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 April 2021

Emmanuel Farhi tragically passed away in July, 2020. He was a one-in-a-lifetime friend and collaborator and we dedicate this paper to his memory. David Baqaee and Kunal Sangani are responsible for any errors that remain. We thank Andy Atkeson, Ariel Burstein, Oleg Itskhoki, 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 April 2021
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, monetary shocks have first-order effects on aggregate TFP. If high-markup firms have lower pass-throughs than low-markup firms, as is consistent with the empirical evidence, then a monetary easing generates an endogenous positive "supply shock" that amplifies the positive "demand shock" on output. The result is a flattening of the Phillips curve. This effect is distinct from another mechanism discussed at length in the real rigidities literature: a monetary easing leads to a reduction in desired markups be-cause of strategic complementarities in pricing. We derive a tractable four-equation dynamic model, disciplined by four sufficient statistics from the firm distribution, and use it to show that a monetary easing generates a procyclical hump-shaped response in aggregate TFP and countercyclical dispersion in firm-level TFPR. A calibration using firm-level pass-throughs suggests that the supply-side channel is quantitatively as important as real rigidities, and amplifies both the impact and persistence of monetary shocks. Moreover, this channel becomes stronger, and the Phillips curve becomes flatter, with increases in industrial concentration.

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
*NA user is deceased

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

A data appendix is available at http://www.nber.org/data-appendix/w28345

1 Introduction

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

Yet, aggregate total factor productivity (TFP), as measured by the Solow residual, is sensitive to nominal demand shocks (see, e.g. Evans, 1992). In fact, variations in monetary and fiscal policy explain between one-quarter and one-half of the observed movements in aggregate TFP at business cycle frequencies. This empirical finding is robust across time and across countries.[1] One interpretation of this result is that the relationship between measured productivity and demand shocks is confounded by capacity utilization or external returns, which bias the measurement of aggregate TFP.[2]

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

The effect of monetary policy on the cross-sectional allocation of resources yields a new channel for the transmission of monetary policy, which we term *the misallocation channel*. Under conditions matching empirical patterns on firms, monetary shocks generate pro-cyclical, hump-shaped movements in aggregate TFP, which match empirical estimates by Evans (1992), Christiano et al. (2005) and others.[3] The endogenous "supply shock" generated by the misallocation channel complements the traditional effects of the "demand shock" on employment and output. Incorporating the misallocation channel heightens the response of output to demand shocks and flattens the Phillips curve.

To a first-order, this supply-side effect only appears when two conditions hold: (1) the initial cross-sectional allocation of resources is inefficient, and (2) monetary policy systematically reallocates resources from low to high marginal revenue product firms. We discuss these two conditions in turn. First, if there is no initial misallocation, the marginal benefit of each input is equated across all competing uses. Therefore, starting at an efficient point, a reallocation of resources triggered by monetary policy has no first-

[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, and replicate the Evans (1992) result in each setting.

[2] See Basu et al. (2006) for a discussion of how capacity utilization can bias measurement of TFP.

[3] Christiano et al. (2005) estimate a positive hump-shaped response of labor productivity to monetary easing. In our one-factor model, labor productivity and aggregate TFP are the same.

order effects on aggregate productivity. Second, even if the initial allocation of resources is distorted, if monetary policy does not differentially affect firms with different marginal revenue products, then reallocations induced by monetary policy do not systematically raise or lower aggregate productivity.

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

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

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

Our first result is that when firms' pass-throughs covary negatively with their initial markups, then a positive demand shock, such as a monetary easing, increases aggregate TFP and moves the economy closer to the efficient frontier. This negative relationship

between markups and pass-throughs has strong empirical support across countries.[4] Intuitively, a monetary easing raises all firms' nominal marginal costs, but high-markup firms, which have lower pass-throughs, raise their prices by less than their low-markup counterparts. This triggers a reallocation toward high-markup firms and away from low-markup firms, which improves allocative efficiency. In principle, this heterogeneity in pass-throughs can be driven either by heterogeneity in desired pass-throughs or heterogeneity in price-stickiness.[5,6]

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

Our model suggests that the misallocation channel constitutes a quantitatively important part of monetary policy transmission mechanism. We use cross-sectional firm data from Belgium (provided by Amiti et al. 2019) to calibrate a simplified version of the model.[7] In our static model, we find that the misallocation channel reduces the slope of the Phillips curve by around 70%, compared to a model with demand-side effects alone. As a point of comparison, we find that real rigidities flatten the price Phillips curve by a similar amount.

Since the strength of real rigidities and the misallocation channel are governed by

[4]See Berman et al. (2012) in France, Chatterjee et al. (2013) in Brazil, Li et al. (2015) in China, Auer and Schoenle (2016) in the United States, and Amiti et al. (2019) in Belgium. We use estimates from Amiti et al. (2019) to calibrate the empirical results presented in this paper.

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

[6]We focus on monetary shocks but other demand shocks, such as discount factor shocks, will have similar effects on TFP.

[7]We follow Baqaee and Farhi (2020a) and solve a series of differential equations to back out the Kimball demand system from data on firm-level sales and pass-throughs. This is a virtue, since pass-throughs can be estimated using weaker assumptions than markups. This approach is also preferable to using an off-the-shelf functional form, since it does not impose the counterfactual restrictions baked in by parametric families of preferences. (We provide an explicit calibration exercise in Appendix F showing that off-the-shelf functional forms are incapable of simultaneously matching all the relevant sufficient statistics in the data.)

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 11%. This increase in the Gini coefficient is in line with the change in the firm employment distribution in the United States from 1978 to 2018.

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

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

Our model's predictions for the movement of both macro and micro measures of productivity have robust empirical support. As mentioned above, at the macroeconomic level, our setup predicts procyclical, hump-shaped responses of aggregate TFP to monetary shocks. At the microeconomic level, our model predicts countercyclical movements of dispersion in firm-level revenue productivity (TFPR).[8] Countercyclical dispersion in firm-level TFPR is documented by Kehrig (2011), among others.

Other related literature. This paper contributes to the large literature on the response of firms to monetary shocks. Our analysis is rooted in models of monopolistic competition

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

with staggered price setting originating in Taylor (1980) and Calvo (1983).

A strand of this literature is devoted to explaining the strength and persistence of monetary policy shocks, which cannot be explained by nominal rigidities alone given the frequency of price adjustment.[9] Ball and Romer (1990) introduce real rigidities, which complement nominal rigidities to increase monetary nonneutrality.[10] A common formulation of real rigidities is incomplete pass-through, where firms are slow to reflect marginal cost shocks in their prices due to strategic complementarities in pricing. Incompleteness of pass-through is documented empirically by Gopinath et al. (2010) and Gopinath and Itskhoki (2011). Our paper complements this literature by showing that incomplete pass-through, when paired with firm-level heterogeneity, provides another mechanism through which monetary policy can affect output by changing allocative efficiency.

