NBER WORKING PAPER SERIES

THE SUPPLY-SIDE EFFECTS OF MONETARY POLICY

David Baqaee
Emmanuel Farhi
Kunal Sangani

Working Paper 28345
http://www.nber.org/papers/w28345

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2021, Revised December 2025

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. We are grateful to Harald Uhlig and three anonymous referees for helpful suggestions. We thank Andy Atkeson, Saki Bigio, Ariel Burstein, Oleg Itskhoki, Ivan Werning, Jon Vogel, and other seminar participants for helpful comments. This paper received support from NSF grant No. 1947611. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

The Supply-Side Effects of Monetary Policy
David Baqaee, Emmanuel Farhi, and Kunal Sangani
NBER Working Paper No. 28345
January 2021, Revised December 2025
JEL No. E0,E12,E24,E3,E4,E5,L11,O4

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, demand shocks such as monetary shocks have a first-order effect on aggregate productivity. If high-markup firms have lower pass-throughs than low-markup firms, as is consistent with empirical evidence, then a monetary easing reallocates resources to high-markup firms and alleviates misallocation. Consequently, positive "demand shocks" are accompanied by endogenous positive "supply shocks" that raise output and productivity, lower inflation, and flatten the Phillips curve. We derive a tractable four-equation dynamic model and use it to show that monetary shocks generate a procyclical hump-shaped response in TFP and endogenous cost-push shocks in the New Keynesian Phillips curve. A calibration of our model suggests that the supply-side effect increases the half-life of a monetary shock's effect on output by about 30% and amplifies the total impact on output by about 70%. Using identified monetary shocks, we provide empirical evidence for both the macro- and micro-level predictions of our model.

David Baqaee
Department of Economics
University of California at Los Angeles
Bunche Hall
Los Angeles, CA 90095
and CEPR
and also NBER
baqaee@econ.ucla.edu

Emmanuel Farhi
Harvard University

Kunal Sangani
Harvard University
Wyss Hall
20 N Harvard St
Boston, MA 02163
ksangani@g.harvard.edu

1 Introduction

How do demand shocks, like monetary shocks, affect an economy's productivity? A common view is that they do not. Instead, aggregate productivity is determined by long-run institutional and technological forces that are orthogonal to short-run demand disturbances.

Yet, aggregate productivity, 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 total factor productivity (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 aggregate productivity is mismeasured, for example due to variable capacity utilization or external returns, resulting in a spurious relationship between measured productivity and shifts in demand.

In this paper, we present an alternative explanation. Rather than being an exogenous primitive, aggregate TFP is an endogenous object that depends on the allocation of resources among producers. We argue that demand shocks, such as monetary or discount factor shocks, can induce changes in aggregate TFP by altering allocations. We provide a model with realistic firm-level heterogeneity where expansionary demand shocks lead to an increase in TFP, not due to mismeasurement or technological changes, but rather due to the beneficial reallocation of resources. Even though the mechanism we propose can apply to any aggregate demand shock that changes nominal marginal costs, we focus on monetary shocks in particular.[2]

The effect of monetary shocks on the allocation of resources yields a new channel through which monetary policy affects real variables, which we call *the misallocation channel*. Under conditions matching empirical patterns on firms, monetary shocks generate procyclical, hump-shaped movements in aggregate TFP. The endogenous "supply shock" caused 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 dampens the response of prices. For example, an expansionary monetary shock boosts aggregate TFP, leading to a larger increase in

[1]The 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.

[2]In our dynamic model, a "demand shock" is a disturbance in the Euler equation (e.g. a monetary or discount factor shock). Other shocks broadly under the umbrella of "demand shocks," such as government spending shocks, can have similar effects on allocative efficiency if they raise nominal marginal costs for all firms, but may also have other distinct effects in a medium-scale model that we abstract from in this paper.

output without as much inflation. Hence, the misallocation channel increases monetary non-neutrality and flattens the Phillips curve.

Monetary shocks increase allocative efficiency if they redirect resources from low to high marginal-revenue-product firms. This presupposes that the initial allocation of resources is inefficient and that the shock has a different impact on firms with different marginal values. Neither condition is satisfied in the workhorse log-linearized New Keynesian model with CES preferences. First, in that model, desired markups are the same for all firms, so the initial cross-sectional allocation of resources is efficient. Since the initial allocation is efficient, optimality implies that demand shocks cannot alter allocative efficiency. Second, even starting at an equilibrium with an initially distorted allocation of resources (i.e. initial markup dispersion), aggregate demand shocks do not differentially affect high and low marginal-revenue-product firms in the standard model, so monetary disturbances do not affect aggregate productivity to a first-order.

In contrast to the benchmark model, the data feature substantial and persistent heterogeneity in markups across firms and systematic differences in how firms pass cost shocks through to their prices. Since firms' desired markups vary, the flexible price equilibrium is generally inefficient: firms with relatively high markups underproduce relative to firms with low markups. Furthermore, since pass-throughs vary systematically with initial markups, demand disturbances that raise or lower marginal costs have differential effects on low- and high-markup firms. In particular, since low-markup firms tend to pass a higher portion of marginal cost changes into prices, an expansionary shock that increases marginal costs causes the prices of low-markup firms to rise relative to high-markup firms. This reallocates resources from low- to high-markup firms and therefore raises aggregate productivity. This misallocation channel 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 incomplete desired pass-through.

To formally analyze these reallocations, we relax the CES demand system in the New Keynesian model using a non-parametric generalized Kimball (1995) demand system.[3] These preferences can accommodate variety-specific downward-sloping residual demand curves of any desired shape while remaining tractable. We couple this demand system with sticky prices using Calvo (1983) frictions.[4] 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

[3]Matsuyama and Ushchev (2017) call this the homothetic with direct implicit additivity (HDIA) demand system.

[4]While Calvo frictions are analytically convenient, we also calibrate a version of our model where nominal rigidities instead take the form of menu costs (see Section 6.5).

desired markups. We consider how TFP and output respond to an aggregate demand shock that raises nominal costs 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 the response of aggregate TFP to a demand shock depends on the cross-sectional covariance of markups and pass-throughs. This covariance can be driven by two factors: heterogeneity in desired pass-through (i.e., pass-through conditional on a price change) or heterogeneity in price stickiness (i.e., the probability of a price change).[5] When markups are negatively correlated with pass-throughs, expansionary monetary shocks that raise nominal marginal costs generate a concomitant increase in aggregate productivity. We argue that this is the empirically relevant case.

Our second result shows that the reaction of output to such shocks can be broken down into distinct demand- and supply-side effects. The demand-side effect is the traditional Keynesian mechanism. It is caused by an increase in labor demand and employment: since nominal rigidities prevent prices from rising one-for-one with spending, increased nominal demand leads to higher labor demand, employment, and output. Real rigidities that dampen the responsiveness of prices to increases in nominal marginal costs enhance this demand-side effect.[6] In contrast, the supply-side effect augments output by raising TFP.

While we illustrate these intuitions in a one-period model, we also extend the benchmark, infinite-horizon New Keynesian model to incorporate these supply-side effects. In the dynamic model, changes in aggregate TFP, output, inflation, and the interest rate satisfy a four-equation system.[7] Relative to the benchmark model, the Taylor rule and the Euler equation are the same but the New Keynesian Phillips curve is different. Our model features a flatter Phillips curve with 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. Our model is 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.

[5]By desired pass-through, we specifically mean the elasticity of the firm's profit-maximizing price with respect to a permanent change in its marginal cost, holding the prices of all competitors constant. In our model, this elasticity depends on the curvature of residual demand curves and is invariant to the source of the marginal cost shock.

[6]In this paper, when we refer to "real rigidities" we specifically mean strategic complementarities in pricing due to variable markups, not real rigidities caused by other forces (like decreasing returns or sticky intermediate input prices).

[7]The four equation system we develop includes two kinds of aggregate demand shocks: monetary shocks and discount factor shocks. One could of course further enrich this framework with other shocks, such as government spending shocks, aggregate productivity shocks, and price-markup shocks.

We calibrate our model using firm data from Belgium (provided by Amiti et al., 2019) and consider the response of economic aggregates to a monetary shock. Our results suggest that the misallocation channel constitutes a quantitatively important part of monetary policy transmission mechanism.[8] In the one-period version of the model, we find that the misallocation channel reduces the slope of 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 Phillips curve by a similar amount. Magnitudes are similar in the dynamic model: the misallocation channel amplifies the cumulative effect of a monetary shock on output by about 70% and increases the half-life of the shock's effect on output by about 30% compared to a model with demand-side effects alone.

As an extension, we show that the misallocation channel is also present and quantitatively similar in a model where nominal rigidities instead take the form of menu costs. In that calibration, changes in the allocation of resources arise due to endogenous differences in the extensive, rather than intensive, margin of price adjustment across firms. In the menu cost model, in response to a monetary expansion, larger firms with higher markups are less likely to adjust their prices than smaller firms with lower markups because they have lower desired pass-through.[9] Hence, monetary expansions reallocate resources from low- to high-markup firms and boost output and productivity.

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%. To put this into context, such an increase in the Gini coefficient is in line with the change in the firm employment distribution in the United States from 1978 to 2018.[10]

Using identified monetary shocks, we provide empirical support for both the macro-

[8]We follow Baqaee et al. (2021) and solve a differential equation to back out the Kimball demand system from data on firm-level sales and pass-throughs. 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 the most popular off-the-shelf functional form, Klenow and Willis (2016), is incapable of simultaneously matching all the relevant sufficient statistics in the data.

[9]See Table 6.

[10]Whether concentration is in fact increasing for relevant market definitions or whether the Phillips curve has indeed flattened over time are topics that are beyond the scope of this paper. See e.g. Rossi-Hansberg et al. (2021), Benkard et al. (2021), Smith and Ocampo (2021) on the former, and e.g., McLeay and Tenreyro (2020), Del Negro et al. (2020), Hooper et al. (2020), Hazell et al. (2020) on the latter.

and micro-level predictions of our model. At the macroeconomic level, we show that aggregate productivity in the U.S.—as measured by labor productivity, the Solow residual, or the cost-based Solow residual—is responsive to Romer and Romer (2004) monetary shocks, in line with the findings of Evans (1992).[11,12] At the microeconomic level, our model ties the increase in aggregate productivity during demand-driven expansions to reallocations towards high-markup firms. Using Compustat data on public firms, we find that expansionary monetary shocks cause high-markup firms to grow relative to low-markup firms in terms of their input usage. This is because firms with high markups cut their markups relative to low-markup firms after a monetary expansion.[13] As a result, both markup dispersion and the dispersion of firm-level revenue productivity (TFPR) fall during demand-driven expansions (as documented by Kehrig, 2011; Meier and Reinelt, 2020). Finally, in keeping with our model's predictions, we show that productivity is more responsive to monetary shocks in industries with higher concentration (measured by the market share of top firms).

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 the real effects of monetary policy shocks, which cannot be explained by nominal rigidities alone given the frequency of price adjustment. Ball and Romer (1990) introduce real rigidities, which complement nominal rigidities to increase monetary nonneutrality.[14] A common formulation of real rigidities is incomplete pass-through, where firms are slow to

[11] Specifically, we use the Wieland and Yang (2020) extension of the Romer and Romer (2004) shocks.

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

[13] We document similar patterns whether we use markups estimated via the user-cost approach from Gutiérrez and Philippon (2017) or from accounting profits; whether we use the updated Romer and Romer (2004) series extended by Wieland and Yang (2020) or monetary shocks identified from high-frequency data by Gorodnichenko and Weber (2016); and whether we consider reallocations across all firms or within industry.

[14] Ball 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 the real effects of monetary shocks. Kimball (1995) formulates a model where real rigidities arise from non-isoelasticity of demand curves. Eichenbaum and Fisher (2004) and Dotsey and King (2005) 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. Mongey (2021) shows that real rigidities can be more powerful, and the extent of pass-through significantly diminished, under dynamic oligopolistic competition.

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, results in another mechanism by which monetary policy affects 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 (2020). For the most part, the misallocation literature is concerned with cross-country 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) show that if price stickiness covaries with markups, then monetary policy affects TFP. The present paper replaces and develops the unpublished analysis in that working paper. In a recent paper, Meier and Reinelt (2020) provide empirical support for this covariance and offer a different microfoundation where 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.

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.

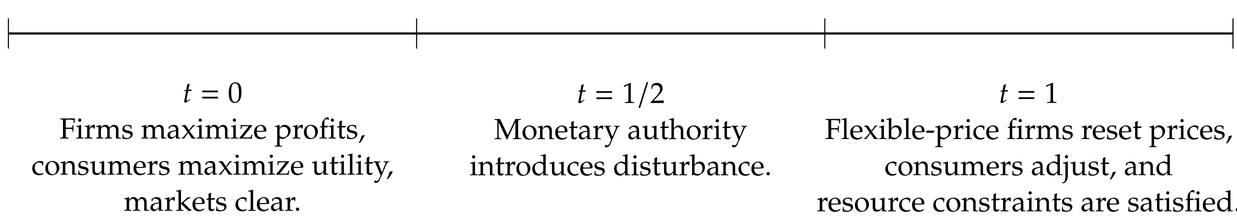
Finally, our paper is also related to a literature on endogenous TFP movements over the business cycle driven by technology change (e.g., Comin and Gertler, 2006; Benigno and Fornaro, 2018; Anzoategui et al., 2019; Bianchi et al., 2019). In this literature, aggregate TFP responds to the business cycle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

Structure of the paper. Section 2 introduces a simple one-period 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. Section 5 generalizes the static model from the previous sections to a fully dynamic setting. Section 6 contains our quantitative results, including an extension with menu costs. 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 an alternative micro-foundation using oligopolistic (rather than monopolistic) competition and versions of the model with multiple sectors, multiple factors, input-output linkages, and sticky wages. Section 9 concludes. All proofs are in Appendix A.

2 Model Setup

To build intuition, we start with a one-period model. Figure 1 shows the timing of the one-period model. At time $t = 0$, the economy is in steady-state: households choose consumption and labor to maximize utility, firms choose prices to maximize profits, and markets clear. The monetary authority then introduces an unexpected disturbance in nominal marginal costs. 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.[15]

Figure 1: One-period model timing.



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

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

[15]We relax the one-period-ahead Calvo friction when we introduce the dynamic model in Section 5. In the infinite-horizon model, each firm changes its price at a constant hazard rate.

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 goods and the consumption bundle Y is defined implicitly by[16]

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

Here, y_θ is the consumption of variety θ, and Φ_θ is an increasing and concave function. CES preferences are the special case when $\Phi_\theta = \Phi$ 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 w is the wage, L is total hours, and Π is profits. Maximization yields the inverse-demand curve for variety θ:

$$\frac{p_\theta}{P} = \Phi'_\theta(\frac{y_\theta}{Y}), \tag{1}$$

where the *price aggregator* P is given by

$$P = \frac{P^Y}{\int_0^1 \Phi'_\theta(\frac{y_\theta}{Y})\frac{y_\theta}{Y}d\theta}, \tag{2}$$

and P^Y is the ideal price index.[17] Equation (1) shows that 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 curves of any desired shape by choosing an appropriate type-specific aggregator Φ_θ.

[16]These preferences are a generalization of Kimball (1995) preferences since the aggregator function Φ_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.

[17]The ideal price index is defined as $\min_{y_\theta}\{\int_0^1 p_\theta y_\theta d\theta : Y = 1\}$. The price aggregator P, which disciplines demand curves, coincides with the ideal price index P^Y if, and only if, preferences are CES. In general, real output Y is given by dividing nominal expenditures by the ideal price index P^Y (and not the price aggregator P). Changes in the ideal price index $d\log P^Y$ are first-order equivalent to changes in the consumer price index (CPI) as calculated by national statistical agencies. Therefore, changes in real output in the data are defined in a way that is consistent with $d\log Y$ in our model.

Firms. Each variety is supplied by a single firm, and a firm of type θ has productivity A_θ. Firms' production technology is linear in labor

$$y_\theta = A_\theta l_\theta.$$

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

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

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:

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

The profit-maximizing price p_θ^{flex} can be written as a desired 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

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

where the markup function is given by the Lerner formula,[18]

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

For CES demand, desired markups $\mu_\theta = \sigma/(\sigma - 1)$ are constant and the same for all firms.

A firm of type θ has a probability δ_θ of being able to reset its price at time $t = 1$. These nominal rigidities may 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.

A firm's desired partial-equilibrium pass-through ρ_θ is the elasticity of its optimal price with respect to its marginal cost, holding the economy-wide aggregates constant.

[18]We 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 $x\Phi'''_\theta(x) + 2\Phi''_\theta(x) < 0$ for every x and θ.

We can express the desired pass-through of firm θ as:

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

Under CES preferences, desired markups do not depend on the firm's position on the demand curve. As a result, desired pass-through is equal to one for all firms, $\rho_\theta \equiv 1$, and firms exhibit "complete desired pass-through." More generally, however, a firm's desired markup may vary with its position on the demand curve and lead to incomplete desired pass-through. For brevity, we refer to ρ_θ simply as the firm's "pass-through" instead of desired partial-equilibrium 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, realized pass-through is de facto equal to zero.

Monetary authority. At time $t = 1/2$, there is an unexpected shock to the nominal wage. We interpret this shock as a disturbance introduced by the monetary authority. We could equivalently 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.[19]

We say that the 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.

[19]For concreteness, 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. In the dynamic version of the model in Section 5, changes in the nominal wage can be caused by either interest rate shocks in the Taylor rule or discount factor shocks in the Euler equation.

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 of firm type θ by[20]

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

and the sales-weighted harmonic average of markups, called the *aggregate markup*, by

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

Log-linearization around initial equilibrium. In what follows, we consider first-order perturbations around an initial equilibrium caused by a change in the nominal wage. For any variable X, we denote its log deviation from its initial value as $d\log X$. More formally, since all variables in this one-period model can be written as implicit functions of the wage w, we use $d\log X$ as a short-hand for $d\log X/d\log w \times \Delta\log w$, where $\Delta\log w$ is a small change in w and the derivatives are evaluated at the initial steady-state.[21]

3 Productivity Response

In this section, we consider how aggregate productivity changes following a monetary shock. Define aggregate productivity A to be aggregate output per unit of labor, so that

$$Y = AL.$$

[20]Without loss of generality, we assume that the type distribution is uniform between [0,1].

[21]$d\log X$ in our notation is the same as the lowercase log deviations used by Galí (2015). We instead opt for $d\log X$ because we use lowercase variables to refer to firm-level variables (e.g., output y_θ and price p_θ) and uppercase variables to refer to economy-wide aggregates (e.g., aggregate output Y and labor L). In the dynamic model in Section 5, these log deviations are instead functions of the entire path of shocks.

Since labor is the sole factor in our model economy, A equals both aggregate TFP and aggregate labor productivity.[22]

Changes in aggregate productivity are closely linked to the distribution of markups across firms. This is because A depends on the efficiency with which workers are divvied up between competing uses. When there is no dispersion in markups, the cross-sectional allocation of resources is efficient. However, when there is heterogeneity in markups, the fraction of labor used by each firm is distorted. Firms with relatively high markups restrict output and use inefficiently too few workers compared to firms with lower markups. Thus, if resources are reallocated to high-markup firms, allocative efficiency improves and output per hour worked rises.

This section shows that the response of aggregate productivity to monetary shocks depends on the cross-sectional covariance of pass-throughs and price elasticities. To establish this, we proceed in steps. First, we show that changes in aggregate productivity are related to changes in markups and then we solve for how markups change in equilibrium.

Lemma 1 applies Theorem 1 from Baqaee and Farhi (2020) to show how changes in aggregate productivity depend on changes in markups.

Lemma 1 (TFP and Changes in Markups)**.** *Following a monetary shock, the change in aggregate productivity is given by*

$$d\log A = d\log\bar{\mu} - \mathbb{E}_\lambda\left[d\log\mu_\theta\right]. \tag{5}$$

Proof. By Shepard's lemma, $d\log P^Y = \mathbb{E}_\lambda[d\log\mu_\theta] + d\log w$. Substitute this into $d\log A = d\log P^Y Y - d\log P^Y - d\log L$ and use the fact that log changes in the labor share of income are negatively related to log changes in the average markup: $d\log(wL/P^Y Y) = -d\log\bar{\mu}$.[23] ■

Lemma 1 demonstrates that when the average markup rises more than individual markups on average ($d\log\bar{\mu} > \mathbb{E}_\lambda[d\log\mu_\theta]$), aggregate productivity A increases due to a composition effect towards firms with higher markups. To see this composition effect explicitly, expand the change in the aggregate markup, $d\log\bar{\mu} = -\mathbb{E}_\lambda\left[(\bar{\mu}/\mu_\theta)d\log(\lambda_\theta/\mu_\theta)\right]$, and substitute it into (5). This yields our next lemma.

[22] Appendix F shows that in a richer economy with multiple factors of production, the relevant measure of A is the distortion-adjusted Solow residual. The distortion-adjusted Solow residual weighs the contributions of primary factors according to their shadow value, rather than their price. See Baqaee and Farhi (2020) for more information on why the distortion-adjusted Solow residual, which generalizes Hall (1990), is the correct object to use in models with misallocation.

[23] Baqaee and Farhi (2020) Theorem 1 states that the change in allocative efficiency in an economy with arbitrary input-output linkages is $d\log A = -\tilde{\Lambda}' d\log\Lambda - \tilde{\lambda}' d\log\mu$, where Λ and $\tilde{\Lambda}$ are vectors of sales- and cost-based factor Domar weights and $\tilde{\lambda}$ is a vector of cost-based Domar weights for firms. In the model developed here, labor is the sole factor, so $\tilde{\Lambda}_L = 1$ and the labor share is the inverse of the aggregate markup $\Lambda_L = 1/\bar{\mu}$. Since there are no intermediates, firms' cost- and sales-based Domar weights coincide ($\tilde{\lambda}_\theta = \lambda_\theta$). Setting $\tilde{\Lambda}_L = 1$, $\tilde{\lambda}_\theta = \lambda_\theta$, and $d\log\Lambda_L = -d\log\bar{\mu}$ yields Equation (5).

Lemma 2 (Reallocations and TFP). *Following a monetary shock, we have*

$$d\log A = -Cov_{\lambda}\left[(\bar{\mu}/\mu_\theta), d\log \lambda_\theta/\mu_\theta\right] = -Cov_{\lambda}\left[(\bar{\mu}/\mu_\theta), d\log Costs_\theta\right], \tag{6}$$

where $Costs_\theta = wl_\theta$ are proportional to $\lambda_\theta/\mu_\theta$.

Aggregate productivity rises when changes in inputs, $d\log \mathrm{Costs}_\theta$, negatively covary with inverse markups $1/\mu_\theta$.[24] In this case, labor is reallocated from low- to high-markup firms. Since high-markup firms are inefficiently too small relative to low-markup firms, such a reallocation boosts aggregate productivity. Lemma 2 is quite general: it continues to hold in the dynamic version of the model (Section 5) and within each sector in a version of the model with intermediate inputs and multiple sectors.[25] A corollary of Lemma 2 is that if initial markups are identical, then a monetary shock has no first-order effect on TFP regardless of how markups change (i.e, regardless of $d\log\mu_\theta$).

To understand how TFP responds to shocks, we must therefore understand how a monetary shock reallocates resources across firms with different initial markups. Whether resources are reallocated toward or away from a firm depends on whether its price rises or falls relative to other firms. The log-linearized residual demand curve is

$$d\log y_\theta - d\log Y = -\sigma_\theta\left[d\log p_\theta - d\log P\right]. \tag{7}$$

Firms that lower their price relative to the market-level price expand in relative terms following a monetary shock. Using the fact that $d\log y_\theta = d\log l_\theta$, we can combine (7) and (6) to get[26]

$$d\log A = (\bar{\mu}/\mathbb{E}_\lambda\left[\sigma_\theta\right])Cov_\lambda\left[\sigma_\theta, d\log p_\theta\right]. \tag{8}$$

[24]A different measure of the change in allocative efficiency relies on the change in markup dispersion and the elasticity of substitution: $\Delta\log TFP = -(\sigma/2)\Delta \mathrm{Var}(\log\mu)$ (see e.g., Hsieh and Klenow, 2009; Meier and Reinelt, 2020). This equation holds only if demand is CES and firm productivities and markups are jointly log-normal, and in general is not the same as the covariance in Lemma 2. When markups are close to one, preferences are CES, and sales shares are symmetric, the two objects approximately coincide: $-(\sigma/2)dVar(\log\mu_\theta) = -\sigma Cov(\log\mu_\theta, d\log\mu_\theta) \approx -\sigma Cov(-1/\mu_\theta, d\log\mu_\theta) \approx Cov(-\bar{\mu}/\mu_\theta, d\log \mathrm{Costs}_\theta)$.

[25]In a multi-sector model, changes in the gross productivity of a sector are given by Lemma 2 as long as all firms within a sector buy inputs at the same prices (see Appendix F).

[26]To get (8), we use $\mathbb{E}_\lambda[\sigma_\theta d\log(p_\theta/P)] = -\mathbb{E}_\lambda[d\log(y_\theta/Y)] = 0$ to rewrite

$$d\log A = Cov_\lambda\left[(\bar{\mu}/\mu_\theta), \sigma_\theta d\log(p_\theta/P)\right] = \bar{\mu}\mathbb{E}_\lambda\left[(\sigma_\theta - 1)\, d\log(p_\theta/P)\right] = -\bar{\mu}\mathbb{E}_\lambda\left[d\log(p_\theta/P)\right].$$

Substitute in $d\log P = \mathbb{E}_{\lambda\sigma}[d\log p_\theta]$ to get

$$d\log A = \bar{\mu}\left(\mathbb{E}_{\lambda\sigma}\left[d\log p_\theta\right] - \mathbb{E}_\lambda\left[d\log p_\theta\right]\right) = \bar{\mu}Cov_\lambda\left[\sigma_\theta/\mathbb{E}_\lambda\left[\sigma_\theta\right], d\log p_\theta\right].$$

In words, the change in aggregate productivity depends on the cross-sectional covariance of price elasticities, σ_θ, and price changes, $d\log p_\theta$. This is because the price elasticity controls the initial markup and the price change controls whether resources flow towards or away from each firm.

