possible that similar distortions could exist in inputs markets. Specifically, if firms have heterogeneous and variable monopsony power in the labor market, then TFP would be procyclical if firms with relatively high markdowns reduce their markdowns during booms and raise them during recessions. Additionally, our analysis abstracts from normative considerations like the optimal conduct of monetary policy. We are pursuing these extensions in ongoing work.

References

- Alvarez, F., H. L. Bihan, and F. Lippi (2016). The real effects of monetary shocks in sticky price models: A sufficient statistic approach. *American Economic Review* 106(10), 2817–2851.
- Amiti, M., O. Itskhoki, and J. Konings (2019). International shocks, variable markups, and domestic prices. *The Review of Economic Studies* 86(6), 2356–2402.
- Andrés, J. and P. Burriel (2018). Inflation and optimal monetary policy in a model with firm heterogeneity and bertrand competition. *European Economic Review* 103, 18–38.
- Anzoategui, D., D. Comin, M. Gertler, and J. Martinez (2019). Endogenous technology adoption and r&d as sources of business cycle persistence. *American Economic Journal: Macroeconomics* 11(3), 67–110.
- Atkeson, A. and A. Burstein (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review* 98(5), 1998–2031.
- Auer, R. A., T. Chaney, and P. Sauré (2018). Quality pricing-to-market. *Journal of International Economics* 110, 87–102.
- Autor, D., D. Dorn, L. F. Katz, C. Patterson, and J. V. Reenen (2020). The fall of the labor share and the rise of superstar firms. *The Quarterly Journal of Economics* 135(2), 645–709.
- Ball, L. and D. Romer (1990). Real rigidities and the non-neutrality of money. *Review of Economic Studies* 57(2), 183–203.
- Baqee, D. R. and E. Farhi (2017). Productivity and misallocation in general equilibrium. Technical Report 24007, National Bureau of Economic Research.

- Baqae, D. R. and E. Farhi (2018). Macroeconomics with heterogeneous agents and input-output networks. Technical Report 24684, National Bureau of Economic Research.
- Baqae, D. R. and E. Farhi (2020a). The darwinian returns to scale. Technical Report 27139, National Bureau of Economic Research.
- Baqae, D. R. and E. Farhi (2020b). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics* 135(1), 105–163.
- Barkai, S. (2020). Declining labor and capital shares. *The Journal of Finance* 75(5), 2421–2463.
- Basu, S., J. G. Fernald, and M. S. Kimball (2006). Are technology improvements contractionary? *American Economic Review* 96(5), 1418–1448.
- Benigno, G. and L. Fornaro (2018). Stagnation traps. *The Review of Economic Studies* 85(3), 1425–1470.
- Berman, N., P. Martin, and T. Mayer (2012). How do different exporters react to exchange rate changes? *The Quarterly Journal of Economics* 127(1), 437–492.
- Bianchi, F., H. Kung, and G. Morales (2019). Growth, slowdowns, and recoveries. *Journey of Monetary Economics* 101, 47–63.
- Burstein, A., V. M. Carvalho, and B. Grassi (2020). Bottom-up markup fluctuations. Technical report, National Bureau of Economic Research.
- Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journey of Monetary Economics* 12(3), 383–398.
- Chatterjee, A., R. Dix-Carneiro, and J. Vichyanond (2013). Multi-product firms and exchange rate fluctuations. *American Economic Journal: Economic Policy* 5(2), 77–110.
- Chen, N. and L. Juvenal (2016). Quality, trade, and exchange rate pass-through. *Journal of International Economics* 100, 61–80.
- Chetty, R., A. Guren, D. Manoli, and A. Weber (2011). Are micro and macro labor supply elasticities consistent? a review of evidence on the intensive and extensive margins. *American Economic Review* 101(3), 471–475.
- Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy* 113(1), 1–45.
- Claus, J. and J. Thomas (2001). Equity premia as low as three percent? evidence from analysts' earnings forecasts for domestic and international stock markets. *The Journal of Finance* 56(5), 1629–1666.
- Comin, D. and M. Gertler (2006). Medium-term business cycles. *American Economic Review* 96(3), 523–551.
- Corhay, A., H. Kung, and L. Schmid (2020). Q: Risk, rents, or growth? Technical report, Working Paper.
- Cozier, B. and R. Gupta (1993). Is productivity exogenous over the business cycle? some

- canadian evidence on the solow residual. Technical report, Bank of Canada.
- Cravino, J. (2017). Exchange rates, aggregate productivity and the currency of invoicing of international trade. Technical report.
- David, J. and D. Zeke (2021, January). Risk-taking, capital allocation and optimal monetary policy. Technical report.
- De Loecker, J. and J. Eeckhout (2018). Global market power. Technical Report 24768, National Bureau of Economic Research.
- De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics* 135(2), 561–644.
- De Loecker, J. and F. Warzynski (2012). Markups and firm-level export status. *American Economic Review* 102(6), 2437–2471.
- Dotsey, M. and R. G. King (2005). Implications of state-dependent pricing for dynamic macroeconomic models. *Journey of Monetary Economics* 52(1), 213–242.
- Eichenbaum, M. and J. Fisher (2004). Evaluating the calvo model of sticky prices. Technical report, National Bureau of Economic Research.
- Etro, F. and L. Rossi (2015). New-keynesian phillips curve with bertrand competition and endogenous entry. *Journal of Economic Dynamics and Control* 51, 318–340.
- Evans, C. L. (1992). Productivity shocks and real business cycles. *Journey of Monetary Economics* 29(2), 191–208.
- Evans, C. L. and F. T. dos Santos (2002). Monetary policy shocks and productivity measures in the g-7 countries. *Portuguese Economic Journal* 1(1), 47–70.
- Fernald, J. (2014). A quarterly, utilization-adjusted series on total factor productivity. Technical report, Federal Reserve Bank of San Francisco.
- Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review* 98(1), 394–425.
- Galí, J. (2015). *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press.
- Gopinath, G. and O. Itskhoki (2011). In search of real rigidities. *NBER Macroeconomics Annual* 25(1), 261–310.
- Gopinath, G., O. Itskhoki, and R. Rigobon (2010). Currency choice and exchange rate pass-through. *American Economic Review* 100(1), 304–336.
- Gutiérrez, G. (2017). Investigating global labor and profit shares. Working Paper.
- Gutiérrez, G. and T. Philippon (2017). Declining competition and investment in the u.s. Technical Report 23583, National Bureau of Economic Research.
- Hall, R. E. (1990). Invariance properties of solow’s productivity residual. In P. Diamond

- (Ed.), *Growth, Productivity, Unemployment: Essays to Celebrate Bob Solow's Birthday*, pp. 71–112. Cambridge: MIT Press.
- Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly Journal of Economics* 124(4), 1403–1448.
- Kehrig, M. (2011). The cyclicalty of productivity dispersion. Technical Report CES-WP-11-15, US Census Bureau Center for Economic Studies.
- Kim, S. and H. Lim (2004). Does solow residual for korea reflect pure technology shocks? Technical report, Seoul, Korea.
- Kimball, M. S. (1995). The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit and Banking* 27(4), 1241–77.
- Klenow, P. J. and J. L. Willis (2016). Real rigidities and nominal price changes. *Economica* 83(331), 443–472.
- Konings, J., P. V. Cayseele, and F. Warzynski (2005). The effects of privatization and competitive pressure on firms' price-cost margins: Micro evidence from emerging economies. *The Review of Economics and Statistics* 87(1), 124–134.
- Levinsohn, J. and A. Petrin (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies* 70(2), 317–341.
- Li, H., H. Ma, and Y. Xu (2015). How do exchange rate movements affect chinese exports? a firm-level investigation. *Journal of International Economics* 97(1), 148–161.
- Martinez, I. Z., E. Saez, and M. Siegenthaler (2018). Intertemporal labor supply substitution? evidence from the swiss income tax holidays. Technical Report 24634, National Bureau of Economic Research.
- Matsuyama, K. and P. Ushchev (2017). Beyond ces: Three alternative classes of flexible homothetic demand systems.
- Meier, M. and T. Reinelt (2020, June). Monetary policy, markup dispersion, and aggregate tfp. *ECB Working Paper No. 2427*.
- Melitz, M. J. (2018). Competitive effects of trade: Theory and measurement. *Review of World Economics* 154(1), 1–13.
- Nakamura, E. and J. Steinsson (2008). Five facts about prices: A reevaluation of menu cost models. *The Quarterly Journal of Economics* 123(4), 1415–1464.
- Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications industry. *Econometrica* 64(6), 1263–1297.
- Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics* 11(4), 707–720.
- Romer, C. D. and D. H. Romer (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review* 94(4), 1055–1084.

- Sigurdsson, J. (2019). Labor supply responses and adjustment frictions: A tax-free year in iceland. Technical Report 3278308, SSRN.
- Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of Political Economy* 88(1), 1–23.
- Taylor, J. B. (1999). Staggered price and wage setting in macroeconomics. *Handbook of Macroeconomics* 1, 1009–1050.
- Wang, O. and I. Werning (2020). Dynamic oligopoly and price stickiness. Technical Report 27536, National Bureau of Economic Research.

Appendix A Proofs for Static Model

In this appendix, we provide derivations for the one-period model. We start by expressing the full set of equilibrium conditions as a function of endogenous equilibrium objects and exogenous parameters. We then characterize the response of firm markups, quantities, and sales shares in response to a nominal wage shock. By aggregating across firms, we derive the changes in aggregate variables, which yields a fixed point in changes in output.

A.1 Equilibrium conditions

Note that we can express each firm's output y_θ and price p_θ as a function of the sales density λ_θ and markup μ_θ :

$$y_\theta = \frac{\bar{\mu}L\lambda_\theta A_\theta}{\mu_\theta} \quad \text{and} \quad \frac{p_\theta}{w} = \frac{\mu_\theta}{A_\theta}. \quad (23)$$

This allows us to write all equilibrium conditions in terms of the endogenous equilibrium variables $\lambda_\theta, \mu_\theta, P, \bar{\mu}, L$, and Y , and the exogenous parameters $A_\theta, \delta_\theta, w, \zeta$, and γ . We restate the equilibrium conditions from the main text.

1. *Consumers maximize utility.* Consumer demand for each good is given by the inverse-demand curve in Equation (1), which becomes

$$\frac{\mu_\theta}{A_\theta} = wP\Upsilon'_\theta\left(\frac{\bar{\mu}L\lambda_\theta A_\theta}{\mu_\theta Y}\right), \quad (24)$$

with the constraint on utility from the consumption bundle, Y , given in Section 2.1, expressed as,

$$1 = \mathbb{E}\left[\Upsilon_\theta\left(\frac{\bar{\mu}L\lambda_\theta A_\theta}{\mu_\theta Y}\right)\right]. \quad (25)$$

The aggregate markup is

$$\bar{\mu} = \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1} = \frac{\Pi}{wL + \Pi}. \quad (26)$$

Finally, the consumer maximizes overall utility from the consumption bundle and labor supply such that

$$\bar{\mu}L^{1+\frac{1}{\zeta}} = Y^{1-\gamma}. \quad (27)$$

2. *Firms maximize profits.* Flexible-price firms set markups according to the Lerner

formula,

$$\mu_{\theta}^{\text{flex}} = \mu_{\theta}(\bar{\mu} \frac{L\lambda_{\theta}A_{\theta}}{\mu_{\theta}Y}), \quad (28)$$

while the markups of sticky-price firms move due to changes in marginal cost that are out of the firm's control,

$$\mu_{\theta}^{\text{sticky}} = \frac{w_{t-1}}{w_t} \frac{A_{\theta,t}}{A_{\theta,t-1}} \mu_{\theta,t-1}. \quad (29)$$

3. *Markets clear.* The price aggregator satisfies

$$\frac{1}{P} = \frac{Y}{wL} \mathbb{E} \left[\frac{L\lambda_{\theta}A_{\theta}}{\mu_{\theta}Y} \Upsilon_{\theta} \left(\frac{\bar{\mu}L\lambda_{\theta}A_{\theta}}{\mu_{\theta}Y} \right) \right]. \quad (30)$$

A.2 Response to monetary policy shocks

We consider a shock to the nominal wage given by $d \log w$. All other exogenous primitives (A_{θ} , δ_{θ} , ζ , and γ) are unchanged by the shock. First, we characterize changes in firm-level markups $d \log \mu_{\theta}$, quantities $d \log y_{\theta}$, and sales shares $d \log \lambda_{\theta}$.

Markups. The markup of a firm of type θ satisfies

$$d \log \mu_{\theta} = d \log p_{\theta} - d \log w. \quad (31)$$

The prices of sticky-price firms do not change, hence

$$d \log \mu_{\theta}^{\text{sticky}} = -d \log w. \quad (32)$$

For flexible-price firms,

$$d \log p_{\theta}^{\text{flex}} = d \log \mu_{\theta} \left(\frac{y_{\theta}}{Y} \right) + d \log w \quad (33)$$

$$= \frac{\frac{y_{\theta}}{Y} \mu'_{\theta} \left(\frac{y_{\theta}}{Y} \right)}{\mu_{\theta}} d \log \left(\frac{y_{\theta}}{Y} \right) + d \log w \quad (34)$$

$$= \frac{1 - \rho_{\theta}}{\rho_{\theta}} \frac{1}{\sigma_{\theta}} d \log \left(\frac{y_{\theta}}{Y} \right) + d \log w \quad (35)$$

$$= (1 - \rho_{\theta}) d \log P + \rho_{\theta} d \log w. \quad (36)$$

This yields,

$$d \log \mu_{\theta}^{\text{flex}} = (1 - \rho_{\theta})(d \log P - d \log w). \quad (37)$$

Quantities. We have

$$d \log\left(\frac{y_\theta}{Y}\right) = -\sigma_\theta (d \log p_\theta - d \log P) \quad (38)$$

$$= -\sigma_\theta (d \log \mu_\theta + d \log w - d \log P). \quad (39)$$

This yields,

$$d \log\left(\frac{y_\theta^{\text{sticky}}}{Y}\right) = \sigma_\theta (d \log P), \quad (40)$$

$$d \log\left(\frac{y_\theta^{\text{flex}}}{Y}\right) = \sigma_\theta \rho_\theta (d \log P - d \log w). \quad (41)$$

Sales shares. We have

$$d \log \lambda_\theta = d \log\left(\frac{y_\theta}{Y}\right) + d \log \mu_\theta + d \log Y - d \log \bar{\mu} - d \log L. \quad (42)$$

This yields,

$$d \log \lambda_\theta^{\text{flex}} = (d \log Y - d \log \bar{\mu} - d \log L) + (1 + (\sigma_\theta - 1)\rho_\theta)(d \log P - d \log w), \quad (43)$$

$$d \log \lambda_\theta^{\text{sticky}} = (d \log Y - d \log \bar{\mu} - d \log L) - d \log w + \sigma_\theta d \log P. \quad (44)$$

Now, we aggregate across firms to get the change in aggregate variables, $d \log \bar{\mu}$, $d \log P$, $d \log Y$, and $d \log L$.