In describing changes in the allocative efficiency of the economy, we also relate to a vast literature on cross-sectional misallocation, which includes Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and Baqaee and Farhi (2020b). For the most part, the misallocation literature is concerned with steady-state or long-run changes in misallocation, whereas we are focused on characterizing short-run changes in misallocation following nominal shocks. Some important exceptions are Cravino (2017), Baqaee and Farhi (2017), and Meier and Reinelt (2020). In an international context, Cravino (2017) shows that heterogeneity in exporters' invoicing currency and desired markups (due to fixed transport costs), coupled with nominal rigidities, implies that exchange rate changes can affect domestic productivity by changing the allocation of resources. Baqaee and Farhi (2017) provide a general framework for how allocative efficiency changes in general equilibrium and apply their results to show that if price-stickiness positively covaries with markups, then monetary policy affects TFP. Meier and Reinelt (2020) provide empirical support for this covariance, and offer a microfoundation where firms have heterogeneous Calvo parameters, so firms with more rigid prices endogenously set higher markups due to a precautionary motive. Our analysis complements, and to some extent unifies, these previous analyses by showing how heterogeneity in realized pass-throughs (driven either by variable stickiness or variable desired pass-throughs) can cause nominal shocks to have effects on productivity.[11]

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

[10] 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 monetary shocks. Eichenbaum and Fisher (2004) and Dotsey and King (2005), for example, investigate how relaxing assumptions of constant elasticities of demand interact with other frictions to generate persistence. Klenow and Willis (2016) compare the predictions of models where real rigidities are generated by a kinked demand curve versus sticky intermediate prices.

[11] Productivity shocks can also affect allocative efficiency: David and Zeke (2021) show that allocative

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

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

Structure of the paper. Section 2 introduces a simple static model and defines the equilibrium. Sections 3 and 4 describe the response of aggregate TFP and output to a monetary shock in the one-period model and elaborate on the findings using a few simple examples. Section 5 generalizes the static model from the previous sections to a fully dynamic setting, which yields a four-equation New Keynesian model with misallocation. Section 6 provides a quantitative illustration of the importance of the misallocation channel. In Section 7, we summarize some extensions discussed in more detail in the appendices, including a model with multiple sectors, multiple factors, input-output linkages, and sticky wages, as well as a calibration in a setting with oligopolistic, rather than monopolistic, competition. Section 8 concludes.

2 Model

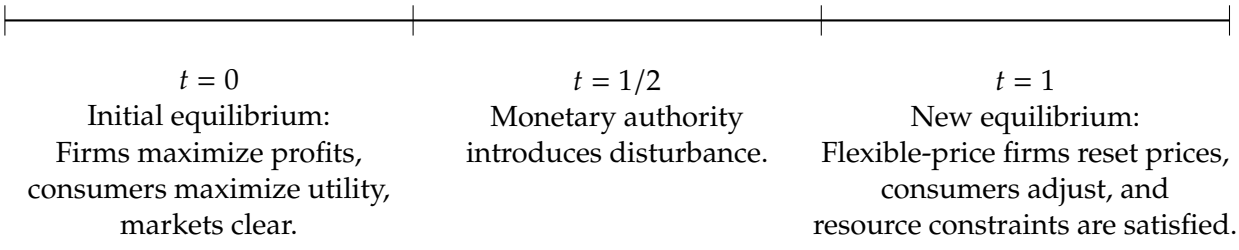
We start with a simple model with a single factor, labor. To build intuition, we first consider a one-period model to highlight the mechanism driving changes in allocative efficiency. In Section 5, we consider the fully dynamic model, which generalizes the findings shown here.

The timing is as follows. At time $t = 0$, the economy is in an equilibrium: households choose consumption and labor to maximize utility, firms choose prices to maximize profits, and markets clear. The monetary authority then introduces an unexpected monetary

efficiency varies over the business cycle when firms have heterogeneous exposure to aggregate shocks.

disturbance into the economy. At time $t = 1$, firms with flexible prices reset prices to maximize profits, while firms with sticky prices keep prices unchanged from the initial equilibrium. Households adjust consumption and labor to maximize utility.

Figure 1: One-period model timing.



2.1 Setup

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

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

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

where $1/\gamma$ is the intertemporal elasticity of substitution, and ζ is the Frisch elasticity of labor supply. The consumption bundle Y consists of different varieties of goods indexed by $\theta \in [0, 1]$. Consumers have homothetic preferences over these final goods, and the utility from the consumption bundle Y is defined implicitly by[12]

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

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

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

The representative consumer maximizes utility subject to the budget constraint

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

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

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

where the *price aggregator* P is defined as

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

and P^Y is the ideal price index.[13] As we can see in Equation (1), relative demand for a variety θ is dictated by the ratio of its price to the price aggregator P. Hence, firms compete with the rest of the market via a single price and quantity aggregator. Equation (1) also illustrates the appeal of these preferences: we can create downward-sloping demand curves of any desired shape by choosing the aggregator Υ_θ.

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

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

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

taking as given its residual inverse-demand curve.

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

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

facing a firm of type θ:

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

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

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

where the markup function is given by the Lerner formula,[14]

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

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

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

where the second subscript denotes the period.

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

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

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

[14] We assume that marginal revenue curves are downward-sloping.

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

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

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

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

We write $\mathbb{E}$ to denote $\mathbb{E}_x$ when $x_\theta = 1$ for all θ. The operator $\mathbb{E}_x$ operates a change of measure by putting more weight on types θ with higher values of x_θ. We denote the sales share density by[15]

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

and define the *aggregate markup* to be

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

In words, this is the harmonic sales-weighted average of markups.

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

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

3 Productivity Response

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

3.1 Allocative Efficiency, TFP, and TFPR

Recall that firms with market power set high prices and markups by restricting output. Compared to an economy with no dispersion in markups, an economy with heterogeneous markups features a distorted allocation of resources across firms: firms with higher markups are inefficiently small, and thus capture a lower share of resources than firms with lower markups. We refer to the change in *allocative efficiency* of an economy as the difference in output due to changes in the cross-sectional allocation of resources across firms, holding fixed the total supply of those resources.[16]

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

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

Here, $d \log A$ is the change in the distortion-adjusted Solow residual, or equivalently, the change in aggregate labor productivity.[17] We refer to $d \log A$ as the change in aggregate TFP. The following lemma, which is an application of the main result in Baqaee and Farhi (2020b), shows how aggregate TFP is related to changes in markups.

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

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

Recall that $\bar{\mu}$ is the (harmonic) average markup. Therefore, Proposition 1 shows that allocative efficiency increases when the average of markups rises more than markups on

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

[17] The distortion-adjusted Solow residual weighs the change in each factor by its share of total factor costs, rather than its share in aggregate income. See Hall (1988) and Baqaee and Farhi (2020b) for more information on why the Solow residual must be corrected in the presence of markups.

average. This can only happen if there is a composition effect whereby high-markup firms expand their use of inputs relative to low-markup firms. Since these reallocations occur due to changes in markups, (5) can also be viewed as a measure of the change in the dispersion in markups. In particular, if high-markup firms expand relative to low-markup firms, this must be because their markups are falling relative to those of low-markup firms — that is, $d \log A$ is positive if the distribution of markups is being compressed, and the allocation is moving closer to the efficient frontier.