Of course, the change in prices in (8) is endogenous. The final step is to express these price changes in terms of primitives. The price charged by firm θ following the monetary shock depends on price stickiness (δ_θ) and desired pass-through (ρ_θ). In particular, the change in the price charged by firms of type θ is

$$d\log p_\theta = \delta_\theta \left[d\log p_\theta^{\text{flex}} \right] = \delta_\theta \left[\rho_\theta d\log w + (1-\rho_\theta) d\log P \right]. \tag{9}$$

The log-linearized optimal reset price is a weighted average of the change in marginal cost and the economy-wide price aggregator. High pass-through firms place a higher weight on marginal cost, while firms with low pass-through instead exhibit "pricing-to-market" behavior and place more weight on the price of competitors (summarized by the price aggregator).

The change in the price aggregator depends on the price changes of all firms. That is,

$$d\log P = \mathbb{E}_{\lambda\sigma}[d\log p_\theta] = \frac{\mathbb{E}_\lambda[\delta_\theta\sigma_\theta\rho_\theta]}{\mathbb{E}_\lambda[\delta_\theta\sigma_\theta\rho_\theta] + \mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]} d\log w, \tag{10}$$

where the second equality uses (9). Let $\kappa \in [0,1]$ denote the elasticity of P with respect to w. Combining (8), (9), and (10) yields an expression for the change in aggregate productivity in terms of primitives:

$$d\log A = (\bar{\mu}/\mathbb{E}_\lambda[\sigma_\theta]) \left(\kappa Cov_\lambda \left[\sigma_\theta, \delta_\theta \right] + (1-\kappa)\, Cov_\lambda \left[\sigma_\theta, \delta_\theta\rho_\theta \right] \right) d\log w. \tag{11}$$

In words, response of productivity to monetary shocks depends on the cross-sectional covariance of price elasticities σ_θ, which control the initial markups, with δ_θ and ρ_θ, which control the change in prices.

Note that the productivity response is zero when prices are either fully flexible or fully rigid. When prices are fully rigid, $\kappa = \delta_\theta = 0$, relative prices cannot change and there are no reallocations due to monetary shocks. When prices are fully flexible, $\kappa = \delta_\theta = 1$, there is complete pass-through of marginal cost shocks into prices in general equilibrium despite the fact that, in partial equilibrium, pass-through is incomplete. Hence, when prices are fully flexible, monetary shocks do not change relative prices or the allocation of resources.

The covariance of price elasticities, σ_θ, and realized pass-throughs, $\delta_\theta\rho_\theta$, can be decomposed into two terms

$$Cov_\lambda\left[\sigma_\theta, \delta_\theta\rho_\theta\right] = \mathbb{E}_\lambda\left[\rho_\theta|\text{flex}\right] Cov_\lambda\left[\sigma_\theta, \delta_\theta\right] + \mathbb{E}_\lambda\left[\delta_\theta\right] Cov_\lambda\left[\sigma_\theta, \rho_\theta|\text{flex}\right], \tag{12}$$

where $\mathbb{E}_\lambda\left[\rho_\theta|\text{flex}\right]$ and $Cov_\lambda\left[\sigma_\theta, \rho_\theta|\text{flex}\right]$ are the average pass-through and the covariance of price elasticities and pass-throughs for firms conditional on having flexible prices.[27] Using (12) with (11) yields the main result of this section in Proposition 1.

Proposition 1 (TFP Response)**.** *Following a monetary shock, the response of aggregate TFP is*

$$d\log A = \Bigg(\underbrace{\kappa_\rho Cov_\lambda\left[\sigma_\theta, \rho_\theta|\text{flex}\right]}_{\substack{\textit{Reallocation due to}\\ \textit{heterogeneous}\\ \textit{pass-through}}} + \underbrace{\kappa_\delta Cov_\lambda\left[\sigma_\theta, \delta_\theta\right]}_{\substack{\textit{Reallocation due to}\\ \textit{heterogeneous}\\ \textit{price stickiness}}} \Bigg) d\log w, \tag{13}$$

and κ_ρ and κ_δ are non-negative constants

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

To build more intuition, we consider the two covariance terms in (13) in isolation.

Mechanism I: heterogeneous desired pass-through. If price stickiness is homogeneous across firms ($\delta_\theta = \delta$), then Proposition 1 simplifies to the following.

Corollary 1 (Heterogeneous Pass-Through)**.** *If price stickiness is homogeneous across firms ($\delta_\theta = \delta$), then*

$$d\log A = \kappa_\rho Cov_\lambda\left[\sigma_\theta, \rho_\theta\right] d\log w, \quad (\kappa_\rho \geq 0).$$

Table 1 illustrates why a positive covariance between price elasticities and pass-throughs leads to an increase in aggregate TFP following an expansionary shock ($d\log w > 0$). Firms with high pass-throughs increase their prices by more than firms with low pass-throughs. When price elasticities positively covary with pass-throughs, firms predominantly lie on the bolded, diagonal axis in Table 1: the relative price of firms with initially high markups fall relative to other firms, reallocating resources towards those firms. By Lemma 2, this boosts aggregate productivity.

In principle, markups may covary with desired pass-throughs for many reasons. One of the most salient is that both markups and pass-throughs vary with firm size. This is formalized below.

[27] That is, $\mathbb{E}_\lambda\left[\rho_\theta|\text{flex}\right] = \mathbb{E}_{\lambda\delta}\left[\rho_\theta\right]$ and $Cov_\lambda\left[\sigma_\theta, \rho_\theta|\text{flex}\right] = Cov_{\lambda\delta}\left[\sigma_\theta, \rho_\theta\right]$.

Table 1: Reallocations due to covariance of desired pass-through ρ and price elasticity σ, in response to an expansionary shock

Price elasticity σ	Low pass-through (ρ)	High pass-through (ρ)
Low σ (high markup)	**Price/markup falls relative to other firms.**	Price/markup rises relative to other firms.
High σ (low markup)	Price/markup falls relative to other firms.	**Price/markup rises relative to other firms.**

Definition 1. *Marshall's third law of demand* states that desired markups are increasing in quantity and desired pass-throughs are decreasing in quantity.[28] That is,

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

If Marshall's third law holds, and firms face the same residual demand curve, then a monetary expansion will raise aggregate productivity. This is because large firms will have higher markups and lower pass-throughs. Marshall's third law of demand has strong empirical support (see, for example, empirical estimates of pass-throughs by firm size from Amiti et al., 2019) and holds in a variety of models. For example, oligopolistic competition models, such as Atkeson and Burstein (2008), satisfy Marshall's third law of demand.[29]

While Marshall's third law is sufficient to generate a positive covariance in Corollary 1, 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 and Auer et al., 2018).

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

Corollary 2 (Heterogeneous Price Rigidity)**.** *If desired pass-through is homogeneous across firms ($\rho_\theta = \rho$),*[30] *then*

$$d\log A = \kappa_\delta Cov_\lambda\left[\sigma_\theta, \delta_\theta\right] d\log w, \quad (\kappa_\delta \geq 0). \tag{14}$$

[28]Marshall's third law of demand is equivalent to requiring that the marginal revenue curve be log-concave. See Melitz (2018), who calls this a stronger version of Marshall's second law, for more information. The name "third" law of demand was coined by Matsuyama and Ushchev (2022).

[29]In Appendix H, we show that our results can also be derived under such an oligopolistic framework.

[30]Homogeneous desired pass-throughs are generated when the Kimball aggregator takes the form, $\Phi(x) = -\text{Ei}(-Ax^{\rho-1})$ where $\text{Ei}(x) = \int_{-x}^{\infty} \frac{e^{-t}}{t} dt$ is the exponential integral function. CES is special case where pass-through is homogenous and equal to one.

Table 2 illustrates why a positive covariance between markups and price stickiness causes an expansionary shock ($d \log w > 0$) to increase TFP. In response to an increase in the nominal wage, firm types with more flexible prices will raise their prices relative to firms with less flexible prices. If high-markup firms are more sticky than low-markup firms, then firms predominantly lie on the highlighted diagonal axis in Table 2. This results in a reallocation of resources towards firms with initially high markups and away from firms with initially low markups, thereby improving allocative efficiency as per Lemma 2.

Table 2: Reallocations due to covariance of price stickiness δ and price elasticity σ, in response to an expansionary shock.

Price elasticity σ	More sticky firms (low δ)	More flex. firms (high δ)
Low σ (high markup)	**Price/markup falls relative to other firms.**	Price/markup rises relative to other firms.
High σ (low markup)	Price/markup falls relative to other firms.	**Price/markup rises relative to other firms.**

This mechanism has recently been analyzed by Meier and Reinelt (2020), who show that in a CES model with heterogeneous price stickiness, firms with more rigid prices endogenously set higher markups due to a precautionary motive. This generates the positive covariance between markups and price stickiness in Corollary 2.

Although we allow for the possibility that price stickiness vary systematically with firm type, we do not pursue this mechanism further and point interested readers to Meier and Reinelt (2020). When we quantify the model, we assume there is no variation in price stickiness and instead focus on heterogeneity in desired pass-through only. This is because whereas there is robust empirical support for Marshall's third law of demand, the covariance of price stickiness with markups is less well documented.[31]

4 Output Response and the Phillips Curve

In the previous section, we showed that aggregate TFP can respond to monetary shocks. In this section, we show how monetary shocks are transmitted to output, taking into account the endogenous response of aggregate productivity. We show that the change in output can be decomposed into three channels: (1) nominal rigidities (as in a CES economy with

[31]For example, see Goldberg and Hellerstein (2011), who find that larger firms, who presumably have higher markups, also have more flexible prices.

sticky prices), (2) real rigidities due to imperfect pass-through (which arise from strategic complementarities in pricing à la Kimball, 1995), and (3) the *misallocation channel*, which is due the endogenous response of aggregate TFP.

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 how real rigidities and the misallocation channel flatten the slope of the Phillips curve relative to the benchmark sticky-price 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 (Output Response)**.** *Following a shock to the nominal wage $d\log w$, the response of output is*

$$d\log Y = \underbrace{\frac{1}{1+\gamma\zeta} d\log A}_{\textit{Supply-side effect}} + \underbrace{\frac{\zeta}{1+\gamma\zeta} \mathbb{E}_{\lambda}\left[-d\log\mu_{\theta}\right],}_{\textit{Demand-side effect}} \tag{15}$$

where $d\log A$ is given by Proposition 1 and

$$\mathbb{E}_{\lambda}\left[-d\log\mu_{\theta}\right] = \left[\underbrace{\mathbb{E}_{\lambda}\left[1-\delta_{\theta}\right]}_{\textit{Nominal rigidities}} + \underbrace{\frac{\mathbb{E}_{\lambda}\left[\delta_{\theta}(1-\rho_{\theta})\right]\mathbb{E}_{\lambda}\left[\sigma_{\theta}(1-\delta_{\theta})\right]}{\mathbb{E}_{\lambda}\left[\left[\delta_{\theta}\rho_{\theta}+(1-\delta_{\theta})\right]\sigma_{\theta}\right]}}_{\textit{Real rigidities}}\right] d\log w. \tag{16}$$

Equation (15) 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 (16) 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 shock. As a result, markups fall for a fraction $\mathbb{E}_{\lambda}\left[1-\delta_{\theta}\right]$ of firms. This reduction in the markups of sticky-price firms boosts labor demand, employment, and ultimately output.

This sticky price effect in (16) is amplified 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, on the other hand, is concerned with the efficiency with which labor is used. Returning to (15), 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. We term this channel the *misallocation channel*.

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, as the Frisch elasticity of labor supply approaches infinity, the supply side effect becomes irrelevant for output. This is because reallocations to high-markup firms, which boost productivity, also have a negative effect on labor demand. When the Frisch is infinite, the positive reallocation benefits are exactly cancelled out by reductions in employment, which contracts due to the expansion of high-markup firms.

4.2 The Misallocation Channel and the Phillips Curve

We now construct the Phillips curve—the relationship between the output gap and inflation generated by a demand shock—in the model and show that the misallocation channel flattens its slope.[32]

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

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

The price and wage Phillips curves are presented in Proposition 3.

Proposition 3 (Wage and Price Phillips Curves)**.** *Let* $\frac{d \log A}{d \log w}$ *and* $\frac{d \log \mu_\theta}{d \log w}$ *denote the total derivatives of* $\log A$ *and* $\log \mu_\theta$ *with respect to the exogenous nominal wage* $\log w$*. The wage Phillips*

[32] In the data, this relationship between the output gap (or unemployment) and inflation is confounded by other shocks that affect output or prices independently. For example, Fratto and Uhlig (2014) show that wage and price markup shocks play an important role in inflation dynamics, thus affecting the empirical Phillips curves constructed from aggregate data. In the dynamic version of our model (Proposition 5), the misallocation channel appears as endogenous cost-push shocks that raise output and lower inflation. These cost-push shocks may show up as exogenous markup shocks when calibrating a model that does not take into account endogenous TFP movements.

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

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

The expressions for $\frac{d\log A}{d\log w}$ *and* $\mathbb{E}_{\lambda}[\frac{d\log\mu_{\theta}}{d\log w}]$ *are provided in Proposition 1 and Proposition 2.*

When $\frac{d\log A}{d\log w} > 0$, the misallocation channel reduces the slope of both the price and wage Phillips curves. We can further quantify the degree to which real rigidities and the misallocation channel each flatten the Phillips curve. To do so, we calculate the flattening of the Phillips curve due to real rigidities by dividing the slope of the Phillips curve with sticky prices alone by the slope of the Phillips curve with sticky prices and real rigidities. If this quantity is, say, 1.5, this means that incorporating real rigidities flattens the slope of the Phillips curve by 50%. Similarly, we calculate the flattening of the Phillips curve due to misallocation channel by dividing the slope of the Phillips curve with sticky prices and real rigidities by the slope of the Phillips curve that also accounts for changes in allocative efficiency.

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 (Flattening of the Phillips Curve)**.** *Suppose* $\delta_{\theta} = \delta$ *for all firms. The flattening of the price Phillips curve due to real rigidities, compared to nominal rigidities alone, is*

$$\textit{Flattening due to real rigidities} = 1 + \frac{\mathbb{E}_{\lambda}[\sigma_{\theta}]\,\mathbb{E}_{\lambda}[1-\rho_{\theta}]}{\delta Cov_{\lambda}[\rho_{\theta},\sigma_{\theta}] + \mathbb{E}_{\lambda}[\rho_{\theta}]\,\mathbb{E}_{\lambda}[\sigma_{\theta}]}. \tag{17}$$

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

$$\textit{Flattening due to the misallocation channel} = 1 + \frac{\bar{\mu}}{\zeta}\frac{\delta Cov_{\lambda}[\rho_{\theta},\sigma_{\theta}]}{\delta Cov_{\lambda}[\rho_{\theta},\sigma_{\theta}] + \mathbb{E}_{\lambda}[\sigma_{\theta}]}. \tag{18}$$

In Equation (17), we see that the flattening of the Phillips curve due to real rigidities increases as average pass-throughs fall (as in Kimball, 1995). The flattening due to real rigidities in (17) 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 Phillips curve due to the misallocation channel depends positively on the covariance of pass-throughs and elasticities ($Cov_{\lambda}[\rho_{\theta}, \sigma_{\theta}]$). The misallocation channel also flattens the Phillips curve more when the Frisch elasticity ζ is low, since the supply-side effect is stronger when labor is inelastically supplied. Finally, since the expansion of high-markup firms relative to low-markup firms occurs only for flexible-price firms, the misallocation channel is relatively more important when prices are more flexible.

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 $\Phi_{\theta}(x) = x^{\frac{\sigma-1}{\sigma}}$, where $\sigma > 1$ is a parameter. 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 and the misallocation channel is absent. Applying Proposition 3, the slope of the price Phillips curve is

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

This is the traditional New Keynesian Phillips Curve.[33] 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 $\Phi'_{\theta} = \Phi'$, and the same productivity ($A_{\theta} = 1$). 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 to be incomplete ($\rho < 1$).

Since markups are uniform, the cross-sectional allocation of resources across firms in the initial equilibrium is still efficient. Hence, as in the CES example, the misallocation channel is still absent. Unlike the CES case, however, incomplete desired pass-through

[33]See, 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.

implies that flexible-price firms will not fully adjust prices to reflect increases in marginal cost from a monetary shock. As noted by Kimball (1995), 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 $1/\rho$.

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 2, the change in aggregate TFP following a nominal shock is

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

where l_H and l_L are employment by H and L firms. Aggregate TFP increases if the growth in employment at high-markup firms outpaces the growth of employment at low-markup firms. For simplicity, again impose homogeneous price stickiness ($\delta_H = \delta_L = \delta$). Proposition 3 implies that the price Phillips curve is

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

This price Phillips curve is flatter than the CES economy if $\rho_H < \rho_L$, i.e., if high-markup firms have lower pass-throughs than low-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.

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 third 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 The New Keynesian Model with Misallocation

This section provides a dynamic model that generalizes the workhorse three-equation model presented in Galí (2015) to account for heterogeneous firms and endogenous aggregate productivity. The static model we used so far is a special case of the dynamic model where the discount factor is equal to zero (i.e., agents are myopic).

5.1 Four-Equation Dynamic Model

In the infinite-horizon model, households choose consumption and leisure to maximize discounted future utility,

$$\max_{\{Y_t, L_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t Z_t u(Y_t, L_t),$$

where the per-period utility function is as in Section 2, the discount factor is β, and Z_t is a discount factor shifter. We allow for the possibility that there may be unanticipated shocks to the discount factor, as in Krugman (1998).

Each firm sets its price to maximize discounted future profits, subject to a Calvo friction. Firm i's profit-maximization problem is

$$\max_{p_{i,t}} \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1}(1+r_{t+j})}(1-\delta_i)^k y_{i,t+k}(p_{i,t} - \frac{w_{t+k}}{A_i})\right], \tag{19}$$

where δ_i is the Calvo parameter and $y_{i,t+k}$ is the quantity firm i sells in period $t+k$ if it last set its price in period t.

The model is closed by the actions of the monetary authority, which we assume follow a Taylor rule,

$$i_t = \left(\frac{P_{t+1}^Y}{P_t^Y}\right)^{\phi_\pi} \left(\frac{Y_t}{\bar{Y}}\right)^{\phi_y} V_t,$$

where i_t is the nominal gross interest rate, $\bar{Y}$ is the steady-state level of output, ϕ_π and ϕ_y are policy parameters that indicate the weight the monetary authority places on inflation and the output gap. The interest rate shifter V_t allows for the possibility of unanticipated shocks to the monetary policy rule.

As in Galí (2015), we log-linearize all variables around the no-inflation steady state. Macroeconomic aggregates like output Y and aggregate productivity A are endogenous outcomes that depend on the path of shocks; for parsimony, we simply write log-linearized variables $d\log Y$ and $d\log A$ with the understanding that these endogenous variables are functions of the entire path of monetary and discount factor shocks.[34]

For expositional simplicity, we present a version with homogeneous price stickiness across firms. Our main result is Proposition 5, which characterizes the movement of aggregate variables up to a first-order approximation.

Proposition 5 (Dynamic Model)**.** *Consider an economy with monetary shocks $v_t = d\log V_t$ and discount factor shocks $\epsilon_t = d\log Z_{t+1} - d\log Z_t$. Log-deviations in endogenous variables in the presence of these shocks satisfy the following four-equation system:*

$$\begin{aligned}
d\log i_t &= \phi_\pi d\log \pi_t + \phi_y d\log Y_t + v_t, && \text{(Taylor rule)}\\
d\log Y_t &= d\log Y_{t+1} - \frac{1}{\gamma}(d\log i_t - d\log \pi_{t+1} + \epsilon_t), && \text{(Euler equation)}\\
d\log \pi_t &= \beta d\log \pi_{t+1} + \varphi \mathbb{E}_\lambda\left[\rho_\theta\right]\frac{1+\gamma\zeta}{\zeta} d\log Y_t - \alpha d\log A_t, && \text{(Phillips curve)}\\
d\log A_t &= \frac{1}{\kappa_A} d\log A_{t-1} + \frac{\beta}{\kappa_A} d\log A_{t+1} + \frac{\varphi}{\kappa_A}\frac{1+\gamma\zeta}{\zeta}\bar{\mu}\frac{Cov_\lambda[\rho_\theta,\sigma_\theta]}{\mathbb{E}_\lambda[\sigma_\theta]} d\log Y_t, && \text{(TFP)}
\end{aligned}$$

where $d\log \pi_t = d\log P_t^Y / P_{t-1}^Y$ is the inflation rate, and φ, α, and κ_A are constants given by $\varphi = \frac{\delta}{1-\delta}(1-\beta(1-\delta))$, $\alpha = \frac{\varphi}{\bar{\mu}}\left(\mathbb{E}_\lambda\left[\rho_\theta\right]\left(1+\frac{\bar{\mu}}{\zeta}\right)-1\right)$, and $\kappa_A = 1+\beta+\varphi\left[1+\frac{Cov_\lambda[\rho_\theta,\sigma_\theta]}{\mathbb{E}_\lambda[\sigma_\theta]}\left(1+\frac{\bar{\mu}}{\zeta}\right)\right]$.

Proposition 5 provides a tractable, four-equation system that can be used to simulate economies with realistic heterogeneity in markups and pass-throughs. In addition to standard parameter values, the model requires four objects from the firm distribution: the average sales-weighted elasticity $\mathbb{E}_\lambda\left[\sigma_\theta\right]$, the average sales-weighted pass-through $\mathbb{E}_\lambda\left[\rho_\theta\right]$, the covariance of elasticities and pass-throughs $Cov_\lambda\left[\sigma_\theta, \rho_\theta\right]$, and the aggregate

[34]Monetary and discount factor shocks are the only shocks that we include in the model. Since both monetary and discount factor shocks show up as disturbances in the Euler equation, they will have similar effects on economic aggregates (as will any shock that shows up solely as a disturbance in the Euler equation). Of course, one could enrich our framework with other sources of exogenous shocks, such as government spending shocks, price- and wage-markup shocks, and productivity shocks (see, e.g., Smets and Wouters 2007, Fratto and Uhlig 2014), which will in general not be isomorphic to monetary and discount factor shocks.

markup $\bar{\mu}$.

Whereas the Taylor rule and Euler equation are the same as in the three-equation model, the last two equations are different. Start by considering the amended Phillips curve. We note two key differences: first, in the standard New Keynesian Phillips Curve (NKPC), the coefficient on $d \log Y_t$ is $\varphi \frac{1+\gamma\zeta}{\zeta}$.[35] In Proposition 5, this coefficient is multiplied by the average pass-through $\mathbb{E}_\lambda[\rho_\theta]$. As in the static version of the model, imperfect pass-through moderates the response of prices to nominal shocks and hence flattens the NKPC. More importantly, changes in aggregate TFP enter the Phillips curve as endogenous, negative cost-push shocks, given by $\alpha d \log A_t$.[36] This means that procyclical movements in aggregate TFP further dampen the response of inflation to an expansionary shock.

The final equation in Proposition 5 pins down the path of aggregate TFP. When markups covary negatively with pass-throughs, output booms, $d \log Y_t > 0$, driven either by monetary shocks or discount factor shocks, are concomitant with improvements in aggregate productivity. Furthermore, unlike the standard New Keynesian model, which consists of only forward-looking terms, the movement of aggregate TFP depends on a backward-looking term. As a result, the augmented four-equation model may generate endogenous 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 3 (Static Model as Special Case)**.** *Suppose output, aggregate TFP, and the price level are in steady state at $t = 0$. 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 Proof Sketch

Before calibrating the model, we provide a high-level walk-through of the derivation for Proposition 5 to highlight the key intuitions; the detailed derivation is in Appendix A. The derivation of the Euler equation is standard, so we focus instead on the Phillips curve and the TFP equations. Start with the firm maximization problem described in Equation (19). The optimal reset price $p_{i,t}^{\text{flex}}$ for profit maximization satisfies

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1}(1+r_{t+j})}(1-\delta_i)^k y_{i,t+k}\left[\frac{dy_{i,t+k}}{dp_{i,t}}\frac{p_{i,t}^{\text{flex}}}{y_{i,t+k}}\frac{p_{i,t}^{\text{flex}}-\frac{w_{t+k}}{A_i}}{p_{i,t}^{\text{flex}}}+1\right]\right]=0. \tag{20}$$

[35]See, e.g., Galí (2015) with constant returns.

[36]We 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.

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 $\frac{1}{\prod_{j=0}^{k-1}(1+r_{t+j})} = \beta^k$.

With some manipulation, the log-linearization of Equation (20) yields,

$$d \log p_{i,t}^{\text{flex}} = [1 - \beta(1 - \delta_i)] \sum_{k=0}^{\infty} \beta^k (1 - \delta_i)^k [\rho_i d \log w_{t+k} + (1 - \rho_i) d \log P_{t+k}]. \tag{21}$$

When prices are fully flexible, this simplifies to the static optimality condition in (9). 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 (21) recursively, and for each firm type θ, as

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

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,

$$\mathbb{E}[d \log p_{\theta,t}] = \delta_\theta \mathbb{E}\left[d \log p_{\theta,t}^{\text{flex}}\right] + (1 - \delta_\theta) \mathbb{E}[d \log p_{\theta,t-1}].$$

Combining the above two equations, and assuming $\delta_\theta = \delta$ for all θ, the expected price of firm θ follows a second-order difference equation,

$$\begin{aligned} &\mathbb{E}[d \log p_{\theta,t} - d \log p_{\theta,t-1}] - \beta \mathbb{E}[d \log p_{\theta,t+1} - d \log p_{\theta,t}] \\ &\qquad = \varphi [-\mathbb{E}[d \log p_{\theta,t}] + \rho_\theta d \log w_t + (1 - \rho_\theta) d \log P_t], \end{aligned} \tag{22}$$

where $\varphi = \delta/(1 - \delta)(1 - \beta(1 - \delta))$. Since Equation (22) pins down type θ firms' average price over time, we can recover the movements of aggregate variables, such as the consumer price index, aggregate TFP, and output, by manipulating this expression and averaging over firm types.

