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 March 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. Baqaee and Farhi 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 March 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 study an economy with heterogeneous firms, sticky prices, and endogenous markups. We show that if, as is consistent with the empirical evidence, bigger firms have higher markups and lower pass-throughs than smaller firms, then a monetary easing endogenously increases aggregate TFP and improves allocative efficiency. This endogenous positive "supply shock" amplifies the effects of the positive "demand shock" on out-put and employment. The result is a flattening of the Phillips curve. This effect is distinct from another mechanism discussed at length in the real rigidities literature: a monetary easing leads to a reduction in desired markups because of strategic complementarities in pricing. We calibrate the model to match firm-level pass-throughs and find that the misallocation channel of monetary policy is quantitatively important, flattening the Phillips curve by about 70% compared to a model with no supply-side effects. We derive a tractable four-equation dynamic model and show that monetary easing generates a procyclical hump-shaped response in aggregate TFP and countercyclical dispersion in firm-level TFPR. The improvements in allocative efficiency amplify both the impact and persistence of interest rate shocks on output.

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@NA.com

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: productivity is 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.

In this paper, we offer an alternative explanation. We argue that 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. In an economy with realistic firm heterogeneity, demand shocks *should* trigger changes in aggregate TFP. These changes do not arise from changes in technical productivity—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.[2] The endogenous "supply shock" generated by the misallocation channel complements the traditional effects of the "demand shock" on employment and output. Incorporating the misallocation channel heightens the response of output to demand shocks and flattens the Phillips curve.

This supply-side effect exists when there is misallocation in the initial allocation, i.e., when the output of the economy can be improved by reallocating resources from some firms to others. If there is no misallocation, that means the cross-sectional allocation of resources equates the marginal benefit of each input across all competing uses. Therefore, starting at an efficient point, a reallocation of resources triggered by monetary policy has no first-order effects on aggregate productivity.

[1] The failed invariance of aggregate TFP to demand shocks is confirmed separately 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] 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.

For this reason, the misallocation channel is absent in the benchmark log-linearized New Keynesian model, which features Constant Elasticity of Substitution (CES) preferences. The CES demand system yields uniform desired markups across firms. As a result, the steady-state allocation of resources is efficient, and the envelope theorem implies that reallocations are irrelevant to aggregate TFP. This logic extends to any model composed of firms with uniform but variable markups.

In contrast to the steady-state of the benchmark model, the data features substantial and persistent heterogeneity in markups across firms. When 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 markups creates differences in the marginal benefit of inputs across firms. As a result, the reallocations triggered by demand shocks, like monetary policy shocks, can have first-order effects on aggregate TFP and output. The direction of this effect depends on whether monetary easing reallocates resources to more or less efficient uses.

To analyze these reallocations, we deviate from the classic CES formulation of the New Keynesian model and adopt a non-parametric 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 a distribution of firms with heterogeneous productivities and sticky prices. Our model is flexible enough to exactly match empirical 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' realized pass-throughs covary negatively with markups, then a positive demand shock, such as a monetary easing, increases aggregate TFP and moves the economy closer to the efficient frontier. 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. Heterogeneity in realized pass-through can be driven either by heterogeneity in desired pass-through or heterogeneity in price-stickiness.[3,4] A

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

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

negative relationship between markups and desired pass-throughs is sometimes called Marshall's strong second law of demand (see Melitz, 2018 or Baqaee and Farhi, 2020a for more information). This relationship between markups and pass-throughs has strong empirical support across countries.[5]

Our second result shows that the response of output to a monetary shock can be decomposed into distinct demand-side and supply-side effects. The demand-side effect of an expansionary shock arises from increases in employment due to increased labor demand. Intuitively, expansionary monetary policy raises 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 Marshall's strong second law of demand holds, an expansionary shock decreases the dispersion in markups across firms, boosting aggregate TFP. We find that the supply-side effect constitutes an important part of monetary policy transmission: when we calibrate our 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.