Therefore, Equation (5) links TFP at the economy level to the dispersion in markups at the firm level. It is useful to note a connection here with a literature that has documented cyclical variation in the dispersion of plant- and firm-level revenue productivity, or TFPR.[18] When production has constant returns to scale, variation in markups is exactly equal to variation in TFPR. Hence, an improvement in aggregate TFP driven by a compression of the markup distribution will also imply that TFPR dispersion should be countercyclical. Specifically, TFPR is usually measured by subtracting input growth from revenue growth. In our model, this is just

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

where changes in the wage $\Delta \log w$ do not vary by firm. As we will see, our model predicts that a monetary easing can simultaneously increase aggregate TFP and reduce dispersion in TFPR.[19]

A corollary of Lemma 1 is that when markups are not dispersed in the initial equilibrium, first-order changes in aggregate productivity are necessarily zero.[20]

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

$$d \log A = 0,$$

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

Corollary 1 can also be derived as a consequence of the envelope theorem: when markups are identical across firms, the cross-sectional allocation of resources is efficient. Hence, changes in aggregate TFP due to reallocations are zero to a first order. This

[18] See Foster et al. (2008) for more on the relationship between TFPR and "physical productivity" (TFPQ, or A_θ in our setting).

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

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

confirms that in models without initial markup dispersion, the response of aggregate TFP to monetary shocks is zero to a first-order, regardless of how markups respond.

3.2 Productivity Response: Two Mechanisms

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

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

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

where $Cov_\lambda\left[\rho_\theta, \sigma_\theta|\text{flex}\right]$ is the covariance of pass-throughs and elasticities for the subset of flexible-price firms,[21] *and κ_ρ and κ_δ are positive constants*

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

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

However, dispersion in markups is not sufficient for monetary policy to affect aggregate productivity. It must also be the case that markups (or equivalently, σ_θ) covary systematically with either desired pass-throughs or price-stickiness. Either of these two mechanisms cause realized pass-throughs to covary with the level of the markups, and as long as realized pass-through covaries negatively with the level of markups, an increase in nominal marginal costs will result in productivity-increasing reallocations.[22] To build more intuition, we now consider each of the two mechanisms mentioned above in isolation.

Mechanism I: heterogeneous pass-through. If price-stickiness is homogeneous across firms ($\delta_\theta = \delta$), then the second covariance term in Proposition 1 is zero, and the productiv-

[21] The reader may note that this is equivalent to $Cov_{\lambda\delta}\left[\rho_\theta, \sigma_\theta\right]$.

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

ity response depends on the covariance between desired pass-throughs ρ_θ and elasticities σ_θ alone:

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

$$\frac{d \log A}{d \log w} = \kappa_\rho Cov_\lambda \left[\rho_\theta, \sigma_\theta\right].$$

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

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

Definition 1. *Marshall's strong second law of demand* requires that desired markups are increasing in quantity and desired pass-throughs are decreasing in quantity. That is,[23]

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

If Marshall's strong second law holds, then markups are increasing and pass-throughs are decreasing in firm size. This guarantees that monetary expansions will raise aggregate productivity. Marshall's second law of demand has strong empirical support (see, for example, empirical estimates of pass-throughs by firm size from Amiti et al. 2019).[24]

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

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

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

where $d \log P$ is the change in the price aggregator defined in Equation (2). The optimal price of a firm with a high pass-through moves closely with shocks to marginal cost from

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

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

the nominal wage. Firms with low pass-through, on the other hand, exhibit "pricing-to-market" behavior: they place less weight on their own marginal cost, and more weight on the expected aggregate price level in the economy. Sticky-price firms, of course, cannot adjust their prices after observing the nominal wage shock.

Following an increase in the nominal wage, flexible-price firms shrink and sticky-price firms, whose prices are kept artificially low, expand[25]:

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

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

The former effect dominates when $Cov_\lambda\left[\rho_\theta, \sigma_\theta\right] > 0$. When this is the case, firms with high markups also have low pass-throughs, allowing them to cut prices and stay competitive. As a result, the allocation of output shifts toward high-markup firms in aggregate. Since high-markup firms are initially too small relative to the efficient cross-sectional allocation, the expansion of high-markup firms boosts allocative efficiency and hence aggregate TFP.

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

Corollary 3. *If desired pass-through is homogeneous across firms ($\rho_\theta = \rho$),[26] then*

$$\frac{d\log A}{d\log w} = \kappa_\delta Cov_\lambda\left[\sigma_\theta, \delta_\theta\right]. \tag{7}$$

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

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

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

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

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

3.3 Discussion

To recap, as long as realized pass-throughs covary negatively with initial markups, aggregate productivity responds procyclically to monetary shocks. Furthermore, expansionary and contractionary monetary shocks lower and raise cross-sectional TFPR dispersion respectively. Indeed, changes in aggregate productivity are the consequences of the changes in the cross-sectional distribution of firm-level TFPR. Both predictions are borne out in the data: Evans (1992), among others, finds a procyclical response of aggregate productivity to demand shocks, and Kehrig (2011) documents countercyclical dispersion in revenue productivity. In our model, both patterns are linked to the reallocation of resources toward high-markup firms triggered by the demand shock.

4 Output Response and the Phillips Curve

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

This section is organized as follows. We first describe the response of output to a monetary shock. Then, we describe the slope of the Phillips curve and formalize two channels—real rigidities and the misallocation channel—which flatten the slope of the

Phillips curve relative to the benchmark model. Finally, to gain intuition, we compute the slope of the Phillips curve in a few simple example economies.

4.1 Output Response

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

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

$$d\log Y = \underbrace{\frac{1}{1+\gamma\zeta} d\log A}_{\textit{Supply-side effect}} - \underbrace{\frac{\zeta}{1+\gamma\zeta}\mathbb{E}_\lambda\left[d\log\mu_\theta\right]}_{\textit{Demand-side effect}}, \tag{8}$$

where

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

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

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

The sticky price effect is exacerbated by real rigidities, which arise from imperfect pass-through. When $\mathbb{E}_\lambda\left[\rho_\theta\right] < 1$, flexible-price firms increase prices less than one-for-one with the marginal cost shock. As a result, the markups of flexible-price firms also fall. Together, the reduction in the markups of both sticky-price and flexible-price firms increase labor demand, which spurs employment and output.

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

Interestingly, whereas the demand-side effect is increasing in the size of the elasticity of labor supply ζ, the supply-side effect is decreasing in ζ. In fact, the supply-side effect is

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

4.2 Flattening of the Phillips Curve

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

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

Both are presented in Proposition 3.

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

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

The price Phillips curve is given by

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

Note that the supply-side effect flattens both the price and wage Phillips curves. The non-parametric decomposition of the output response in Proposition 2 allows us to quantify the amount of flattening caused by real rigidities and the misallocation channel relative to the CES baseline.

We calculate the flattening of the Phillips curve due to real rigidities by dividing the slope of the Phillips curve due to sticky-prices alone and by the slope of the Phillips curve due to sticky prices and real rigidities. Since real rigidities flatten the Phillips curve, this ratio is greater than. If this quantity is, say, 1.5, this means that incorporating real rigidities decreases the responsiveness of output to the price level by 50%. Similarly, we calculate the flattening of the Phillips curve due to misallocation channel by comparing the slope of the Phillips curve in due to sticky prices and real rigidities by the slope of the Phillips curve where we account for changes in allocative efficiency. As long as $\frac{d \log A}{d \log w} > 0$, this

quantity is also greater than one.