For instance, by taking the sales-weighted expectation of both sides in Equation (22), we recover the movement of the consumer price index.[37]

$$d \log \pi_t - \beta d \log \pi_{t+1} = \varphi \left[\mathbb{E}_\lambda [\rho_\theta] (d \log w_t - d \log P_t) + (d \log P_t - d \log P_t^Y)\right]. \tag{23}$$

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

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 $d\log P_t^Y$ —can be re-expressed in more familiar terms using the following identities:

$$d\log P_t - d\log P_t^Y = \bar{\mu}^{-1}\, d\log A_t, \tag{24}$$

$$d\log P_t^Y - d\log w_t = \frac{1}{\zeta}\left[d\log A_t - (1+\gamma\zeta) d\log Y_t\right]. \tag{25}$$

Equation (24) can be derived by log-linearizing and rearranging the expression for the price aggregator in (2),[38] and (25) comes from rearranging (15) for the average change in markups. Substituting these identities into (23) yields the Phillips curve in Proposition 5.

Movements in TFP also come from rearranging (22). From (5), we have

$$d\log A_t = d\log\bar{\mu} - \mathbb{E}_\lambda\left[d\log\mu_\theta\right] = \bar{\mu}\left(\mathbb{E}_{\lambda\sigma}\left[d\log\mu_{\theta,t}\right] - \mathbb{E}_\lambda\left[d\log\mu_{\theta,t}\right]\right). \tag{26}$$

The changes in markups can in turn be derived from (22) 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 (26).

6 Quantitative Results

We now calibrate the model to assess the quantitative importance of the misallocation channel. This section is organized as follows. Section 6.1 describes how to calibrate the model without relying on an off-the-shelf functional form for preferences. Section 6.2 calibrates the model using empirical pass-through estimates from Amiti et al. (2019) with Belgian firm-level data. Section 6.3 reports results from the static model, and Section 6.4 presents impulse response functions from the dynamic model. Finally, Section 6.5 shows that similar aggregate responses result in a model where nominal rigidities take the form of menu costs instead of Calvo frictions, though with some differences in the underlying patterns of price adjustment.

[38] Using the fact that $d\log P = \mathbb{E}_{\lambda\sigma}[d\log p_\theta]$, we get $\bar{\mu}(d\log P - d\log P^Y) = \bar{\mu}\left(\mathbb{E}_{\lambda\sigma}[d\log p_\theta] - \mathbb{E}_\lambda[d\log p_\theta]\right) = d\log A$ from (26).

6.1 Non-parametric Calibration Procedure

It may be tempting to use an off-the-shelf functional form for Φ 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.[39] Instead, we follow Baqaee et al. (2021) and back out the shape of the Kimball aggregator non-parametrically from the data. We summarize this approach below.

Assume that Φ_θ take the form

$$\Phi_\theta\left(\frac{y_\theta}{Y}\right) = \Phi\left(B_\theta \frac{y_\theta}{Y}\right).$$

Hence, firms differ in their productivities A_θ and taste shifters B_θ. Allowing for taste shifters is important since, in practice, two firms that charge the same price in the data can have very different sales and taste shifters allow us to accommodate this possibility.

We order firms by their size and let $\theta \in [0,1]$ be firm θ's quantile in the size distribution. Baqaee et al. (2021) show that, in the cross-section, markups and sales must satisfy the following differential equation[40]

$$\frac{d\log\mu_\theta}{d\theta} = (\mu_\theta - 1)\frac{1-\rho_\theta}{\rho_\theta}\frac{d\log\lambda_\theta}{d\theta}. \tag{27}$$

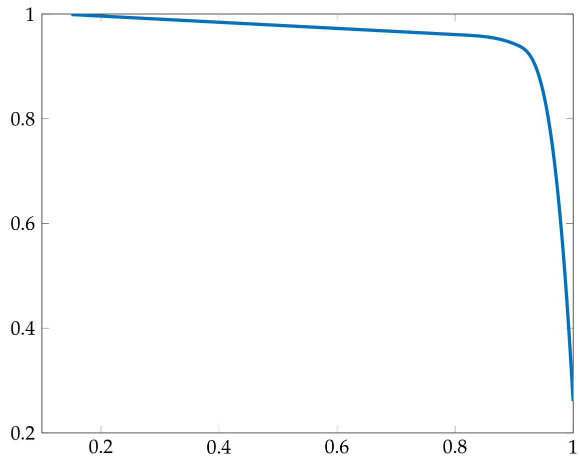
Given data on sales shares λ_θ and pass-throughs ρ_θ, we can use this differential equation to solve for markups μ_θ up to a boundary condition. We choose the boundary condition to target a given value of the (harmonic) sales-weighted average markup, $\bar{\mu}$. We then use $\sigma_\theta = 1/(1-1/\mu_\theta)$ 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.[41]

[39] As 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.

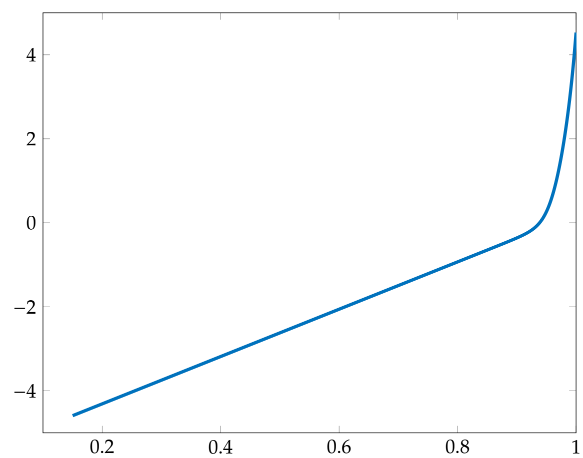
[40] This follows from combining the following two differential equations: $\frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log(A_\theta B_\theta)}{d\theta}$, and $\frac{d\log\mu_\theta}{d\theta} = (1-\rho_\theta)\frac{d\log(A_\theta B_\theta)}{d\theta}$. The first differential equation uses the fact that the firm of type $\theta + d\theta$ will have lower "taste-adjusted" price, $\log p_{\theta+d\theta} - \log p_\theta = \rho_\theta d\log(A_\theta B_\theta)/d\theta$, and higher sales $d\log\lambda_{\theta+d\theta} - \log\lambda_\theta = (\sigma_\theta - 1)\rho_\theta d\log(A_\theta B_\theta)/d\theta$, with $\sigma_\theta - 1 = 1/(\mu_\theta - 1)$. The second differential equation uses the fact that the relationship of desired markups to productivity is $d\log\mu_\theta/d\log(A_\theta B_\theta) = 1-\rho_\theta$.

[41] Our calibration imposes that markups and pass-throughs vary only as a function of market share. In Appendix I, we characterize how arbitrary noise in markups and pass-throughs unrelated to firm size affects the strength of the TFP response. We show that noise that moves markups and pass-throughs in the same direction will result in a stronger negative correlation between markups and pass-throughs and thus magnify the TFP response.

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



(a) Pass-through ρ_θ



(b) Log sales share density $\log \lambda_\theta$

6.2 Data and Parameter Values

We follow the implementation in Baqaee et al. (2021) (and refer interested readers to Appendix A of that paper for details). To calibrate the model, we need data on pass-throughs ρ_θ and the sales density λ_θ. For pass-throughs, we use estimates of (partial equilibrium) pass-throughs by firm size for manufacturing firms in Belgium from Amiti et al. (2019).[42] We interpolate between their point estimates with smooth splines and assume that pass-throughs go to 1 for the smallest firms (they find that the average pass-through for the smallest 75% of firms is already 0.97). Figure 2 shows the pass-through ρ_θ and log sales share density $\log \lambda_\theta$ as a function of θ. Pass-throughs are strictly decreasing with firm size, which means that Marshall's third law holds.

To compute the distribution of markups and elasticities from this data using equation (27), we must take a stance on the average markup. We assume that the average markup $\bar{\mu} = \mathbb{E}_\lambda \left[\mu_\theta^{-1}\right]^{-1} = 1.15$, in line with estimates from micro-data.[43]

[42] Amiti et al. (2019) use exchange rate shocks as instruments for changes in marginal cost and control for changes in competitors' prices. This identifies the partial equilibrium pass-through by firm size under assumptions consistent with our model. Note that standard exchange rate pass-through regressions that do not control for competitors' prices measure a general equilibrium object that is not the same as firms' partial equilibrium desired pass-through. See Proposition 3 in Amiti et al. (2014) for more detail.

[43] The resulting markup function μ_θ is shown in Figure G.1 in Appendix G. The markup distribution we recover is consistent with direct estimates from the literature. Konings 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 the working paper version of Amiti et al. (2019), they report that small firms in their calibration have a markup of around 14%, and large firms have markups of around 30%. These micro-estimated average markups are also broadly in line with macro estimates from Gutiérrez and

To calibrate the rest of the model, we use standard values from the literature. We 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) and set the intertemporal elasticity of substitution $\gamma = 1$. We consider a time period of one quarter, and set the Calvo parameter $\delta_\theta = \delta = 0.5$ according to an average price duration of about six months (Nakamura and Steinsson, 2008). We specify the coefficients on the Taylor rule, ϕ_π and ϕ_y, to match the calibration in Galí (2015). For the dynamic model, we set the discount factor $\beta = 0.99$, corresponding to a 4% annual interest rate. We assume that monetary disturbances follow an AR(1) process $v_t = \rho_v v_{t-1} + \epsilon_t$, 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. These parameter values are listed in Table 3.

Table 3: Calibrated parameter values for the static and dynamic versions of the model.

Static model			*Additional parameters for dynamic model*		
Parameter	Description	Value	Parameter	Description	Value
$\bar{\mu}$	Aggregate markup	1.15	ϕ_y	Output gap coefficient	0.5 / 4
$1/\gamma$	IES	1	ϕ_π	Inflation coefficient	1.5
ζ	Frisch elasticity	0.2	β	Discount factor	0.99
δ	Calvo friction	0.5	ρ_v	Shock persistence	0.7

6.3 Results from Static Model

Table 4 reports the estimated flattening of the Phillips curve due to real rigidities and the misallocation channel in the static model (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%.

Table 4: Flattening of the Phillips curve 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

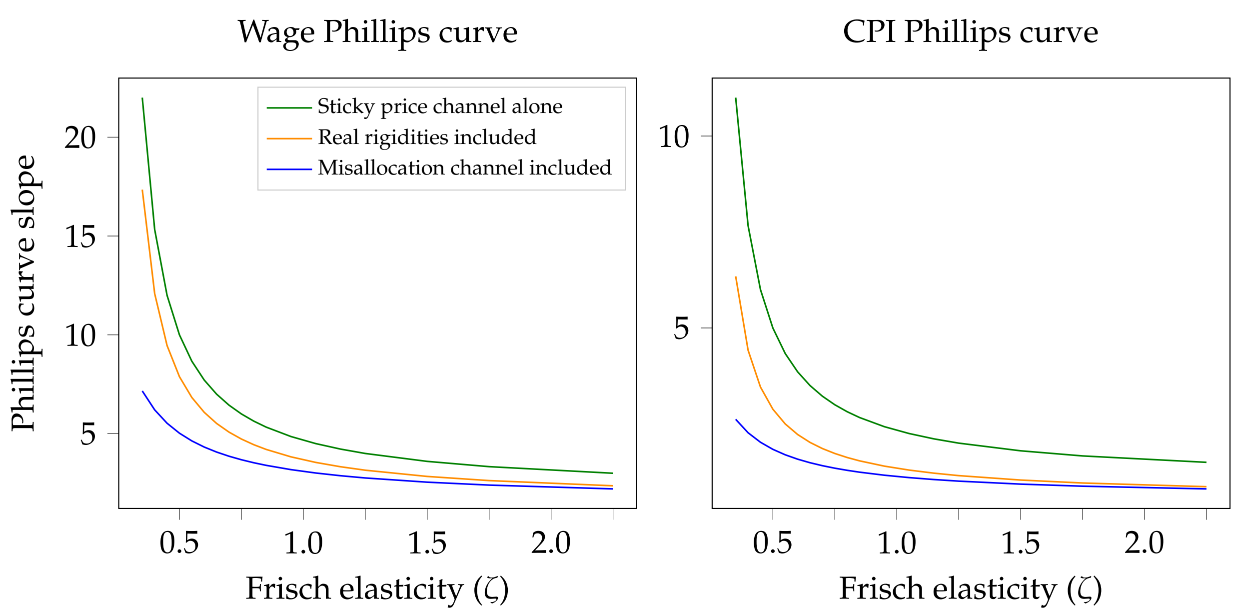
Philippon (2017) and Barkai (2020), who estimate average markups on the order of 10-20%. Edmond et al. (2018) also choose $\bar{\mu} = 1.15$.

To highlight the key forces at play in this calibration, we consider how these estimates change as we vary the Frisch elasticity and the degree of industrial concentration.[44]

The Frisch elasticity. The discussion following Proposition 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 more monetary non-neutrality at lower levels of the Frisch elasticity.

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



[44] Additional comparative statics with respect to the average markup and the price-stickiness parameter can be found in Appendix D.

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.[45] We 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.

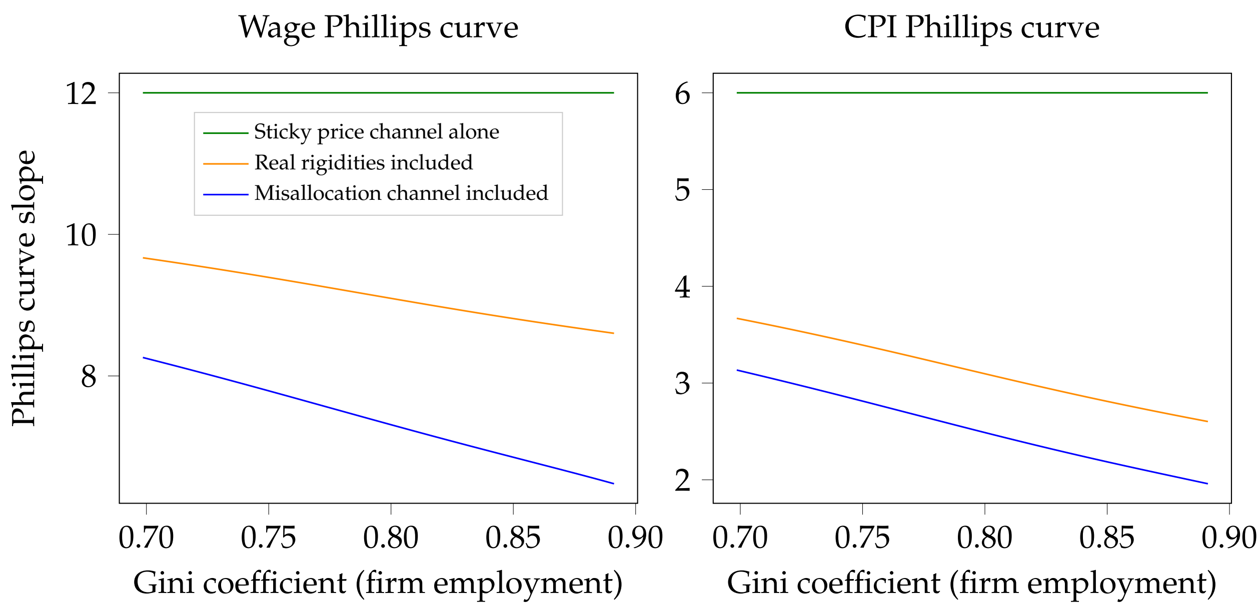
This exercise suggests that increasing the Gini coefficient from 0.80 to 0.85 flattens the price Phillips curve by an additional 14%. To put these numbers into context, such a change in the Gini coefficient is in line with the increase in the Gini coefficient in firm employment from 1978 to 2018 in the 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%.

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 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. As mentioned earlier, discount factor shocks are isomorphic to monetary shocks in our model, so the results can equally be taken to be the

[45]We choose the beta distribution since, as a bounded distribution, it allows us to remain within the 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.



response of the model to discount factor shocks.[46]

In the CES and homogeneous firms case, aggregate TFP does not react to the monetary shock, as implied by Lemma 2. In contrast, when firms have heterogeneous markups, aggregate TFP falls in response to the contractionary shock. The fall in aggregate TFP dampens the extent of disinflation caused by the monetary contraction and deepens the immediate response of output to the shock. The reduction in aggregate TFP coincides with an increase in the cross-sectional dispersion of firm-level TFPR since high-markup firms are raising their markups relative to low-markup firms.[47] The magnitude of the increase in TFPR dispersion is broadly consistent with Kehrig (2011), who finds that TFPR dispersion increases about 10% during a typical recession and increased over 20% from 2007 to the trough of the recession in 2009.

We quantify how the misallocation channel affects real output in Table 5. The contraction in output in the full model is about 45% deeper on impact than in the homogeneous firm model. The persistence of the shock's effect on real output also increases: while the

[46]To see that discount factor shocks and monetary shocks enter the four-equation system identically, combine the Taylor rule and the Euler equation in Proposition 5: $\gamma\left(d\log Y_{t+1} - d\log Y_t\right) = \phi_\pi d\log\pi_t + \phi_y d\log Y_t - d\log\pi_{t+1} + (v_t + \epsilon_t)$.

[47]Under constant returns to scale, like our model, changes in TFPR are equal to changes in firm markups: $\Delta\log\text{TFPR} = \Delta\log p_\theta y_\theta - \Delta\log l_\theta = \Delta\log\mu_\theta$. (See Foster et al. (2008) for a discussion of the relationship between TFPR and physical productivity A_θ.) Meier and Reinelt (2020) also provide corroborating evidence that markup dispersion rises following monetary contractions.

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.

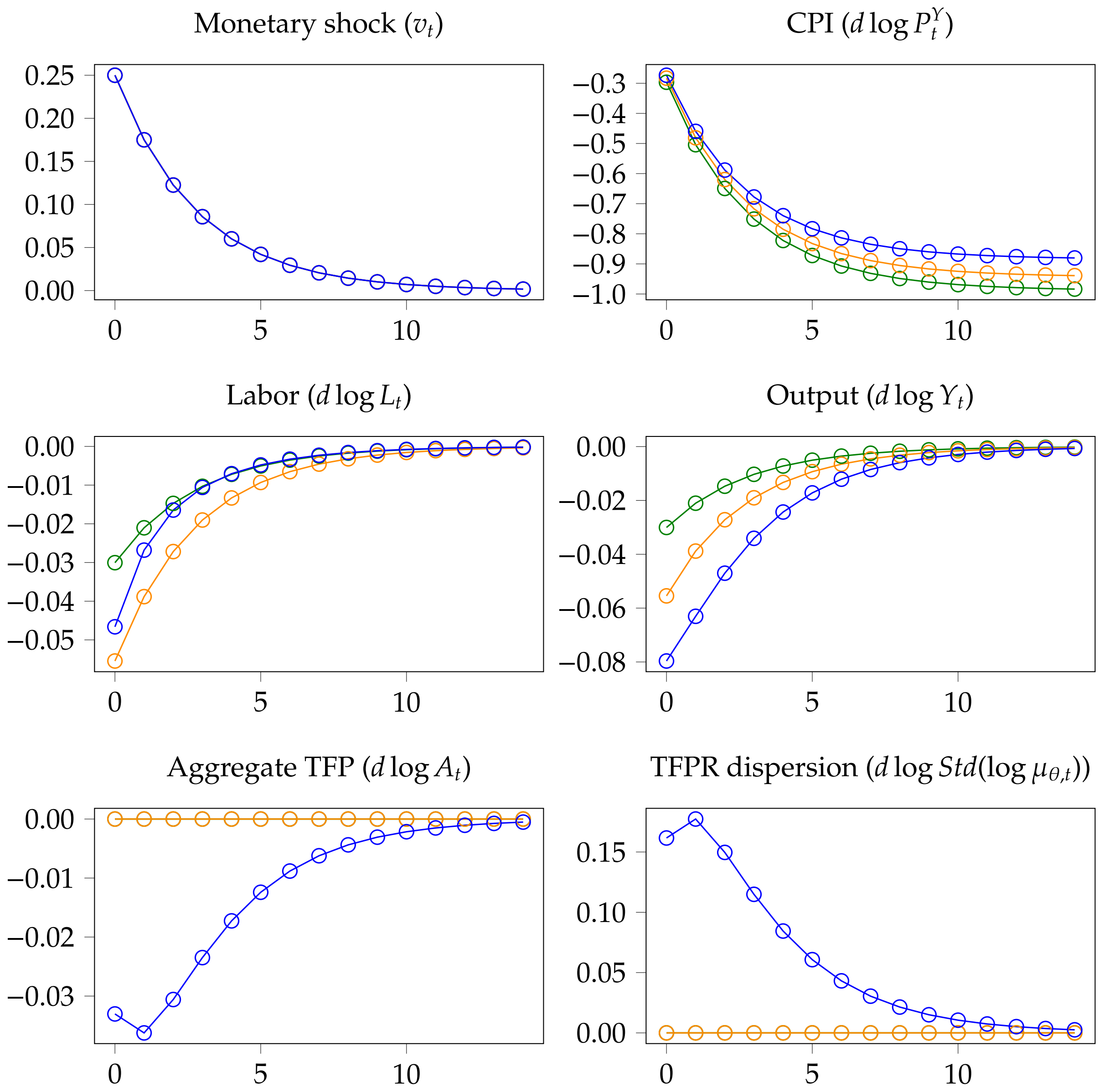
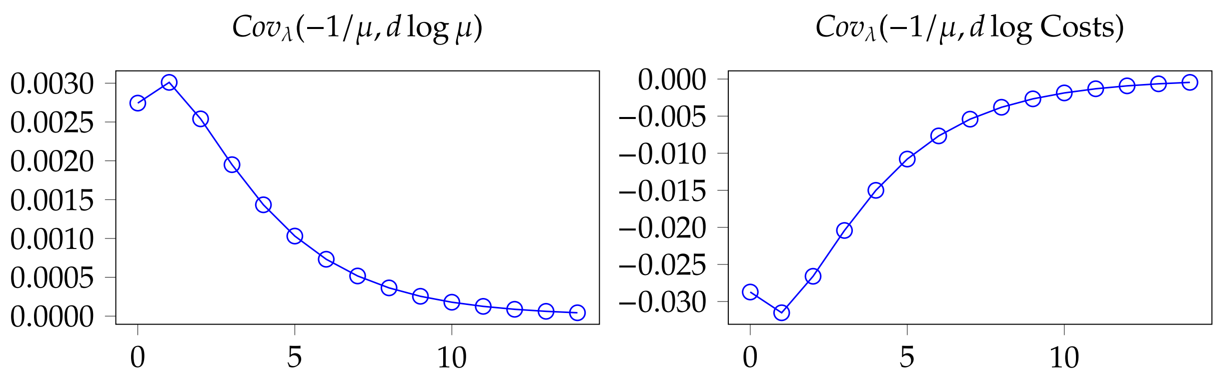


Figure 6: Covariance of firms' inverse markups with changes in markups and costs following a 25bp monetary shock. The contractionary shock leads high-markup firms to increase their markups relative to low-markup firms (left), causing a reallocation of resources away from high-markup firms (right).



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 about 30% to about 2.6 quarters.[48] In full, the misallocation channel increases the cumulative impact on output of the monetary shock by around 70%.

Table 5: Effect of monetary policy shock on output.

Model	Output effect at $t = 0$	Half life	Cumulative output impact
CES	-0.030	1.95	-0.10
Homogeneous Firms	-0.055	1.95	-0.18
Heterogeneous Firms	-0.080	2.56	-0.31

Figure 6 shows the covariance between firms' inverse markups and their change in markups (left) and change in total input costs (right). Following Lemma 2, the contractionary monetary shock reallocates inputs to low-markup firms, generating the fall in TFP. This is a directly testable prediction of the model that we return to in Section 7.

We provide additional calibration results in Appendix D. In particular, we report the change in sales shares for firms at different percentiles of the size distribution. The sales shares of small firms are about as volatile as aggregate output, whereas the sales shares of the largest firms are less volatile. 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

[48]Due to the second-order difference equation in aggregate TFP, the full model no longer features a constant half-life. We report the half-life at period zero.

Table 6: Extensive and intensive margins of price adjustment in Calvo and menu cost model for one year after money supply shock.

	Calvo model		*Menu cost model*	
Quintile of initial size	Share of firms with price change	Average size of price change	Share of firms with price change	Average size of price change
1	0.938	0.0359	0.921	0.0374
2	0.938	0.0358	0.841	0.0408
3	0.938	0.0357	0.766	0.0446
4	0.938	0.0356	0.719	0.0458
5	0.938	0.0345	0.676	0.0495

Note: Response to a 4 basis point money supply shock in both models. The share of firms with price change reports the fraction of firms with at least one price change within one year of the initial shock. The average size of price change is the average magnitude of the first price change by firms in each quintile.

that the misallocation channel is as powerful as the real rigidities channel in affecting the transmission of monetary policy.

6.5 Menu Cost Calibration

The price rigidities we have explored so far take the form of Calvo frictions. A natural question is whether the effects we identify would also arise under a different model of nominal rigidities. In Appendix C, we nonlinearly solve and provide impulse response functions for a quantitative model with menu costs instead.[49] We first calibrate the model under CES preferences, and then replace those preferences with the Kimball demand system estimated in the Belgian data. In response to a money supply shock, the Kimball calibration generates a procyclical TFP response that increases the effect of the shock on output. Similar to our baseline results, roughly half of the movement of output on impact is due to the supply-side effect. Accordingly, the response of output on impact is more than twice as large in the Kimball calibration relative to the CES calibration.