We provide both a static model, which highlights the key intuitions driving the productivity response, and a fully dynamic model. In the dynamic model, changes in aggregate TFP are an endogenous outcome of reallocations across high- and low-markup firms. We describe the movement of aggregate TFP, output, inflation, and the interest rate using a four-equation system that augments the classic three-equation model with realistic firm heterogeneity and endogenous changes in allocative efficiency. The Taylor rule and the Euler equation are unchanged in our model. However, the New Keynesian Phillips curve in our model is different. Compared to the workhorse New Keynesian model, 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

[5]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.

average desired pass-through, and the covariance of markups and desired pass-throughs.

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), as has been documented by the literature (see, e.g. Kehrig, 2011).[6]

To calibrate our model, 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 approach is preferable to using an off-the-shelf functional form, since it does not impose the counterfactual restrictions baked in by parametric families of preferences.[7] Furthermore, we can calibrate our model by relying on estimates of firm-level pass-throughs rather than firm-level markups. This is a virtue since pass-throughs can be estimated using weaker assumptions than markups.

We use cross-sectional firm data from Belgium (provided by Amiti et al. 2019) to quantify the importance of the misallocation channel. We find that it substantially flattens the Phillips curve compared to a model without supply-side effects. In the dynamic model, the misallocation channel deepens the loss in output following a contractionary shock by 20% on impact. The role of misallocation 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%.

Since the strength of real rigidities and the misallocation channel are governed by moments of the firm distribution, our analysis ties the strength of monetary policy to the industrial organization of the economy. In particular, we show that an increase in industrial concentration can increase the potency of both the real rigidities and misallocation channels. While the standard New Keynesian model is silent on the role of industrial concentration, in our setup increasing the Gini coefficient of firm employment from 0.80 to 0.85 flattens the Phillips curve by an additional 11%.[8]

[6]When firms have constant returns to scale, as in our model, firm-level TFPR is equal to the firm's markup. Increased dispersion in these markups generate a greater degree of misallocation, as commented on by Hsieh and Klenow (2009).

[7]For example CES, quadratic, or Klenow and Willis (2016) preferences. Under monopolistic competition, CES preferences imply complete desired pass-throughs and constant markups, Klenow and Willis (2016) preferences imply pass-throughs are too low for small firms and markups are too high for large firms, and quadratic preferences imply desired pass-through starts at 1/2 and declines to zero. We provide an explicit calibration exercise in Appendix E.

[8]This 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 G.2.

Other related literature.

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

A strand of this literature is devoted to explaining the strength and persistence of monetary policy shocks, which cannot be explained by nominal rigidities alone given the frequency of price adjustment.[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). Amiti et al. (2019) show that pass-throughs are decreasing in firm size, so that larger firms tend to exhibit a greater degree of "pricing-to-market" behavior.[11] 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 in an economy, 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

[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] Amiti et al. (2019) provide the cross-section of firm pass-throughs in Belgium that we use to calibrate the model in this paper. This pattern is also documented across countries: see Footnote 5.

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 real productivity.[12]

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

Finally, our paper is also related to a recent 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/or 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.

The structure of the paper is as follows. Section 2 introduces a simple, one-period 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 analyze the findings in a couple of simple example economies. 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 then takes our findings to data to quantify the importance of the misallocation channel. In Section 7, we summarize various extensions provided in the appendix, 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. All proofs are included in the appendix.

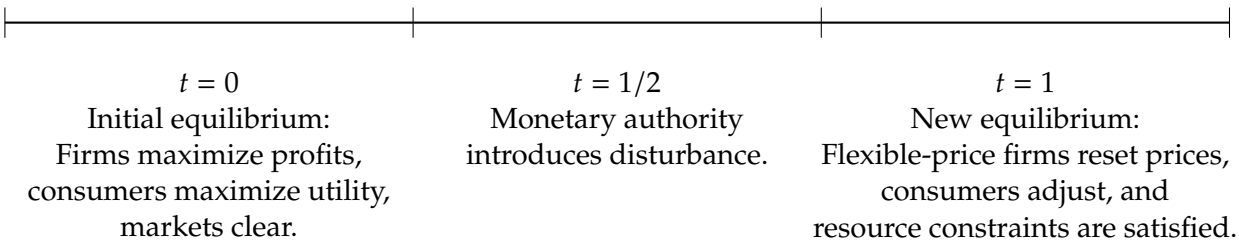
[12]Productivity shocks can also affect allocative efficiency: David and Zeke (2021) show that allocative efficiency varies over the business cycle when firms have heterogeneous exposure to aggregate shocks.