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

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

$$\textit{Flattening due to real rigidities} = 1 + \frac{\mathbb{E}_\lambda[\sigma_\theta]\,\mathbb{E}_\lambda[1-\rho_\theta]}{\delta Cov_\lambda[\rho_\theta,\sigma_\theta] + \mathbb{E}_\lambda[\rho_\theta]\,\mathbb{E}_\lambda[\sigma_\theta]}. \tag{10}$$

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

$$\textit{Flattening due to the misallocation channel} = 1 + \frac{\bar{\mu}}{\zeta}\frac{\delta Cov_\lambda[\rho_\theta,\sigma_\theta]}{\delta Cov_\lambda[\rho_\theta,\sigma_\theta] + \mathbb{E}_\lambda[\sigma_\theta]}. \tag{11}$$

In Equation (10), we see that the flattening of the price Phillips curve due to real rigidities increases as average pass-throughs fall.The flattening due to real rigidities in (10) is also decreasing in the price flexibility δ. As price flexibility increases, the price aggregator moves more closely with shocks to marginal cost; hence the "pricing-to-market" effect from incomplete pass-throughs is less powerful.

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

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

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

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

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

This is the traditional New Keynesian Phillips Curve.[27] Nominal rigidities, captured by the Calvo parameter $\delta < 1$, flatten the Phillips curve. As δ approaches one, prices become perfectly flexible and the Phillips curve becomes vertical.

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

The homogeneous firms in this economy have identical markups, $\mu_\theta = \mu$, and pass-throughs, $\rho_\theta = \rho$. By deviating from CES, however, we allow firms' desired pass-throughs to be incomplete, i.e., $\rho < 1$.

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

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

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

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

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

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

[27] 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.

Reallocations in inputs (labor in our single factor model) across high- and low-markup firms, paired with the initial distribution of sales and markups, determines the change in aggregate TFP. In particular, aggregate TFP increases if the growth in employment at high-markup firms outpaces the growth of employment at low-markup firms.[28]

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

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

This price Phillips curve is flatter than the CES economy if $\rho_L > \rho_H$, i.e., if low-markup firms have higher pass-throughs than high-markup firms. An increase in the covariance of elasticities and pass-throughs, $\left(\sigma_L-\sigma_H\right)\left(\rho_L-\rho_H\right)$, 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.[29] This means that industrial concentration plays a role in shaping the Phillips curve. We consider this in more detail in our empirical calibration, 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

[28] As an illustration, suppose markups are positively correlated with firm size, and consider the cyclical differences in employment growth recorded by Moscarini and Postel-Vinay (2012). They observe two groups of firms with comparable shares of total employment (firms with fewer than 50 employees and greater than 1000 employees) and find that the differential growth rate of big and small firms is about 5% over the business cycle. For a back-of-the-envelope calculation: if small firms have 10% markups, large firms have 40% markups, and the total sales of both groups are equal ($\lambda_S = \lambda_B = 0.5$), then this estimate implies aggregate productivity moves by 0.3% over the business cycle due to these reallocations. Of course, we can do much better than this illustrative calculation by using moments from firm size distribution; we undertake a detailed calibration in Section 6.

[29] When nominal rigidities across firms are homogeneous, the sales-weighted average elasticity $\mathbb{E}_\lambda\left[\sigma_\theta\right]$, the sales-weighted average desired pass-through $\mathbb{E}_\lambda\left[\rho_\theta\right]$, the covariance of elasticities and desired pass-throughs $Cov_\lambda\left[\sigma_\theta, \rho_\theta\right]$, and the aggregate markup $\bar{\mu}$ are sufficient statistics of the firm size distribution to compute all results.

Marshall's second law of demand, and hence yields similar implications (we show this in Appendix G). In the body of the paper we focus on the monopolistic competition model because monopolistic competition is much more tractable in a fully dynamic environment.

5 Four-Equation Dynamic Model

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

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

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

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

where r_{t+j} is the interest rate, δ_i is the Calvo parameter, $y_{i,t+k}$ is the quantity firm i sells in period $t+k$ if it last set its price in period t. The solution to the firms' maximization problem describes how prices move in the economy. We can describe the movement of inflation, output, and aggregate TFP by aggregating across firms.

5.1 The New Keynesian Model with Misallocation

The standard model, augmented to include real-rigidities and endogenous TFP, is presented in Proposition 5.

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

Taylor rule.

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

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

Dynamic IS equation.

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

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

New Keynesian Phillips curve with misallocation.

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

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

Endogenous TFP equation.

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

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

Note that the actions of the monetary authority and of households are unchanged vis á vis the standard model. Hence, the Taylor Rule and Dynamic IS equations are the same as the workhorse three-equation model.

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

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

[31] We find that $\alpha > 0$ when $\mathbb{E}_\lambda\left[\rho_\theta\right] > \frac{\bar{\mu}^{-1}\zeta}{1+\bar{\mu}^{-1}\zeta}$. 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.

Finally, the path of aggregate TFP is pinned down by (14). Under Marshall's Second Law of Demand, the covariance term in the equation is positive, and changes in output $d \log Y$ are positively associated with aggregate TFP. Note that, unlike the standard New Keynesian model's equations, which are first-order difference equations, aggregate TFP follows a second-order difference equation. As a result, the augmented four-equation model can generate hump-shaped impulse responses to monetary shocks.

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

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

5.2 Solution Strategy

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

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

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

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

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

When prices are fully flexible, this simplifies to just

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

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

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

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

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

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

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

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

where

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

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

For instance, by taking the sales-weighted expectation of both sides in Equation (17), we can recover the movement of the consumer price index.[32] We get,

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

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

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

$$d\log P^Y_t - d\log w_t = \frac{1}{\zeta}\left[d\log A_t - (1+\gamma\zeta)d\log Y_t\right] \quad (20)$$

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

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

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

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

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

5.3 Discussion

The model presented in Proposition 5 provides a tractable, four-equation system that can be used to calibrate economies with realistic heterogeneity in markups and pass-throughs. This model incorporates real rigidities and the misallocation channel while being parsimoniously governed by four objects from the firm distribution: the average sales-weighted elasticity $\mathbb{E}_\lambda[\sigma_\theta]$, the average sales-weighted pass-through $\mathbb{E}_\lambda[\rho_\theta]$, the covariance of elasticities and pass-throughs $Cov_\lambda[\sigma_\theta, \rho_\theta]$, and the aggregate markup $\bar{\mu}$. We present one such calibration in the section below.

The second-order difference equation for aggregate TFP generates hump-shaped patterns for $d\log A$ and potentially other aggregate variables. Empirical estimates of the impulse response of labor productivity to monetary shocks (see, e.g., Christiano et al. 2005) exhibit this shape. This may be preferable to models that rely on habit persistence to achieve hump-shaped impulse responses, as habit persistence generates counterfactually large swings in labor supply, wages, and the real interest rate following large shocks (see the discussion in Jappelli and Pistaferri 2010).

For the purposes of our discussion, we have focused on monetary policy shocks. However, the four-equation model introduced here can also accommodate other demand shocks, such as discount rate shocks or expansionary fiscal policy. These demand shocks

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

would trigger similar reallocations across firms and hence also raise aggregate TFP, like the monetary shocks highlighted here.