As in the Calvo model, aggregate TFP rises in response to monetary expansions because high-markup firms have lower realized pass-through than low-markup firms. However,

[49] In the menu cost calibration in Appendix C, we also include idiosyncratic productivity shocks, resulting in large, frequent, and symmetric price changes, which matches the facts documented by Bils and Klenow (2004). Our Calvo model does not match these micro pricing facts. However, this could be remedied by adding idiosyncratic demand shocks. These demand shocks could generate large, frequent, and symmetric price changes. Such demand shocks generate price changes in our Kimball model but they do not in a CES model because in our model the desired markup is not the same at every point of the residual demand curve. The addition of such idiosyncratic demand shocks would have no aggregate implications in the Calvo model, but would allow us to match micro pricing facts better.

unlike the Calvo model, the differences in realized pass-throughs comes from the extensive rather than the intensive margin of price changes. Table 6 shows the intensive and extensive margins of price adjustment for firms in the Calvo and menu cost models in response to a similar-sized money supply shock.[50] In the Calvo model, we assume that all firms have the same degree of price stickiness δ_θ, so that all differences in realized pass-through come from intensive margin differences in the degree to which firms adjust their prices. On the other hand, in the menu cost model, high-markup firms endogenously choose to keep their prices unchanged for longer due to lower desired pass-through. As a result, large firms are less likely to change their prices in the first year after the shock. However, conditional on changing their price, large firms make slightly larger adjustments. This is because of a selection effect where large firms that choose to adjust their prices are those that have been buffeted by large idiosyncratic shocks. Lower realized pass-through of large firms—due to differences in the extensive margin of price adjustment in the menu cost model—generates the misallocation channel.[51]

Berger and Vavra (2019) find a positive correlation between (reduced-form, general equilibrium) exchange rate pass-through and dispersion in price changes in the time series. They attribute this to the intensive margin—variation in pass-through conditional on a price change—rather than the extensive margin—variation in the frequency of price adjustment. Our baseline Calvo model is able to better match this pattern in the sense that increases in dispersion of price changes are due to differences in desired pass-through across firms, rather than variations in the probability of price adjustment caused by a monetary shock.

7 Empirical Evidence

In this section, we provide empirical evidence in support of the reallocation mechanism described in this paper. We first present macro-level evidence on the response of aggregate TFP to identified monetary shocks. We then show, at the micro-level, that contractionary monetary shocks lead high-markup firms to increase their markups relative

[50] Appendix E provides impulse responses of the Calvo model to a money supply shock, and Appendix C provides impulse responses for the menu cost model.

[51] The fact that large firms make slightly larger price adjustments conditional on a price change is not inconsistent with the evidence from Amiti et al. (2019). For idiosyncratic shocks that overcome the menu cost, a large firm in our calibration makes a smaller price adjustment than a small firm. However, this pattern flips in our calibration for aggregate monetary shocks because these shocks are small relative to the idiosyncratic shocks hitting firms. Therefore, large firms that have lower desired pass-through change their prices only if they are being buffeted by large idiosyncratic shocks. This is why the pass-through conditional on a price change for monetary shocks is higher for large firms in Table 6.

to low-markup firms, leading to a deleterious reallocation of inputs across firms. Finally, we provide evidence that the contraction in productivity following monetary tightening is greater in more concentrated industries, as in Figure 4.

Macro-level evidence. To see the response of aggregate TFP and output to identified monetary shocks, we compute local projections à la Jordà (2005) using the specification,

$$Y_{t+h} = a + \sum_{k=0}^{4} b_k^h \cdot \text{MonetaryShock}_{t-k} + \sum_{k=1}^{4} c_k^h \cdot Y_{t-k} + \epsilon_t,$$

where Y_t is the aggregate outcome of interest and MonetaryShock_t are exogenous monetary shocks.

For monetary shocks, we use an extended version of the Romer and Romer (2004) monetary shock series constructed by Wieland and Yang (2020) for 1969–2007. We use three different measures of aggregate productivity—labor productivity, the Solow residual, and the cost-based Solow residual (see Hall, 1990).[52] We do not use utilization-adjusted TFP (e.g. Basu et al., 2006; Fernald, 2014). 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.

Figure 7 plots the estimated coefficients b_0^h for horizons up to sixteen quarters. Following a contractionary shock, there is a significant contraction in aggregate productivity and output. The magnitude of the decline in aggregate productivity is more than half of the effect on output. This movement in aggregate productivity relative to output is moderately larger than that predicted by our model, which suggests that allocative effects explain part but perhaps not all of the aggregate productivity response.[53,54]

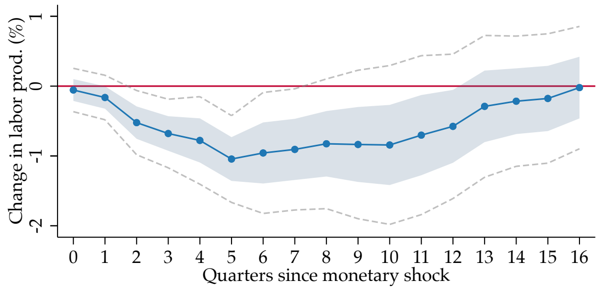
Micro-level evidence. In our model, aggregate TFP responds to monetary shocks due to systematic reallocations among firms with different markups. We now turn to micro-level evidence on these reallocations. To do so, we use estimates of markups for publicly

[52] We 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.

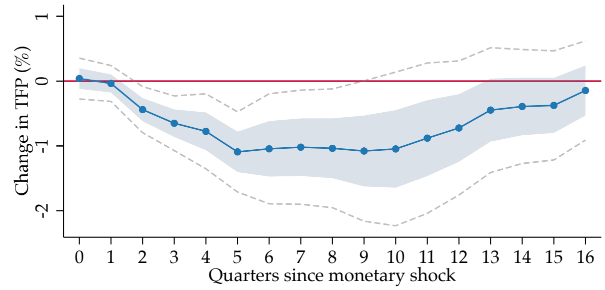
[53] In Appendix B.2, we also show that labor productivity, the Solow residual, and the cost-based Solow residual are unconditionally procyclical over the period 1948–2020.

[54] The dynamic calibration in Section 6 predicts that a 1% change in output due to a monetary shock is accompanied by a 0.4% change in aggregate productivity. In Figure 7, our point estimates suggest that a 1% change in output due to a monetary shock is accompanied by a 0.7% change in aggregate productivity. So, the relative size of the productivity response in our model is roughly half of that in the data.

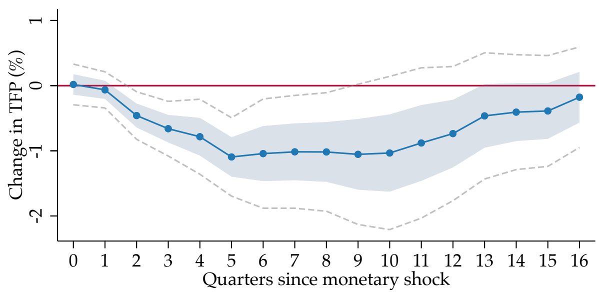
Figure 7: Local projection of a contractionary Romer and Romer (2004) shock (using extension by Wieland and Yang, 2020) on aggregate productivity and output.



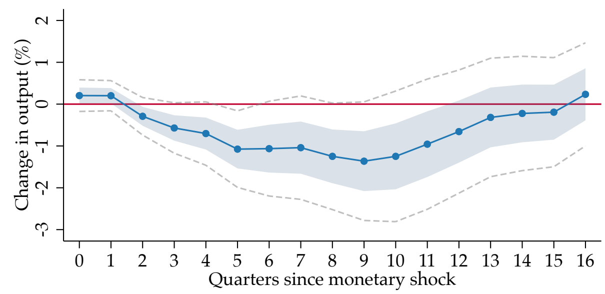
(a) Labor productivity



(b) Solow residual



(c) Cost-based Solow residual



(d) Output

Notes: The shaded region indicates Newey-West standard errors. Dashed lines are 95% confidence intervals. Sample covers 1969–2007.

traded firms in Compustat. Of course, this exercise must be interpreted with caution since measuring markups accurately at high frequency is challenging and Compustat is not a representative sample of all US producers. Nevertheless, our empirical results are supportive of the basic mechanism underlying the misallocation channel.

We study the response of firm-level markup changes and input reallocations across firms to identified exogenous monetary shocks.[55] For our baseline estimate of firm markups, we follow the user-cost approach of Gutiérrez and Philippon (2017) and Gutiérrez (2017). That is, we estimate each firm's capital stock and user-cost of capital. To estimate the user-cost of capital, we use industry-specific depreciation rates and industry-level risk premia. We estimate profits by subtracting total estimated costs from total revenues, and we back out the markup by assuming firms have constant returns to scale. Appendix B describes the data sources and assumptions underlying our markup estimation procedure

[55]In the body of the paper, we focus only on responses conditional on identified monetary shocks. Figure B.1 in Appendix B shows that, unconditionally, high-markup firms are more procyclical than low-markup firms in Compustat. This is consistent with a view that recessions are primarily demand-driven and that the misallocation channel is active.

in more detail.

We then estimate the following local projections:

$$Cov_\lambda(-1/\mu_t, \Delta \log \mu_{t\to t+h}) = a^h + \sum_{k=0}^{4} b_k^h \cdot \text{MonetaryShock}_{t-k} + \sum_{k=1}^{4} c_k^h \cdot Cov_\lambda(-1/\mu_t, \Delta \log \mu_{t-k\to t}) + \epsilon_t^h,$$

$$Cov_\lambda(-1/\mu_t, \Delta \log \text{Costs}_{t\to t+h}) = \tilde{a}^h + \sum_{k=0}^{4} \tilde{b}_k^h \cdot \text{MonetaryShock}_{t-k} + \sum_{k=1}^{4} \tilde{c}_k^h \cdot Cov_\lambda(-1/\mu_t, \Delta \log \text{Costs}_{t-k\to t}) + \epsilon_t^h,$$

where $Cov_\lambda(-1/\mu_t, \Delta \log \mu_{t\to t+h})$ is the sales-weighted covariance between inverse markups at time t and the change in markups from time t to time $t+h$, $Cov_\lambda(-1/\mu_t, \Delta \log \text{Costs}_{t\to t+h})$ is the sales-weighted covariance between inverse markups at t and the change in total costs, and MonetaryShock_t is the (extended) Romer and Romer (2004) shock in quarter t.[56] This is a direct test of the model, as in Lemma 2. Figure 6 shows that in our calibrated model, a contractionary shock leads to relative increases in the markups of high-markup firms ($Cov_\lambda(-1/\mu, \Delta \log \mu) > 0$) and a reallocation of resources toward low-markup firms ($Cov_\lambda(-1/\mu, \Delta \log \text{Costs}) < 0$).[57]

Figure 8 shows estimates of b_0^h and $\tilde{b}_0^h$ following a monetary shock. As the top left panel shows, a contractionary shock leads high-markup firms to increase their markups relative to low-markup firms; the result, in the top right panel, is a reallocation of resources away from high-markup firms and toward low-markup firms. In the bottom panels, we estimate a panel version of the above specifications across 3-digit NAICS industries with industry fixed effects.[58] Both the direction and magnitude of the impulse responses are similar, suggesting that within-sector reallocations play an important role.

In terms of magnitudes, we find that the ratio of $Cov_\lambda(-1/\mu, \Delta \log \mu)$ to the response of output is similar in the model and in the data. However, the resulting covariance of initial markups with the change in costs, $Cov_\lambda(-1/\mu, \Delta \log \text{Costs})$, is smaller in the Compustat data than predicted by the model. One reason for the difference could be that Compustat is a subsample of very large firms. In particular, since public firms tend to be much larger than the average firm, the demand elasticities of the firms in our sample are likely to be lower than the average, resulting in less reallocation given changes in markups.

In Appendix B, we show that our results are robust to using firm accounting profits to measure markups (Figure B.2) and to including intangible capital when estimating user-

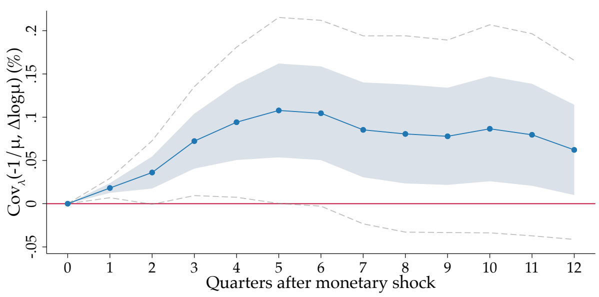
[56] We measure these covariances for firms that report earnings in both quarter t and $t+h$. Sales in quarter t are used to weight the covariances.

[57] Our results are unlikely to be driven by procyclicality of capital intensive firms since our estimate of profits, and hence markups, do not include capital costs. At any rate, Jaimovich et al. (2019) provide evidence that cyclicality is negatively correlated with capital intensity among firms in our sample.

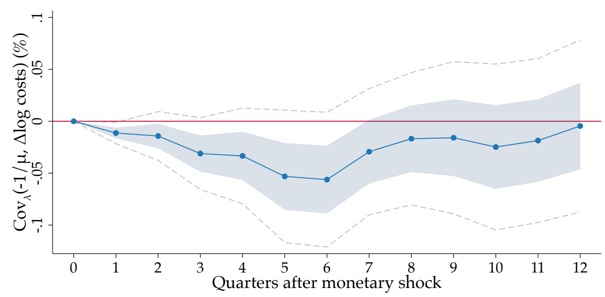
[58] See Appendix B for the estimating equations for the industry-level specifications.

cost markups (Figure B.3). Our results are also robust to using monetary shocks identified using high-frequency methods by Gorodnichenko and Weber (2016) (Figure B.7).

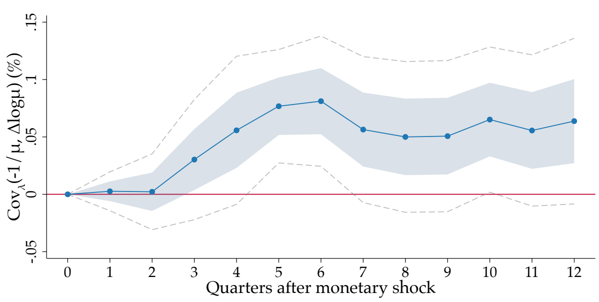
Figure 8: Local projection of contractionary Romer and Romer (2004) shock (using extension by Wieland and Yang, 2020) on $Cov_{\lambda}(-1/\mu, \Delta \log \mu)$ and $Cov_{\lambda}(-1/\mu, \Delta \log \text{Costs})$.



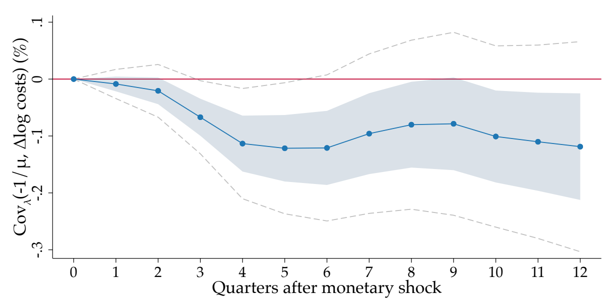
(a) $Cov_{\lambda}(-1/\mu, \Delta \log \mu)$



(b) $Cov_{\lambda}(-1/\mu, \Delta \log \text{Costs})$



(c) $Cov_{\lambda}(-1/\mu, \Delta \log \mu)$ within NAICS-3



(d) $Cov_{\lambda}(-1/\mu, \Delta \log \text{Costs})$ within NAICS-3

Notes: The shaded region indicates Newey-West standard errors in panels (a)-(b) and Driscoll-Kraay standard errors in panels (c)-(d). Dashed lines are 95% confidence intervals.

Cross-sector evidence. Figure 4 suggests that industrial concentration may play a role in how productivity responds to monetary shocks. All things being equal, higher industrial concentration is likely to be accompanied by greater heterogeneity in pass-through and hence a greater response of productivity to monetary shocks.

To see whether this prediction is borne out in the data, we use annual estimates of multifactor productivity across 4-digit NAICS manufacturing industries from the Bureau of Labor Statistics and data on the concentration of sales from the Economic Census of Manufacturing. We estimate the following local projection:

$$\Delta \log \text{TFP}_{i,t} = \beta \left(\text{Concentration}_i \times \text{MonetaryShock}_t \right) + \sum_{k=1}^{2} \gamma_k^p \log \text{TFP}_{i,t-k} + \delta_i + \alpha_t + \epsilon_{i,t},$$

where i is the 4-digit NAICS industry, t is the year, δ_i are industry fixed effects, and α_t are year fixed effects.[59] The coefficient of interest is β, which indicates whether multifactor productivity in concentrated industries is differentially responsive to the monetary shock. Our calibration suggests that a contractionary monetary shock leads to a greater reduction in multifactor productivity in concentrated industries, and hence $\beta < 0$.

Table 7 shows the estimated coefficient β using three measures of industrial concentration—the sales share of the industry's top eight, twenty, and fifty firms in the 2002 Economic Census for Manufacturing—and using the extended Romer and Romer (2004) shock series. For all three measures, we observe that the estimated $\beta < 0$, which suggests that the productivity effects of a monetary shock are more pronounced in concentrated industries.

Table 7: Differential response of industry multifactor productivity to monetary shocks in concentrated manufacturing industries.

	$\Delta \log \text{MultifactorProductivity}_{i,t}$ (1)	(2)	(3)
Top 8 Firms Share$_i$ $\times$ MonetaryShock$_t$	-0.0185**		
	(0.00906)		
Top 20 Firms Share$_i$ $\times$ MonetaryShock$_t$		-0.0183**	
		(0.00762)	
Top 50 Firms Share$_i$ $\times$ MonetaryShock$_t$			-0.0176**
			(0.00699)
Industry FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes
N	1634	1634	1634

Notes: The sales shares of the top 8, 20, and 50 firms in each 4-digit NAICS industry are from the 2002 Economic Census for Manufacturing. Monetary shocks are from the extension of the Romer and Romer (2004) shock series by Wieland and Yang (2020). ** indicates significance at 5%.

In Appendix B, we show that these results are robust to using concentration data from the 2007 Economic Census (Table B.2) and to using monetary shocks identified from high frequency data by Gorodnichenko and Weber (2016) (Table B.3).

8 Extensions

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

[59] We do not include industry-level concentration or the monetary shock as regressors since these would be collinear with the industry-fixed effect and the time-fixed effect respectively.

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. It has only one sector and only one factor of production. This means that 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.[60] 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 markups and pass-throughs at the initial equilibrium only vary as a function of firm size. While 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, 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.

Oligopoly calibration. In the main text, we model a continuum of firms in monopolistic competition where the positive covariance between price elasticities and pass-throughs is due to the shape of the residual demand curve. An alternative micro-foundation for this covariance is an oligopoly model like the one in Atkeson and Burstein (2008). In Appendix H, we develop a static oligopoly version of our model and compute the flattening of the Phillips curve due to real rigidities and the misallocation channel. The results are qualitatively and quantitatively similar to the calibration in Section 6.

[60]For the importance of sectoral heterogeneity and intermediate inputs in monetary models, see recent papers by Rubbo (2020), Castro (2019), La'O and Tahbaz-Salehi (2022), and Pasten et al. (2020).

Klenow and Willis (2016) calibration. In the main text, 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 (see Figure 2), 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, standard calibrations of these preferences fail to capture the cross-sectional covariance between pass-throughs and markups and hence imply counterfactually small supply-side effects.

9 Conclusion

We analyze the transmission of aggregate demand shocks, like monetary policy shocks, 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 for output, we find that in this richer model aggregate demand 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 that aggregate TFP moves procyclically in response to monetary and fiscal shocks. We link these pieces of evidence and show how monetary shocks can generate both effects.

While we focus on heterogeneous markups in product markets, it is possible that similar distortions could exist in input markets. Specifically, if firms have heterogeneous and variable monopsony power in the labor market, then TFP would increase if firms with relatively high markdowns reduce their markdowns following an expansionary shock. Finally, our analysis is purely positive, and we leave the normative implications for optimal policy for future 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 (2014). Importers, exporters, and exchange rate disconnect. *American Economic Review 104*(7), 1942–78.

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.

Baqaee, D. R. and E. Farhi (2017). Productivity and misallocation in general equilibrium. Technical Report 24007, National Bureau of Economic Research.

Baqaee, D. R. and E. Farhi (2018). Macroeconomics with heterogeneous agents and input-output networks. Technical Report 24684, National Bureau of Economic Research.

Baqaee, D. R. and E. Farhi (2020). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics 135*(1), 105–163.

Baqaee, D. R., E. Farhi, and K. Sangani (2021). The darwinian returns to scale. Technical Report 27139, National Bureau of Economic Research.

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.

Benkard, C. L., A. Yurukoglu, and A. L. Zhang (2021). Concentration in product markets. Technical Report 28745, National Bureau of Economic Research.

Berger, D. and J. Vavra (2019). Shocks versus responsiveness: What drives time-varying dispersion? *Journal of Political Economy 127*(5), 2104–2142.

Bianchi, F., H. Kung, and G. Morales (2019). Growth, slowdowns, and recoveries. *Journey of Monetary Economics 101*, 47–63.

Bils, M. and P. J. Klenow (2004). Some evidence on the importance of sticky prices. *Journal of Political Economy 112*(5), 947–985.

Burstein, A., V. M. Carvalho, and B. Grassi (2020). Bottom-up markup fluctuations. Technical report, National Bureau of Economic Research.

Burstein, A. T. (2006). Inflation and output dynamics with state-dependent pricing decisions. *Journal of monetary Economics 53*(7), 1235–1257.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journey of Monetary Economics 12*(3), 383–398.

Castro, N. (2019). The importance of production networks and sectoral heterogeneity for monetary policy. *Available at SSRN 3499902*.

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.

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.

Del Negro, M., M. Lenza, G. E. Primiceri, and A. Tambalotti (2020). What's up with the phillips curve? Technical Report 27003, National Bureau of Economic Research.

Dotsey, M. and R. G. King (2005). Implications of state-dependent pricing for dynamic macroeconomic models. *Journey of Monetary Economics 52*(1), 213–242.

Edmond, C., V. Midrigan, and D. Y. Xu (2018). How costly are markups? Technical Report 24800, National Bureau of Economic Research.

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.

Fratto, C. and H. Uhlig (2014). Accounting for post-crisis inflation and employment: A retro analysis. Technical Report 20707, National Bureau of Economic Research.

Galí, J. (2015). *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press.

Goldberg, P. and R. Hellerstein (2011). How rigid are producer prices? Technical Report 407, Federal Reserve Bank of New York.

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.

Gorodnichenko, Y. and M. Weber (2016). Are sticky prices costly? evidence from the stock market. *American Economic Review 106*(1), 165–99.

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.

Hazell, J., J. Herreno, E. Nakamura, and J. Steinsson (2020). The slope of the phillips curve: Evidence from us states. Technical Report 28005, National Bureau of Economic Research.

Hooper, P., F. S. Mishkin, and A. Sufi (2020). Prospects for inflation in a high pressure economy: Is the phillips curve dead or is it just hibernating? *Research in Economics 74*(1), 26–62.

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.

Jaimovich, N., S. Rebelo, and A. Wong (2019). Trading down and the business cycle.

Journal of Monetary Economics 102, 96–121.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review 95*(1), 161–182.

Kehrig, M. (2011). The cyclicality 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.

Krugman, P. R. (1998). It's baaack: Japan's slump and the return of the liquidity trap. *Brookings papers on economic activity 1998*(2), 137–205.

La'O, J. and A. Tahbaz-Salehi (2022). Optimal monetary policy in production networks. *Econometrica 90*(3), 1295–1336.

Levy, D., M. Bergen, S. Dutta, and R. Venable (1997). The magnitude of menu costs: Direct evidence from large u.s. supermarket chains. *The Quarterly Journal of Economics 112*(3), 791–825.

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.

Matsuyama, K. and P. Ushchev (2022). Selection and sorting of heterogeneous firms through competitive pressures.

McLeay, M. and S. Tenreyro (2020). Optimal inflation and the identification of the phillips curve. *NBER Macroeconomics Annual 34*(1), 199–255.

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.

Midrigan, V. (2011). Menu costs, multiproduct firms, and aggregate fluctuations. *Econometrica 79*(4), 1139–1180.

Mongey, S. (2021). Market structure and monetary non-neutrality. Technical Report 29233, National Bureau of Economic Research.

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.

Nakamura, E. and J. Steinsson (2010). Monetary non-neutrality in a multisector menu cost model. *The Quarterly Journal of Economics 125*(3), 961–1013.

Pasten, E., R. Schoenle, and M. Weber (2020). The propagation of monetary policy shocks in a heterogeneous production economy. *Journal of Monetary Economics 116*, 1–22.

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.

Rossi-Hansberg, E., P.-D. Sarte, and N. Trachter (2021). Diverging trends in national and local concentration. *NBER Macroeconomics Annual 35*(1), 115–150.

Rubbo, E. (2020). Networks, phillips curves and monetary policy. *Unpublished manuscript*.

Sigurdsson, J. (2019). Labor supply responses and adjustment frictions: A tax-free year in iceland. Technical Report 3278308, SSRN.

Smets, F. and R. Wouters (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *American Economic Review 97*(3), 586–606.

Smith, D. and S. Ocampo (2021). The evolution of us retail concentration. Working paper.

Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of Political Economy 88*(1), 1–23.

Wang, O. and I. Werning (2020). Dynamic oligopoly and price stickiness. Technical Report 27536, National Bureau of Economic Research.

Wieland, J. F. and M.-J. Yang (2020). Financial dampening. *Journal of Money, Credit and Banking 52*(1), 79–113.

Online Appendix to *The Supply-Side Effects of Monetary Policy*

David Rezza Baqaee   Emmanuel Farhi   Kunal Sangani

[Not for publication]

[*]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.

[†]We are grateful to Harald Uhlig and three anonymous referees for helpful suggestions. We thank Andy Atkeson, Saki Bigio, Ariel Burstein, Oleg Itskhoki, Ivan Werning, Jon Vogel, and other seminar participants for helpful comments. This paper received support from NSF grant No. 1947611.