Aggregate markup. We start with the equation for the change in the aggregate markup,

$$d \log \bar{\mu} = \mathbb{E}_{\lambda_{\mu^{-1}}} [d \log \mu_\theta] - \mathbb{E}_{\lambda_{\mu^{-1}}} [d \log \lambda_\theta]. \quad (45)$$

We use

$$\begin{aligned} \mathbb{E}_{\lambda_{\mu^{-1}}} [d \log \lambda_\theta] &= (d \log Y - d \log \bar{\mu} - d \log L) \\ &\quad + \mathbb{E}_{\lambda_{\mu^{-1}}} [\delta_\theta [(1 + (\sigma_\theta - 1)\rho_\theta)(d \log P - d \log w)]] \\ &\quad + \mathbb{E}_{\lambda_{\mu^{-1}}} [(1 - \delta_\theta)[-d \log w + \sigma_\theta d \log P]], \end{aligned} \quad (46)$$

which simplifies to

$$\begin{aligned} \mathbb{E}_{\lambda_{\mu^{-1}}} [d \log \lambda_\theta] &= (d \log Y - d \log L - d \log \bar{\mu}) \\ &\quad + \mathbb{E}_{\lambda_{\mu^{-1}}} [\delta_\theta (1 + (\sigma_\theta - 1)\rho_\theta) + (1 - \delta_\theta)\sigma_\theta] d \log P \end{aligned}$$

$$+ \mathbb{E}_{\lambda\mu^{-1}} [(-1 - \delta_\theta(\sigma_\theta - 1)\rho_\theta)] d \log w, \quad (47)$$

and

$$\mathbb{E}_{\lambda\mu^{-1}} [d \log \mu_\theta] = \mathbb{E}_{\lambda\mu^{-1}} [\delta_\theta [(1 - \rho_\theta)(d \log P - d \log w)]] + \mathbb{E}_{\lambda\mu^{-1}} [(1 - \delta_\theta) [-d \log w]] \quad (48)$$

$$= \mathbb{E}_{\lambda\mu^{-1}} [\delta_\theta(1 - \rho_\theta)] d \log P - \mathbb{E}_{\lambda\mu^{-1}} [1 - \delta_\theta\rho_\theta] d \log w. \quad (49)$$

Combining yields,

$$d \log Y - d \log L = -\mathbb{E}_{\lambda\mu^{-1}} [[\delta_\theta\rho_\theta + (1 - \delta_\theta)]\sigma_\theta] d \log P + \mathbb{E}_{\lambda\mu^{-1}} [\delta_\theta\sigma_\theta\rho_\theta] d \log w. \quad (50)$$

Price index. We have

$$d \log P = -d \log Y + d \log L + d \log \bar{\mu} + d \log w - \mathbb{E}_\lambda \left[\left(1 - \frac{1}{\sigma_\theta}\right) d \log \left(\frac{y_\theta}{Y}\right) \right]. \quad (51)$$

Using

$$\mathbb{E}_\lambda \left[\left(1 - \frac{1}{\sigma_\theta}\right) d \log \left(\frac{y_\theta}{Y}\right) \right] = \mathbb{E}_\lambda \left[\delta_\theta \left(1 - \frac{1}{\sigma_\theta}\right) \sigma_\theta \rho_\theta (d \log P - d \log w) \right] \quad (52)$$

$$+ \mathbb{E}_\lambda \left[(1 - \delta_\theta) \left(1 - \frac{1}{\sigma_\theta}\right) \sigma_\theta (d \log P) \right] \quad (53)$$

$$= \mathbb{E}_\lambda [[1 + \delta_\theta(\rho_\theta - 1)](\sigma_\theta - 1)] d \log P - \mathbb{E}_\lambda [\delta_\theta\rho_\theta(\sigma_\theta - 1)] d \log w, \quad (54)$$

we get

$$d \log P = \frac{-(d \log Y - d \log L - d \log \bar{\mu}) + \mathbb{E}_\lambda [1 + \delta_\theta\rho_\theta(\sigma_\theta - 1)] d \log w}{1 + \mathbb{E}_\lambda [(\sigma_\theta - 1)[(1 - \delta_\theta) + \delta_\theta\rho_\theta]}. \quad (55)$$

Output. We have

$$d \log Y = d \log L + d \log \bar{\mu} - \mathbb{E}_\lambda [d \log \mu_\theta] \quad (56)$$

$$= d \log L + d \log \bar{\mu} - \mathbb{E}_\lambda [\delta_\theta(1 - \rho_\theta)] d \log P + \mathbb{E}_\lambda [1 - \delta_\theta\rho_\theta] d \log w. \quad (57)$$

Labor. We have

$$d \log \bar{\mu} + \left(1 + \frac{1}{\zeta}\right) d \log L = (1 - \gamma) d \log Y, \quad (58)$$

which we rearrange to get

$$d \log L = \frac{\zeta(1-\gamma)}{1+\zeta} d \log Y - \frac{\zeta}{1+\zeta} d \log \bar{\mu}. \quad (59)$$

Solving the fixed point. We have a system of four equations in four unknowns, $d \log Y$, $d \log L$, $d \log \bar{\mu}$, and $d \log P$:

$$d \log Y = d \log L + d \log \bar{\mu} - \mathbb{E}_\lambda [\delta_\theta(1-\rho_\theta)] d \log P + \mathbb{E}_\lambda [1-\delta_\theta\rho_\theta] d \log w \quad (60)$$

$$d \log L = \frac{\zeta(1-\gamma)}{1+\zeta} d \log Y - \frac{\zeta}{1+\zeta} d \log \bar{\mu} \quad (61)$$

$$d \log Y - d \log L = -\mathbb{E}_{\lambda\mu^{-1}} [[\delta_\theta\rho_\theta + (1-\delta_\theta)]\sigma_\theta] d \log P + \mathbb{E}_{\lambda\mu^{-1}} [\delta_\theta\sigma_\theta\rho_\theta] d \log w \quad (62)$$

$$d \log P = \frac{-(d \log Y - d \log L - d \log \bar{\mu}) + \mathbb{E}_\lambda [1 + \delta_\theta\rho_\theta(\sigma_\theta - 1)] d \log w}{1 + \mathbb{E}_\lambda [(\sigma_\theta - 1)[(1 - \delta_\theta) + \delta_\theta\rho_\theta]}. \quad (63)$$

Rearranging yields,

$$d \log P = \frac{\mathbb{E}_\lambda [\delta_\theta\rho_\theta\sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} d \log w. \quad (64)$$

Proof of Proposition 1. Using our equations for $d \log P$ and $d \log Y - d \log L$, we solve for the response of aggregate TFP

$$d \log A = d \log Y - d \log L \quad (65)$$

$$= -\mathbb{E}_{\lambda\mu^{-1}} [[\delta_\theta\rho_\theta + (1-\delta_\theta)]\sigma_\theta] d \log P + \mathbb{E}_{\lambda\mu^{-1}} [\delta_\theta\sigma_\theta\rho_\theta] d \log w \quad (66)$$

$$= \left[\mathbb{E}_{\lambda\mu^{-1}} [\delta_\theta\sigma_\theta\rho_\theta] - \frac{\mathbb{E}_{\lambda\mu^{-1}} [[\delta_\theta\rho_\theta + (1-\delta_\theta)]\sigma_\theta] \mathbb{E}_\lambda [\delta_\theta\rho_\theta\sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} \right] d \log w \quad (67)$$

$$= \bar{\mu} \left[\mathbb{E}_\lambda [\delta_\theta\rho_\theta(\sigma_\theta - 1)] - \left[\frac{\mathbb{E}_\lambda [[\delta_\theta\rho_\theta + (1-\delta_\theta)](\sigma_\theta - 1)] \mathbb{E}_\lambda [\delta_\theta\rho_\theta\sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} \right] \right] d \log w \quad (68)$$

$$= \bar{\mu} \frac{\mathbb{E}_\lambda [1-\delta_\theta] \mathbb{E}_\lambda [\delta_\theta\rho_\theta\sigma_\theta] - \mathbb{E}_\lambda [\sigma_\theta(1-\delta_\theta)] \mathbb{E}_\lambda [\delta_\theta\rho_\theta]}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} d \log w \quad (69)$$

$$= \bar{\mu} \frac{\mathbb{E}_\lambda [\delta_\theta\rho_\theta\sigma_\theta] - \mathbb{E}_\lambda [\delta_\theta] \mathbb{E}_\lambda [\delta_\theta\rho_\theta\sigma_\theta] - \mathbb{E}_\lambda [\sigma_\theta] \mathbb{E}_\lambda [\delta_\theta\rho_\theta] + \mathbb{E}_\lambda [\sigma_\theta\delta_\theta] \mathbb{E}_\lambda [\delta_\theta\rho_\theta]}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} d \log w \quad (70)$$

$$= \bar{\mu} \frac{\mathbb{E}_\lambda [\delta_\theta] (1 - \mathbb{E}_\lambda [\delta_\theta]) \mathbb{E}_{\lambda\delta} [\rho_\theta\sigma_\theta] + \mathbb{E}_{\lambda\delta} [\rho_\theta] (\mathbb{E}_\lambda [\sigma_\theta\delta_\theta] \mathbb{E}_\lambda [\delta_\theta] - \mathbb{E}_\lambda [\sigma_\theta] \mathbb{E}_\lambda [\delta_\theta])}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} d \log w \quad (71)$$

$$= \bar{\mu} \frac{\mathbb{E}_\lambda [\delta_\theta] (1 - \mathbb{E}_\lambda [\delta_\theta]) \text{Cov}_{\lambda\delta} [\rho_\theta, \sigma_\theta] + \mathbb{E}_{\lambda\delta} [\rho_\theta] \text{Cov}_\lambda [\sigma_\theta, \delta_\theta]}{\mathbb{E}_\lambda [\sigma_\theta[(1-\delta_\theta) + \delta_\theta\rho_\theta]]} d \log w, \quad (72)$$

which concludes the proof of Proposition 1.

Proof of Proposition 2. We use the change in firm markups to calculate

$$\mathbb{E}_\lambda [d \log \mu_\theta] = \mathbb{E}_\lambda [\delta_\theta(1 - \rho_\theta)] d \log P - \mathbb{E}_\lambda [1 - \delta_\theta \rho_\theta] d \log w \quad (73)$$

$$= \left[\frac{\mathbb{E}_\lambda [\delta_\theta(1 - \rho_\theta)] \mathbb{E}_\lambda [\delta_\theta \rho_\theta \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta [(1 - \delta_\theta) + \delta_\theta \rho_\theta]]} - \mathbb{E}_\lambda [1 - \delta_\theta \rho_\theta] \right] d \log w \quad (74)$$

$$= \left[\mathbb{E}_\lambda [\delta_\theta(1 - \rho_\theta)] \left(\frac{\mathbb{E}_\lambda [\delta_\theta \rho_\theta \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta [(1 - \delta_\theta) + \delta_\theta \rho_\theta]]} - 1 \right) - \mathbb{E}_\lambda [1 - \delta_\theta] \right] d \log w \quad (75)$$

$$= \left[- \frac{\mathbb{E}_\lambda [\delta_\theta(1 - \rho_\theta)] \mathbb{E}_\lambda [\sigma_\theta (1 - \delta_\theta)]}{\mathbb{E}_\lambda [\sigma_\theta [(1 - \delta_\theta) + \delta_\theta \rho_\theta]]} - \mathbb{E}_\lambda [1 - \delta_\theta] \right] d \log w, \quad (76)$$

which yields Equation (9). We have

$$d \log L = \frac{\zeta(1 - \gamma)}{1 + \zeta} d \log Y - \frac{\zeta}{1 + \zeta} d \log \bar{\mu}, \quad (77)$$

and

$$d \log A = d \log \bar{\mu} - \mathbb{E}_\lambda [d \log \mu_\theta], \quad (78)$$

which yields,

$$\frac{1 + \gamma \zeta}{1 + \zeta} d \log Y = \frac{1}{1 + \zeta} d \log A - \frac{\zeta}{1 + \zeta} \mathbb{E}_\lambda [d \log \mu_\theta]. \quad (79)$$

Rearranging yields Equation (8), which concludes the proof of Proposition 2.

General form for Phillips curve flattening. Proposition 3 follows immediately from dividing Equation (8) by $d \log w$ and rearranging.

Proposition 6 generalizes Proposition 4 to the case where both price-stickiness and pass-throughs are allowed to be heterogeneous.