6 Quantitative Illustration

We now calibrate both the static and dynamic versions of the model to assess the quantitative importance of the misallocation channel. This section is organized as follows. First, we describe how a model like ours can be calibrated. Second, we calibrate the model using empirical pass-through estimates from Amiti et al. (2019) with Belgian firm-level data. We then report results from this calibration exercise: we compute the flattening of the Phillips curve due to real rigidities and misallocation in the static model and discuss comparative statics with respect to the Frisch elasticity and the degree of industrial concentration. At the end of the section, we turn to the dynamic model, where we present impulse response functions following a monetary policy shock.

6.1 Non-parametric Estimation Procedure

It may be tempting to use an off-the-shelf functional form for the Kimball aggregator and tune parameters to match moments from the data. However, there is no guarantee that parametric specifications of preferences are able to match the relevant features of the data required for generating correct aggregate properties.[34] Instead, we follow Baqaee and Farhi (2020a) and back out the shape of the Kimball aggregator non-parametrically from the data. We summarize this approach below.

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

[34] As an example, see Section 7 for a discussion of the unsuitability of the popular parametric family of preferences considered by Klenow and Willis (2016) for our application.

in the cross-section, markups and sales must satisfy the following differential equation[35]

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

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

6.2 Data and Parameter Values

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

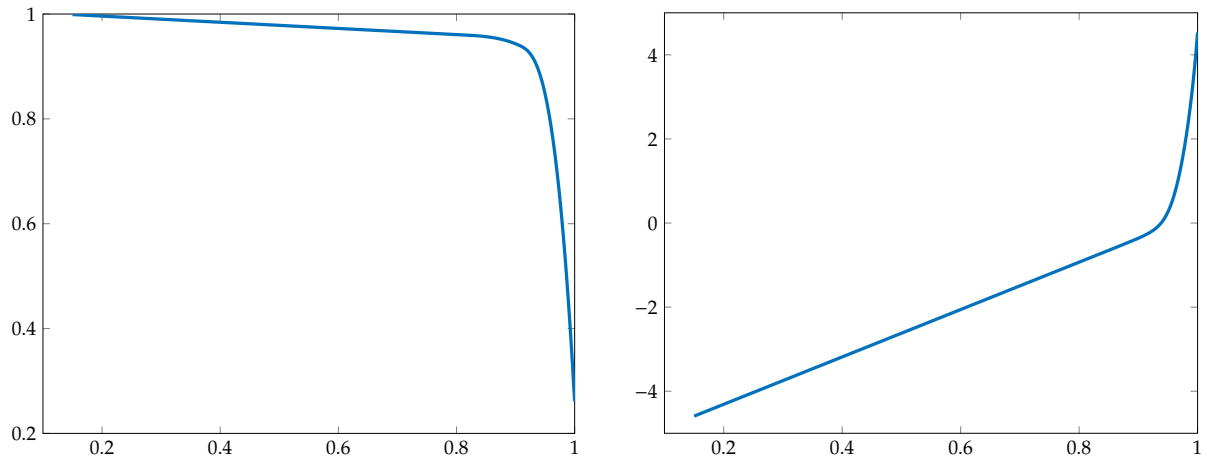
To impute the distribution of markups and elasticities from this data, our estimation procedure requires that we take a stance on the average markup. We assume that the average markup $\bar{\mu} = \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1} = 1.15$, in line with estimates from micro-data.[36]

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

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

[36] 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 an earlier version of their paper, Amiti et al. (2019) 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 Philippon (2017) and Barkai (2020), who estimate average markups on the order of 10-20%, but lower than estimates by De Loecker et al. (2020), who estimate the average markup for public firms at 61%.

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 $\log \lambda_\theta$

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

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

6.3 Results from Static Model

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

To highlight the key forces at play in this calibration, we consider how these estimates

Table 1: Parameters for empirical calibration.

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

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

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

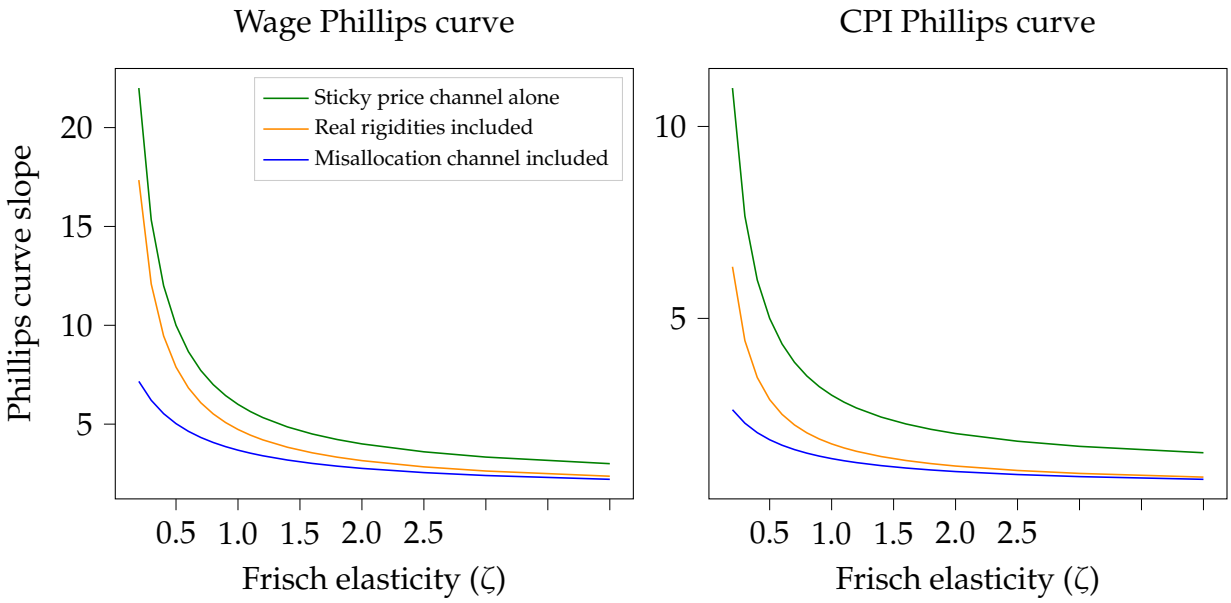
change as we vary the Frisch elasticity, the degree of industrial concentration, the average markup, and the level of price-stickiness.

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

The introduction of the misallocation channel—and its increased strength at low Frisch elasticities—helps explain the discrepancy between micro-evidence on the Frisch elasticity and those required to explain the slope of the Phillips curve in traditional models. The standard New Keynesian model requires a large Frisch elasticity (e.g., $\zeta \approx 2$) to explain the magnitude of employment and output fluctuations over the business cycle. Evidence accumulated from quasi-experimental studies, however, suggests that labor supply is

much lower in reality, on the order of 0.1-0.4. To achieve realistic business cycles at these levels of the Frisch elasticity requires counterfactually large swings in wages and prices or else large, exogenous TFP shocks.

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



Incorporating the misallocation channel allows us to generate flatter Phillips curves at lower levels of the Frisch elasticity. 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 $\zeta \approx 1$. The procyclical movement of aggregate TFP takes some of the burden from fluctuations in labor to explain observed fluctuations in output. Furthermore, the movements in aggregate TFP do not require technological regress, or any change in technical primitives for that matter; the movements arise simply out of changes in the allocation of resources across firms.