A Proofs

Proof of Lemma 2. The aggregate markup is

$$\bar{\mu} = \left(\int_0^1 \frac{\lambda_\theta}{\mu_\theta} d\theta \right)^{-1}.$$

Log-linearizing, the change in the aggregate markup is

$$\begin{aligned} d \log \bar{\mu} &= -\bar{\mu} \int_0^1 \frac{\lambda_\theta}{\mu_\theta} (d \log \lambda_\theta - d \log \mu_\theta) d\theta \\ &= \mathbb{E}_{\lambda \mu^{-1}} [d \log \mu_\theta - d \log \lambda_\theta]. \end{aligned}$$

We can rewrite this as

$$\begin{aligned} d \log \bar{\mu} &= \mathbb{E}_{\lambda \mu^{-1}} [d \log \mu_\theta] - \mathbb{E}_{\lambda \mu^{-1}} [d \log \lambda_\theta] \\ &= \mathbb{E}_\lambda [(\bar{\mu}/\mu_\theta) \, d \log \mu_\theta] - \mathbb{E}_\lambda [(\bar{\mu}/\mu_\theta) \, d \log \lambda_\theta] \\ &= Cov_\lambda [\bar{\mu}/\mu_\theta, d \log \mu_\theta] + \mathbb{E}_\lambda [d \log \mu_\theta] - Cov_\lambda [\bar{\mu}/\mu_\theta, d \log \lambda_\theta], \end{aligned}$$

so we get

$$d \log A = -Cov_\lambda [\bar{\mu}/\mu_\theta, d \log (\lambda_\theta / \mu_\theta)].$$

Note that

$$\frac{\lambda_\theta}{\mu_\theta} = \frac{p_\theta y_\theta / I}{\mu_\theta} = \frac{w}{I} \frac{y_\theta}{A_\theta} = \frac{w l_\theta}{I}.$$

Thus we have,

$$\begin{aligned} d \log A &= -Cov_\lambda [\bar{\mu}/\mu_\theta, d \log (\lambda_\theta / \mu_\theta)] \\ &= -Cov_\lambda [\bar{\mu}/\mu_\theta, d \log w l_\theta - d \log I] \\ &= -Cov_\lambda [\bar{\mu}/\mu_\theta, d \log w l_\theta] \\ &= -Cov_\lambda [\bar{\mu}/\mu_\theta, d \log \mathrm{Costs}_\theta], \end{aligned}$$

which concludes the proof of Lemma 2. ∎

Proof of Proposition 1. We can alternatively write Lemma 1 as

$$\begin{aligned} d \log A &= -Cov_\lambda [\bar{\mu}/\mu_\theta, d \log w l_\theta] \\ &= -\bar{\mu} Cov_\lambda [1 - 1/\sigma_\theta, d \log y_\theta] \end{aligned}$$

$$= -\bar{\mu} Cov_{\lambda}\left[1/\sigma_{\theta}, \sigma_{\theta} d\log\frac{p_{\theta}}{P}\right]$$

Recall from the implicit definition of Y that

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

Log-linearizing, we get

$$0 = \int_0^1 \frac{\Phi_{\theta}(\frac{y_{\theta}}{Y})\frac{y_{\theta}}{Y}}{\int_0^1 \Phi_{\theta'}(\frac{y_{\theta'}}{Y})\frac{y_{\theta'}}{Y}d\theta'} \frac{\frac{y_{\theta}}{Y}\Phi'_{\theta}(\frac{y_{\theta}}{Y})}{\Phi_{\theta}(\frac{y_{\theta}}{Y})} d\log\frac{y_{\theta}}{Y}d\theta = \mathbb{E}_{\lambda}\left[d\log\frac{y_{\theta}}{Y}\right].$$

Using $\mathbb{E}_{\lambda}[\sigma_{\theta} d\log(p_{\theta}/P)] = -\mathbb{E}_{\lambda}[d\log(y_{\theta}/Y)] = 0$, we can rewrite

$$\begin{aligned} d\log A &= Cov_{\lambda}\left[(\bar{\mu}/\mu_{\theta}), \sigma_{\theta} d\log(p_{\theta}/P)\right] \\ &= \bar{\mu}\mathbb{E}_{\lambda}\left[(\sigma_{\theta}-1)\, d\log(p_{\theta}/P)\right] \\ &= -\bar{\mu}\mathbb{E}_{\lambda}\left[d\log(p_{\theta}/P)\right]. \end{aligned}$$

It will also be useful to calculate the expression for the change in the price aggregator, $d\log P$. Recall from (2) that

$$P = \frac{P^Y}{\int_0^1 \Phi'_{\theta}(\frac{y_{\theta}}{Y})\frac{y_{\theta}}{Y}d\theta}.$$

Log-linearizing yields

$$\begin{aligned} d\log P &= d\log P^Y - \int_0^1 \frac{\Phi'_{\theta}(\frac{y_{\theta}}{Y})\frac{y_{\theta}}{Y}}{\int_0^1 \Phi'_{\theta'}(\frac{y_{\theta'}}{Y})\frac{y_{\theta'}}{Y}d\theta'}\left(1 + \frac{\frac{y_{\theta}}{Y}\Phi''_{\theta}(\frac{y_{\theta}}{Y})}{\Phi'_{\theta}(\frac{y_{\theta}}{Y})}\right) d\log\frac{y_{\theta}}{Y}d\theta \\ &= \mathbb{E}_{\lambda}[d\log p_{\theta}] - \mathbb{E}_{\lambda}\left[\left(1 - \frac{1}{\sigma_{\theta}}\right) d\log\frac{y_{\theta}}{Y}\right] \\ &= \mathbb{E}_{\lambda}[d\log p_{\theta}] - \mathbb{E}_{\lambda}\left[(\sigma_{\theta}-1) d\log\frac{p_{\theta}}{P}\right] \\ &= \mathbb{E}_{\lambda\sigma}\left[d\log p_{\theta}\right]. \end{aligned}$$

So, using $d\log P = \mathbb{E}_{\lambda\sigma}[d\log p_{\theta}]$, our expression for $d\log A$ becomes

$$d\log A = \bar{\mu}\left(\mathbb{E}_{\lambda\sigma}\left[d\log p_{\theta}\right] - \mathbb{E}_{\lambda}\left[d\log p_{\theta}\right]\right) = \bar{\mu} Cov_{\lambda}\left[\sigma_{\theta}/\mathbb{E}_{\lambda}\left[\sigma_{\theta}\right], d\log p_{\theta}\right].$$

Recall that

$$
\begin{aligned}
d\log p_\theta^{\text{sticky}} &= 0 \\
d\log p_\theta^{\text{flex}} &= \rho_\theta d\log w + (1-\rho_\theta)\, d\log P
\end{aligned}
$$

Solving the fixed point for change in the price aggregator yields:

$$
\begin{aligned}
d\log P &= \frac{\mathbb{E}_\lambda\left[\sigma_\theta d\log p_\theta\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\right]} \\
&= \frac{\mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\left(\rho_\theta d\log w + (1-\rho_\theta)\, d\log P\right)\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\right]} \\
&= \frac{\mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\rho_\theta\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} d\log w
\end{aligned}
$$

Returning to $d\log A$, we get:

$$
\begin{aligned}
d\log A &= \bar{\mu} Cov_\lambda\left[\frac{\sigma_\theta}{\mathbb{E}_\lambda\left[\sigma_\theta\right]}, \delta_\theta\rho_\theta d\log w + \delta_\theta\left(1-\rho_\theta\right) d\log P\right] \\
&= \frac{\bar{\mu} Cov_\lambda\left[\sigma_\theta, \delta_\theta\rho_\theta\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right] + \delta_\theta\left(1-\rho_\theta\right)\mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\rho_\theta\right]\right] d\log w}{\mathbb{E}_\lambda\left[\sigma_\theta\right]\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} \\
&= \frac{\bar{\mu} Cov_\lambda\left[\sigma_\theta, \delta_\theta\rho_\theta\mathbb{E}_\lambda\left[\sigma_\theta\left(1-\delta_\theta\right)\right] + \delta_\theta\mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\rho_\theta\right]\right] d\log w}{\mathbb{E}_\lambda\left[\sigma_\theta\right]\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} \\
&= \frac{\bar{\mu}\left(\left(\mathbb{E}_\lambda\left[\sigma_\theta\right]-\mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\right]\right)\left(-\mathbb{E}_\lambda\left[\sigma_\theta\right]\mathbb{E}_\lambda\left[\delta_\theta\rho_\theta\right]\right) + \mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\rho_\theta\right]\left(-\mathbb{E}_\lambda\left[\sigma_\theta\right]\mathbb{E}_\lambda\left[\delta_\theta\right]+\mathbb{E}_\lambda\left[\sigma_\theta\right]\right)\right) d\log w}{\mathbb{E}_\lambda\left[\sigma_\theta\right]\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} \\
&= \frac{\bar{\mu}\left(\left(\mathbb{E}_\lambda\left[\sigma_\theta\right]-\mathbb{E}_\lambda\left[\sigma_\theta\delta_\theta\right]\right)\left(-\mathbb{E}_\lambda\left[\delta_\theta\right]\mathbb{E}_{\lambda\delta}\left[\rho_\theta\right]\right) + \mathbb{E}_\lambda\left[\delta_\theta\right]\left(1-\mathbb{E}_\lambda\left[\delta_\theta\right]\right)\mathbb{E}_{\lambda\delta}\left[\sigma_\theta\rho_\theta\right]\right)}{\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} d\log w \\
&= \frac{\bar{\mu}}{\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]}\left(\mathbb{E}_{\lambda\delta}\left[\rho_\theta\right] Cov_\lambda\left[\sigma_\theta,\delta_\theta\right] + \mathbb{E}_\lambda\left[\delta_\theta\right]\left(1-\mathbb{E}_\lambda\left[\delta_\theta\right]\right) Cov_{\lambda\delta}\left[\sigma_\theta,\rho_\theta\right]\right) d\log w.
\end{aligned}
$$

■

Proof of Propositions 2 and 3. We use the change in firm markups to calculate

$$
\begin{aligned}
\mathbb{E}_\lambda\left[d\log\mu_\theta\right] &= \mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right] d\log P - \mathbb{E}_\lambda\left[1-\delta_\theta\rho_\theta\right] d\log w \\
&= \left[\frac{\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda\left[\delta_\theta\rho_\theta\sigma_\theta\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} - \mathbb{E}_\lambda\left[1-\delta_\theta\rho_\theta\right]\right] d\log w \\
&= \left[-\frac{\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda\left[\sigma_\theta\left(1-\delta_\theta\right)\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\left[(1-\delta_\theta)+\delta_\theta\rho_\theta\right]\right]} - \mathbb{E}_\lambda\left[1-\delta_\theta\right]\right] d\log w,
\end{aligned}
$$

which yields Equation (16). Combining the log-linearized labor-leisure condition and

Equation (5) yields

$$
\begin{aligned}
d\log L &= \frac{\zeta(1-\gamma)}{1+\zeta} d\log Y - \frac{\zeta}{1+\zeta} d\log \bar{\mu}, \\
d\log A &= d\log \bar{\mu} - \mathbb{E}_\lambda \left[d\log \mu_\theta \right], \\
\Rightarrow \frac{1+\gamma\zeta}{1+\zeta} d\log Y &= \frac{1}{1+\zeta} d\log A - \frac{\zeta}{1+\zeta} \mathbb{E}_\lambda \left[d\log \mu_\theta \right].
\end{aligned}
$$

Rearranging yields Equation (15), which concludes the proof of Proposition 2.

Proposition 3 follows immediately from dividing Equation (15) 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*

$$
\frac{\textit{Phillips curve slope w/ nominal rigidities only}}{\textit{Phillips curve slope w/ real rigidities}} = 1 + \frac{1}{\mathbb{E}_\lambda \left[1-\delta_\theta\right]} \frac{\mathbb{E}_\lambda \left[\delta_\theta(1-\rho_\theta)\right] \mathbb{E}_\lambda \left[\sigma_\theta(1-\delta_\theta)\right]}{\mathbb{E}_\lambda \left[\delta_\theta\right] \mathbb{E}_\lambda \left[\delta_\theta \rho_\theta \sigma_\theta\right] + \mathbb{E}_\lambda \left[\delta_\theta \rho_\theta\right] \mathbb{E}_\lambda \left[\sigma_\theta(1-\delta_\theta)\right]}. \tag{28}
$$

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

$$
\frac{\textit{Phillips curve slope w/ real rigidities}}{\textit{Phillips curve slope w/ misallocation}} = 1 + \frac{\bar{\mu}}{\zeta} \frac{\mathbb{E}_\lambda[\delta_\theta]\mathbb{E}_\lambda \left[1-\delta_\theta\right] Cov_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \mathbb{E}_{\lambda\delta} \left[\rho_\theta\right] Cov_\lambda[\sigma_\theta, \delta_\theta]}{\mathbb{E}_\lambda \left[1-\delta_\theta\right] \mathbb{E}_\lambda \left[\delta_\theta \rho_\theta \sigma_\theta\right] + \mathbb{E}_\lambda \left[1-\delta_\theta \rho_\theta\right] \mathbb{E}_\lambda \left[\sigma_\theta(1-\delta_\theta)\right]}. \tag{29}
$$

Proof. The flattening due to the misallocation channel is,

$$
\begin{aligned}
&\text{Flattening due to the misallocation channel} \\
&= \frac{\frac{d\log A}{d\log w} - \zeta \mathbb{E}_\lambda \left[\frac{d\log \mu_\theta}{d\log w}\right]}{-\zeta \mathbb{E}_\lambda \left[\frac{d\log \mu_\theta}{d\log w}\right]} \\
&= 1 + \frac{\frac{d\log A}{d\log w}}{-\zeta \mathbb{E}_\lambda \left[\frac{d\log \mu_\theta}{d\log w}\right]} \\
&= 1 + \frac{1}{\zeta} \frac{\kappa_\rho Cov_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \kappa_\delta Cov_\lambda[\sigma_\theta, \delta_\theta]}{\frac{\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda\left[\left[\delta_\theta\rho_\theta+(1-\delta_\theta)\right]\sigma_\theta\right]} + \mathbb{E}_\lambda \left[1-\delta_\theta\right]}
\end{aligned}
$$

$$= 1 + \frac{\bar{\mu}}{\zeta} \frac{\mathbb{E}_\lambda[\delta_\theta]\mathbb{E}_\lambda\left[1-\delta_\theta\right] Cov_{\lambda\delta}[\rho_\theta, \sigma_\theta] + \mathbb{E}_{\lambda\delta}\left[\rho_\theta\right] Cov_\lambda[\sigma_\theta, \delta_\theta]}{\mathbb{E}_\lambda\left[1-\delta_\theta\rho_\theta\right]\mathbb{E}_\lambda\left[\sigma_\theta(1-\delta_\theta)\right] + \mathbb{E}_\lambda\left[1-\delta_\theta\right]\mathbb{E}_\lambda\left[\delta_\theta\rho_\theta\sigma_\theta\right]}.$$

The flattening due to real rigidities is,

$$\begin{aligned}
&\text{Flattening due to real rigidities}\\
&= \left(\underbrace{\frac{1 - \mathbb{E}_\lambda\left[1-\delta_\theta\right] - \frac{\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda\left[\left[\delta_\theta\rho_\theta+(1-\delta_\theta)\right]\sigma_\theta\right]}}{-\zeta\left[-\mathbb{E}_\lambda\left[1-\delta_\theta\right] - \frac{\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda\left[\left[\delta_\theta\rho_\theta+(1-\delta_\theta)\right]\sigma_\theta\right]}\right]}}_{\text{Slope with real rigidities}} \right)^{-1} \underbrace{\frac{1-\mathbb{E}_\lambda\left[1-\delta_\theta\right]}{\zeta\mathbb{E}_\lambda\left[1-\delta_\theta\right]}}_{\text{Slope in CES model}}\\
&= \frac{\mathbb{E}_\lambda\left[\delta_\theta\right]\mathbb{E}_\lambda\left[1-\delta_\theta\right] + \frac{\mathbb{E}_\lambda[\delta_\theta]\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda\left[\left[\delta_\theta\rho_\theta+(1-\delta_\theta)\right]\sigma_\theta\right]}}{\mathbb{E}_\lambda\left[\delta_\theta\right]\mathbb{E}_\lambda\left[1-\delta_\theta\right] - \frac{\mathbb{E}_\lambda[1-\delta_\theta]\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda[\sigma_\theta(1-\delta_\theta)]}{\mathbb{E}_\lambda\left[\left[\delta_\theta\rho_\theta+(1-\delta_\theta)\right]\sigma_\theta\right]}}\\
&= 1 + \frac{1}{\mathbb{E}_\lambda\left[1-\delta_\theta\right]} \frac{\mathbb{E}_\lambda\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_\lambda\left[\sigma_\theta(1-\delta_\theta)\right]}{\mathbb{E}_\lambda\left[\delta_\theta\right]\mathbb{E}_\lambda\left[\delta_\theta\rho_\theta\sigma_\theta\right] + \mathbb{E}_\lambda\left[\delta_\theta\rho_\theta\right]\mathbb{E}_\lambda\left[\sigma_\theta(1-\delta_\theta)\right]}.
\end{aligned}$$

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

Proof of Proposition 5. Firms choose reset prices to maximize future discounted profits,

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

The first order condition is

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

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

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

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:

$$
\begin{aligned}
\frac{p_{i,t}^{\text{flex}} A_i}{w_t} &= \frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t+k}\sigma_{i,t+k}\left(d\log\left(\frac{w_{t+k}}{w_t}\right)+1\right)\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t+k}\left(\sigma_{i,t+k}-1\right)\right]} \\
&= \frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t}\sigma_{i,t}(1+d\log(y_{i,t+k}\sigma_{i,t+k}))\left(d\log\left(\frac{w_{t+k}}{w_t}\right)+1\right)\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k y_{i,t}\left(\sigma_{i,t}-1\right)(1+d\log(y_{i,t+k}(\sigma_{i,t+k}-1)))\right]} \\
&= \mu_{i,t}\frac{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k\right] + \mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d\log\left(\frac{w_{t+k}}{w_t}\right)\right] + \mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d\log(y_{i,t+k}\sigma_{i,t+k})\right]}{\mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k\right] + \mathbb{E}\left[\sum_{k=0}^{\infty} \beta^k (1-\delta_i)^k d\log(y_{i,t+k}(\sigma_{i,t+k}-1))\right]}.
\end{aligned}
$$

Using $\mu_{i,t}^{\text{flex}}/\mu_{i,t} = 1 + d\log\mu_{i,t}$ and removing second order terms, we get:

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

At the time of a price reset, we know that

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

Then,

$$
\begin{aligned}
d\log(\mu(\frac{y_{i,t+k}}{Y_{t+k}})) &= \frac{\frac{y_{i,t}}{Y_t}\mu_i'(\frac{y_{i,t}}{Y_t})}{\mu_{i,t}} d\log(\frac{y_{i,t+k}}{Y_{t+k}}) \\
&= \frac{1-\rho_{i,t}}{\rho_{i,t}}\frac{1}{\sigma_{i,t}} d\log(\frac{y_{i,t+k}}{Y_{t+k}}) \\
&= \frac{1-\rho_{i,t}}{\rho_{i,t}}\left(d\log P_{t+k} - d\log w_{t+k} - d\log\mu_{i,t+k}\right) \\
&= \frac{1-\rho_{i,t}}{\rho_{i,t}}\left(d\log P_{t+k} - d\log w_{t+k} - d\log\mu_{i,t}^{\text{flex}} + d\log\frac{w_{t+k}}{w_t}\right),
\end{aligned}
$$

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 ($d\log\mu_{i,t}^{\text{flex}}$) and subsequent changes due to the shifts in the nominal wage.

Plugging this in yields,

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

Finally, since $d\log\mu^{\text{flex}}_{i,t} = d\log p^{\text{flex}}_{i,t} - d\log w_t$, we get

$$d\log p^{\text{flex}}_{i,t} = [1-\beta(1-\delta_i)]\sum_{k=0}^{\infty}\beta^k(1-\delta_i)^k\left[\rho_i d\log w_{t+k} + (1-\rho_i)d\log P_{t+k}\right].$$

We can write this equation recursively as

$$d\log p^{\text{flex}}_{i,t} = (1-\beta(1-\delta_i))\left[\rho_i d\log w_t + (1-\rho_i)d\log P_t\right] + \beta(1-\delta_i)p^{\text{flex}}_{i,t+1},$$

or in terms of firm types as,

$$d\log p^{\text{flex}}_{\theta,t} = (1-\beta(1-\delta_\theta))\left[\rho_\theta d\log w_t + (1-\rho_\theta)d\log P_t\right] + \beta(1-\delta_\theta)p^{\text{flex}}_{\theta,t+1}.$$

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,

$$\mathbb{E}\left[d\log p_{\theta,t+1}\right] = \delta_\theta d\log p^{\text{flex}}_{\theta,t+1} + (1-\delta_\theta)d\log p_{\theta,t},$$

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

$$\begin{aligned}&\mathbb{E}[d\log p_{\theta,t} - d\log p_{\theta,t-1}] - \beta\mathbb{E}[d\log p_{\theta,t+1} - d\log p_{\theta,t}]\\ &\qquad= \frac{\delta_\theta}{1-\delta_\theta}(1-\beta(1-\delta_\theta))\left[-\mathbb{E}[d\log p_\theta] + \rho_\theta d\log w_t + (1-\rho_\theta)d\log P_t\right].\end{aligned} \tag{30}$$

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 be rearranging the above.

$$\begin{aligned}d\log P_t - d\log P^Y_t &= \bar{\mu}^{-1}\, d\log A_t\\ d\log P^Y_t - d\log w_t &= \mathbb{E}_\lambda\left[d\log\mu_\theta\right]\end{aligned}$$

$$\begin{aligned} d\log A_t &= d\log\bar{\mu}_t - \mathbb{E}_\lambda\left[d\log\mu_{\theta,t}\right] \\ d\log Y_t &= \frac{1}{1+\gamma\zeta}\left(d\log A_t - \zeta\mathbb{E}_\lambda\left[d\log\mu_{\theta,t}\right]\right) \\ -\mathbb{E}_\lambda\left[d\log\mu_{\theta,t}\right] &= \left(\frac{1+\gamma\zeta}{\zeta}\right)d\log Y_t - \frac{1}{\zeta}d\log A_t \\ d\log w_t - d\log P_t &= \frac{1+\gamma\zeta}{\zeta}d\log Y_t - \left(\frac{1}{\zeta}+\frac{1}{\bar{\mu}}\right)d\log A_t. \end{aligned}$$

We now take the sales-weighted expectation of Equation (30) to get:

$$\begin{aligned} d\log\pi_t - \beta d\log\pi_{t+1} &= \varphi\left[-d\log P_t^Y + \mathbb{E}_\lambda\left[\rho_\theta\right]d\log w_t + \left(1-\mathbb{E}_\lambda\left[\rho_\theta\right]\right)d\log P_t\right] \\ &= \varphi\left[\left(d\log P_t - d\log P_t^Y\right) + \mathbb{E}_\lambda\left[\rho_\theta\right]\left(d\log w_t - d\log P_t\right)\right] \\ &= \varphi\left[\left(\bar{\mu}^{-1}d\log A_t\right) + \mathbb{E}_\lambda\left[\rho_\theta\right]\left(\frac{1+\gamma\zeta}{\zeta}d\log Y_t - \left(\frac{1}{\zeta}+\frac{1}{\bar{\mu}}\right)d\log A_t\right)\right] \\ &= \varphi\mathbb{E}_\lambda\left[\rho_\theta\right]\frac{1+\gamma\zeta}{\zeta}d\log Y_t + \varphi\left[\frac{1}{\bar{\mu}} - \mathbb{E}_\lambda\left[\rho_\theta\right]\left(\frac{1}{\zeta}+\frac{1}{\bar{\mu}}\right)\right]d\log A_t, \end{aligned}$$

which is the NKPC equation.

Endogenous TFP equation. Start by subtracting $\mathbb{E}\left[d\log w_t - d\log w_{t-1}\right] - \beta\mathbb{E}\left[d\log w_{t+1} - d\log w_t\right]$ from both sides of Equation (30). This yields,

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

We can write

$$d\log A_t = d\log\bar{\mu} - \mathbb{E}_\lambda\left[d\log\mu_\theta\right] = \bar{\mu}\left(\frac{\mathbb{E}_\lambda\left[\sigma_\theta d\log\mu_{\theta,t}\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\right]} - \mathbb{E}_\lambda\left[d\log\mu_\theta\right]\right). \tag{31}$$

Now, we take Equation (31) 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 (31); and multiply (1) − (2) by $\bar{\mu}$. This yields,

$$\begin{aligned} &\left(d\log A_t - d\log A_{t-1}\right) - \beta\left(d\log A_{t+1} - d\log A_t\right) \\ &\quad = \varphi\left[-d\log A_t + \bar{\mu}\left(1 - \frac{\mathbb{E}_\lambda\left[\sigma_\theta\rho_\theta\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\right]} - \left(1-\mathbb{E}_\lambda\left[\rho_\theta\right]\right)\right)\left(d\log P_t - d\log w_t\right)\right] \end{aligned}$$

$$
\begin{aligned}
&= \varphi \left[-d \log A_t + \bar{\mu} \left(\frac{Cov_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \right) (d \log w_t - d \log P_t) \right] \\
&= \varphi \left[-d \log A_t + \bar{\mu} \left(\frac{Cov_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \right) \left(\frac{1 + \gamma\zeta}{\zeta} d \log Y_t - \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) d \log A_t \right) \right] \\
&= \varphi \left[- \left(1 + \bar{\mu} \left(\frac{Cov_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \right) \left(\frac{1}{\zeta} + \frac{1}{\bar{\mu}} \right) \right) d \log A_t + \bar{\mu} \frac{Cov_\lambda [\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda [\sigma_\theta]} \frac{1 + \gamma\zeta}{\zeta} d \log Y_t \right].
\end{aligned}
$$

■

B Empirical Evidence Appendix

This appendix describes the data and procedures used in Section 7. First, section B.1 describes how we construct firm-level markup data. Section B.2 explores the unconditional relationship between aggregate productivity and the business cycle, and Section B.3 provides the unconditional relationship between cyclicality of high- and low-markup firms in our sample. Section B.4 provides additional detail and robustness for the estimation of procyclical reallocations to high-markup firms following identified monetary shocks.