2 Model

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

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

Figure 1: One-period model timing.



2.1 Setup

Households.

We have a population of identical consumers. Each consumer experiences utility from a consumption bundle Y and disutility from labor L given by

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

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 $\omega \in [0, \bar{\omega}]$. Consumers have homothetic Kimball (1995) preferences over these final goods, so that the utility from the consumption bundle Y is defined implicitly by

$$\int_0^{\bar{\omega}} \Upsilon(\frac{y_\omega}{Y}) d\omega = 1. \tag{2}$$

Here, y_ω is the consumption of variety ω, and Υ is an increasing and concave function. CES preferences are a special case of the general preferences above, when $\Upsilon(\cdot)$ is a power function.

The representative consumer maximizes subject to the budget constraint

$$\int_0^{\bar{\omega}} p_\omega y_\omega d\omega = wL + \Pi, \tag{3}$$

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

$$\frac{p_\omega}{P} = \Upsilon'\left(\frac{y_\omega}{Y}\right), \tag{4}$$

where the *price aggregator* P is defined as

$$P = \frac{P^Y}{\int_0^{\bar{\omega}} \Upsilon'\left(\frac{y_\omega}{Y}\right)\frac{y_\omega}{Y} d\omega}, \tag{5}$$

and P^Y is the ideal price index.[13] As we can see in Equation (4), the demand for a variety ω is dictated by the ratio of its price to the price aggregator P. Hence, firms compete with the rest of the market via this price aggregator. Equation (4) also illustrates the flexibility of Kimball preferences: by choosing the aggregator $\Upsilon(\cdot)$, we can create downward-sloping demand curves of any desired shape.

Firms.

Each variety is supplied by a single firm. We order firms according to their productivity and index them by $\theta \in [0, 1]$. A firm of type θ and has productivity A_θ, where A is increasing in 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,

$$p_\theta^{\text{flex}} = \underset{p_\theta}{\operatorname{argmax}}\, \mathbb{E}\left(p_\theta y_\theta - \frac{w}{A_\theta} y_\theta\right), \tag{6}$$

[13] Recall that the ideal price index is defined as $\min_{y_\omega}\{\int p_\omega y_\omega : Y = 1\}$. Since Kimball 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).

taking as given its residual inverse-demand curve,

$$\frac{p_\theta}{P} = \Upsilon'(\frac{y_\theta}{Y}). \tag{7}$$

Unlike the CES demand system, which imposes a constant price elasticity of demand, Kimball preferences allow the price elasticity facing a firm to vary with the firm's position on the demand curve. We can use the inverse-demand function in (7) to solve for the price elasticity of demand facing a firm of type θ:

$$\sigma_\theta = -\frac{d \log y_\theta}{d \log p_\theta} = \frac{\Upsilon'(\frac{y_\theta}{Y})}{-\frac{y_\theta}{Y}\Upsilon''(\frac{y_\theta}{Y})}. \tag{8}$$

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, markup, and quantity are determined together by

$$p_\theta^{\text{flex}} = \mu_\theta^{\text{flex}} \frac{w}{A_\theta}, \qquad y_\theta = y(p_\theta^{\text{flex}}), \qquad \text{and} \qquad \mu_\theta^{\text{flex}} = \mu(\frac{y_\theta}{Y}), \tag{9}$$

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

$$\mu(\frac{y}{Y}) = \frac{1}{1 - \frac{1}{\sigma(\frac{y}{Y})}}. \tag{10}$$

Using Equation (8), a firm of type θ has the desired markup

$$\mu_\theta^{\text{flex}} = \frac{1}{1 - \frac{-\frac{y_\theta}{Y}\Upsilon''(\frac{y_\theta}{Y})}{\Upsilon'(\frac{y_\theta}{Y})}}. \tag{11}$$

Following Calvo (1983), we add nominal rigidities by assuming 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 $t = 1$ after observing the monetary shock according to the optimal price and markup formulas above. However, sticky-price firms keep their prices unchanged. As a result, the prices and markups of sticky-price firms at $t = 1$ are given by

$$p_{\theta,1}^{\text{sticky}} = p_{\theta,0} \qquad \text{and} \qquad \mu_{\theta,1}^{\text{sticky}} = \frac{w_1}{w_0}\mu_{\theta,0}, \tag{12}$$

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

where the second subscript denotes the period.