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.

The Belgian data and pass-through estimates offer us a single cross-sectional view of firm observables. In order to illustrate the role of industrial concentration, we must consider counterfactual firm distributions. To do so, we note that the distribution of firm size (measured by employment) approximates a Pareto distribution (as discussed in Axtell 2001 and Gabaix 2011). We match this empirical description by modeling an economy in

which the distribution of firm employment follows a truncated Pareto,[37]

$$\frac{y_\theta}{A_\theta} \sim \text{Truncated Pareto}(\xi, H). \tag{23}$$

The tail parameter ξ controls firm concentration: as ξ decreases, the employment distribution becomes increasingly fat-tailed, leading to greater concentration. H imposes a maximum ratio between the size of the largest firm and smallest firm in the bounded distribution. We use $H = 7000$, which we calibrate from the Belgian ProdCom data. By pairing this employment distribution with the Kimball aggregator function estimated from the Belgian data, we pin down the distributions of markups, sales shares, and pass-throughs.

In Figure 4, we plot slope of the Phillips curve against the Gini coefficient of the firm employment distribution as we vary the Pareto tail parameter ξ. As the Pareto tail parameter ξ falls, the employment distribution becomes more unequal, and the Gini coefficient becomes larger. As expected, the slope of the Phillips curve under nominal rigidities alone (as in the CES demand system) is unchanged as we vary the employment distribution parameter over this range. However, the strength of real rigidities and the misallocation channel do depend on the firm size distribution: the strength of both channels increases as we increase concentration.

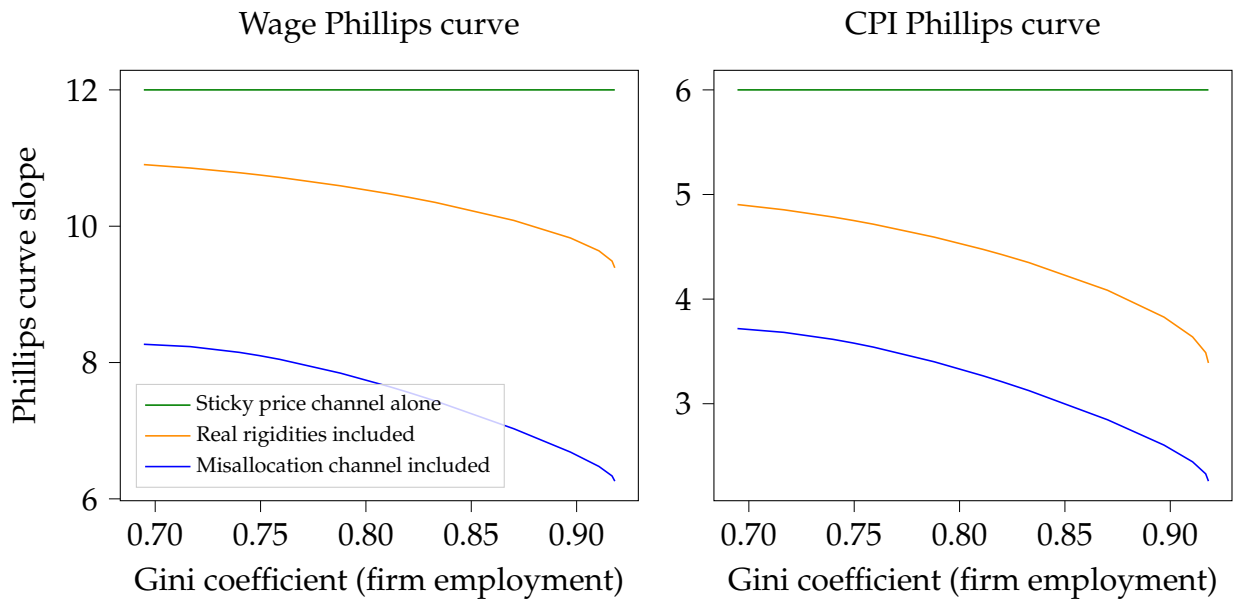
To put these numbers into context, our model predicts that increasing the Gini coefficient from 0.80 to 0.85 flattens the price Phillips curve by an additional 11%.[38] This experiment is in line with the increase in the Gini coefficient in firm employment from 1978 to 2018 measured in data from the Census Business Dynamics Statistics, as we show in Appendix I. 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 26%.

Other parameters. We show how the estimated slope of the Phillips curve changes as we vary the average markup $\bar{\mu}$ and the price-stickiness δ in Appendix C. We briefly summarize the results: Increasing the average markup $\bar{\mu}$ has no effect on the flattening due to real rigidities, but increases the flattening due to misallocation channel linearly. Taking the De Loecker et al. (2020) estimated average markups at face value would imply a large role for the misallocation channel relative to our baseline calibration.

[37] We follow studies such as Helpman et al. (2008) and Feenstra (2018), who use the bounded Pareto to describe the firm distribution. Feenstra (2018) provides more detail on the distribution's implications. A truncated Pareto is useful since it means we do not have to extrapolate pass-throughs outside of the sample range.

[38] Increasing the Gini coefficient from 0.80 to 0.85 is equivalent in our model to decreasing the employment distribution parameter ξ from 0.97 to 0.86.

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



Increasing the price flexibility parameter δ increases the flattening of the price Phillips curve due to the misallocation channel, and decreases the flattening due to real rigidities, for reasons explained after the statement of Proposition 4.

6.4 Results from Dynamic Model

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

In the CES and homogeneous firms case, aggregate TFP does not react to the monetary shock, as implied by Corollary 1. In contrast, when firms have heterogeneous markups, the dispersion in TFPR across firm types increases by around 30 basis points following the contractionary shock, and the response of aggregate TFP is procyclical and hump-shaped.[39] The fall in aggregate TFP dampens the extent of disinflation following the contractionary monetary shock and deepens the immediate response of output to the shock.

Figure 6 zooms in on the output response following the shock. The panel on the right computes the ratio of the output response across models: the orange line is the ratio of the output response in the homogeneous firm Kimball model relative to the CES