B.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 two approaches: (1) accounting profits (AP), (2) user cost (UC). We broadly use the same approaches described in Baqaee and Farhi (2020); the following text provides a brief overview.

B.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:

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

B.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). 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).

B.2 Unconditional Cyclicality of Aggregate TFP

In the main text, we show that aggregate TFP is responsive to identified monetary shocks. Table B.1 shows the unconditional association between the same three measures of aggregate TFP used in the main text—labor productivity, the Solow residual, and the cost-based Solow residual—and three different measures of the business cycle: the unemployment rate, NBER recession dates, and real GDP growth. For all measures, we find that productivity covaries significantly with business cycle indicators.

Table B.1: Procyclical aggregate productivity.

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

Notes: *Unemp.* is the average unemployment rate in year $t+1$, $\%\Delta GDP$ 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%.

B.3 Differential Cyclicality of Low- and High-Markup Firms

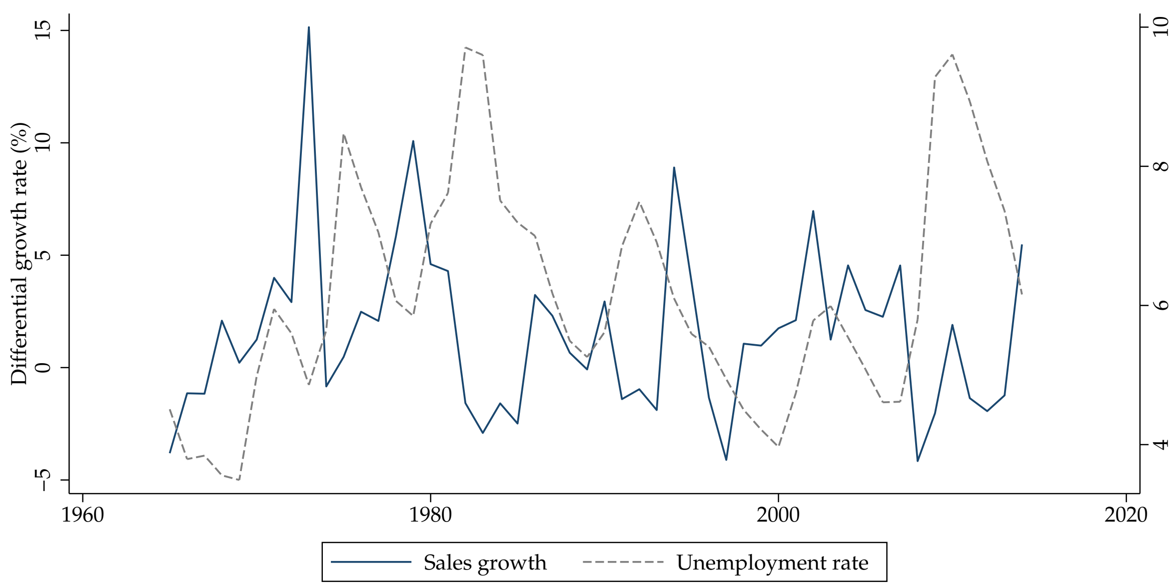


Figure B.1: Procyclical reallocations to high-markup firms. A firm is categorized as high-markup (low-markup) if its markup is above (below) median in year t. The solid line shows the difference in sales growth of high- and low-markup firms from t to $t+1$.

Figure B.1 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.[1] We then calculate the difference in the sales growth of both groups from year t to $t + 1$.[2] As shown in Figure B.1, the differential growth rate shows substantial variance over the sample and is correlated with the business cycle (here, captured by the unemployment rate).

B.4 Reallocations to High-Markup Firms

For within-industry local projection estimates, we use the following panel specification:

$$\begin{aligned}
Cov_{\lambda}(-1/\mu_{f,t}, \Delta \log \mu_{f,t \to t+h}) &= a_i^h + \sum_{k=0}^{4} b_k^h \cdot \text{MonetaryShock}_{t-k} \\
&\quad + \sum_{k=1}^{4} c_k^h \cdot Cov_{\lambda}(-1/\mu_{f,t}, \Delta \log \mu_{f,t-k \to t}) + \epsilon_{i,t}^h, \\
Cov_{\lambda}(-1/\mu_{f,t}, \Delta \log \text{Costs}_{f,t \to t+h}) &= \tilde{a}_i^h + \sum_{k=0}^{4} \tilde{b}_k^h \cdot \text{MonetaryShock}_{t-k} \\
&\quad + \sum_{k=1}^{4} \tilde{c}_k^h \cdot Cov_{\lambda}(-1/\mu_{f,t}, \Delta \log \text{Costs}_{f,t-k \to t}) + \epsilon_{i,t}^h,
\end{aligned}$$

where the subscript i denotes a NAICS-3 industry, the subscript $f \in \mathcal{F}(i)$ denotes a firm f in industry i, and a_i^h and $\tilde{a}_i^h$ are industry fixed effects. We limit our analysis to industries with at least five public firms in year t and weight the regression by NAICS-3 industry sales at time t. Confidence intervals use Driscoll-Kraay standard errors.

The impulse responses in the main text use user-cost markups and the extension of the Romer and Romer (2004) shock series by Wieland and Yang (2020). Figure B.2 shows that our results are robust to instead using accounting profits, and Figure B.3 shows that our results are robust to including intangible capital in our measure of total firm capital when calculating user-cost markups. For completeness, we also provide impulse responses using the covariance of firms' initial markups (rather than inverse markups) with the change in firms' markups and costs. Our results continue to hold when we use firms' initial user-cost markups (Figure B.4), markups measured using accounting

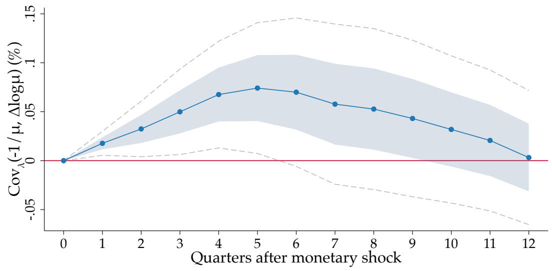
[1] Specifically, we assume operating income minus depreciation is profit and infer the markup by assuming firms have constant returns to scale. We use accounting profits in Figure B.1 since that allows us to plot the series for the longest sample.

[2] For each year t, we limit our analysis to firms in the sample in both years t and $t + 1$. The high-markup and low-markup group are constructed in year t by comparing each firm's markup to the median markup in that year. The differential growth rate is then calculated as the growth rate of total sales for the high-markup group minus the growth rate of total sales for the low-markup group.

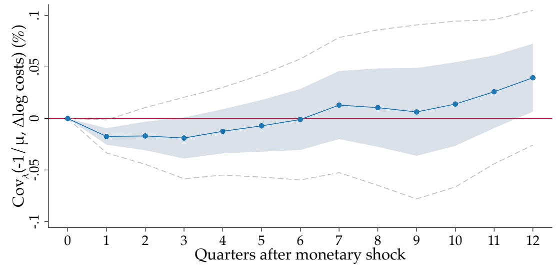
profits (Figure B.5), and user-cost markups measured using both tangible and intangible capital (Figure B.6). Finally, in Figure B.7 and Figure B.8, we show that our results also hold using an alternate monetary shock series identified with high-frequency methods by Gorodnichenko and Weber (2016).

In the main text, we show that more concentrated manufacturing industries (using measures of concentration from the 2002 Economic Census for Manufacturing) experience a greater contraction in multifactor productivity following Romer and Romer (2004) shocks. In Table B.2, we show these results are robust to using measures of concentration from the 2007 Economic Census for Manufacturing. In Table B.3, we show these results are robust to instead using monetary shocks identified using high-frequency methods from Gorodnichenko and Weber (2016).

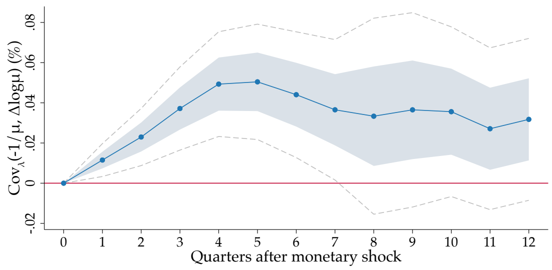
Figure B.2: Local projections using accounting profits markups.



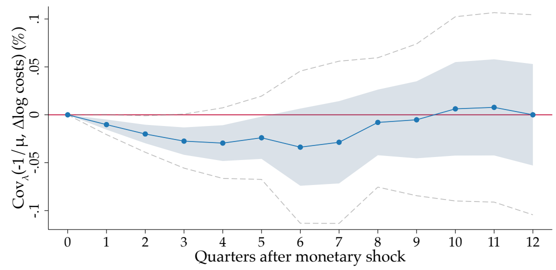
(a) $Cov_\lambda(-1/\mu, \Delta \log \mu)$



(b) $Cov_\lambda(-1/\mu, \Delta \log \text{Costs})$



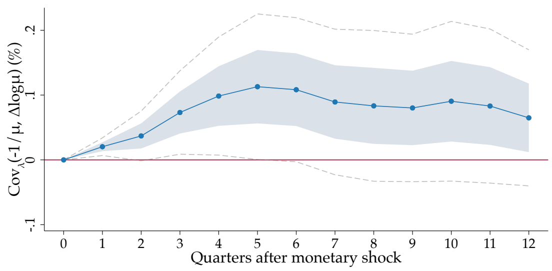
(c) $Cov_\lambda(-1/\mu, \Delta \log \mu)$ within NAICS-3



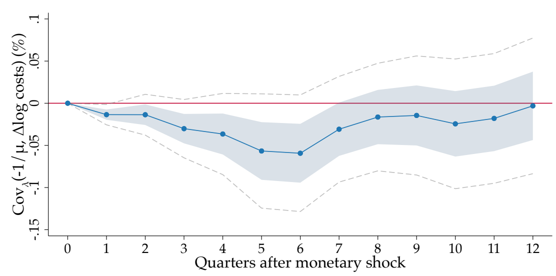
(d) $Cov_\lambda(-1/\mu, \Delta \log \text{Costs})$ within NAICS-3

Notes: The shaded region indicates Newey-West standard errors in panels (a)-(b) and Driscoll-Kraay standard errors in panels (c)-(d). Dashed lines are 95% confidence intervals.

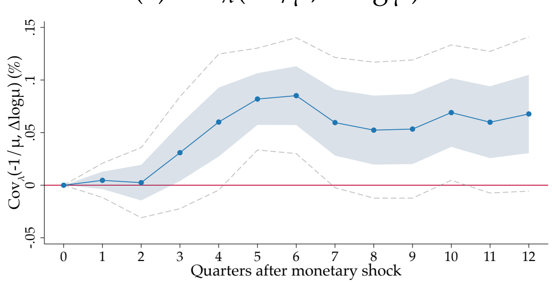
Figure B.3: Local projections including intangible capital in user-cost markup estimates.



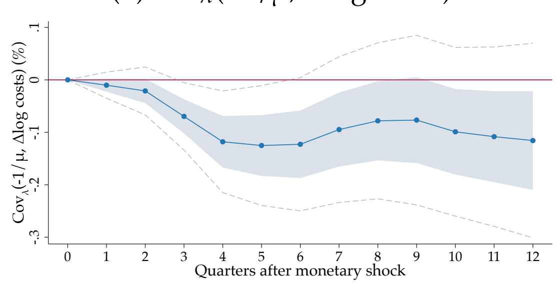
(a) $Cov_\lambda(-1/\mu, \Delta \log \mu)$



(b) $Cov_\lambda(-1/\mu, \Delta \log \text{Costs})$



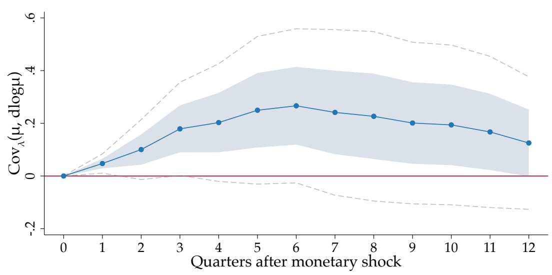
(c) $Cov_\lambda(-1/\mu, \Delta \log \mu)$ within NAICS-3



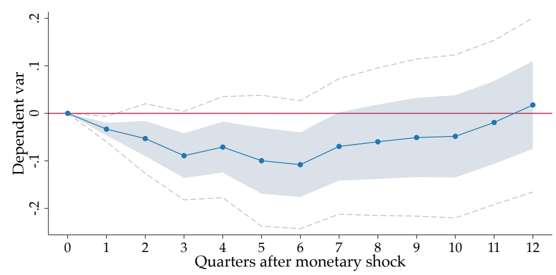
(d) $Cov_\lambda(-1/\mu, \Delta \log \text{Costs})$ within NAICS-3

Notes: The shaded region indicates Newey-West standard errors in panels (a)-(b) and Driscoll-Kraay standard errors in panels (c)-(d). Dashed lines are 95% confidence intervals.

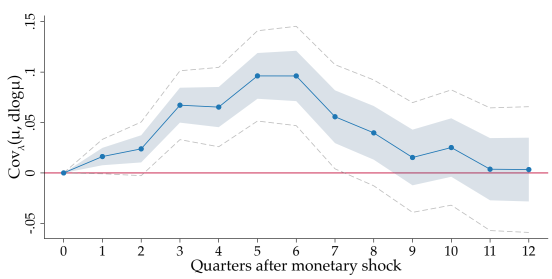
Figure B.4: Local projections using covariance with initial markups.



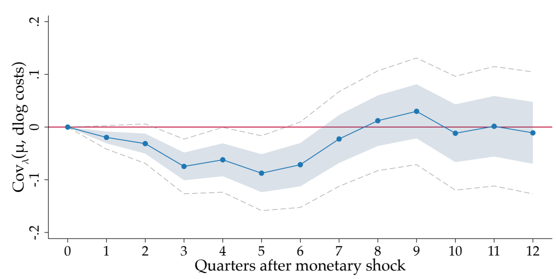
(a) $Cov_\lambda(\mu, \Delta \log \mu)$



(b) $Cov_\lambda(\mu, \Delta \log \text{Costs})$



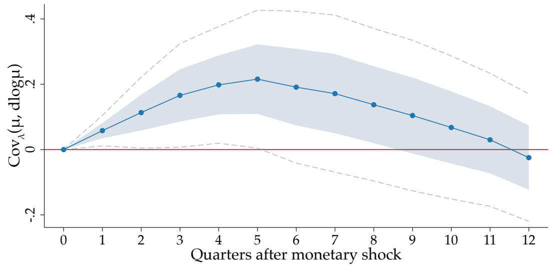
(c) $Cov_\lambda(\mu, \Delta \log \mu)$ within NAICS-3



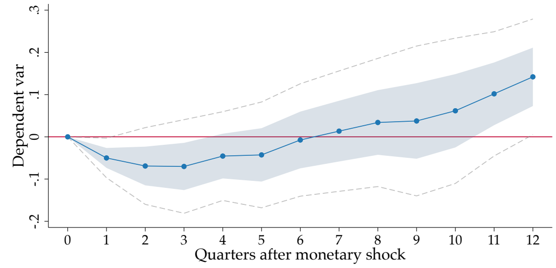
(d) $Cov_\lambda(\mu, \Delta \log \text{Costs})$ within NAICS-3

Notes: The shaded region indicates Newey-West standard errors in panels (a)-(b) and Driscoll-Kraay standard errors in panels (c)-(d). Dashed lines are 95% confidence intervals.

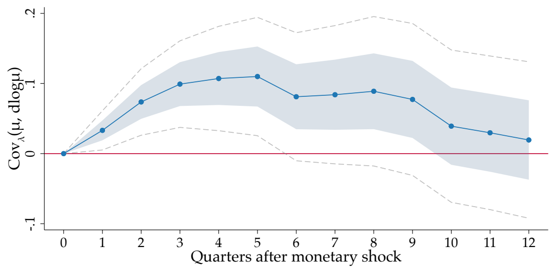
Figure B.5: Local projections using covariance with initial markups, using accounting profits markups.



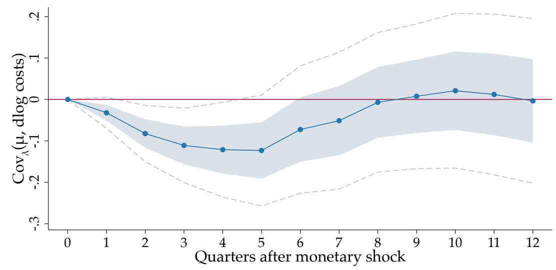
(a) $Cov_\lambda(\mu, \Delta \log \mu)$



(b) $Cov_\lambda(\mu, \Delta \log \text{Costs})$



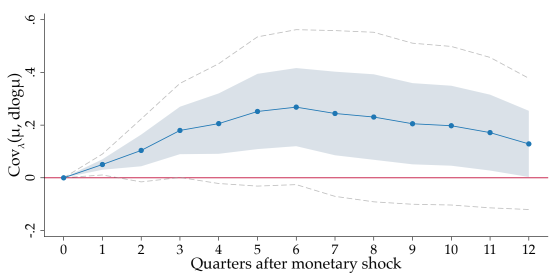
(c) $Cov_\lambda(\mu, \Delta \log \mu)$ within NAICS-3



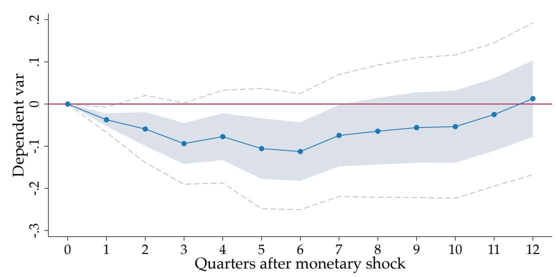
(d) $Cov_\lambda(\mu, \Delta \log \text{Costs})$ within NAICS-3

Notes: The shaded region indicates Newey-West standard errors in panels (a)-(b) and Driscoll-Kraay standard errors in panels (c)-(d). Dashed lines are 95% confidence intervals.

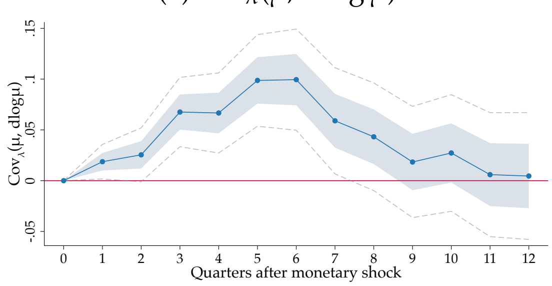
Figure B.6: Local projections using covariance with initial markups, including intangible capital in user-cost markup estimates.



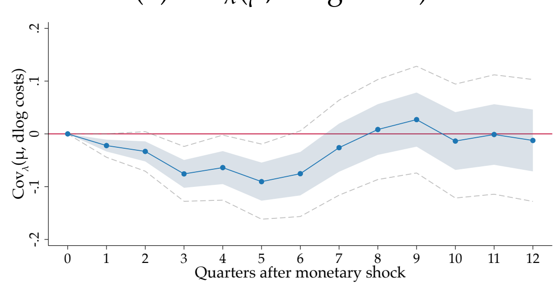
(a) $Cov_\lambda(\mu, \Delta \log \mu)$



(b) $Cov_\lambda(\mu, \Delta \log \text{Costs})$



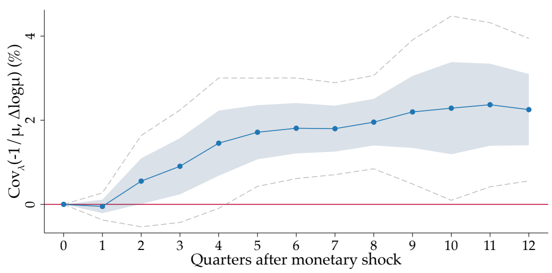
(c) $Cov_\lambda(\mu, \Delta \log \mu)$ within NAICS-3



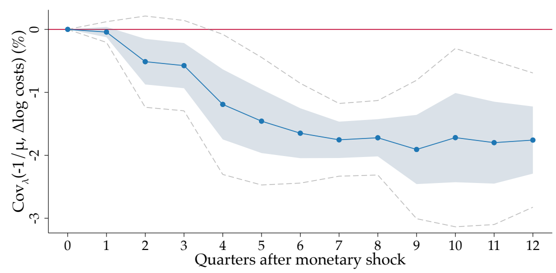
(d) $Cov_\lambda(\mu, \Delta \log \text{Costs})$ within NAICS-3

Notes: The shaded region indicates Newey-West standard errors in panels (a)-(b) and Driscoll-Kraay standard errors in panels (c)-(d). Dashed lines are 95% confidence intervals.

Figure B.7: Local projections using high-frequency monetary shock series from Gorodnichenko and Weber (2016).



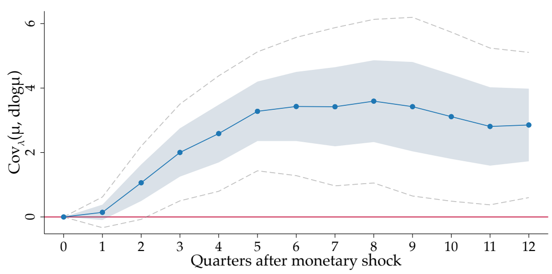
(a) $Cov_\lambda(-1/\mu, \Delta \log \mu)$



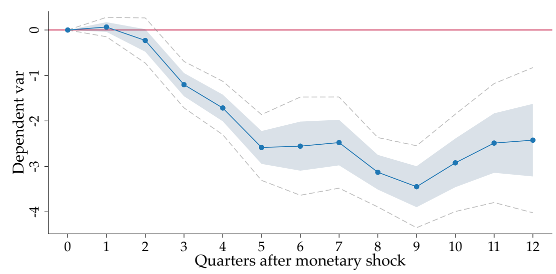
(b) $Cov_\lambda(-1/\mu, \Delta\text{Costs})$

Notes: The shaded region indicates Newey-West standard errors. Dashed lines are 95% confidence intervals.

Figure B.8: Local projections using covariance with initial markups and high-frequency monetary shock series from Gorodnichenko and Weber (2016).



(a) $Cov_\lambda(\mu, \Delta \log \mu)$



(b) $Cov_\lambda(\mu, \Delta\text{Costs})$

Notes: The shaded region indicates Newey-West standard errors. Dashed lines are 95% confidence intervals.

Table B.2: Differential response of industry multifactor productivity to monetary shocks in concentrated manufacturing industries, using concentration measures from 2007.

	Δ log MultifactorProductivity$_{i,t}$		
	(1)	(2)	(3)
Top 8 Firms Share$_i$ $\times$ MonetaryShock$_t$	-0.0155*		
	(0.00909)		
Top 20 Firms Share$_i$ $\times$ MonetaryShock$_t$		-0.0156**	
		(0.00748)	
Top 50 Firms Share$_i$ $\times$ MonetaryShock$_t$			-0.0155**
			(0.00702)
Industry FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes
N	1634	1634	1634

Notes: The sales shares of the top 8, 20, and 50 firms in each 4-digit NAICS industry are from the 2007 Economic Census for Manufacturing. Monetary shocks are from the extension of the Romer and Romer (2004) shock series by Wieland and Yang (2020). * indicates significance at 10%, ** at 5%.

Table B.3: Differential response of industry multifactor productivity to monetary shocks in concentrated manufacturing industries, using Gorodnichenko and Weber (2016) monetary shocks.

	Δ log MultifactorProductivity$_{i,t}$		
	(1)	(2)	(3)
Top 8 Firms Share$_i$ $\times$ MonetaryShock$_t$	-0.127**		
	(0.0383)		
Top 20 Firms Share$_i$ $\times$ MonetaryShock$_t$		-0.120**	
		(0.0343)	
Top 50 Firms Share$_i$ $\times$ MonetaryShock$_t$			-0.117**
			(0.0331)
Industry FEs	Yes	Yes	Yes
Year FEs	Yes	Yes	Yes
N	1204	1204	1204

Notes: The sales shares of the top 8, 20, and 50 firms in each 4-digit NAICS industry are from the 2002 Economic Census for Manufacturing. Monetary shocks are from Gorodnichenko and Weber (2016). * indicates significance at 10%, ** at 5%.

C Menu Cost Model

In our baseline model, price rigidities take the form of Calvo frictions. An alternative is to use menu costs, which are incurred by firms that choose to change their prices. In this appendix, we calibrate a version of the model where firms face menu costs and idiosyncratic productivity shocks. The calibration yields results that are quantitatively similar to those in our baseline calibration while matching empirical evidence on firm price changes documented by the literature.[3]

The strategy for the calibration is as follows. We start by calibrating a model with a CES demand system, menu costs, and idiosyncratic productivity shocks in line with recent work. We consider how this economy responds to an MIT shock to the money supply. Then, we replace the CES demand system with the Kimball demand system estimated from the Belgian data and simulate the economy's response to the money supply shock keeping all other factors constant.

Following the insight by Midrigan (2011) that fat-tailed productivity shocks are required to generate sufficient nonneutrality in menu costs models, we choose a fat-tailed, symmetric productivity process. Figure C.1 compares the shock process we use to a normal distribution. The standard deviation of productivity shocks is 0.025 log points. To preserve the steady-state distribution of firm productivity levels, we assume that firms that receive productivity shocks to points outside the productivity grid exit and are replaced by new firms at the productivity grid boundary.

We choose menu costs to generate a mean frequency of price adjustment in steady-state of 11% per month, in line with Nakamura and Steinsson (2008).[4] The result is menu costs that are 2% of monthly steady-state revenue, which means that firms spend about 0.24% of annual revenue on menu costs. This cost is moderate relative to Levy et al. (1997), who measure menu costs equal to 0.7% of revenue, and Midrigan (2011), who sets menu costs to 0.34% of annual revenue. We use fine grids to discretize both prices and productivities: the price grid consists of 2,800 points with spacing of 0.001 log price points, and the productivity grid consists of 71 points with spacing of 0.02 log productivity points. Our results do not change significantly if we further discretize the grids.