Proposition 6. *The flattening of the price Phillips curve due to real rigidities, compared to nominal rigidities alone, is*

$$\frac{\text{Phillips curve slope w/ nominal rigidities only}}{\text{Phillips curve slope w/ real rigidities}} = 1 + \frac{1}{\mathbb{E}_\lambda [1 - \delta_\theta]} \frac{\mathbb{E}_\lambda [\delta_\theta(1 - \rho_\theta)] \mathbb{E}_\lambda [\sigma_\theta(1 - \delta_\theta)]}{\mathbb{E}_\lambda [\delta_\theta] \mathbb{E}_\lambda [\delta_\theta \rho_\theta \sigma_\theta] + \mathbb{E}_\lambda [\delta_\theta \rho_\theta] \mathbb{E}_\lambda [\sigma_\theta(1 - \delta_\theta)]}. \quad (80)$$

The flattening of the price Phillips curve due to the misallocation channel is

Phillips curve slope w/ real rigidities
Phillips curve slope w/ misallocation

$$= 1 + \frac{\bar{\mu} \mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[1 - \delta_\theta] \text{Cov}_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \mathbb{E}_{\lambda\delta}[\rho_\theta] \text{Cov}_\lambda[\sigma_\theta, \delta_\theta]}{\zeta \mathbb{E}_\lambda[1 - \delta_\theta] \mathbb{E}_\lambda[\delta_\theta \rho_\theta \sigma_\theta] + \mathbb{E}_\lambda[1 - \delta_\theta \rho_\theta] \mathbb{E}_\lambda[\sigma_\theta(1 - \delta_\theta)]}. \quad (81)$$

Proof. The flattening due to the misallocation channel is,

Flattening due to the misallocation channel (82)

$$= \frac{\frac{d \log A}{d \log w} - \zeta \mathbb{E}_\lambda \left[\frac{d \log \mu_\theta}{d \log w} \right]}{-\zeta \mathbb{E}_\lambda \left[\frac{d \log \mu_\theta}{d \log w} \right]} \quad (83)$$

$$= 1 + \frac{\frac{d \log A}{d \log w}}{-\zeta \mathbb{E}_\lambda \left[\frac{d \log \mu_\theta}{d \log w} \right]} \quad (84)$$

$$= 1 + \frac{1}{\zeta} \frac{\kappa_\rho \text{Cov}_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \kappa_\delta \text{Cov}_\lambda[\sigma_\theta, \delta_\theta]}{\frac{\mathbb{E}_\lambda[\delta_\theta(1-\rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda[[\delta_\theta \rho_\theta + (1-\delta_\theta)] \sigma_\theta]} + \mathbb{E}_\lambda[1 - \delta_\theta]} \quad (85)$$

$$= 1 + \frac{\bar{\mu} \mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[1 - \delta_\theta] \text{Cov}_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \mathbb{E}_{\lambda\delta}[\rho_\theta] \text{Cov}_\lambda[\sigma_\theta, \delta_\theta]}{\zeta \mathbb{E}_\lambda[\delta_\theta(1 - \rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1 - \delta_\theta)] + \mathbb{E}_\lambda[1 - \delta_\theta] \mathbb{E}_\lambda[[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \sigma_\theta]} \quad (86)$$

$$= 1 + \frac{\bar{\mu} \mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[1 - \delta_\theta] \text{Cov}_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \mathbb{E}_{\lambda\delta}[\rho_\theta] \text{Cov}_\lambda[\sigma_\theta, \delta_\theta]}{\zeta \mathbb{E}_\lambda[1 - \delta_\theta \rho_\theta] \mathbb{E}_\lambda[\sigma_\theta(1 - \delta_\theta)] + \mathbb{E}_\lambda[1 - \delta_\theta] \mathbb{E}_\lambda[\delta_\theta \rho_\theta \sigma_\theta]}. \quad (87)$$

The flattening due to real rigidities is,

Flattening due to real rigidities (88)

$$= \left(\frac{1 - \mathbb{E}_\lambda[1 - \delta_\theta] - \frac{\mathbb{E}_\lambda[\delta_\theta(1-\rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda[[\delta_\theta \rho_\theta + (1-\delta_\theta)] \sigma_\theta]}}{-\zeta \left[-\mathbb{E}_\lambda[1 - \delta_\theta] - \frac{\mathbb{E}_\lambda[\delta_\theta(1-\rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda[[\delta_\theta \rho_\theta + (1-\delta_\theta)] \sigma_\theta]} \right]} \right)^{-1} \underbrace{\frac{1 - \mathbb{E}_\lambda[1 - \delta_\theta]}{\zeta \mathbb{E}_\lambda[1 - \delta_\theta]}}_{\text{Slope in CES model}} \quad (89)$$

$$= \frac{\mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[1 - \delta_\theta] + \frac{\mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[\delta_\theta(1-\rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda[[\delta_\theta \rho_\theta + (1-\delta_\theta)] \sigma_\theta]}}{\mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[1 - \delta_\theta] - \frac{\mathbb{E}_\lambda[1-\delta_\theta] \mathbb{E}_\lambda[\delta_\theta(1-\rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda[[\delta_\theta \rho_\theta + (1-\delta_\theta)] \sigma_\theta]}} \quad (90)$$

$$= 1 + \frac{1}{\mathbb{E}_\lambda[1 - \delta_\theta]} \frac{\mathbb{E}_\lambda[\delta_\theta(1 - \rho_\theta)] \mathbb{E}_\lambda[\sigma_\theta(1 - \delta_\theta)]}{\mathbb{E}_\lambda[\delta_\theta] \mathbb{E}_\lambda[\delta_\theta \rho_\theta \sigma_\theta] + \mathbb{E}_\lambda[\delta_\theta \rho_\theta] \mathbb{E}_\lambda[\sigma_\theta(1 - \delta_\theta)]}. \quad (91)$$

Setting $\delta_\theta = \delta$ in both equations yields Proposition 4.

Appendix B Proofs for Dynamic Model

Firms choose reset prices to maximize future discounted profits,

$$\max_{p_{i,t}} \mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} (1-\delta_i)^k y_{i,t+k} \left(p_{i,t} - \frac{w_{t+k}}{A_i} \right) \right]. \quad (92)$$

The first order condition yields,

$$\mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} (1-\delta_i)^k y_{i,t+k} \left[\frac{dy_{i,t+k}}{dp_{i,t}} \frac{p_{i,t}^*}{y_{i,t+k}} \frac{p_{i,t}^* - \frac{w_{t+k}}{A_i}}{p_{i,t}^*} + 1 \right] \right] = 0. \quad (93)$$

Using $\sigma_{i,t} = -\frac{p_i}{y_{i,t}} \frac{dy_{i,t}}{dp_i}$,

$$\mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} (1-\delta_i)^k y_{i,t+k} \left(-\sigma_{i,t+k} \left(1 - \frac{w_{t+k}}{p_{i,t}^* A_i} \right) + 1 \right) \right] = 0. \quad (94)$$

Rearranging gives us

$$\frac{p_{i,t}^* A_i}{w_t} = \frac{\mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} (1-\delta_i)^k y_{i,t+k} \left(-\sigma_{i,t+k} \frac{w_{t+k}}{w_t} \right) \right]}{\mathbb{E} \left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1} (1+r_{t+j})} (1-\delta_i)^k y_{i,t+k} (1 - \sigma_{i,t+k}) \right]}. \quad (95)$$

We now log-linearize around a perfect foresight, no-inflation steady state. This steady state is characterized by a constant discount factor such that $\left[\prod_{j=0}^{k-1} (1+r_{t+j}) \right]^{-1} = \beta^k$. After removing all second-order terms, we get:

$$\frac{p_{i,t}^* A_i}{w_t} = \frac{\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t+k} \sigma_{i,t+k} \left(d \log \left(\frac{w_{t+k}}{w_t} \right) + 1 \right) \right]}{\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t+k} (\sigma_{i,t+k} - 1) \right]} \quad (96)$$

$$= \frac{\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t} \sigma_{i,t} (1 + d \log(y_{i,t+k} \sigma_{i,t+k})) \left(d \log \left(\frac{w_{t+k}}{w_t} \right) + 1 \right) \right]}{\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t} (\sigma_{i,t} - 1) (1 + d \log(y_{i,t+k} (\sigma_{i,t+k} - 1))) \right]} \quad (97)$$

$$= \mu_{i,t} \frac{\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k \right] + \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log \left(\frac{w_{t+k}}{w_t} \right) \right] + \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(y_{i,t+k} \sigma_{i,t+k}) \right]}{\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k \right] + \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(y_{i,t+k} (\sigma_{i,t+k} - 1)) \right]}. \quad (98)$$

Rearranging, we get

$$d \log \mu_{i,t}^* = \frac{\frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log\left(\frac{w_{t+k}}{w_t}\right)\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k\right]} + \frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(y_{i,t+k} \sigma_{i,t+k})\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k\right]} - \frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(y_{i,t+k} (\sigma_{i,t+k} - 1))\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k\right]}}{1 + \frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(y_{i,t+k} (\sigma_{i,t+k} - 1))\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k\right]}} \quad (99)$$

$$= \frac{\Gamma^{-1} \left[\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log\left(\frac{w_{t+k}}{w_t}\right) \right] + \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(\mu_{i,t+k}) \right] \right]}{1 + \Gamma^{-1} \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(y_{i,t+k} (\sigma_{i,t+k} - 1)) \right]}, \quad (100)$$

where $\Gamma = \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k \right] = [1 - \beta(1-\delta_i)]^{-1}$. We multiply both sides by the denominator and remove second order terms to get

$$d \log \mu_{i,t}^* = \Gamma^{-1} \left[\mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log\left(\frac{w_{t+k}}{w_t}\right) \right] + \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log(\mu_{i,t+k}) \right] \right]. \quad (101)$$

At the time of a price reset, we know that

$$\mu_{i,t} = \mu_i \left(\frac{y_{i,t}}{Y_t} \right). \quad (102)$$

Then,

$$d \log \left(\mu \left(\frac{y_{i,t+k}}{Y_{t+k}} \right) \right) = \frac{\frac{y_{i,t}}{Y_t} \mu_i' \left(\frac{y_{i,t}}{Y_t} \right)}{\mu_{i,t}} d \log \left(\frac{y_{i,t+k}}{Y_{t+k}} \right) \quad (103)$$

$$= \frac{1 - \rho_{i,t}}{\rho_{i,t}} \frac{1}{\sigma_{i,t}} d \log \left(\frac{y_{i,t+k}}{Y_{t+k}} \right) \quad (104)$$

$$= \frac{1 - \rho_{i,t}}{\rho_{i,t}} (d \log P_{t+k} - d \log w_{t+k} - d \log \mu_{i,t+k}) \quad (105)$$

$$= \frac{1 - \rho_{i,t}}{\rho_{i,t}} \left(d \log P_{t+k} - d \log w_{t+k} - d \log \mu_{i,t}^* + d \log \frac{w_{t+k}}{w_t} \right), \quad (106)$$

where in the last line, we use the fact that the change in the markup $d \log \mu_{i,t+k}$ includes changes that occur at the time of the price change ($d \log \mu_{i,t}^*$) and subsequent changes due to the shifts in the nominal wage.

Plugging this in yields,

$$\frac{1}{\rho_{i,t}} d \log \mu_{i,t}^* = \Gamma^{-1} \left[\frac{1}{\rho_{i,t}} \mathbb{E} \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d \log \left(\frac{w_{t+k}}{w_t} \right) \right] + E \left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k \frac{1 - \rho_{i,t}}{\rho_{i,t}} (d \log P_{t+k} - d \log w_{t+k}) \right] \right]. \quad (107)$$

Finally, since $d \log \mu_{i,t}^* = d \log p_{i,t}^* - d \log w_t$, we get

$$d \log p_{i,t}^* = [1 - \beta(1 - \delta_i)] \sum_{k=0}^{\infty} \beta^k (1 - \delta_i)^k [\rho_i d \log w_{t+k} + (1 - \rho_i) d \log P_{t+k}]. \quad (108)$$

We can also write this equation recursively as

$$d \log p_{i,t}^* = (1 - \beta(1 - \delta_i)) [\rho_i d \log w_t + (1 - \rho_i) d \log P_t] + \beta(1 - \delta_i) p_{i,t+1}^*, \quad (109)$$

or in terms of firm types as,

$$d \log p_{\theta,t}^* = (1 - \beta(1 - \delta_{\theta})) [\rho_{\theta} d \log w_t + (1 - \rho_{\theta}) d \log P_t] + \beta(1 - \delta_{\theta}) p_{\theta,t+1}^*. \quad (110)$$

Now that we have a recursive formulation for the optimal reset price, we can solve for the movement in the expected price for firms of type θ . Here, we use \mathbb{E} to indicate the expectation over a continuum of identical firms of type θ , some of which will have the opportunity to change their prices and the remainder of which will not. The expected price for a firm of type θ follows,

$$\mathbb{E}[d \log p_{\theta,t+1}] = \delta_{\theta} d \log p_{\theta,t+1}^* + (1 - \delta_{\theta}) d \log p_{\theta,t}, \quad (111)$$

since with probability δ_{θ} the firm is able to change its price to the optimal reset price at time $t + 1$. Combining this with the recursive formula for optimal reset prices above, we get

$$\begin{aligned} & \mathbb{E}[d \log p_{\theta,t} - d \log p_{\theta,t-1}] - \beta \mathbb{E}[d \log p_{\theta,t+1} - d \log p_{\theta,t}] \\ &= \frac{\delta_{\theta}}{1 - \delta_{\theta}} (1 - \beta(1 - \delta_{\theta})) [-\mathbb{E}[d \log p_{\theta}] + \rho_{\theta} d \log w_t + (1 - \rho_{\theta}) d \log P_t]. \end{aligned} \quad (112)$$

We can then aggregate this equation over firm types to get the modified New Keynesian Phillips curve and to get the Endogenous TFP equation.