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

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

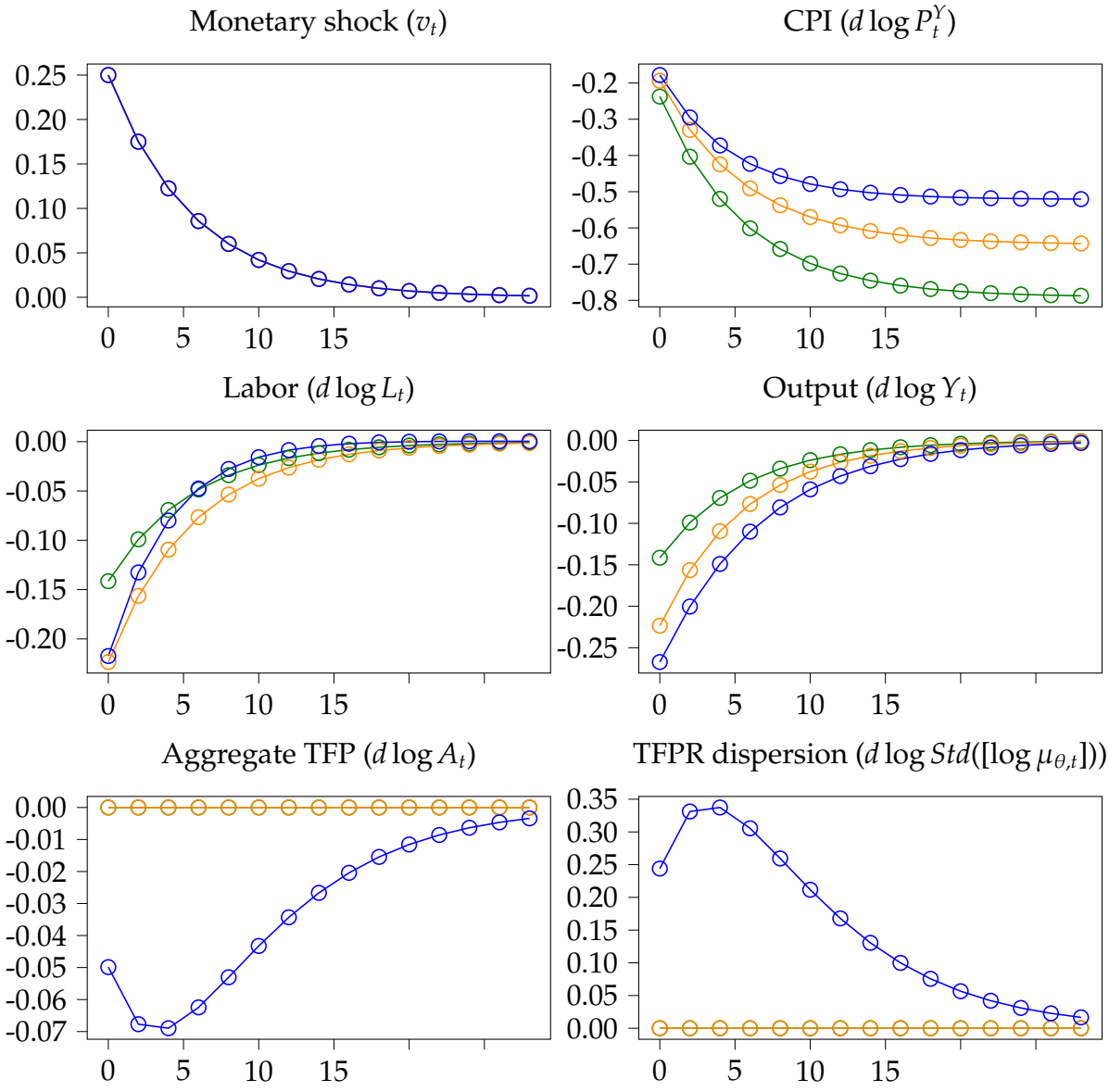
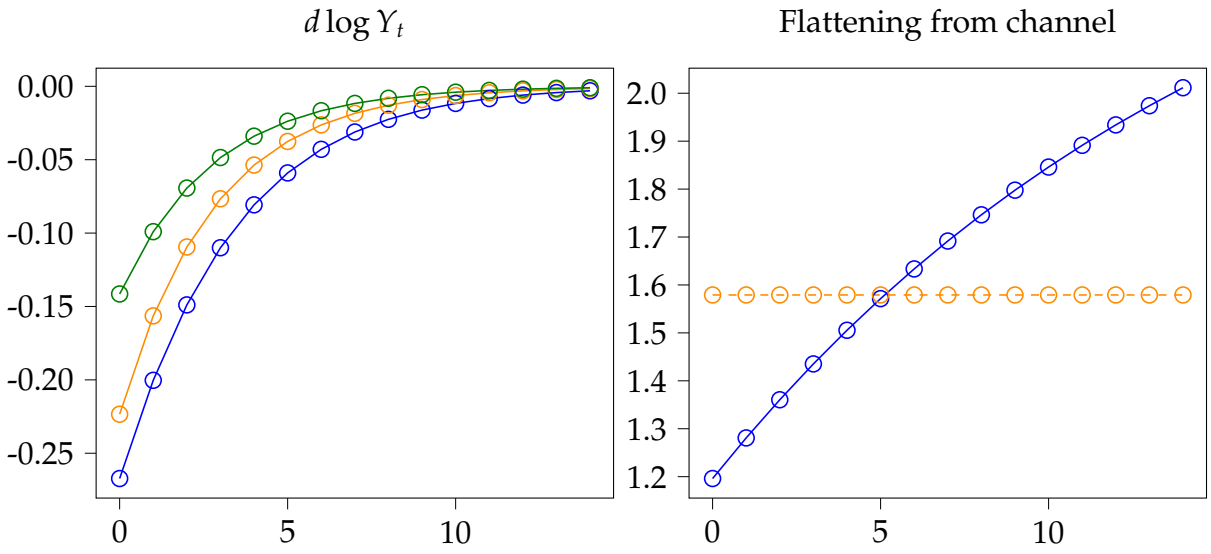


Figure 6: Impulse response function of output following a 25bp monetary policy shock. In the left panel, green, orange, and blue IRFs indicate the CES, homogeneous firms, and heterogeneous firms models respectively. In the right panel, the orange and blue lines are the ratios of the output response of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



model, and the blue line is the ratio of the output response in the full model relative to the homogeneous firm Kimball model. We find that the contraction in output in the full model is about 20% deeper on impact than in the homogeneous firm model. The difference widens over time, because the shock's effect on output is more persistent in the full model.

We quantify this effect on persistence by calculating the half-life of the shock on output. The CES and homogeneous firm models feature a constant half-life of just under two quarters; the misallocation channel increases the half-life of the shock by 23% to about 2.4 quarters.[40] Alvarez et al. (2016) provide a single statistic, the cumulative output impact, that summarizes the total effect of the shock. We report this statistic in Table 3. The misallocation channel increases the cumulative output impact of the monetary shock by 37% compared to the model with real rigidities alone.

Quantitatively similar results are found for other implementations of monetary policy. For example, in Appendix D, we show that results are similar when monetary policy is implemented via changes in money supply (with a cash-in-advance constraint) rather than an interest rate rule.

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

Table 3: Effect of monetary policy shock on output.

Model	Output effect at $t = 0$	Half life	Cumulative output impact
CES	-0.14	1.95	-0.47
With real rigidities	-0.22	1.95	-0.75
Full model	-0.27	2.39	-1.02

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

7 Extensions

Before concluding, we describe some extensions of our results, which are developed in detail in the appendices.

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

Variation in markups and pass-throughs unrelated to size. In our calibrations, we assume that there are no firm-specific taste shifters and all firms face the same residual demand curve. Furthermore, we also calibrate the model starting at a zero-inflation steady-state where all firms charge their desired markups. This means that, at the initial point, markups are a monotone function of firm-size in our model. Whilst markups and pass-throughs do vary as a function of firm size (e.g. see Burstein et al., 2020 or Amiti et al., 2019), in practice, firm markups and pass-throughs also vary for reasons unrelated to size, such as firm-specific shifters in demand curves, quality differences, or

markup dispersion inherited from previous periods. In Appendix H, we show how our baseline results change if there is variation in markups and pass-throughs unrelated to size. We show that the supply-side effects of monetary policy are strengthened if the excess variation in markups is negatively correlated with the excess variation in pass-throughs, and weakened if this correlation is positive. When excess variation in markups and pass-throughs are orthogonal, then the presence of the noise does not affect the strength of supply-side effects of monetary policy relative to our benchmark calibration.