A firm's desired pass-through, ρ_θ, is the degree to which a firm's optimal price changes in response to a shock to marginal cost, holding the aggregate price index P constant. For the remainder of the text, we will refer to ρ_θ simply as the firm's "pass-through." Keep in mind, however, that this pass-through is conditional on the firm's ability to change its price; nominal rigidities also alter the relationship between price and marginal cost, so in practice, realized pass-through of marginal cost into prices depends on both desired pass-through and price rigidity.

We can express pass-throughs as $\rho_\theta = \rho(y_\theta/Y)$, where

$$\rho(\frac{y}{Y}) = \frac{\partial \log p^{\text{flex}}}{\partial \log mc} = \frac{1}{1 + \frac{d \log \mu^{\text{flex}}}{d \log \frac{y}{Y}} \frac{-d \log y}{d \log p}} = \frac{1}{1 + \frac{\frac{y}{Y}\mu'(\frac{y}{Y})}{\mu(\frac{y}{Y})}\sigma(\frac{y}{Y})}. \tag{13}$$

Note that the elasticity of desired markups to marginal cost shocks is given by $\frac{\partial \log \mu_\theta^{\text{flex}}}{\partial \log mc_\theta} = \rho_\theta - 1$.

Under CES preferences, desired markups $\mu_\theta = \mu = \frac{\sigma}{\sigma-1}$ are constant across firms. This means that firms exhibit "complete pass-through": $\rho_\theta = 1$ for all firms. Idiosyncratic shocks to a firm's marginal cost result in one-for-one changes in the firm's desired price.

Kimball preferences allow markups and pass-throughs to vary across firms. The distributions of markups and pass-throughs will depend on the shape of $\Upsilon(\cdot)$, which determines $\mu(\cdot)$and $\rho(\cdot)$ (as in Equations 11 and 13).

Given this flexibility, it is useful to keep in mind the following set of restrictions that, when imposed on Υ, generate reasonable patterns in firms' markups and pass-throughs. We do not impose these restrictions to derive our theoretical results, but they are useful to keep in mind when discussing our results.

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

$$\mu'(\frac{y}{Y}) > 0 \qquad \text{and} \qquad \rho'(\frac{y}{Y}) < 0. \tag{14}$$

Since productivity A_θ is increasing in type θ and marginal revenue curves are downward-sloping, a firm's relative size y_θ/Y is increasing in θ. Hence, Marshall's strong second law is equivalent to requiring in our context that the markup function $\mu(\cdot)$ is upward sloping and the pass-through function $\rho(\cdot)$ is downward-sloping.[15]

[15]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 $\mu'(\frac{y}{Y}) \geq 0$

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). Workhorse models with heterogeneous markups, such as the nested CES oligopoly model of Atkeson and Burstein (2008), also satisfy Marshall's second law of demand. Flexible Kimball preferences do not presuppose that these relationships hold, and in general we do not impose them.

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 price level (CPI) or money supply. The nominal wage is especially convenient as it directly affects the marginal cost of every firm.

We consider a small perturbation in the wage introduced by the monetary authority, $d \log w$. If $d \log w$ is positive, we say that the monetary shock is expansionary, as it decreases markups for firms whose prices cannot adjust, and this reduction in markups boosts labor demand and hence output.

Equilibrium Conditions.

In equilibrium, consumers maximize utility taking prices as given, firms set prices to maximize profits taking other firms' prices and their residual demand curves as given, and resource constraints are satisfied.

The equilibrium conditions are summarized below.

1. *Consumers maximize utility.* Consumers choose labor supply and demand for each good taking the wage and prices as given. The inverse-demand curve in Equation (4) determines the demand for each good, y_θ, given the good's price and the price aggregator. Equation (2) determines the utility from the aggregate consumption bundle, Y, given the consumption of each good. Finally, the household chooses labor supply so that the ratio of the marginal disutility from labor to the marginal utility from consumption equals the ratio of the wage to the price level; in equilibrium, this yields

$$L^{\frac{1}{\zeta}} = \frac{w}{P^Y} Y^{-\gamma}. \tag{15}$$

(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.