New Keynesian Phillips curve with misallocation. We list a few identities that will be helpful in the subsequent derivations. The first four are derived in the main text, and the latter two can be formed by rearranging the above.

$$d \log P_t - d \log P_t^Y = \bar{\mu}^{-1} d \log A_t \quad (113)$$

$$d \log P_t^Y - d \log w_t = \mathbb{E}_{\lambda} [d \log \mu_{\theta}] \quad (114)$$

$$d \log A_t = d \log \bar{\mu}_t - \mathbb{E}_\lambda [d \log \mu_{\theta,t}] \quad (115)$$

$$d \log Y_t = \frac{1}{1 + \gamma \zeta} (d \log A_t - \zeta \mathbb{E}_\lambda [d \log \mu_{\theta,t}]) \quad (116)$$

$$-\mathbb{E}_\lambda [d \log \mu_{\theta,t}] = \left(\frac{1 + \gamma \zeta}{\zeta} \right) d \log Y_t - \frac{1}{\zeta} d \log A_t \quad (117)$$

$$d \log w_t - d \log P_t = \frac{1 + \gamma \zeta}{\zeta} d \log Y_t - \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) d \log A_t. \quad (118)$$

We now take the sales-weighted expectation of Equation (112) to get:

$$d \log \pi_t - \beta d \log \pi_{t+1} = \varphi \left[-d \log P_t^Y + \mathbb{E}_\lambda [\rho_\theta] d \log w_t + (1 - \mathbb{E}_\lambda [\rho_\theta]) d \log P_t \right] \quad (119)$$

$$= \varphi \left[(d \log P_t - d \log P_t^Y) + \mathbb{E}_\lambda [\rho_\theta] (d \log w_t - d \log P_t) \right] \quad (120)$$

$$= \varphi \left[\left(\bar{\mu}^{-1} d \log A_t \right) + \mathbb{E}_\lambda [\rho_\theta] \left(\frac{1 + \gamma \zeta}{\zeta} d \log Y_t - \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) d \log A_t \right) \right] \quad (121)$$

$$= \varphi \mathbb{E}_\lambda [\rho_\theta] \frac{1 + \gamma \zeta}{\zeta} d \log Y_t + \varphi \left[\frac{1}{\bar{\mu}} - \mathbb{E}_\lambda [\rho_\theta] \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) \right] d \log A_t, \quad (122)$$

which concludes the proof.

Endogenous TFP equation. Start by subtracting $\mathbb{E} [d \log w_t - d \log w_{t-1}] - \beta \mathbb{E} [d \log w_{t+1} - d \log w_t]$ from both sides of Equation (112). This yields,

$$\begin{aligned} & \mathbb{E} [d \log \mu_{\theta,t} - d \log \mu_{\theta,t-1}] - \beta \mathbb{E} [d \log \mu_{\theta,t+1} - d \log \mu_{\theta,t}] \\ &= - [\mathbb{E} [d \log w_t - d \log w_{t-1}] - \beta \mathbb{E} [d \log w_{t+1} - d \log w_t]] \\ & \quad + \varphi [-\mathbb{E} [d \log \mu_{\theta,t}] + (\rho_\theta - 1) d \log w_t + (1 - \rho_\theta) d \log P_t]. \end{aligned} \quad (123)$$

We can write

$$d \log A_t = d \log \bar{\mu} - \mathbb{E}_\lambda [d \log \mu_\theta] = \bar{\mu} \left(\frac{\mathbb{E}_\lambda [\sigma_\theta d \log \mu_{\theta,t}]}{\mathbb{E}_\lambda [\sigma_\theta]} - \mathbb{E}_\lambda [d \log \mu_\theta] \right). \quad (124)$$

Now, we take Equation (124) and (1) multiply all terms by σ_θ , take the sales-weighted expectation, and divide by $\mathbb{E}_\lambda [\sigma_\theta]$; (2) take the sales-weighted expectation of (124); and multiply (1) – (2) by $\bar{\mu}$. This yields,

$$(d \log A_t - d \log A_{t-1}) - \beta (d \log A_{t+1} - d \log A_t) \quad (125)$$

$$= \varphi \left[-d \log A_t + \bar{\mu} \left(1 - \frac{\mathbb{E}_\lambda [\sigma_\theta \rho_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} - (1 - \mathbb{E}_\lambda [\rho_\theta]) \right) (d \log P_t - d \log w_t) \right] \quad (126)$$

$$= \varphi \left[-d \log A_t + \bar{\mu} \left(\frac{\text{Cov}_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \right) (d \log w_t - d \log P_t) \right] \quad (127)$$

$$= \varphi \left[-d \log A_t + \bar{\mu} \left(\frac{\text{Cov}_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \right) \left(\frac{1 + \gamma \zeta}{\zeta} d \log Y_t - \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) d \log A_t \right) \right] \quad (128)$$

$$= \varphi \left[- \left(1 + \bar{\mu} \left(\frac{\text{Cov}_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \right) \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) \right) d \log A_t + \bar{\mu} \frac{\text{Cov}_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \frac{1 + \gamma \zeta}{\zeta} d \log Y_t \right], \quad (129)$$

which concludes the proof.

Appendix C Empirical Calibration: Additional Details

In this appendix, we provide additional results from our calibration exercise. C.1 provides additional comparative statics from the calibration of the static model as we change the average markup and the degree of price-stickiness. C.2 shows additional impulse responses for the dynamic calibration of a 25bp interest rate shock.

Our procedure for extracting pass-throughs over the firm distribution from estimates provided by Amiti et al. (2019) is described in Appendix A of Baqaee and Farhi (2020a). We refer interested readers to that appendix.

C.1 Static model: Additional results

We vary the average markup $\bar{\mu}$ from just over one to 1.60 in Figure C.1. We do so by re-calculating markups of all firms according to the differential equation in Equation (22) according to the boundary condition implied by $\bar{\mu}$. As expected, the average markup does not affect the CES or real rigidities models, but the strength of the misallocation channel increases in $\bar{\mu}$. This reflects the dependence of the productivity response on $\bar{\mu}$.

In Figure C.2, we vary the degree of price-stickiness between zero (complete rigidity) and one (complete flexibility). We find that the flattening of the price Phillips curve due to real rigidities increases as the price becomes more rigid, and the flattening of the price Phillips curve due to the misallocation channel decreases as the price becomes more rigid. These comparative statics match the intuitions provided in the main text (see the discussion of Proposition 4).

Figure C.1: Decomposition of Phillips curve slope, varying the average markup $\bar{\mu}$.

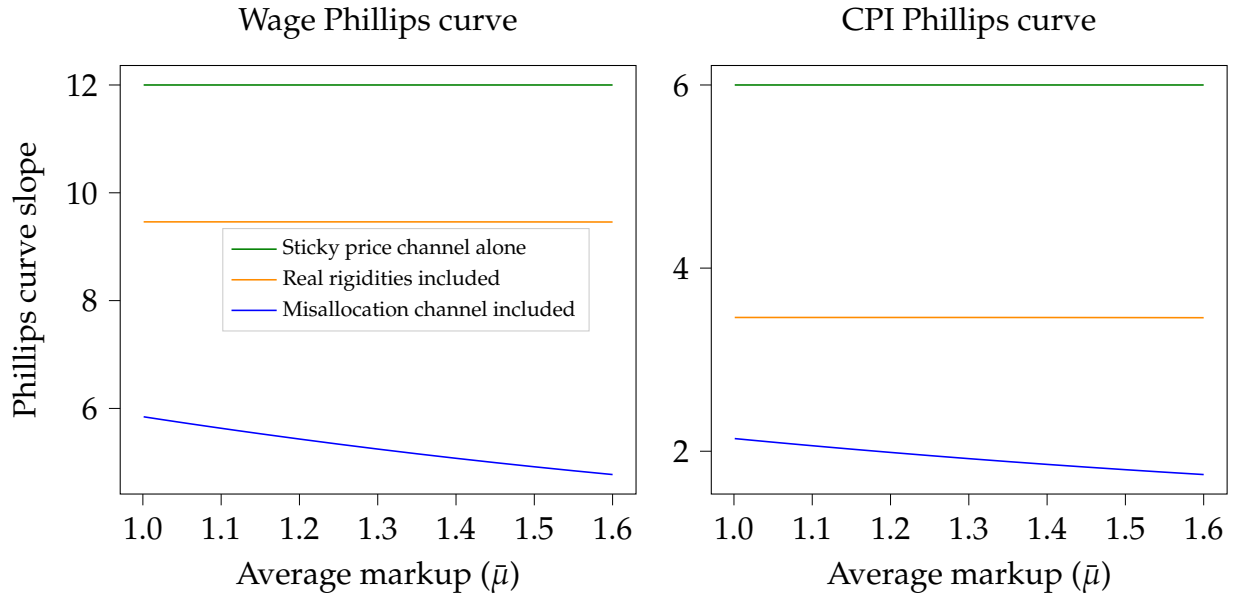
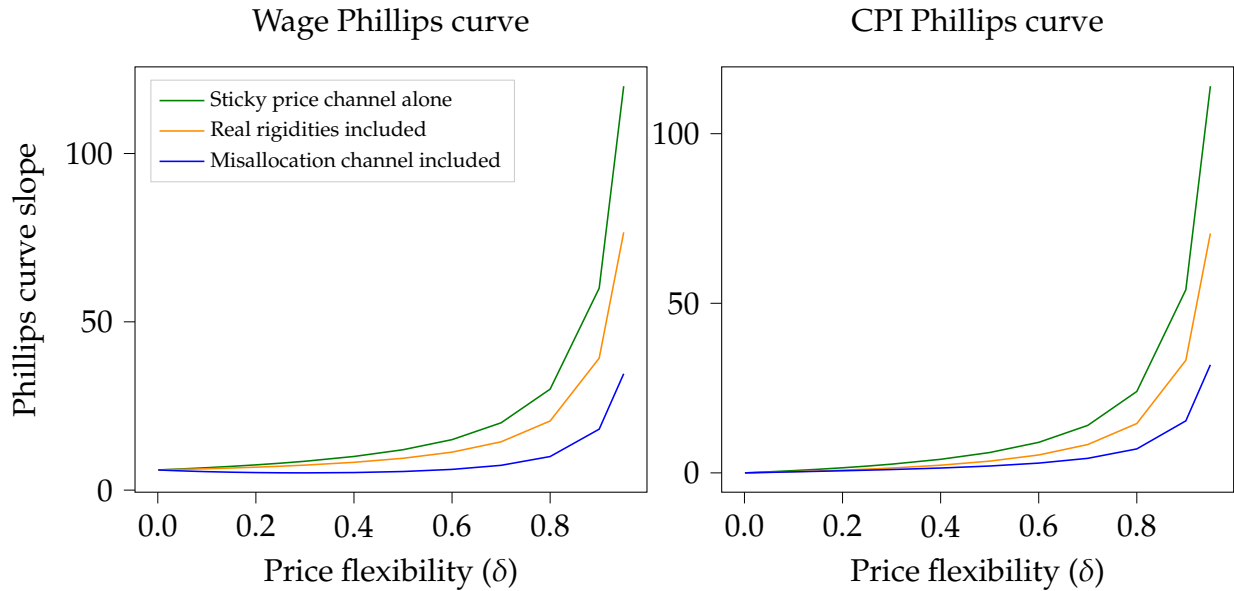


Figure C.2: Decomposition of Phillips curve slope, varying the degree of price-stickiness δ .



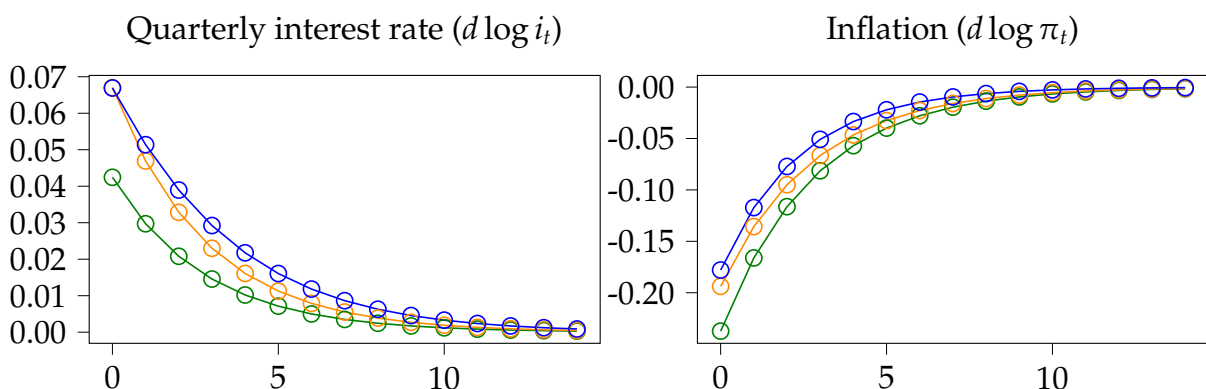
C.2 Dynamic model: Additional results

Figure C.3 shows the impulse response of the nominal interest rate and inflation following the 25bp contractionary monetary policy shock calibrated in the main text. The nominal interest rate differs across models since the monetary authority responds to the contemporaneous output and inflation gap. Compared to the CES and homogeneous firm models,

the full model predicts less deflation following the shock.

Figure C.4 shows the change in sales shares of different firm types following the 25bp contractionary monetary policy shock calibrated in the main text. The contractionary shock leads to an expansion in the sales of smaller firms and a contraction in the sales of larger firms.

Figure C.3: Impulse response functions (IRFs) following a 25bp monetary shock. Green, orange, and blue IRFs indicate the CES, homogeneous firms, and heterogeneous firms models respectively.