The remaining parameters are set in line with our baseline calibration. The Frisch elasticity is set to $\zeta = 0.2$; the intertemporal elasticity of substitution is set to $\gamma = 1$; the elasticity of substitution in the CES model is set to $\sigma = 5$, corresponding to a static profit-

[3]This calibration is also fully nonlinear and hence is not limited to first-order effects captured in the log-linearized model.

[4]Nakamura and Steinsson (2008) estimate the median frequency of nonsale price changes is 9–12% per month.

maximizing markup of 1.25; and the magnitude of the money supply shock is set to 4 basis points.[5]

As in the data, our model generates large, frequent, and symmetric price changes in steady state (Bils and Klenow 2004). The steady-state median price change is 0.079 log points in the CES calibration, which is close to the median regular price change in BLS data of 0.07 log points reported by Midrigan (2011). When we instead apply the Kimball aggregator from the Belgian data, the median price change is moderately smaller at 0.033 log points.

To simulate the response of the economy to an MIT shock to money supply, we use the algorithm in Burstein (2006). We conjecture that the distribution of prices and productivities T periods after the shock is identical to the steady-state distribution, but with all prices increased by the size of the money shock. Given a set of conjectured path of wages, output, and the price aggregator, we calculate the pricing decisions of firms by backward iteration from period T. Then, we use forward iteration from the initial steady-state distribution to calculate the distribution of firms across the price-productivity grid in each period and the resulting path of all aggregates. We iterate this procedure until firms' pricing decisions from backward iteration and the path of aggregates are mutually consistent.

Figure C.2 shows the response of the CES and Kimball economies to the money supply shock. Menu costs generate significant non-neutrality: in the CES calibration, 8% of the initial money shock loads on real output, while in the Kimball economy 20% of the money shock loads on real output. Compared to the CES economy, the Kimball economy has less inflation and a greater output effect on impact. The procyclical increase in aggregate productivity accounts for half of the output response. The CES model also generates a procyclical, albeit much smaller, response of aggregate productivity to the money shock. This response is driven by inherited dispersion in markups in the steady-state.

Unlike the calibration in the main text, where the Calvo friction was assumed to be identical across firm types, the frequency of price adjustment in the menu cost model is endogenous. Accordingly, the aggregate productivity effect in this menu cost model may be driven by differences across firms in both the extensive margin and intensive margin of price adjustment.

To investigate the correlation between firms' initial markups and the extensive margin of price changes, Figure C.3 plots the percent of firms changing their price t months after the shock separately for small firms (defined as the below the 70th percentile of firm

[5]Note that most menu cost calibrations (e.g., Midrigan 2011 and Nakamura and Steinsson 2010) assume an infinite Frisch elasticity to generate sufficient nonneutrality in output. With our Kimball calibration, we are able to generate significant nonneutrality even with a much lower Frisch elasticity of $\zeta = 0.2$, in line with the empirical evidence.

Figure C.1: Fat-tailed productivity shocks.

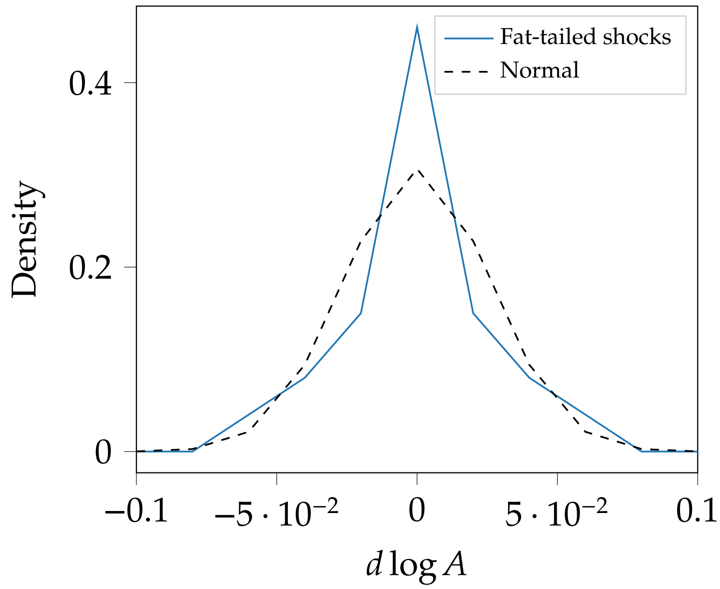


Figure C.2: Impulse response functions (IRFs) following a 4bp money supply shock in the menu cost model. Green and blue IRFs indicate the CES and Kimball models respectively.

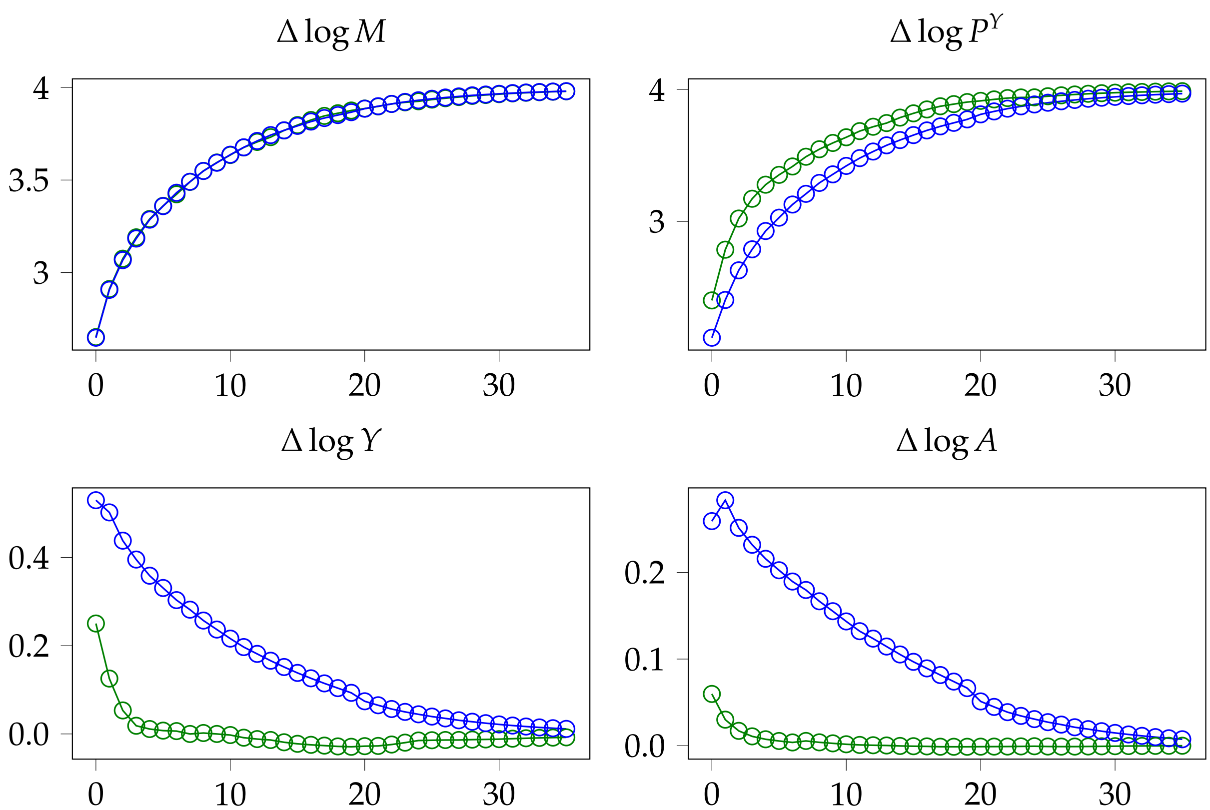
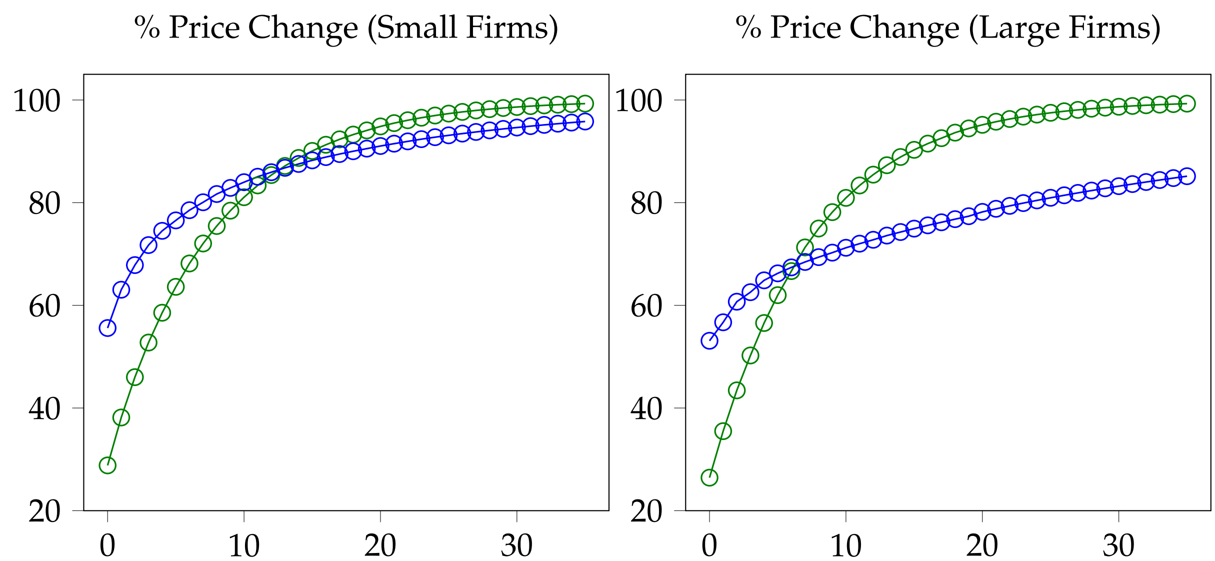


Figure C.3: Extensive margin of price changes across large (above 70th percentile productivity) and small firms. Green and blue lines indicate the CES and Kimball models respectively.



productivity) and large firms (above the 70th percentile of firm productivity). Unlike in the CES calibration, in the Kimball calibration there are clear differences in the likelihood of a price change between small and large firms. About the same fraction of large firms as small firms change their price on impact, but a smaller fraction of large firms change their prices in subsequent periods.

Intuitively, in our calibration, large firms are unwilling to pay the menu cost while others' prices are low due to strategic complementarities. As a result, the prices of large firms are endogenously more rigid, their markups fall by more after the expansionary money supply shock, and the reallocation of resources to these high-markup firms leads to an increase in aggregate productivity. As shown in Table 6 in the main text, these endogenous differences in the extensive margin of price adjustment are responsible for the misallocation channel in the menu cost model.

D Additional Calibrated Results

In this appendix, we provide additional results from our calibration exercise. D.1 provides additional comparative statics from the calibration of the static model as we change the average markup and the degree of price-stickiness. D.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 et al. (2021). We refer interested readers to that appendix.

D.1 Static model: Additional results

We vary the average markup $\bar{\mu}$ from just over one to 1.60 in Figure D.1. We do so by re-calculating markups of all firms according to the differential equation in Equation (27) 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 D.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 D.1: Decomposition of Phillips curve slope, varying the average markup $\bar{\mu}$.

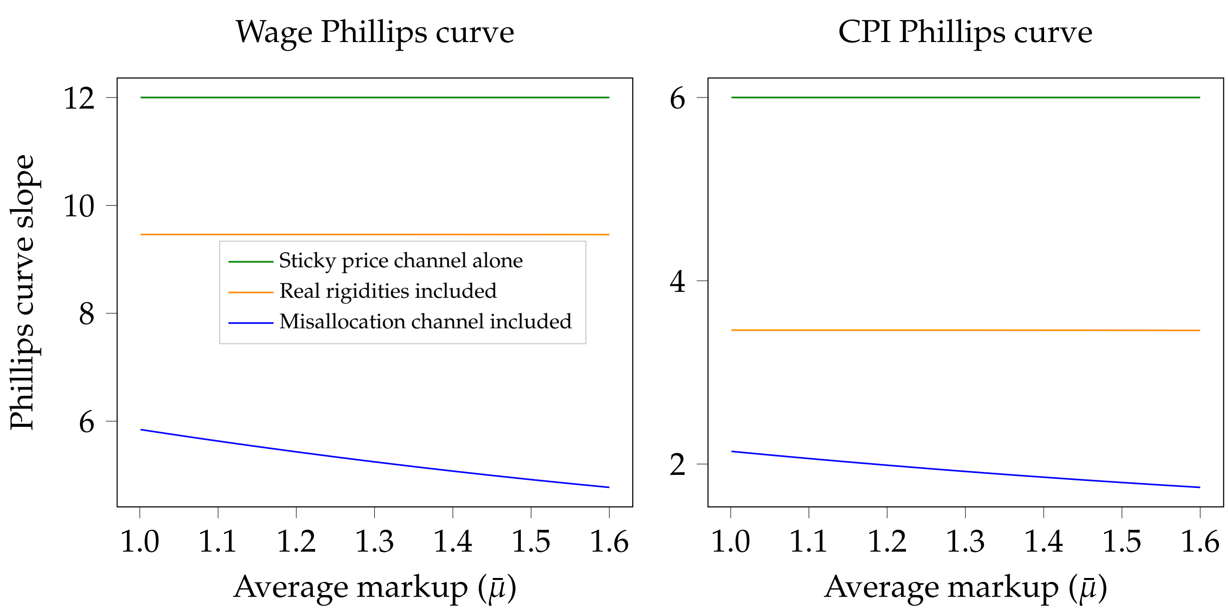
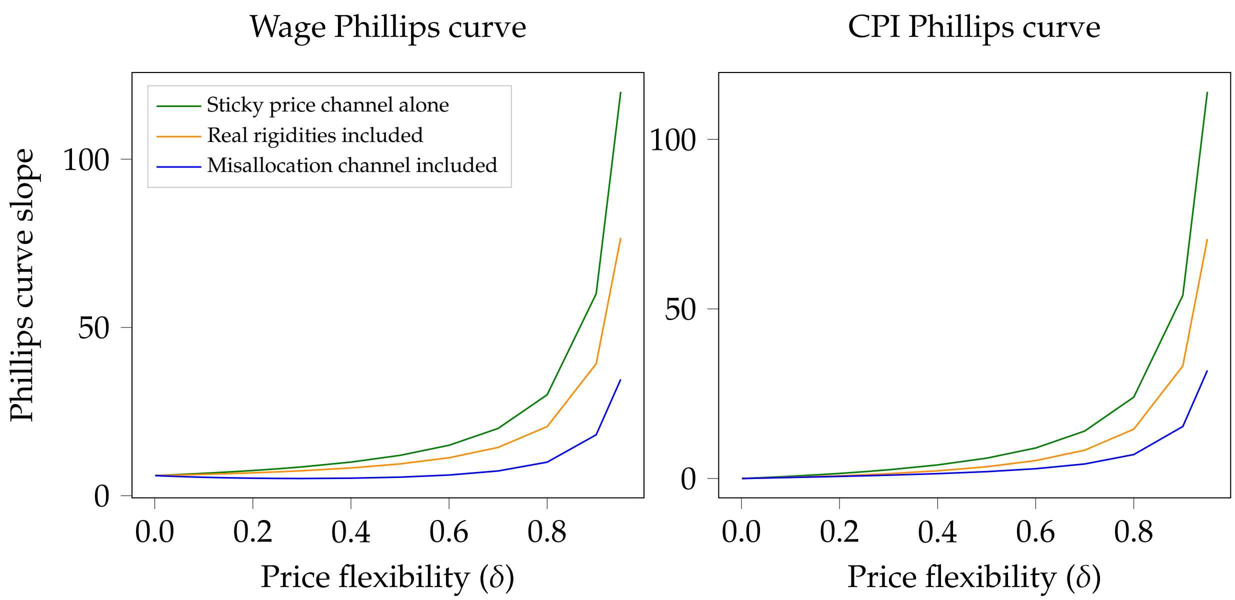


Figure D.2: Decomposition of Phillips curve slope, varying the degree of price stickiness δ.



D.2 Dynamic model: Additional results

Figure D.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 D.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.

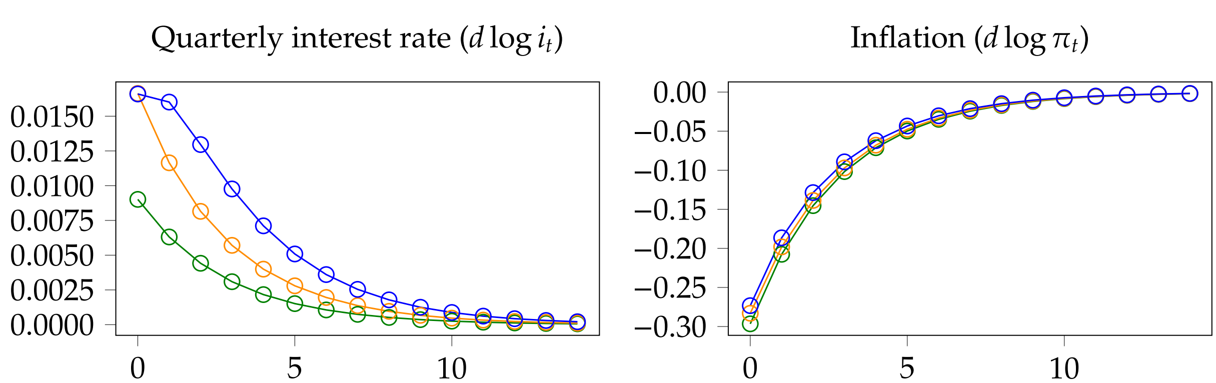
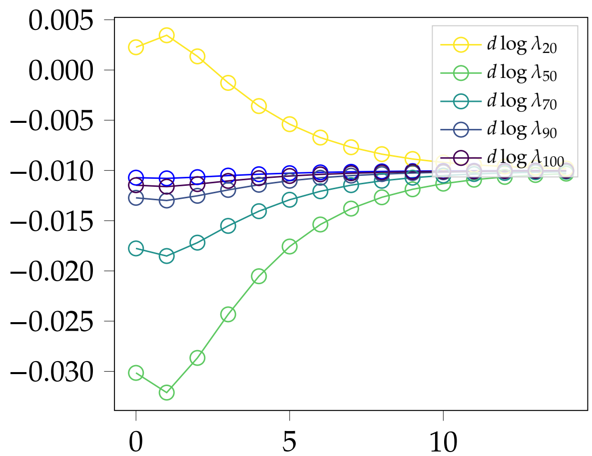


Figure D.4 shows the change in sales shares of different firm types following the 25bp contractionary monetary policy shock calibrated in the main text. The contractionary shock leads to an expansion in the sales of smaller firms and a contraction in the sales of larger firms.

Figure D.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.



E Money Supply Shocks

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

$$d \log M = d \log P^Y + d \log Y. \tag{32}$$

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

$$\Delta d \log M_t = \rho_m \Delta d \log M_{t-1} + \epsilon_t^m. \tag{33}$$

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 TFPR. 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 35% and increases the total output impact by 60% compared to the model with real rigidities alone.

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

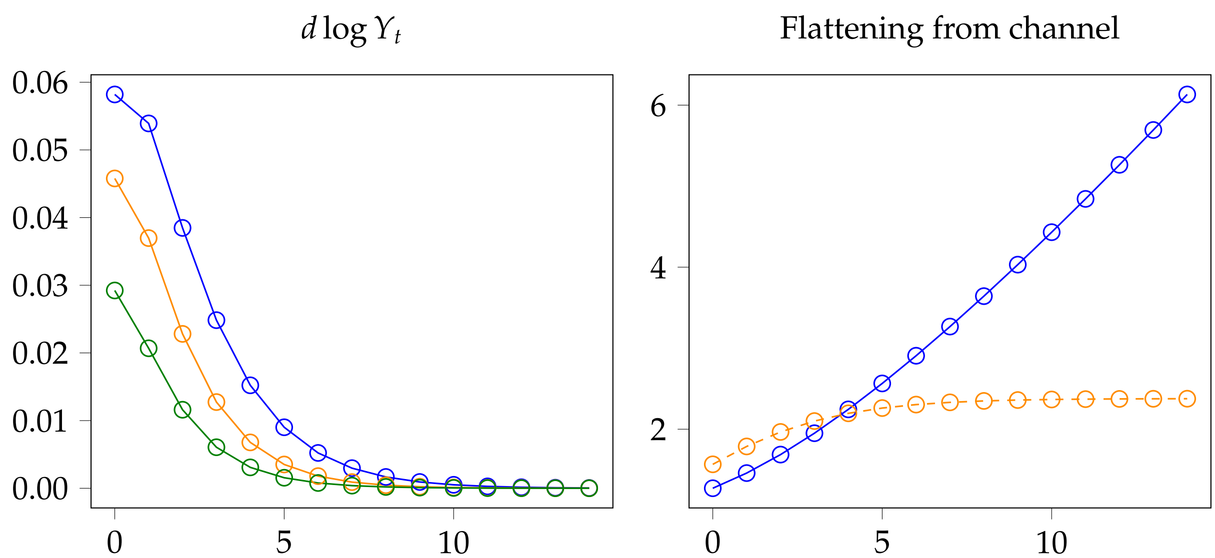
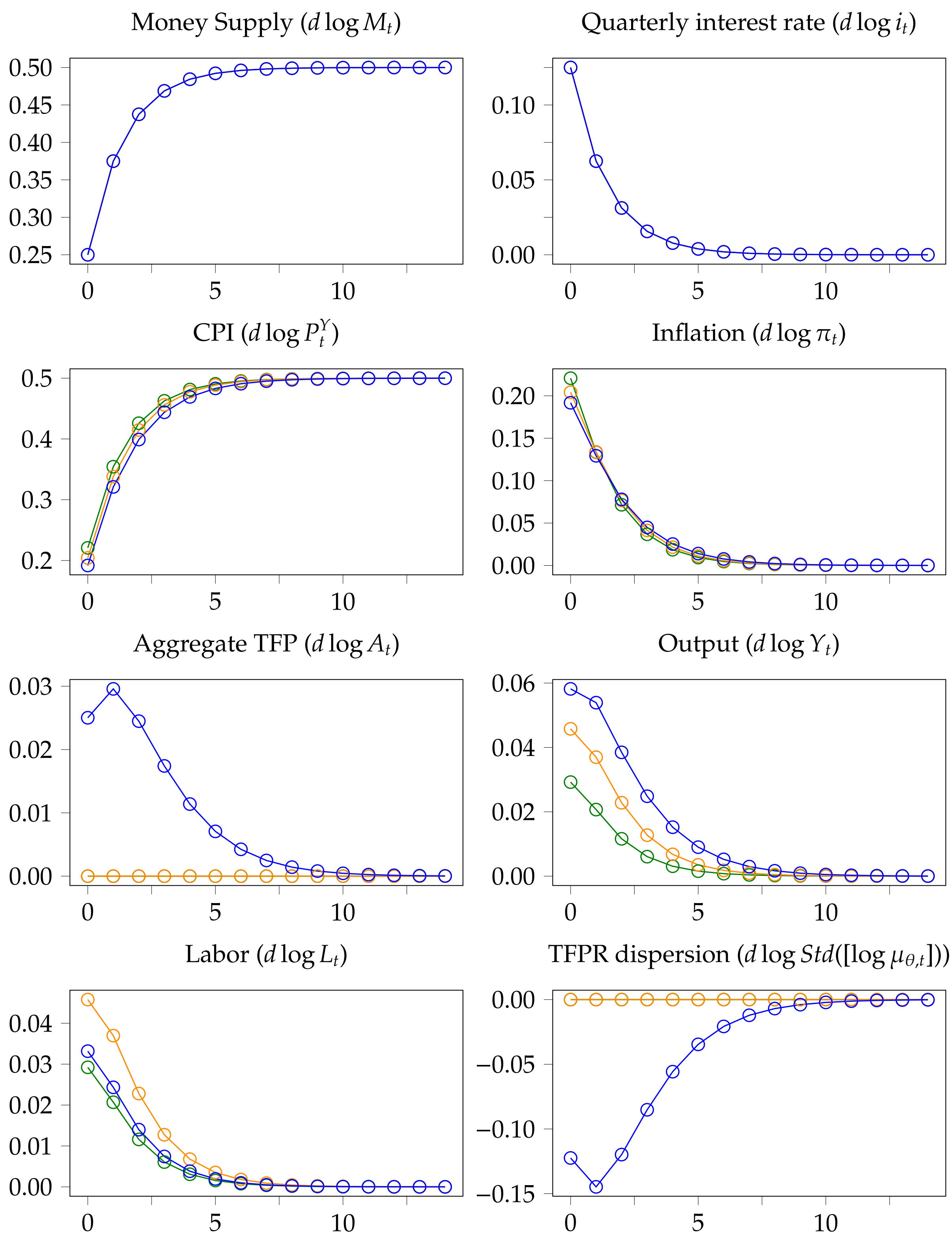


Table E.1: Effect of exogenous money supply shock on output. The cumulative output impact is calculated as in Alvarez et al. (2016).

Model	Output effect at $t = 0$	Half life	Cumulative output impact
CES	0.029	1.67	0.074
Homogeneous Firms	0.046	1.99	0.132
Heterogeneous Firms	0.058	2.69	0.212

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.



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,

$$\Omega_{ij} = \frac{p_j x_{ij}}{p_i y_i}, \tag{34}$$

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,

$$\tilde{\Omega}_{ij} = \frac{p_j x_{ij}}{\sum_l p_l x_{il}}, \tag{35}$$

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,

$$\Psi = (1 - \Omega)^{-1}, \tag{36}$$

$$\tilde{\Psi} = (1 - \tilde{\Omega})^{-1}. \tag{37}$$

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 $\lambda_{\mathcal{I}}$ to refer to the sales share of sector $\mathcal{I}$. 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 $\theta_{\mathcal{I}}$ is the elasticity of substitution between inputs for sector $\mathcal{I}$. We use the notation of the covariance operator $Cov_{\Omega^{(j)}}$ as defined in Baqaee and Farhi (2018).