2. *Firms maximize profits.* Firms set prices and meet whatever demand they face. The behavior of firms can be described by their markups. The markups of all firms in the initial equilibrium are set according to the Lerner formula (10). After the monetary shock, the markups of flexible-price firms are again set according to the Lerner formula, while the markups of sticky-price firms move due to changes in marginal cost that are out of the firm's control,

$$\mu_\theta^{\text{sticky}} = \frac{w_1}{w_0}\mu_{\theta,0}. \tag{16}$$

3. *Markets equilibrium.* In equilibrium, resource constraints are satisfied and the labor market clears.

Notation.

For convenience, throughout the rest of the paper, we use the following notation. For two variables $x_\theta > 0$ and z_θ, define

$$\mathbb{E}_x[z_\theta] = \frac{\int_0^1 z_\theta x_\theta d\theta}{\int_0^1 x_\theta d\theta}. \tag{17}$$

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.

Define the sales share density as[16]

$$\lambda_\theta = \frac{p_\theta y_\theta}{\int_0^1 p_\theta y_\theta d\theta}. \tag{18}$$

The *aggregate markup* is the harmonic sales-weighted average of firm markups, that is,

$$\bar{\mu} = \mathbb{E}_\lambda\left[\mu_\theta^{-1}\right]^{-1}. \tag{19}$$

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 discrete changes in a variable, we write $\Delta\log X$ instead.

[16]The type distribution is uniform between [0,1] because a firm's type is defined by the fraction of firms whose productivity is less than that firm. This is without loss of generality as long as the productivity distribution is continuous.

3 Productivity Response

In this section, we consider the movement of aggregate total factor productivity (TFP) from the initial allocation to the period following the 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 that follows the monetary shock changes aggregate TFP. The response of aggregate TFP relies on two potential sources of heterogeneity, which we discuss in detail.

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 *allocative efficiency* of an economy as the efficiency of the cross-sectional allocation of resources across firms, holding the supply of those resources fixed.[17]

We can explicitly relate movements in output to shifts in the allocative efficiency of the economy:

$$d\log Y = d\log L + \underbrace{[d\log Y - d\log L]}_{d\log A} \tag{20}$$

Here, A is the distortion-adjusted Solow residual (Hall 1988, Baqaee and Farhi 2020b), which measures aggregate total factor productivity (TFP).[18] The following proposition is an application of the main result in Baqaee and Farhi (2020b).

Proposition 1. *To a first order, the change in aggregate TFP is given by*

$$d\log A = d\log \bar{\mu} - \mathbb{E}_{\lambda}\left[d\log \mu_{\theta}\right], \tag{21}$$

where $\bar{\mu}$ is the (harmonic) average markup.

[17] 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 D.

[18] 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. Baqaee and Farhi (2020b) discuss in detail how this modification makes the measurement of aggregate TFP robust to changes in factor levels, where the traditional Solow residual may detect spurious changes in aggregate TFP.

Proposition 1 shows that allocative efficiency increases when the average markup rises more than markups on average. Intuitively, Equation (21) is a measure of the change in the dispersion in markups. Note that, for a set of numbers, a mean-preserving spread does not change the arithmetic mean, but does decrease the harmonic mean. A similar logic applies here: when the (harmonic) average markup rises more than markups on average, this means that high-markup firms are expanding relative to low-markup firms, and the distribution of markups compresses, moving the allocation closer to the efficient frontier.

At its core, Equation (21) links TFP at the economy level to the dispersion in markups at the firm level. It is useful here to make the connection to an industrial organization literature which has documented variation in the dispersion of plant- and firm-level revenue productivity, or TFPR. Kehrig (2011), for example, finds that the dispersion in plant-level TFPR is countercyclical, increasing about 10% during recessions compared to booms. When production has constant returns to scale, like in our model, variation in TFPR is exactly equal to variation in markups.[19] To see this, note that TFPR is usually measured by subtracting input growth from revenue growth. In our model, this is just

$$\Delta \log \text{TFPR} = \Delta \log p_\theta y_\theta - \Delta \log l_\theta = \Delta \log \frac{\mu_\theta w}{A_\theta} A_\theta l_\theta - \Delta \log l_\theta = \Delta \log \mu_\theta + \Delta \log w,$$

where changes in the wage $\Delta \log w$ do not vary by firm. Hence, an improvement in aggregate TFP driven by a compression of the markup distribution will also imply that TFPR dispersion should be countercyclical. As we will see, our model predicts that a monetary easing will simultaneously increase aggregate TFP and reduce dispersion in TFPR.