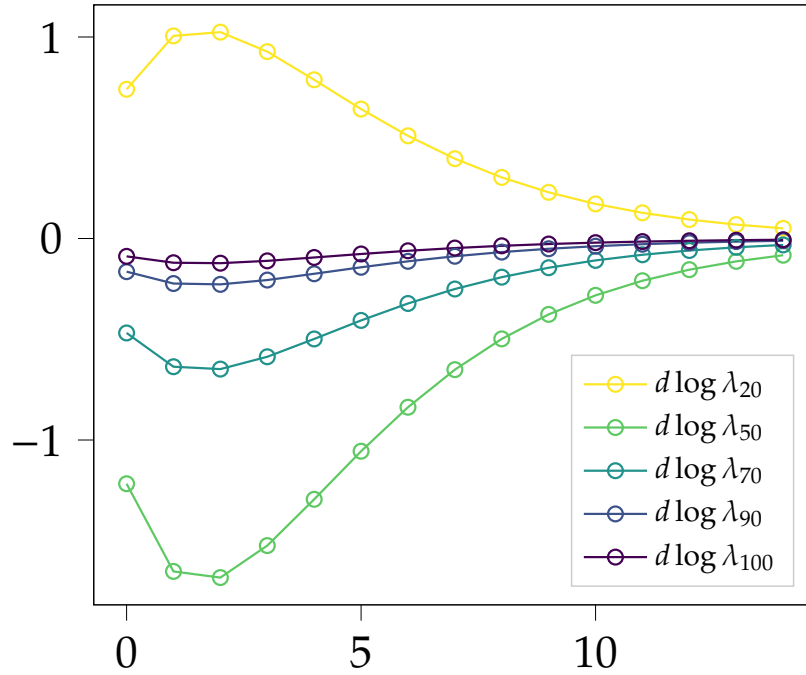
Appendix D Details of Empirical Evidence

This appendix describes the data and procedures used in Section 7. First, section D.1 describes how we construct firm-level markup data. Then, section D.2 provides additional detail for the estimation of procyclical reallocations to high-markup firms.

D.1 Estimates of Markups

We construct firm-level estimates of markups using data from Compustat, which includes all public firms in the U.S. We exclude Farm and Agriculture (SIC codes 0100-0999), Construction (SIC codes 1500-1799), Financials (SIC codes 6000-6999), Real Estate (SIC codes 5300-5399), Utilities (SIC codes 4900-4999), and other (SIC codes 9000-9999). We also exclude firm-year observations with assets less than 1 million, negative revenues, negative book or market value, or missing year, assets, or book liabilities. Our analysis is over the period from 1965-2015. Firm-level markups are estimated using three approaches: (1) accounting profits (AP), (2) user cost (UC), and (3) production function estimation (PF).

Figure C.4: Change in sales shares following a 25bp contractionary monetary policy shock by firm type. In the legend, $d \log \lambda_j$ refers to the change in the sales share of a firm at the j 'th percentile of cumulative sales.



We broadly use the same approaches described in Baqaee and Farhi (2020b); the following text provides a brief overview.

D.1.1 Accounting Profits Approach

The accounting profits approach estimates accounting profits as operating income before depreciation minus depreciation. Operating income before depreciation comes directly from Compustat. For depreciation, we use the industry-level depreciation rate from the BEA's investment series. BEA depreciation rates are better than the Compustat depreciation measures, since the latter are influenced by accounting rules and tax incentives. Markups are estimated as:

$$\text{Accounting Profits}_i = \left(1 - \frac{1}{\mu_i}\right) \text{Sales}_i.$$

D.1.2 User Cost Approach

The user-cost approach accounts for the user cost of capital more carefully. We rely on replication files from Gutiérrez and Philippon (2017) provided by German Gutierrez. We

assume that the operating surplus of each firm consists of payments of capital and rents:

$$OS_{i,t} - r_{i,t}K_{i,t} = \left(1 - \frac{1}{\mu_i}\right) \text{Sales}_i,$$

where $OS_{i,t}$ is operating income after depreciation and minus income taxes, $r_{i,t}$ is the user-cost of capital to firm i , and $K_{i,t}$ is the quantity of capital used by firm i . Following Gutiérrez and Philippon (2017), the user cost of capital is given by

$$r_{i,t} = r_t^f + RP_{j,t} - (1 - \delta_{j,t})E[\Pi_{j,t+1}],$$

where r_t^f is the risk-free rate, $RP_{j,t}$ is the industry-level capital risk premium, $\delta_{j,t}$ is the industry-level BEA depreciation rate, and $E[\Pi_{j,t+1}]$ is the expected growth in the relative price of capital. For the risk-free rate, we use the yield on the 10-year TIPS starting in 2003 and the 10-year yield on nominal Treasuries minus the average nominal-TIPS spread before 2003. Following Gutiérrez (2017), we calculate the industry-level risk premium from equity risk premia as in Claus and Thomas (2001). Due to limitations on this data, markup estimates from the user cost approach are limited to 1985-2015. We assume expected capital gains are equal to realized capital gains, measured as the growth in the relative price of capital compared to the PCE deflator. Finally, for a measure of the capital stock, we use either net property, plant, and equipment (UC1) or net property, plant, and equipment plus intangibles (UC2).

D.1.3 Production Function Estimation Approach

To estimate markups using the production function estimation approach, we follow the PF1 procedure described by De Loecker et al. (2020) with the minor differences described in Baqaee and Farhi (2020b). In particular, we estimate the the elasticity of output with respect to a variable input using Olley and Pakes (1996) rather than Levinsohn and Petrin (2003), which we implement using the *prodest* Stata package. We refer interested readers to Appendix C.2 of Baqaee and Farhi (2020b) for more detail. Because we use KLEMS sector-level price indices to deflate COGS for each firm, our markup estimates using this approach are limited to 1979-2015.

Once we have output-elasticities, the markup is

$$\mu_i = \frac{\partial \log F_i}{\partial \log COGS_i} \frac{1}{\Omega_{i,COGS}},$$

where $\frac{\partial \log F_i}{\partial \log COGS_i}$ is the elasticity of output to the variable input (in our case, COGS) and $\Omega_{i,COGS}$ is the firm's expenditures on COGS relative to its turnover.

D.2 Estimating Reallocation to High-Markup Firms

As noted in the main text, we estimate cyclical reallocations to high-markup firms using the specification:

$$SalesGrowth_{i,t \rightarrow t+1} = \beta_1 \cdot (HighMarkup_{i,t} \times Cycle_t) + \delta_1 Cycle_t + \mu_1 HighMarkup_{i,t} + \lambda_t + \epsilon_{i,t},$$

$$CostGrowth_{i,t \rightarrow t+1} = \beta_2 \cdot (HighMarkup_{i,t} \times Cycle_t) + \delta_2 Cycle_t + \mu_2 HighMarkup_{i,t} + \lambda_t + \epsilon_{i,t}.$$

Cost growth is the log change in total costs. We measure total costs by dividing sales by markups. In the regressions above, we winsorize both sales- and cost-growth at the 5-95th percentile and markup estimates at the 10-90th percentile.

Since markups are observed once per year with firms' 10K filings, we use Q4 as our cutoff for constructing annual business cycle measures. Our $Cycle_t$ variables are as follows: $Unemployment_t$ (lagged unemployment) is the average monthly unemployment rate from Q4 of year t to the end of Q3 of year $t + 1$. $\% \Delta GDP_t$ (GDP growth) is the year-over-year growth in the four-quarter sum of real GDP to Q4 of year t . $Recession_t$ is an indicator equal to one if any quarter in the previous year was marked an NBER recession.

Our reported results use the full time series available for each set of markup estimates. Limiting all markup series to the same timeframe (1985-2015) yields similar results. Table D.1 replicates the analysis, but instead partitions firms as above or below median *within* industry (at the 2-digit NAICS level). The results are similar to those presented in the main text, suggesting reallocations within industry play an important role.

Appendix E Dynamic Calibration: Money Supply Shock

Suppose the monetary shock takes the form of an exogenous shock to the money supply, rather than the interest rate rule. We calibrate the impulse response functions for the dynamic model, as in Section 6.4, for such a shock.

Money supply is linked to real variables via a cash-in-advance constraint, so that

$$d \log M = d \log P^Y + d \log Y. \quad (130)$$

Table D.1: Reallocations to high-markup firms within industry (2-digit NAICS).

Cycle measure	Markup series	Sales ($\% \Delta \lambda_{t \rightarrow t+1}$)		Costs ($\% \Delta (\lambda/\mu)_{t \rightarrow t+1}$)	
		Estimate	Standard error	Estimate	Standard error
<i>Unemployment</i>	UC1	-1.399**	(0.362)	-1.226**	(0.341)
<i>Unemployment</i>	UC2	-1.104**	(0.404)	-0.902**	(0.356)
<i>Unemployment</i>	PF	0.0881	(0.572)	0.056	(0.606)
<i>Unemployment</i>	AP	-0.594**	(0.212)	-0.469**	(0.164)
<i>Recession</i>	UC1	-4.103**	(1.683)	-3.656**	(1.665)
<i>Recession</i>	UC2	-1.134	(1.220)	-0.778	(1.152)
<i>Recession</i>	PF	-3.190**	(1.554)	-3.177*	(1.779)
<i>Recession</i>	AP	-1.100	(0.968)	-0.868	(0.684)
$\% \Delta GDP$	UC1	0.991**	(0.233)	0.817**	(0.210)
$\% \Delta GDP$	UC2	0.612*	(0.330)	0.451	(0.334)
$\% \Delta GDP$	PF	0.342	(0.380)	0.427	(0.398)
$\% \Delta GDP$	AP	0.514**	(0.111)	0.338**	(0.113)

As in Galí (2015), we assume that the money supply follows an exogenous AR(1) process,

$$\Delta d \log M_t = \rho_m \Delta d \log M_{t-1} + \epsilon_t^m. \quad (131)$$

where $\Delta d \log M_t = d \log M_t - d \log M_{t-1}$ and ϵ_t^m is white noise. We choose $\rho_m = 0.5$ and calibrate impulse response functions for an expansionary money supply shock where $\epsilon_t^m = 0.25$ for $t = 0$ and zero in all subsequent periods.

Figure E.1 shows the response of output to the money supply shock, and Figure E.2 shows the response of other variables. Like an interest rate shock, the money supply shock generates procyclical aggregate TFP and countercyclical dispersion in firm-level TFP. Real rigidities and the misallocation channel both increase the responsiveness of output to the monetary shock.

The effects on output are summarized in Table E.1. The misallocation channel increases the half-life of the shock by 28% and increases the total output impact by 45% compared to the model with real rigidities alone.

Figure E.1: Impulse response function of output following an expansionary money supply shock.

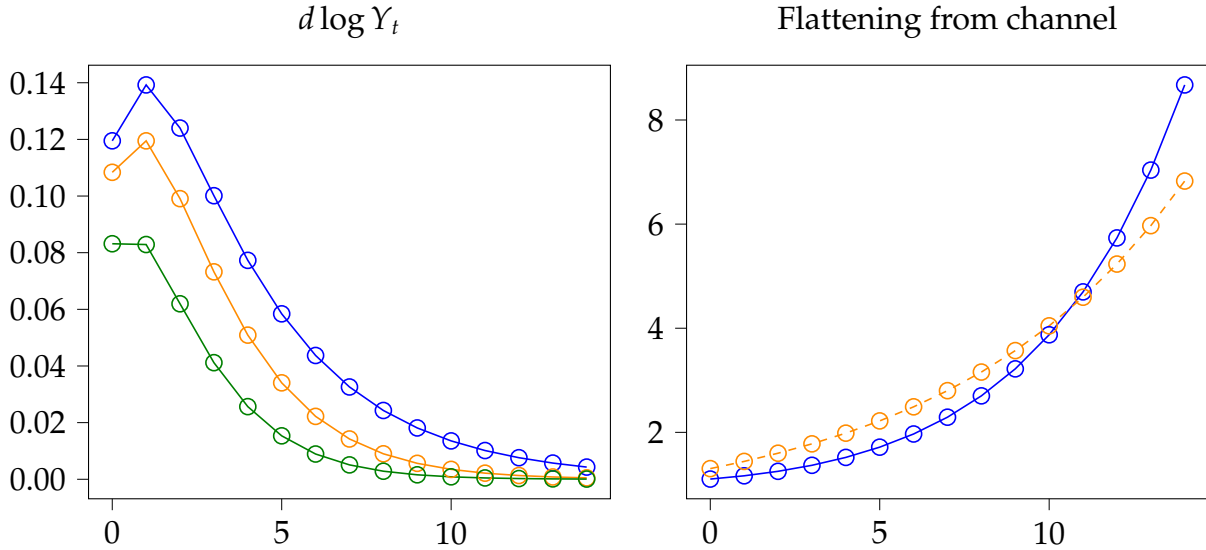


Table E.1: Effect of exogenous money supply shock on output. The cumulative output impact is calculated as in Alvarez et al. (2016).