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

$$d \log A_{\mathcal{I}} = \mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}} \left[\mu_\theta^{-1} \right] \frac{\mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}} [\delta_\theta] \left(1 - \mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}} [\delta_\theta] \right) Cov_{\frac{\lambda}{\lambda_{\mathcal{I}}}\delta} [\rho_\theta, \sigma_\theta] + \mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}\delta} [\rho_\theta] \, Cov_{\frac{\lambda}{\lambda_{\mathcal{I}}}} [\sigma_\theta, \delta_\theta]}{\mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}} [[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \, \sigma_\theta]} \cdot \left[\sum_{\mathcal{J}} \tilde{\Omega}_{\mathcal{I}J} \, d \log \frac{p_{\mathcal{J}}}{P} + d \log P \right]. \tag{38}$$

$$d\log\mu_{\mathcal{I}} = -\left[\frac{\mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}}\left[\delta_\theta(1-\rho_\theta)\right]\mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}}\left[\sigma_\theta(1-\delta_\theta)\right]}{\mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}}\left[\left[\delta_\theta\rho_\theta+(1-\delta_\theta)\right]\sigma_\theta\right]} + \mathbb{E}_{\frac{\lambda}{\lambda_{\mathcal{I}}}}\left[1-\delta_\theta\right]\right]\left[\sum_{\mathcal{J}}\tilde{\Omega}_{\mathcal{I}J}\, d\log\frac{p_{\mathcal{J}}}{P} + d\log P\right] + d\log A_{\mathcal{I}}. \quad (39)$$

The remaining aggregation equations follow directly from Baqaee and Farhi (2018). The change in output is:

$$d\log Y = \frac{1}{\sum_f \tilde{\Lambda}_f \frac{1+\gamma_f\zeta_f}{1+\zeta_f}}\left[\sum_{\mathcal{I}}\tilde{\lambda}_{\mathcal{I}}\left(d\log A_{\mathcal{I}} - d\log\mu_{\mathcal{I}}\right) - \frac{1}{1+\zeta_f}\sum_f \tilde{\Lambda}_f d\log\Lambda_f\right]. \quad (40)$$

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

$$\begin{aligned} d\log\lambda_{\mathcal{K}} = &\sum_{\mathcal{I}}\left(\delta_{\mathcal{K}\mathcal{I}} - \lambda_{\mathcal{I}}\frac{\Psi_{\mathcal{I}\mathcal{K}}}{\lambda_{\mathcal{K}}}\right)d\log\mu_{\mathcal{I}} \\ &+ \sum_{\mathcal{J}}(\theta_{\mathcal{J}}-1)\lambda_{\mathcal{J}}\mu_{\mathcal{J}}^{-1}Cov_{\tilde{\Omega}^{(\mathcal{J})}}\left(\sum_{\mathcal{I}}\tilde{\Psi}_{(\mathcal{I})}\left(d\log A_{\mathcal{I}} - d\log\mu_{\mathcal{I}}\right), \frac{\Psi_{(\mathcal{K})}}{\lambda_{\mathcal{K}}}\right) \\ &- \sum_{\mathcal{J}}(\theta_{\mathcal{J}}-1)\lambda_{\mathcal{J}}\mu_{\mathcal{J}}^{-1}Cov_{\tilde{\Omega}^{(\mathcal{J})}}\left(\sum_g \frac{\tilde{\Psi}_{(g)}}{1+\zeta_g}\left(d\log\Lambda_g + (\gamma_g\zeta_g-\zeta_g)\log Y\right), \frac{\Psi_{(\mathcal{K})}}{\lambda_{\mathcal{K}}}\right). \quad (41) \end{aligned}$$

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

$$\begin{aligned} d\log\Lambda_f = &-\sum_{\mathcal{I}}\lambda_{\mathcal{I}}\frac{\Psi_{\mathcal{I}f}}{\Lambda_f}d\log\mu_{\mathcal{I}} + \sum_{\mathcal{J}}(\theta_{\mathcal{J}}-1)\lambda_{\mathcal{J}}\mu_{\mathcal{J}}^{-1}Cov_{\tilde{\Omega}^{(\mathcal{J})}}\left(\sum_{\mathcal{I}}\tilde{\Psi}_{(\mathcal{I})}\left(d\log A_{\mathcal{I}} - d\log\mu_{\mathcal{I}}\right), \frac{\Psi_{(f)}}{\Lambda_f}\right) \\ &- \sum_{\mathcal{J}}(\theta_{\mathcal{J}}-1)\lambda_{\mathcal{J}}\mu_{\mathcal{J}}^{-1}Cov_{\tilde{\Omega}^{(\mathcal{J})}}\left(\sum_g \frac{\tilde{\Psi}_{(g)}}{1+\zeta_g}\left(d\log\Lambda_g + (\gamma_g\zeta_g-\zeta_g)\log Y\right), \frac{\Psi_{(f)}}{\Lambda_f}\right). \quad (42) \end{aligned}$$

Factor and sector prices follow:

$$d\log\frac{w_f}{P} = \frac{1}{1+\zeta_f}d\log\Lambda_f + \frac{1+\gamma_f\zeta_f}{1+\zeta_f}d\log Y, \quad (43)$$

$$d\log\frac{p_{\mathcal{I}}}{P} = -\sum_{\mathcal{K}}\tilde{\Psi}_{\mathcal{I}\mathcal{K}}\left(d\log A_{\mathcal{K}} - d\log\mu_{\mathcal{K}}\right) + \sum_f \tilde{\Psi}_{\mathcal{I}f}d\log\frac{w_f}{P}. \quad (44)$$

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.

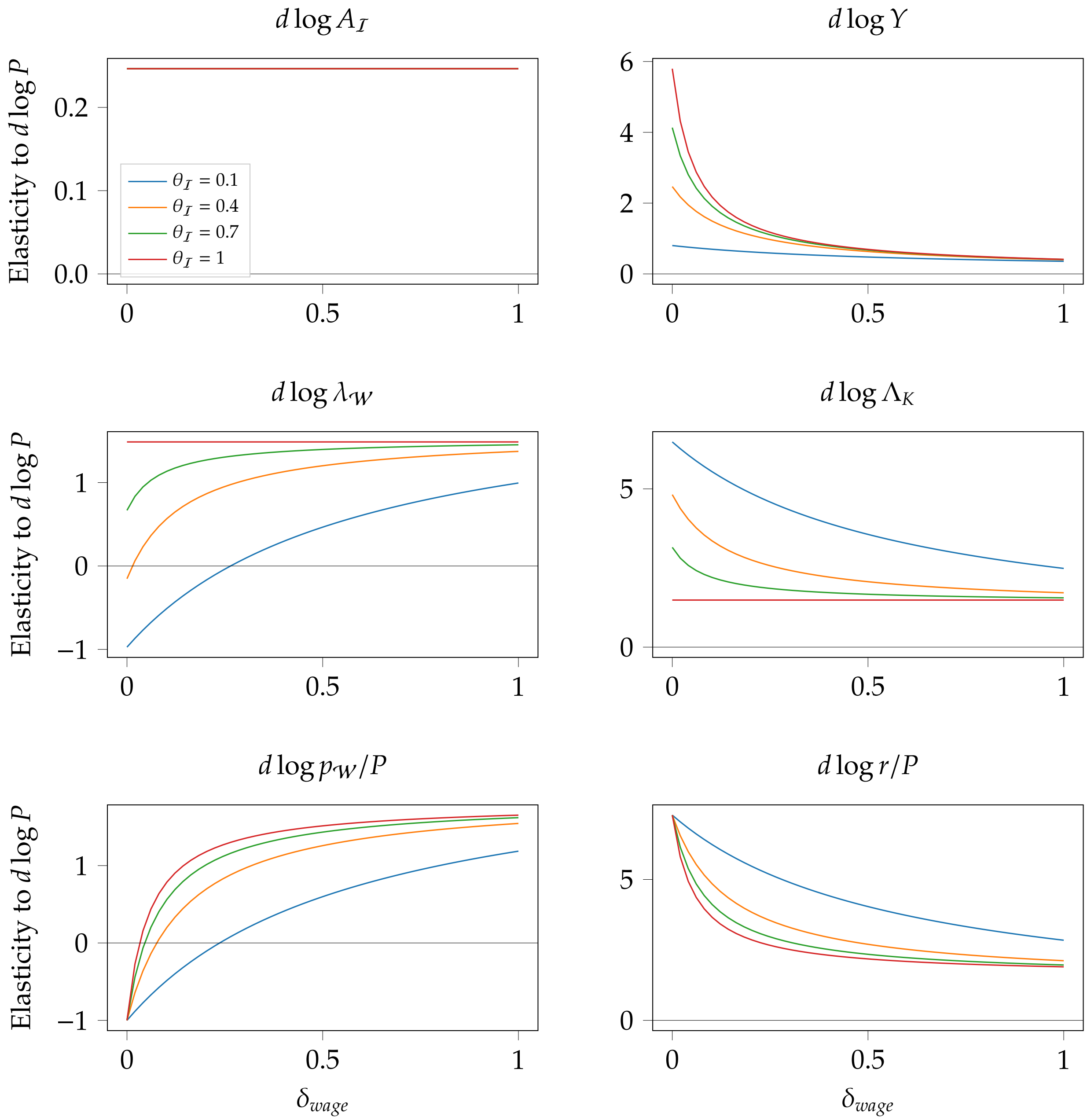
The industry sector consists of firms in monopolistic competition who use capital and labor provided by the labor union to produce varieties. Just as in the main text, firms in the industry sector have heterogeneous productivities and endogenous markups and pass-throughs; we use the same parameters and 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 $\theta_{\mathcal{I}}$, 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, ($d\log A_{\mathcal{I}}$), the change in output ($d\log Y$), the change in the shares of income to labor and capital ($d\log \lambda_{\mathcal{W}}$ and $d\log \Lambda_K$), and the real price of labor and capital ($d\log p_{\mathcal{W}/P}$ and $d\log r/P$) following a shock to the price level ($d\log P$).[6]

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).

[6]We 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.



G Klenow-Willis Calibration

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

$$\mu_\theta = \mu(\frac{y_\theta}{Y}) = \frac{1}{1 - \frac{1}{\sigma}(\frac{y_\theta}{Y})^{\frac{\epsilon}{\sigma}}}, \tag{45}$$

$$\rho_\theta = \rho(\frac{y_\theta}{Y}) = \frac{1}{1 + \frac{\epsilon}{\sigma - (\frac{y_\theta}{Y})^{\frac{\epsilon}{\sigma}}}} = \frac{1}{1 + \frac{\epsilon}{\sigma}\mu_\theta}. \tag{46}$$

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

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

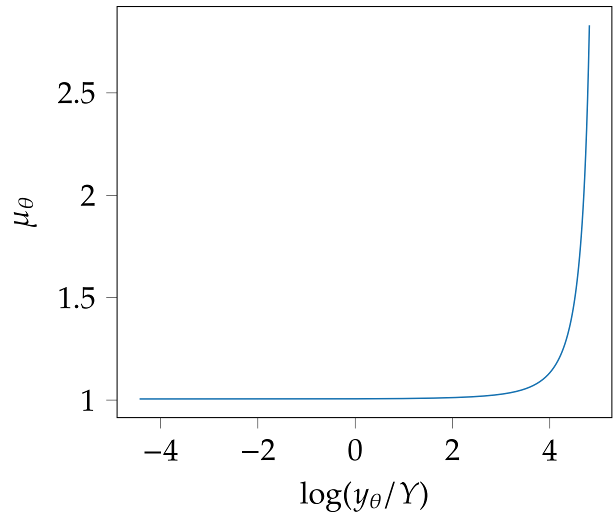
Amiti et al. (2019) estimate the average pass-through for the smallest 75% of firms in ProdCom is 0.97. In order to match the nearly complete pass-through for small firms, we must choose ϵ/σ to be around $0.01 - 0.03$.

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 $\Phi'(\cdot)$, incorporating additional modeling elements (such as demand shifters correlated with firm productivity) does not avoid the counterfactual properties shown here.

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.[7] The simulated distributions of firm pass-throughs and sales shares are shown in Figure G.2. Over the range of drawn

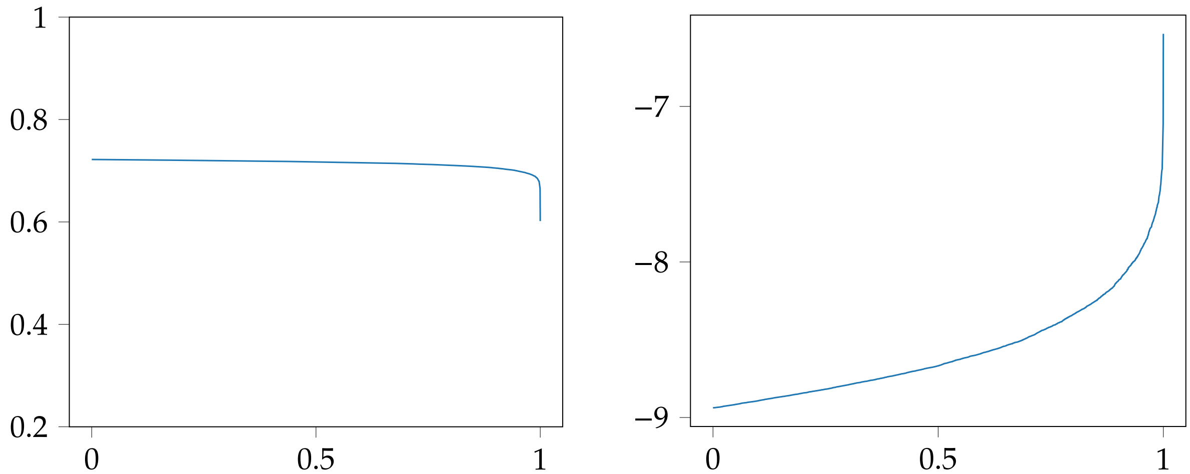
[7]We calibrate the model by drawing 1000 firms and finding a fixed point in output. Since the Pareto distribution is unbounded, we could theoretically draw firms with zero pass-throughs and infinite sales shares; the simulated distributions are bounded away from these extremes.

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



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. Because the model does not generate sufficient variation in pass-throughs, we find that the parametric specification dramatically understates the misallocation channel, compared to the nonparametric approach adopted in the main text.

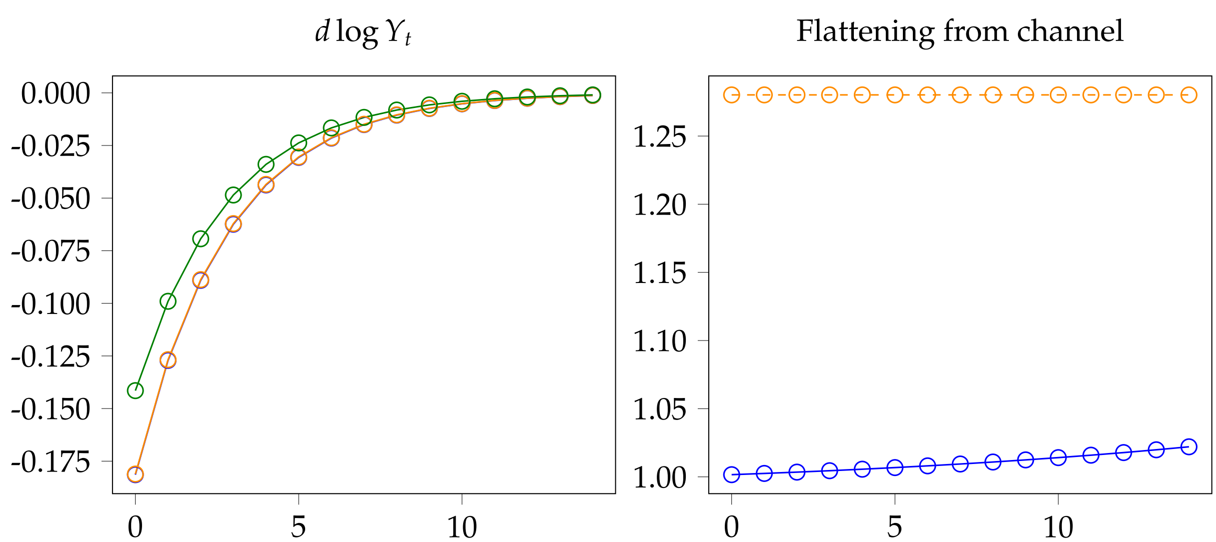
Figure G.2: Pass-through ρ_θ and sales share density $\log \lambda_\theta$ for Klenow and Willis (2016) calibration.



(a) Pass-through ρ_θ

(b) Log sales share $\log \lambda_\theta$

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.



H Oligopoly Model

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. With some abuse of notation, we now define the *effective expected equilibrium pass-through* of firm i, which we denote ρ_i, and which depends on desired pass-through ρ_i^{flex}, price stickiness δ_i, and industry market share s_i.

Lemma 3 (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 $\mu_{\mathcal{I}}$ 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*

$$d\log A = -\sum_{\mathcal{I}} \lambda_{\mathcal{I}} \sigma_{\mathcal{I}} Cov_{\frac{\lambda_i}{\lambda_{\mathcal{I}}}} \left(1 - \frac{\mu_i^{-1}}{\mu_{\mathcal{I}}^{-1}}, \mathbb{E}\left[d\log \mu_i\right] \right) - Cov_{\lambda_{\mathcal{I}}} \left(1 - \frac{\mu_{\mathcal{I}}^{-1}}{\mu^{-1}}, \mathbb{E}\left[d\log \mu_{\mathcal{I}}\right] \right),$$

where

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

and

$$\mathbb{E}\left[d\log \mu_{\mathcal{I}}\right] = -\mathbb{E}_{\frac{\lambda_i}{\lambda_{\mathcal{I}}}}\left[1-\rho_i\right] + \sigma_{\mathcal{I}} \mu_{\mathcal{I}} Cov_{\frac{\lambda_i}{\lambda_{\mathcal{I}}}} \left(\mu_i^{-1}, \rho_i \right) d\log w.$$

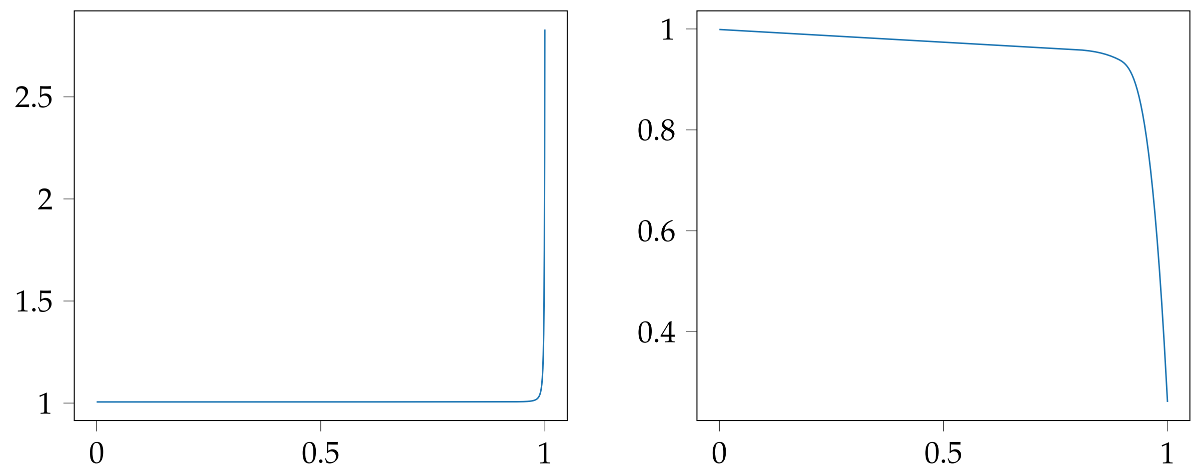
The first set of summands in $d\log A$ are changes in allocative efficiency due to reallocations within sectors, and the second set of summands are changes in allocative efficiecy due to reallocations across sectors. If sectoral markups are the same across all sectors, the second set of summands in $d\log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

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

$$\frac{d\log Y}{d\log w} = \rho \left[d\log A - \frac{\zeta}{1+\zeta} \sum_{\mathcal{I}} \frac{\lambda_{\mathcal{I}} \mu_{\mathcal{I}}^{-1}}{\mu^{-1}} \mathbb{E}\left[\frac{d\log \mu_{\mathcal{I}}}{d\log w} \right] \right].$$

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.

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



(a) Markup μ_i (b) Pass-through ρ_i

H.2 Calibration

To calibrate the model, we follow Amiti 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.[8]

We order firms by market share within sector, and plot the markups and pass-throughs of firms in Figure H.1.[9] The markups and pass-throughs generated by the nested CES 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%.

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

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

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

I Markups and Pass-through Variation Unrelated to Size

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,

$$\sigma_i = \underbrace{\mathbb{E}\left[\sigma_i|\lambda_i\right]}_{\sigma_\lambda} + \epsilon_i,$$

$$\rho_i = \underbrace{\mathbb{E}\left[\rho_i|\lambda_i\right]}_{\rho_\lambda} + \nu_i,$$

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,

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

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

$$\begin{aligned}
Cov_\lambda\left[\sigma_i, \rho_i\right] &= Cov_\lambda\left(\sigma_\lambda + \epsilon_i, \rho_\lambda + \nu_i\right) \\
&= Cov_\lambda\left(\sigma_\lambda, \rho_\lambda\right) + Cov_\lambda\left(\epsilon_i, \nu_i\right)
\end{aligned}$$

$$= Cov_\lambda\left(\sigma_\lambda, \rho_\lambda\right) + \underbrace{\sqrt{Var_\lambda\left(\epsilon_i\right) Var_\lambda\left(\nu_i\right)}\, Corr_\lambda\left(\epsilon_i, \nu_i\right)}_{\text{Bias}}.$$

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 demand shifters B_i (i.e., $\Phi_i(\cdot) = B_i\Phi(\cdot)$):

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

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,

$$\frac{d\log\sigma_i}{d\log B_i} = \frac{\partial\log\sigma(\frac{y}{Y})}{\partial\log\frac{y}{Y}}\rho_i\sigma_i$$
$$\frac{d\log\rho_i}{d\log B_i} = \frac{\partial\log\rho(\frac{y}{Y})}{\partial\log\frac{y}{Y}}\rho_i\sigma_i$$

Under Marshall's strong second law, $\frac{\partial\log\sigma(\frac{y}{Y})}{\partial\log\frac{y}{Y}} < 0$ and $\frac{\partial\log\rho(\frac{y}{Y})}{\partial\log\frac{y}{Y}} < 0$, hence $Corr(\epsilon_i, \nu_i) > 0$, 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 from Proposition 1 (assuming $\delta_i = \delta$ across firms):

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

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

$$\frac{d\log A^{true}}{d\log A} = 1 + \frac{1 - d\log A}{d\log A + \frac{Cov_\lambda\left(\sigma_\lambda, \rho_\lambda\right)}{\sqrt{Var_\lambda(\epsilon_i) Var_\lambda(\nu_i)} Corr_\lambda(\epsilon_i, \nu_i)}}.$$

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 $\frac{d\log A^{true}}{d\log A} \in (0.69, 1.27)$; 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%.

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.

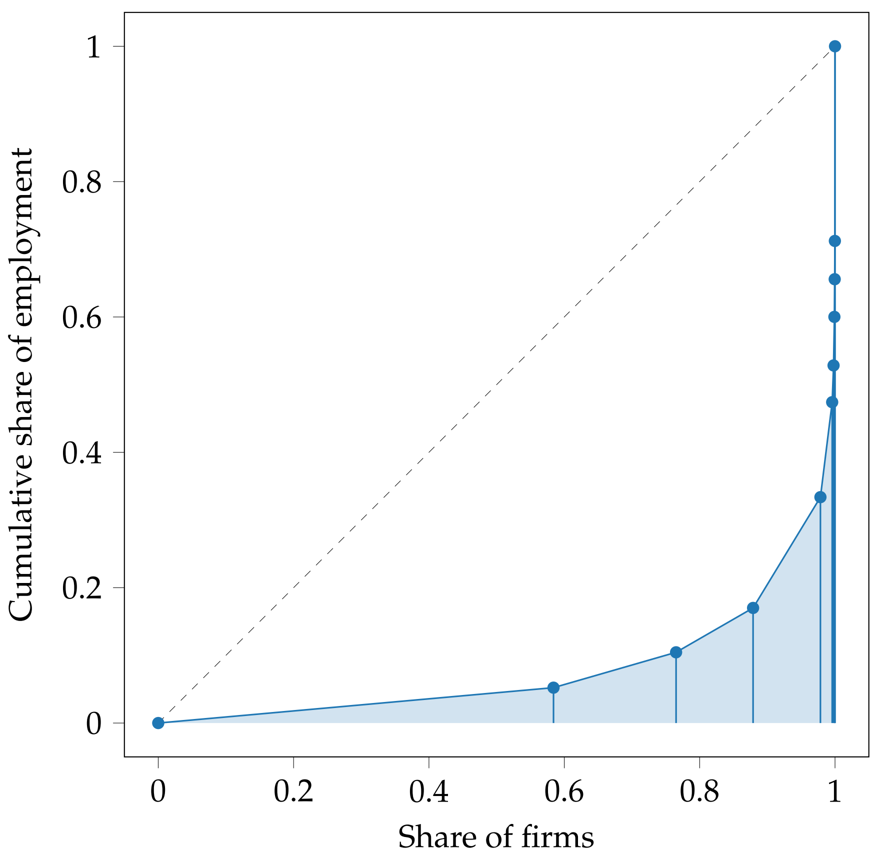


Figure J.2: Estimated Gini coefficients in Census BDS data from 1978-2018.