Proposition 1 suggests that aggregate TFP increases when resources are shifted toward high-markup firms, since those firms are inefficiently small to begin with. At first blush, this may run counter to the reader's intuition, since greater market power (i.e., higher markups) constricts employment and output below their optimal levels. We stress that these are two separate sources of inefficiency. One has to do with the *level* of markups: higher markups depress the supply of labor, and hence lower output. The other has to do with the *dispersion* of markups: holding the supply of labor fixed, more dispersed markups increase the degree to which resources are misallocated across firms. Since changes in productivity hold the supply of resources fixed, $d \log A$ is linked to the latter distortion arising from markup dispersion.

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

A corollary of Proposition 1 is that when markups are uniform in the initial equilibrium, first-order changes in aggregate productivity are necessarily zero.

Lemma 1. *If $\mu_\theta = \mu$ in the initial equilibrium, then*

$$d \log A = 0. \tag{22}$$

To show Lemma 1 formally, note that we can write $d \log \bar{\mu} = \mathbb{E}_{\lambda\mu^{-1}} [d \log \mu_\theta] - \mathbb{E}_{\lambda\mu^{-1}} [d \log \lambda_\theta]$. When markups are initially uniform, this term is exactly $\mathbb{E}_\lambda [d \log \mu_\theta]$.

Lemma 1 is a consequence of the Envelope Theorem: when markups are identical across firms, the cross-sectional allocation of resources is efficient. Hence, changes in aggregate TFP due to reallocations are zero to a first order. This confirms that in models with representative firms or CES demand, where markups are the same for all firms in the initial equilibrium, aggregate TFP does not respond to monetary policy to a first-order.

3.2 Productivity response: Two mechanisms

When markups are initially dispersed, the monetary disturbance triggers a change in the economy's allocative efficiency. Proposition 2 describes the response of aggregate TFP to the shock.

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

$$\frac{d \log A}{d \log w} = \underbrace{\kappa_\rho Cov_\lambda [\rho_\theta, \sigma_\theta | \text{flex}]}_{\substack{\textit{Reallocation due to} \\ \textit{heterogeneous pass-through}}} + \underbrace{\kappa_\delta Cov_\lambda [\sigma_\theta, \delta_\theta]}_{\substack{\textit{Reallocation due to} \\ \textit{heterogeneous price-stickiness}}}, \tag{23}$$

where $Cov_\lambda [\rho_\theta, \sigma_\theta | \text{flex}]$ is the covariance of pass-throughs and elasticities for the subset of flexible-price firms,[20] *and*

$$\kappa_\rho = \frac{\bar{\mu} \mathbb{E}_\lambda [\delta_\theta] \mathbb{E}_\lambda [1 - \delta_\theta]}{\mathbb{E}_\lambda [[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \sigma_\theta]} > 0,$$
$$\kappa_\delta = \frac{\bar{\mu} \mathbb{E}_{\lambda\delta} [\rho_\theta]}{\mathbb{E}_\lambda [[\delta_\theta \rho_\theta + (1 - \delta_\theta)] \sigma_\theta]} > 0.$$

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 (23) are zero.

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

When markups are dispersed, two potential mechanisms drive a change in aggregate TFP in response to a monetary disturbance. The first is the change in allocative efficiency resulting from heterogeneous desired pass-through across firms, and the second is a change in allocative efficiency resulting from heterogeneous price-stickiness across firms. 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.[21] To build more intuition, we now consider each of the two mechanisms mentioned above in isolation.

3.2.1 Heterogeneous pass-through

If price-stickiness is homogeneous across firms ($\delta_\theta = \delta$), then the latter covariance term in Proposition 2 is zero, and the productivity response depends on the covariance between desired pass-throughs ρ_θ and elasticities σ_θ alone:

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

$$\frac{d\log A}{d\log w} = \kappa_\rho Cov_\lambda\left[\rho_\theta, \sigma_\theta\right]. \tag{24}$$

Under Marshall's strong second law of demand, $\frac{d\log A}{d\log w} > 0$.