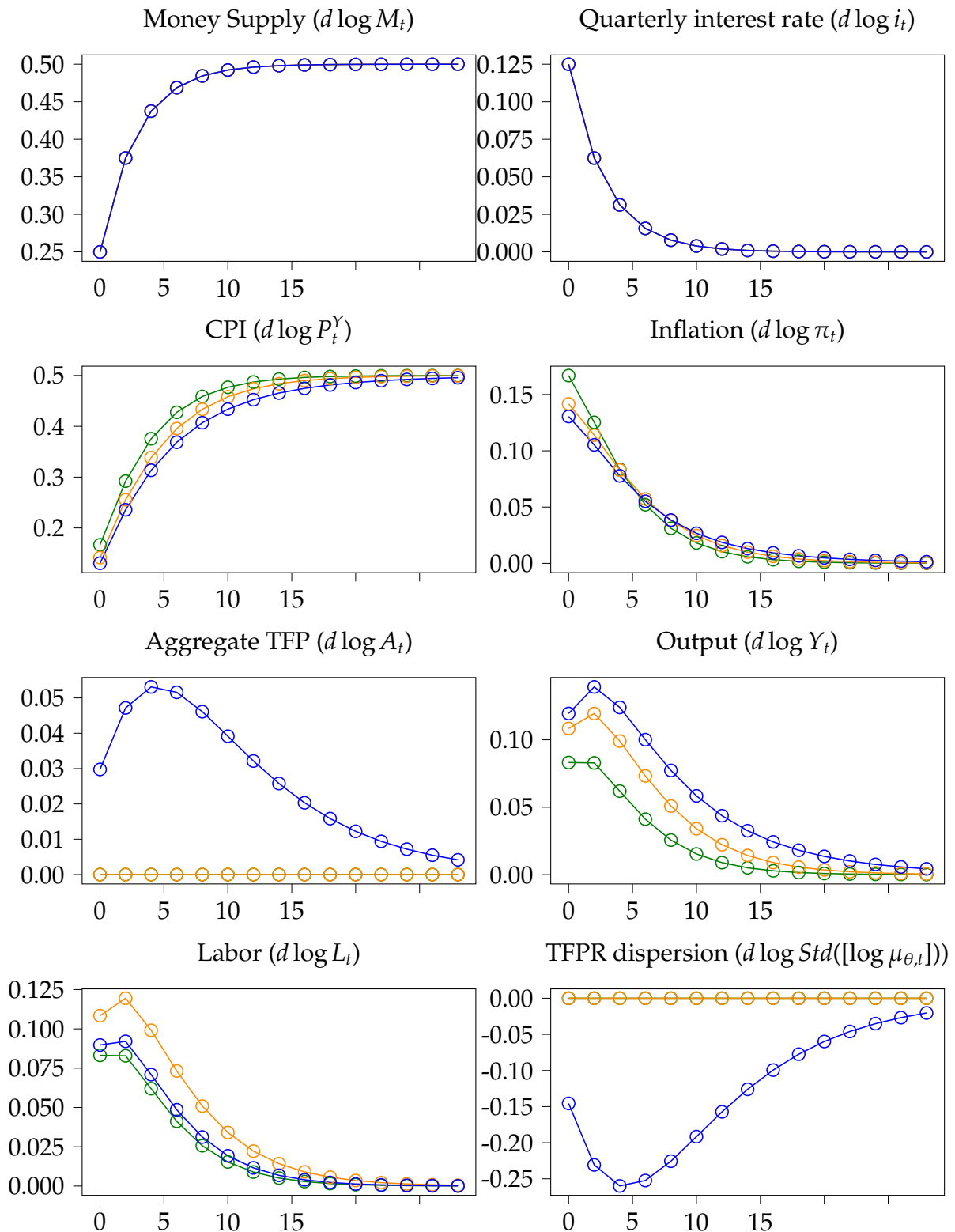
Model	Output effect at $t = 0$	Half life	Cumulative output impact
CES	0.083	2.98	0.330
Homogeneous Firms	0.108	3.85	0.545
Heterogeneous Firms	0.119	4.93	0.792

Appendix F Multiple Sectors, Multiple Factors, and Sticky Wages

In this appendix, we provide an extension of the model to multiple sectors and multiple factors, following the general network production structure provided by Baqaee and Farhi (2018). We use Ω to refer to the revenue-based input-output matrix,

$$\Omega_{ij} = \frac{p_j x_{ij}}{p_i y_i}, \quad (132)$$

Figure E.2: Impulse response functions (IRFs) following an expansionary money supply shock. Green, orange, and blue IRFs indicate the CES, homogeneous firms, and heterogeneous firms models respectively.



where Ω_{ij} is share of producer i 's costs spent on good j as a fraction of producer i 's total revenue. Similarly, the cost-based input-output matrix,

$$\tilde{\Omega}_{ij} = \frac{p_j x_{ij}}{\sum_l p_l x_{il}}, \quad (133)$$

describes producer i 's spending on good j as a fraction of producer i 's total costs. The revenue-based Leontief inverse matrix and cost-based Leontief inverse matrix are defined as,

$$\Psi = (1 - \Omega)^{-1}, \quad (134)$$

$$\tilde{\Psi} = (1 - \tilde{\Omega})^{-1}. \quad (135)$$

Some additional notation: We use $\tilde{\Lambda}_f$ and Λ_f to refer to the share of factor f as a fraction of nominal GDP and as a fraction of total factor costs, respectively, and use λ_I to refer to the sales share of sector I . The parameter ζ_f is the elasticity of factor f to its real price (or wage, in the case of labor), and $\gamma_f \zeta_f$ is the elasticity of factor f to income. The parameter θ_I is the elasticity of substitution between inputs for sector I . We use the notation of the covariance operator $Cov_{\Omega(\theta)}$ as defined in Baqaee and Farhi (2018).

We can now derive the aggregate productivity and markup of any sector I just as in the one-sector model:

$$d \log A_I = \mathbb{E}_{\frac{\lambda}{\lambda_I}} \left[\mu_{\theta}^{-1} \right] \frac{\mathbb{E}_{\frac{\lambda}{\lambda_I}} [\delta_{\theta}] \left(1 - \mathbb{E}_{\frac{\lambda}{\lambda_I}} [\delta_{\theta}] \right) Cov_{\frac{\lambda}{\lambda_I} \delta} [\rho_{\theta}, \sigma_{\theta}] + \mathbb{E}_{\frac{\lambda}{\lambda_I} \delta} [\rho_{\theta}] Cov_{\frac{\lambda}{\lambda_I}} [\sigma_{\theta}, \delta_{\theta}]}{\mathbb{E}_{\frac{\lambda}{\lambda_I}} [[\delta_{\theta} \rho_{\theta} + (1 - \delta_{\theta})] \sigma_{\theta}]} \cdot \left[\sum_{\mathcal{J}} \tilde{\Omega}_{IJ} d \log \frac{p_{\mathcal{J}}}{P} + d \log P \right]. \quad (136)$$

$$d \log \mu_I = - \left[\frac{\mathbb{E}_{\frac{\lambda}{\lambda_I}} [\delta_{\theta} (1 - \rho_{\theta})] \mathbb{E}_{\frac{\lambda}{\lambda_I}} [\sigma_{\theta} (1 - \delta_{\theta})]}{\mathbb{E}_{\frac{\lambda}{\lambda_I}} [[\delta_{\theta} \rho_{\theta} + (1 - \delta_{\theta})] \sigma_{\theta}]} + \mathbb{E}_{\frac{\lambda}{\lambda_I}} [1 - \delta_{\theta}] \right] \left[\sum_{\mathcal{J}} \tilde{\Omega}_{IJ} d \log \frac{p_{\mathcal{J}}}{P} + d \log P \right] + d \log A_I. \quad (137)$$

The remaining aggregation equations follow directly from Baqaee and Farhi (2018). The

change in output is:

$$d \log Y = \frac{1}{\sum_f \tilde{\Lambda}_f \frac{1+\gamma_f \zeta_f}{1+\zeta_f}} \left[\sum_I \tilde{\lambda}_I (d \log A_I - d \log \mu_I) - \frac{1}{1+\zeta_f} \sum_f \tilde{\Lambda}_f d \log \Lambda_f \right]. \quad (138)$$

The change in the sales share of sector \mathcal{K} is:

$$\begin{aligned} d \log \lambda_{\mathcal{K}} &= \sum_I \left(\delta_{\mathcal{K}I} - \lambda_I \frac{\Psi_{I\mathcal{K}}}{\lambda_{\mathcal{K}}} \right) d \log \mu_I \\ &\quad + \sum_{\mathcal{J}} (\theta_{\mathcal{J}} - 1) \lambda_{\mathcal{J}} \mu_{\mathcal{J}}^{-1} \text{Cov}_{\tilde{\Omega}(\mathcal{J})} \left(\sum_I \tilde{\Psi}_{(I)} (d \log A_I - d \log \mu_I), \frac{\Psi_{(\mathcal{K})}}{\lambda_{\mathcal{K}}} \right) \\ &\quad - \sum_{\mathcal{J}} (\theta_{\mathcal{J}} - 1) \lambda_{\mathcal{J}} \mu_{\mathcal{J}}^{-1} \text{Cov}_{\tilde{\Omega}(\mathcal{J})} \left(\sum_g \frac{\tilde{\Psi}_{(g)}}{1+\zeta_g} (d \log \Lambda_g + (\gamma_g \zeta_g - \zeta_g) \log Y), \frac{\Psi_{(\mathcal{K})}}{\lambda_{\mathcal{K}}} \right). \end{aligned} \quad (139)$$

The change in the share of income going to factor f is:

$$\begin{aligned} d \log \Lambda_f &= - \sum_I \lambda_I \frac{\Psi_{If}}{\Lambda_f} d \log \mu_I + \sum_{\mathcal{J}} (\theta_{\mathcal{J}} - 1) \lambda_{\mathcal{J}} \mu_{\mathcal{J}}^{-1} \text{Cov}_{\tilde{\Omega}(\mathcal{J})} \left(\sum_I \tilde{\Psi}_{(I)} (d \log A_I - d \log \mu_I), \frac{\Psi_{(f)}}{\Lambda_f} \right) \\ &\quad - \sum_{\mathcal{J}} (\theta_{\mathcal{J}} - 1) \lambda_{\mathcal{J}} \mu_{\mathcal{J}}^{-1} \text{Cov}_{\tilde{\Omega}(\mathcal{J})} \left(\sum_g \frac{\tilde{\Psi}_{(g)}}{1+\zeta_g} (d \log \Lambda_g + (\gamma_g \zeta_g - \zeta_g) \log Y), \frac{\Psi_{(f)}}{\Lambda_f} \right). \end{aligned} \quad (140)$$

Factor and sector prices follow:

$$d \log \frac{w_f}{P} = \frac{1}{1+\zeta_f} d \log \Lambda_f + \frac{1+\gamma_f \zeta_f}{1+\zeta_f} d \log Y, \quad (141)$$

$$d \log \frac{p_I}{P} = - \sum_{\mathcal{K}} \tilde{\Psi}_{I\mathcal{K}} (d \log A_{\mathcal{K}} - d \log \mu_{\mathcal{K}}) + \sum_f \tilde{\Psi}_{If} d \log \frac{w_f}{P}. \quad (142)$$

To illustrate the results, we consider a simple example with two factors (capital and labor) and sticky wages.

F.1 Example: Two factors and sticky wages

We apply the multiple factor and multiple sector model above. Consider an economy with two factors, labor and capital. Labor is elastic, with a Frisch elasticity of 0.2, as in the model considered in the main text, while capital is inelastic. We allow for sticky wages by introducing a ‘‘labor union sector’’: this sector buys all labor, and then supplies labor

to firms in the industry sector at a price which is subject to nominal rigidities.

The industry sector consists of firms in monopolistic competition who use capital and labor provided by the labor union to produce varieties. Just as in the main text, firms in the industry sector have heterogeneous productivities and endogenous markups and pass-throughs; we use the same parameters and objects from the firm distribution given in the main text for this calibration. Additionally, we set the share of labor to $\tilde{\Lambda}_L = 2/3$ and the share of capital to $\tilde{\Lambda}_K = 1/3$. We allow both the elasticity of substitution between labor and capital used by firms in the industry sector, denoted θ_I , and the degree of wage-stickiness, denoted δ_w , to vary across calibrations.

We show the results of this model in Figure F.1. The plot shows the change in aggregate productivity in the firm sector, $(d \log A_I)$, the change in output $(d \log Y)$, the change in the shares of income to labor and capital $(d \log \lambda_W$ and $d \log \lambda_K)$, and the real price of labor and capital $(d \log p_{W/P}$ and $d \log r/P)$ following a shock to the price level $(d \log P)$.⁴⁶

One immediate implication of this exercise is that the productivity response in the firm sector is independent of frictions upstream, such as sticky wages or complementarity in inputs. As a result, the importance of the misallocation channel in transmitting monetary shocks is robust to the addition of wage rigidities or deviating from Cobb-Douglas production. Furthermore, note that the cyclicality of labor's share of income is, in general, ambiguous. With sufficiently rigid wages, it is possible to make the labor share countercyclical (and the share of income accruing to profits and capital procyclical).

Appendix G Klenow-Willis Calibration

Under Klenow and Willis (2016) preferences, the markup and pass-through functions are

$$\mu_\theta = \mu\left(\frac{y_\theta}{Y}\right) = \frac{1}{1 - \frac{1}{\sigma}\left(\frac{y_\theta}{Y}\right)^{\frac{\epsilon}{\sigma}}}, \quad (143)$$

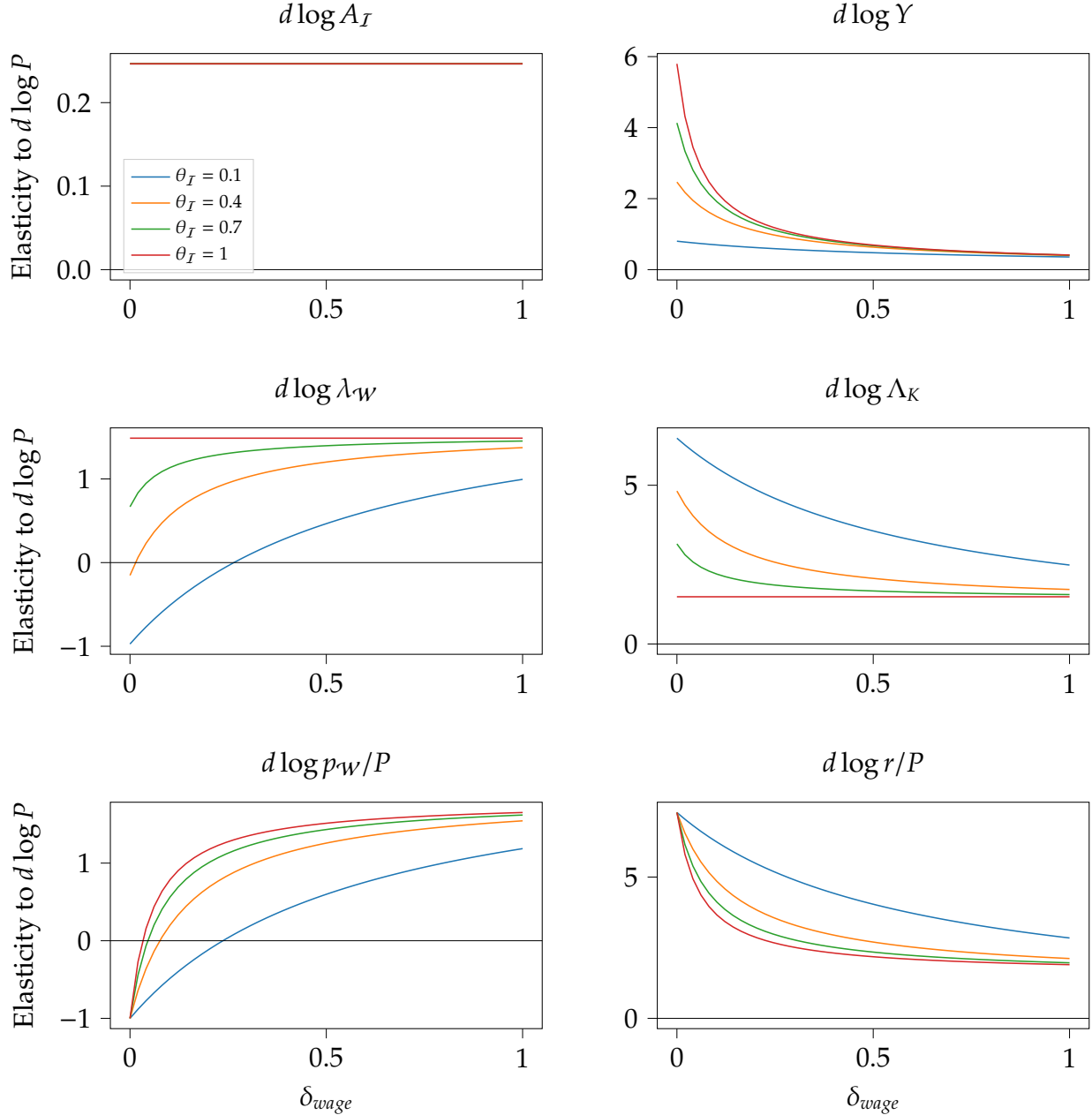
$$\rho_\theta = \rho\left(\frac{y_\theta}{Y}\right) = \frac{1}{1 + \frac{\epsilon}{\sigma - \left(\frac{y_\theta}{Y}\right)^{\frac{\epsilon}{\sigma}}}} = \frac{1}{1 + \frac{\epsilon}{\sigma}\mu_\theta}. \quad (144)$$