Klenow and Willis (2016) calibration. In the main text of the paper, we caution against using off-the-shelf functional forms for preferences. We illustrate this by calibrating our model with the commonly used Klenow and Willis (2016) specification in Appendix F. We show that to match the observed relationship between pass-through and firm-size, with near complete pass-through for small firms and very incomplete pass-through for large firms, large firms must have markups that are on the order of $10,000\%$. Under standard calibrations, which do not produce astronomically large markups for large firms, the implied pass-through function does not vary much as a function of firm-size. Therefore, under standard calibrations, these preferences fail to capture the cross-sectional covariance between pass-throughs and markups, and hence imply counterfactually small supply-side effects.

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

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

8 Conclusion

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

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

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

A key implication of our analysis is that the knife-edge conditions implied by the standard model mute important channels by which monetary policy, or demand shocks more generally, may be affecting economic aggregates. In this paper, we do not address normative questions like the optimal conduct of monetary policy. We are pursuing this extension in ongoing work.

References

Alvarez, F., H. L. Bihan, and F. Lippi (2016). The real effects of monetary shocks in sticky price models: A sufficient statistic approach. *American Economic Review 106*(10), 2817–2851.

Amiti, M., O. Itskhoki, and J. Konings (2019). International shocks, variable markups, and domestic prices. *The Review of Economic Studies 86*(6), 2356–2402.

Andrés, J. and P. Burriel (2018). Inflation and optimal monetary policy in a model with firm heterogeneity and bertrand competition. *European Economic Review 103*, 18–38.

Anzoategui, D., D. Comin, M. Gertler, and J. Martinez (2019). Endogenous technology adoption and r&d as sources of business cycle persistence. *American Economic Journal: Macroeconomics 11*(3), 67–110.

Atkeson, A. and A. Burstein (2008). Pricing-to-market, trade costs, and international relative prices. *American Economic Review 98*(5), 1998–2031.

Auer, R. A., T. Chaney, and P. Sauré (2018). Quality pricing-to-market. *Journal of International Economics 110*, 87–102.

Auer, R. A. and R. Schoenle (2016). Market structure and exchange. *Journal of International Economics 98*, 60–77.

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.

Axtell, R. (2001). Zipf distribution of us firm sizes. *Science 293*(5536), 1818–1820.

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 (2020a). The darwinian returns to scale. Technical Report 27139, National Bureau of Economic Research.

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

Barkai, S. (2020). Declining labor and capital shares. *The Journal of Finance 75*(5), 2421–2463.

Basu, S., J. G. Fernald, and M. S. Kimball (2006). Are technology improvements contractionary? *American Economic Review 96*(5), 1418–1448.

Berman, N., P. Martin, and T. Mayer (2012). How do different exporters react to exchange rate changes? *The Quarterly Journal of Economics 127*(1), 437–492.

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

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

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

Chatterjee, A., R. Dix-Carneiro, and J. Vichyanond (2013). Multi-product firms and exchange rate fluctuations. *American Economic Journal: Economic Policy 5*(2), 77–110.

Chen, N. and L. Juvenal (2016). Quality, trade, and exchange rate pass-through. *Journal of International Economics 100*, 61–80.

Chetty, R., A. Guren, D. Manoli, and A. Weber (2011). Are micro and macro labor supply elasticities consistent? a review of evidence on the intensive and extensive margins.

American Economic Review 101(3), 471–475.

Christiano, L. J., M. Eichenbaum, and C. L. Evans (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy 113*(1), 1–45.

Comin, D. and M. Gertler (2006). Medium-term business cycles. *American Economic Review 96*(3), 523–551.

Corhay, A., H. Kung, and L. Schmid (2020). Q: Risk, rents, or growth? Technical report, Working Paper.

Cozier, B. and R. Gupta (1993). Is productivity exogenous over the business cycle? some canadian evidence on the solow residual. Technical report, Bank of Canada.

Cravino, J. (2017). Exchange rates, aggregate productivity and the currency of invoicing of international trade. Technical report.

David, J. and D. Zeke (2021, January). Risk-taking, capital allocation and optimal monetary policy. Technical report.

De Loecker, J., J. Eeckhout, and G. Unger (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics 135*(2), 561–644.

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

Eichenbaum, M. and J. Fisher (2004). Evaluating the calvo model of sticky prices. Technical report, National Bureau of Economic Research.

Etro, F. and L. Rossi (2015). New-keynesian phillips curve with bertrand competition and endogenous entry. *Journal of Economic Dynamics and Control 51*, 318–340.

Evans, C. L. (1992). Productivity shocks and real business cycles. *Journey of Monetary Economics 29*(2), 191–208.

Evans, C. L. and F. T. dos Santos (2002). Monetary policy shocks and productivity measures in the g-7 countries. *Portuguese Economic Journal 1*(1), 47–70.

Feenstra, R. C. (2018). Restoring the product variety and pro-competitive gains from trade with heterogeneous firms and bounded productivity. *Journal of International Economics 110*, 16–27.

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.

Gabaix, X. (2011). Power laws in economics and finance. *Annual Review of Economics 1*(1), 255–294.

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

Gopinath, G. and O. Itskhoki (2011). In search of real rigidities. *NBER Macroeconomics*

Annual 25(1), 261–310.

Gopinath, G., O. Itskhoki, and R. Rigobon (2010). Currency choice and exchange rate pass-through. *American Economic Review 100*(1), 304–336.

Gutiérrez, G. and T. Philippon (2017). Declining competition and investment in the u.s. Technical Report 23583, National Bureau of Economic Research.

Hall, R. E. (1988). The relation between price and marginal cost in us industry. *Journal of Political Economy 96*(5), 921–947.

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.

Helpman, E., M. J. Melitz, and Y. Rubinstein (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics 123*(2), 441–487.

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.

Jappelli, T. and L. Pistaferri (2010). The consumption response to income changes. *Annual Review of Economics 2*, 479–506.

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.

Li, H., H. Ma, and Y. Xu (2015). How do exchange rate movements affect chinese exports? a firm-level investigation. *Journal of International Economics 97*(1), 148–161.

Martinez, I. Z., E. Saez, and M. Siegenthaler (2018). Intertemporal labor supply substitution? evidence from the swiss income tax holidays. Technical Report 24634, National Bureau of Economic Research.

Matsuyama, K. and P. Ushchev (2017). Beyond ces: Three alternative classes of flexible homothetic demand systems.

Meier, M. and T. Reinelt (2020, June). Monetary policy, markup dispersion, and aggregate tfp. *ECB Working Paper No. 2427*.

Moscarini, G. and F. Postel-Vinay (2012). The contribution of large and small employers to job creation in times of high and low unemployment. *American Economic Review 102*(6), 2509–39.

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.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic Dynamics 11*(4), 707–720.

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

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

Taylor, J. B. (1999). Staggered price and wage setting in macroeconomics. *Handbook of Macroeconomics 1*, 1009–1050.

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