where the parameters σ and ϵ are the elasticity and superelasticity (i.e., the rate of change in the elasticity) that firms would face in a symmetric equilibrium. This functional form imposes a maximum output of $(y_\theta/Y)^{\max} = \sigma^{\frac{\sigma}{\epsilon}}$, at which markups approach infinity.

Unfortunately, these preferences are unable to match the empirical distribution of firm pass-throughs without counterfactually large markups. To see why, note that the

⁴⁶We focus on the labor share and the real wage of the labor union sector, since these are the labor share and real wage that would be observed.

Figure F.1: Response to shock to price level ($d \log P$) in one period model with capital, labor, and sticky wages. The degree of wage-stickiness varies along the x-axis, from complete rigidity (zero) to complete flexibility (one). Lines indicate calibrations with different elasticities of substitution between capital and labor.



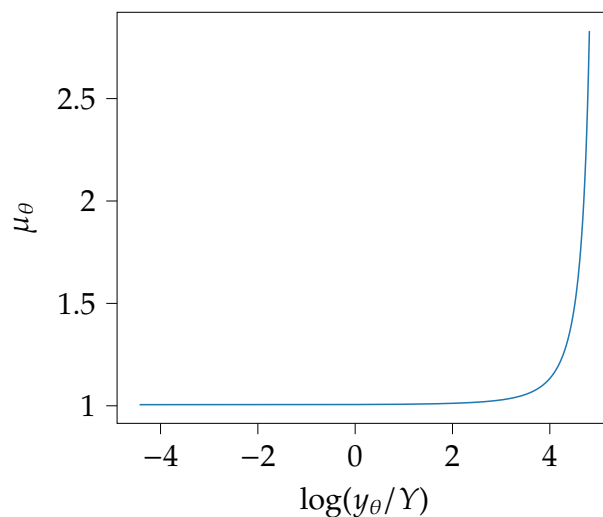
pass-through function $\rho(\cdot)$ is strictly decreasing, and that the maximum pass-through admissible (for a firm with $y_\theta/Y = 0$) is

$$\rho^{\max} = \frac{1}{1 + \epsilon/\sigma}.$$

Amiti et al. (2019) estimate the average pass-through for the smallest 75% of firms in ProdCom is 0.97. In order to match the nearly complete pass-through for small firms, we must choose ϵ/σ to be around 0.01 – 0.03.

This makes it difficult, however, to match the incomplete pass-throughs estimated for the largest firms. To match a pass-through of $\rho_\theta = 0.3$ with $\epsilon/\sigma \in [0.01, 0.03]$, for example, we need a markup of $\mu_\theta \in [78, 233]$ for the largest firms. In contrast, our non-parametric procedure matches the pass-through distribution with moderate markups for the largest firms, shown in Figure G.1. Importantly, since markups and pass-throughs depend on the elasticity of $Y'(\cdot)$, incorporating additional modeling elements (such as demand shifters correlated with firm productivity) does not avoid the counterfactual properties shown here.

Figure G.1: Firm markups μ_θ estimated using nonparametric approach with $\bar{\mu} = 1.15$.



Rather than attempting to match the empirical pass-through distribution, suppose we used a set of parameters from the literature. We adopt the calibration from Appendix D of Amiti et al. (2019): $\sigma = 5$, $\epsilon = 1.6$, and firm productivities are drawn from a Pareto distribution with shape parameter equal to 8.⁴⁷ The simulated distributions of firm pass-throughs and sales shares are shown in Figure G.2. Over the range of drawn productivities, we see little variation in pass-through. Figure G.3 shows the response of output to an interest rate shock, calibrated with the same parameters as in Section 6.4. We find that the parametric specification dramatically understates the misallocation channel, compared to the nonparametric approach adopted in the main text.

⁴⁷We calibrate the model by drawing 1000 firms and finding a fixed point in output. Since the Pareto distribution is unbounded, we could theoretically draw firms with zero pass-throughs and infinite sales shares; the simulated distributions are bounded away from these extremes.

Figure G.2: Pass-through ρ_θ and sales share density $\log \lambda_\theta$ for Klenow and Willis (2016) calibration.

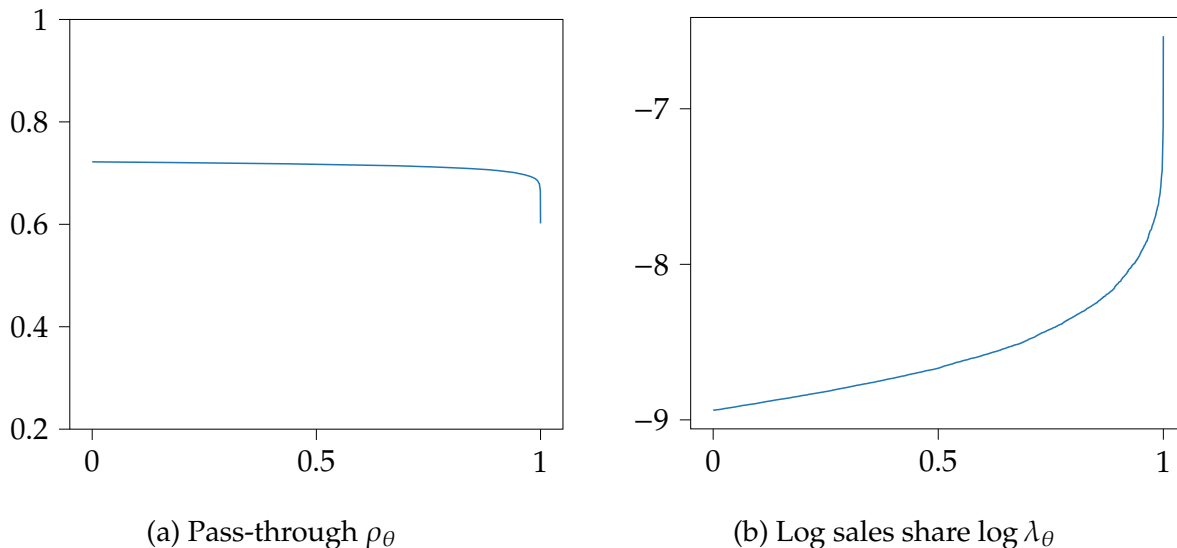
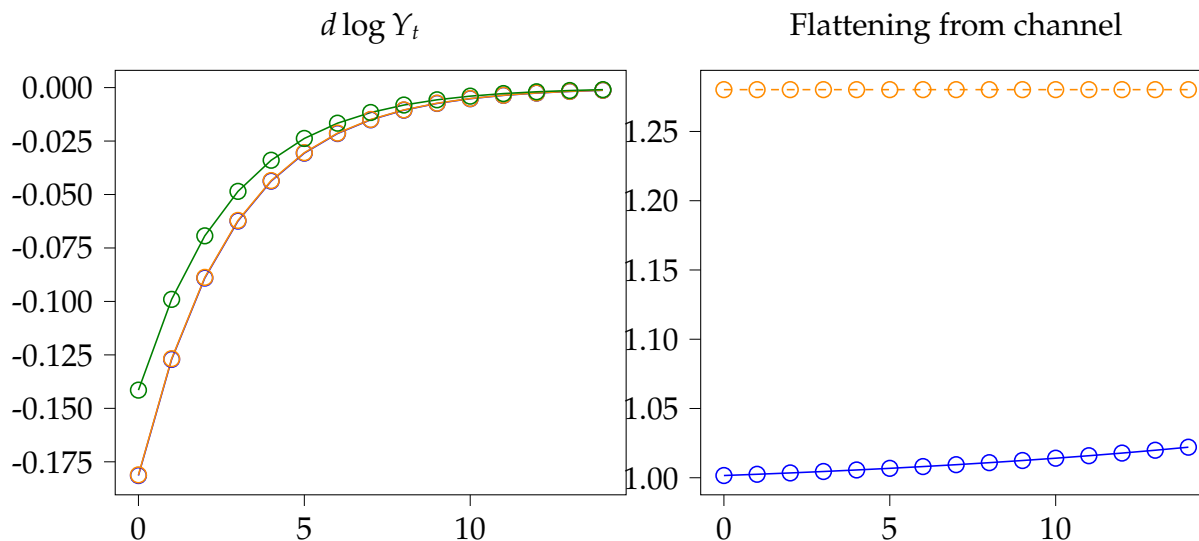


Figure G.3: Impulse response function of output following a monetary policy shock, calibrated using Klenow and Willis (2016) preferences. The real rigidities model IRF and full model IRF coincide in the left panel.



Appendix H Oligopoly Calibration

An alternative to using the monopolistic competition framework is analyzing monetary policy through the lens of oligopoly. We describe the model set up first, and then show our calibrated results. We find that both qualitatively and quantitatively, the misallocation

channel behaves similarly to the model with monopolistic competition.

H.1 Model Setup

We show how Propositions 1 and 2 can be rederived in an environment with oligopolistic competition. To do so, we adopt the nested CES model of Atkeson and Burstein (2008). Assume that there is a continuum of sectors indexed by \mathcal{I} . The representative agent has Cobb-Douglas preferences across sectors. There is a finite number of heterogenous firms in each sector. The representative agent has CES preferences with an elasticity $\sigma_{\mathcal{I}}$ over varieties within a sector. We denote by γ and ζ the income and Frisch elasticities of labor supply.

Each firm $i \in \mathcal{I}$ has a probability δ_i of being able to change its price, and a probability $1 - \delta_i$ of its price remaining fixed. The realizations are independent across firms. It will simplify the analysis to assume that when the firms that get to change their price make their pricing decision, they do not know which other firms will get to change their prices. We assume throughout that firms take the prices of inputs and other firms as given (Bertrand competition). Let λ_i be the sales share of firm i and $\lambda_{\mathcal{I}}$ be the sales share of sector \mathcal{I} .

Desired pass-through is given by

$$\rho_i^{flex} = 1 - s_i \frac{\sigma_{\mathcal{I}} - 1}{\sigma_{\mathcal{I}}},$$

where $s_i = \lambda_i / \lambda_{\mathcal{I}}$ is the market share of firm i . Hence, larger firms will have lower desired pass-throughs.

Lemma 2 (Effective pass-through). *The effective expected equilibrium pass-through of firm i is given by*

$$\rho_i = 1 - \frac{1}{1 + \delta_i \frac{1 - \rho_i^{flex}}{1 - s_i} s_i} \left[\delta_i \frac{1 - \rho_i^{flex}}{1 - s_i} \frac{\sum_{j \in \mathcal{I}} s_j \frac{1 - \delta_j}{1 + \delta_j \frac{1 - \rho_j^{flex}}{1 - s_j} s_j}}{1 - \sum_{j \in \mathcal{I}} \frac{\delta_j \frac{1 - \rho_j^{flex}}{1 - s_j} s_j}{1 + \delta_j \frac{1 - \rho_j^{flex}}{1 - s_j} s_j}} + (1 - \delta_i) \right].$$

This is how much the price of firm i is expected to change in response to an aggregate shock to nominal marginal cost, taking into account the nominal rigidities and the responses of other firms in the sector. Effective expected equilibrium pass-through of firm i is increasing in desired pass-through. Note that when there are no nominal rigidities,

effective equilibrium pass-through is complete. Define the sectoral markup μ_I and the aggregate markup μ to be market-share weighted harmonic averages.

Proposition 7 (TFP in Oligopoly Model). *Following a monetary shock, the response of aggregate TFP at $t = 1$ is*

$$d \log A = - \sum_I \lambda_I \sigma_I \text{Cov}_{\frac{\lambda_i}{\lambda_I}} \left(1 - \frac{\mu_i^{-1}}{\mu_I^{-1}}, \mathbb{E} [d \log \mu_i] \right) - \text{Cov}_{\lambda_I} \left(1 - \frac{\mu_I^{-1}}{\mu^{-1}}, \mathbb{E} [d \log \mu_I] \right),$$

where

$$\mathbb{E} [d \log \mu_i] = (1 - \rho_i) d \log w$$

and

$$\mathbb{E} [d \log \mu_I] = -\mathbb{E}_{\frac{\lambda_i}{\lambda_I}} [1 - \rho_i] + \sigma_I \mu_I \text{Cov}_{\frac{\lambda_i}{\lambda_I}} (\mu_i^{-1}, \rho_i) d \log w.$$

The first set of summands in $d \log A$ are changes in allocative efficiency due to reallocations within sectors, and the second set of summands are changes in allocative efficiency due to reallocations across sectors. If sectoral markups are the same across all sectors, the second set of summands in $d \log A$ drop out.

Proposition 8 (Output in Oligopoly Model). *Following a monetary shock, the response of aggregate output at $t = 1$ is*

$$\frac{d \log Y}{d \log w} = \rho \left[d \log A - \frac{\zeta}{1 + \zeta} \sum_I \frac{\lambda_I \mu_I^{-1}}{\mu^{-1}} \mathbb{E} \left[\frac{d \log \mu_I}{d \log w} \right] \right].$$

Using these expressions we can recover the price and wage Phillips curve, and calibrate the amount of flattening due to the misallocation channel and due to real rigidities respectively.

H.2 Calibration

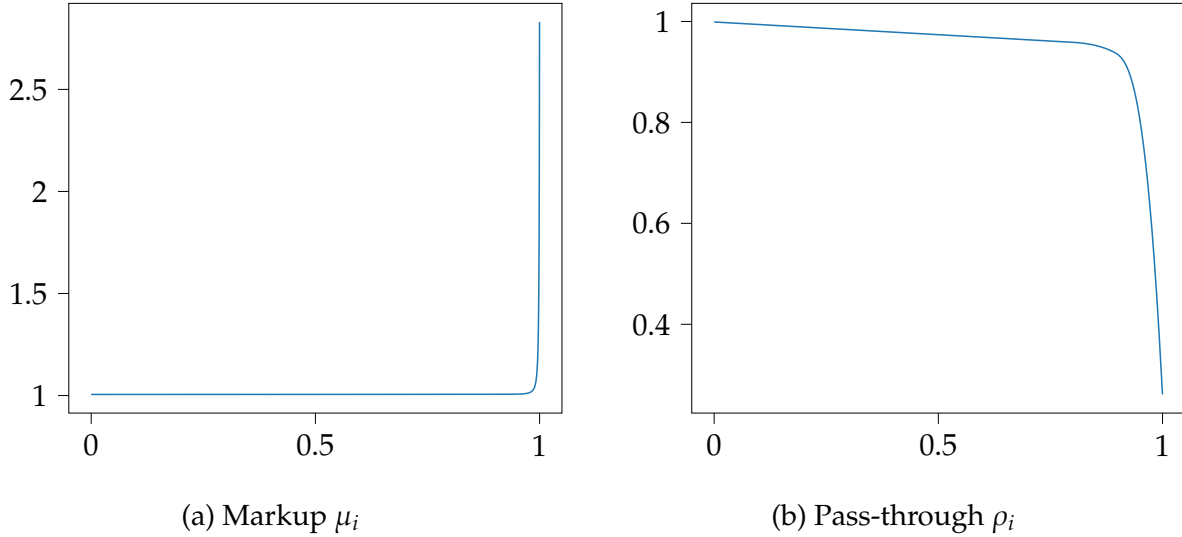
To calibrate the model, we follow Amity et al. (2019) and set the elasticity of substitution across sectors to one, and the elasticity within sectors to 10. We draw firm productivities from a Pareto distribution with shape parameter equal to 8.⁴⁸

We order firms by market share within sector, and plot the markups and pass-throughs of firms in Figure H.1.⁴⁹ The markups and pass-throughs generated by the nested CES

⁴⁸These parameters are chosen by Amity et al. (2019) to match moments of the empirical distribution. We refer readers to Appendix D of their paper for more detail.

⁴⁹If we instead plot markups and pass-throughs against firm market shares, we exactly replicate Figure A3 from Amity et al. (2019).

Figure H.1: Markups μ_i and pass-throughs ρ_i for firms in the oligopoly calibration, ordered by market share.



model satisfy Marshall’s strong second law of demand: markups are increasing in firm productivity, and pass-throughs are decreasing in productivity. Both the markup and pass-through function are quantitatively similar to the ones we derived for the monopolistic competition version of the model used in the main text.

We calculate the slope of the wage and price Phillips curves in a one-period setting, mirroring the timing of the one-period model presented in the main text. The flattening of the Phillips curves due to real rigidities and the misallocation channel are presented in Table H.1. In this setting, as in the setting with monopolistic competition, we find that the misallocation channel is quantitatively important: the misallocation channel flattens both the wage and price Phillips curves by 31%, compared to real rigidities, which flatten the wage Phillips curve by 17% and the price Phillips curve by 42%.

Table H.1: Estimates of Phillips curve flattening due to real rigidities and the misallocation channel in oligopoly calibration.

Flattening	Wage Phillips curve	Price Phillips curve
Real rigidities	1.17	1.42
Misallocation channel	1.31	1.31

Appendix I Variation in markups and pass-throughs

The calibration in the main text assumes that firm markups and pass-throughs are vary only as a function of firm size. In practice, other factors unrelated to firm size may also influence markups and pass-throughs, however. Suppose that we allow the demand elasticity and desired pass-throughs of a firm i to vary due to factors unrelated to firm size,

$$\begin{aligned}\sigma_i &= \underbrace{\mathbb{E}[\sigma_i|\lambda_i]}_{\sigma_\lambda} + \epsilon_i, \\ \rho_i &= \underbrace{\mathbb{E}[\rho_i|\lambda_i]}_{\rho_\lambda} + \nu_i,\end{aligned}$$

where ϵ_i and ν_i are orthogonal to λ_i (and hence to σ_λ and ρ_λ), but may be correlated with each other ($\mathbb{E}[\epsilon_i\nu_i] \neq 0$). We can microfound this by perturbing the Kimball aggregator by firm. We consider how this flexibility changes the sales-weighted elasticity, sales-weighted pass-through, and covariance of elasticities and pass-throughs, which are sufficient to determine the model's results.

Introducing variation unrelated to firm size does not change the sales-weighted average elasticity and pass-through, due to the law of iterated expectations,

$$\begin{aligned}\mathbb{E}_\lambda[\sigma_i] &= \mathbb{E}[\mathbb{E}[\lambda_i\sigma_i|\lambda_i]]/\mathbb{E}[\lambda_i] \\ &= \mathbb{E}[\lambda_i\sigma_\lambda]/\mathbb{E}[\lambda_i] \\ &= \mathbb{E}_\lambda[\sigma_\lambda].\end{aligned}$$

The covariance of elasticities and pass-throughs may change, however:

$$\begin{aligned}\text{Cov}_\lambda[\sigma_i, \rho_i] &= \text{Cov}_\lambda(\sigma_\lambda + \epsilon_i, \rho_\lambda + \nu_i) \\ &= \text{Cov}_\lambda(\sigma_\lambda, \rho_\lambda) + \text{Cov}_\lambda(\epsilon_i, \nu_i) \\ &= \text{Cov}_\lambda(\sigma_\lambda, \rho_\lambda) + \underbrace{\sqrt{\text{Var}_\lambda(\epsilon_i)\text{Var}_\lambda(\nu_i)}\text{Corr}_\lambda(\epsilon_i, \nu_i)}_{\text{Bias}}.\end{aligned}$$

Whether the bias attenuates or magnifies the supply-side effects in the model depends on the correlation between ϵ_i and ν_i , and the magnitude of the bias is bounded by the sales-weighted variance of both errors.

For example, consider the case where the consumer bundle aggregator includes de-

mand shifters B_i (i.e., $\Upsilon_i(\cdot) = B_i\Upsilon(\cdot)$):

$$\int_0^1 B_i \Upsilon\left(\frac{y_i}{Y}\right) di = 1.$$

Suppose we perturb B_i for some firm i away from one, and hold $B_j = 1$ for all $j \neq i$. To a first order, the changes in the elasticity and pass-through of firm i are,

$$\begin{aligned} \frac{d \log \sigma_i}{d \log B_i} &= \frac{\partial \log \sigma\left(\frac{y}{Y}\right)}{\partial \log \frac{y}{Y}} \rho_i \sigma_i \\ \frac{d \log \rho_i}{d \log B_i} &= \frac{\partial \log \rho\left(\frac{y}{Y}\right)}{\partial \log \frac{y}{Y}} \rho_i \sigma_i \end{aligned}$$

Under Marshall's strong second law, $\frac{\partial \log \sigma\left(\frac{y}{Y}\right)}{\partial \log \frac{y}{Y}} < 0$ and $\frac{\partial \log \rho\left(\frac{y}{Y}\right)}{\partial \log \frac{y}{Y}} < 0$, hence $\text{Corr}(\epsilon_i, v_i) > 0$, and the supply-side effects are magnified, rather than attenuated.

More generally, we can bound the bias in the supply-side effects using the result from Proposition 1 (assuming $\delta_i = \delta$ across firms):

$$d \log A = \bar{\mu} \left(\frac{\delta (1 - \delta) \text{Cov}_\lambda [\sigma_i, \rho_i]}{(1 - \delta) \mathbb{E}_\lambda [\sigma_i] + \delta (\text{Cov}_\lambda [\sigma_i, \rho_i] + \mathbb{E}_\lambda [\sigma_i] \mathbb{E}_\lambda [\rho_i])} \right) d \log w.$$

The true supply-side effect, $d \log A^{\text{true}}$ (calculated using $\text{Cov}_\lambda [\sigma_i, \rho_i]$) is related to the supply-side effect calculated using variation due to sales share alone, $d \log A$ (calculated using $\text{Cov}_\lambda [\sigma_\lambda, \rho_\lambda]$), by

$$\frac{d \log A^{\text{true}}}{d \log A} = 1 + \frac{1 - d \log A}{d \log A + \frac{\text{Cov}_\lambda(\sigma_\lambda, \rho_\lambda)}{\sqrt{\text{Var}_\lambda(\epsilon_i) \text{Var}_\lambda(v_i) \text{Corr}_\lambda(\epsilon_i, v_i)}}}.$$

To illustrate, suppose 90% of variation in elasticities and pass-throughs comes from sales share, and 10% from other factors. For the calibration exercise given in the main paper, we find $\frac{d \log A^{\text{true}}}{d \log A} \in (0.69, 1.27)$; i.e., if variation not due to sales share in elasticities and pass-throughs is perfectly negatively correlated, the supply-side effect is attenuated by 31%, and if this variation is perfectly positively correlated, the supply-side effect is magnified by 27%.

Appendix J Gini coefficient in US data

We use Business Dynamic Statistics (BDS) data from the US Census to calculate the Gini coefficient in firm employment. Figure J.1 shows the Lorenz curve in employment for the firm distribution in 2018. We calculate the ratio of the shaded area (approximated using trapezoids) to the area under the 45-degree line to measure the Gini coefficient.

Figure J.2 plots the estimated Gini coefficients from 1978-2018 for all firms, as well as within sectors provided by the BDS. The trends by sector are consistent with the trends described in Figure A.1 of Autor et al. (2020), who measure HHI across sectors: we find increasing concentration in retail, wholesale trade, utilities, and finance, and flat or decreasing concentration in manufacturing. We use the beginning and end of the time series for all firms and for the retail sector for calibrations in the main text.

Figure J.1: Lorenz curve of cumulative firm employment by share of firms in 2018.

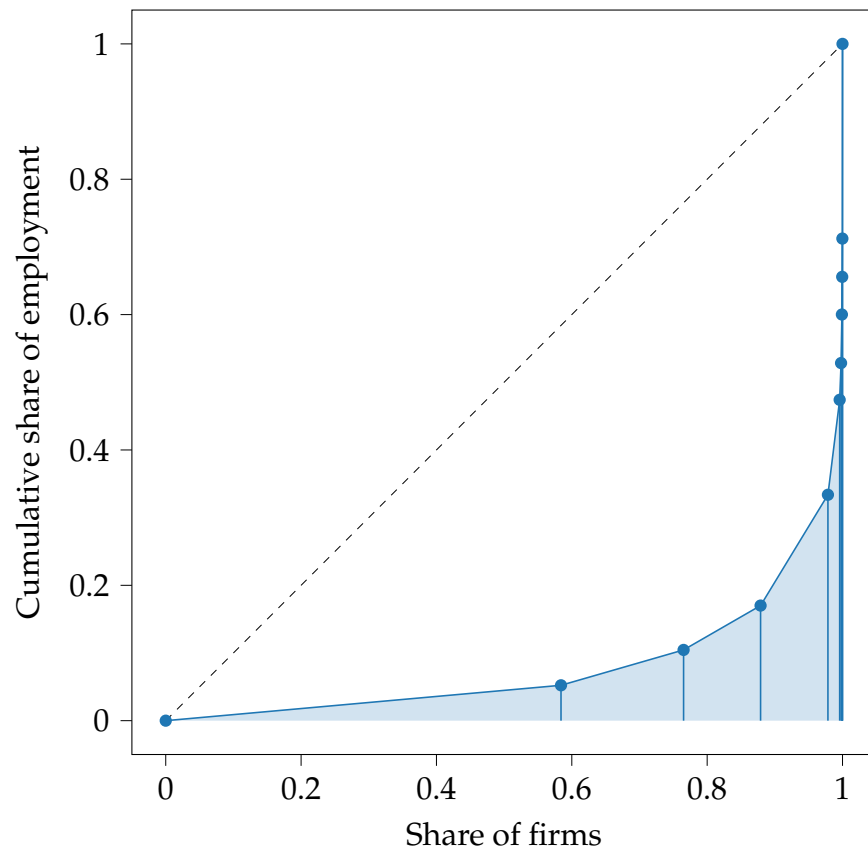


Figure J.2: Estimated Gini coefficients in Census BDS data from 1978-2018.