Intuition. 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

$$d\log p_\theta^{\text{flex}} = (1-\rho_\theta) d\log P + \rho_\theta d\log w, \tag{25}$$

where $d\log P$ is the change in the price aggregator defined in Equation (5). The optimal price of a firm with a high pass-through (i.e., ρ_θ close to one) moves closely with shocks to marginal cost from the nominal wage. Firms with low pass-through, on the other hand, exhibit "pricing-to-market" behavior: they place less weight on 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

[21] 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.

wage shock:

$$d \log p_\theta^{\text{sticky}} = 0. \tag{26}$$

Following an increase in the nominal wage, flexible-price firms shrink and sticky-price firms, whose prices are kept artificially low, expand[22]:

$$d \log(\frac{y_\theta^{\text{flex}}}{Y}) = -\sigma_\theta \rho_\theta \left(d \log w - d \log P \right), \tag{27}$$

$$d \log(\frac{y_\theta^{\text{sticky}}}{Y}) = \sigma_\theta d \log P. \tag{28}$$

Among flexible-price firms, those where the product of elasticities and pass-throughs ($\sigma_\theta \rho_\theta$) is highest will shrink the most. Firms with low pass-throughs and high markups shrink less, and thus expand relative to their flexible-price counterparts. Among sticky-price firms, the opposite effect prevails: firms with low markups (and higher elasticities σ_θ) 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.

Recall that under Marshall's second law of demand, markups are increasing and pass-throughs are decreasing in firm type θ. This means that the covariance between elasticities σ_θ and pass-throughs ρ_θ will be positive, and thus aggregate TFP moves procyclically with the nominal wage.

3.2.2 Heterogeneous price-stickiness

We now consider the case where desired pass-through is homogeneous, but price-stickiness is not.

Corollary 2. *If pass-through is homogeneous across firms ($\rho_\theta = \rho$),*[23] *then*

$$\frac{d \log A}{d \log w} = \kappa_\delta Cov_\lambda \left[\sigma_\theta, \delta_\theta \right]. \tag{29}$$

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

[23]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.

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

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

The endogenous movement in aggregate TFP is driven by the interaction of heterogeneous markups with heterogeneity in either pass-throughs or price-stickiness. Either dimension of heterogeneity will yield a shift in the dispersion of markups following a shock, thereby changing allocative efficiency.

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.[24] Our findings in Proposition 2 contextualize this channel as one of two sources of heterogeneity that contribute to the endogenous response of aggregate TFP.

For the remainder of this paper, we focus on the interaction of heterogeneous markups with heterogeneous desired pass-throughs and abstract away from the heterogeneous price-stickiness channel. While there is little consensus on the drivers of heterogeneity in price-stickiness, there is robust empirical support for Marshall's second law of demand, and we can calibrate our model's predictions to empirical estimates of pass-throughs from the literature.

4 Output Response and the Phillips Curve

In the previous section, we showed that aggregate TFP endogenously responds to a monetary shock. In this section, we incorporate this endogenous "supply shock" into the response of aggregate output.

As we will see, 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

[24] Recent research may provide further micro-foundations for how price-stickiness correlates with other firm characteristics. Gorodnichenko and Weber (2016) find significant heterogeneity in the degree of price-stickiness across companies in the S&P500 and document that this heterogeneity contributes to differences in the volatility of stock market returns following a monetary policy shock, though their focus is on exogenous, sectoral differences in price-stickiness.

pass-through (which arise, for example, 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 two simple example economies.

4.1 Output response

Proposition 3 describes the response of output to a monetary shock.

Proposition 3. *Following a shock to the nominal wage* $d \log w$*, the response of output at* $t = 1$ *is*

$$d \log Y = \underbrace{\frac{1}{1+\gamma\zeta} d \log A}_{\textit{Supply-side effect}} - \underbrace{\frac{\zeta}{1+\gamma\zeta} \mathbb{E}_\lambda \left[d \log \mu_\theta \right]}_{\textit{Demand-side effect}}, \tag{30}$$

where

$$\frac{\mathbb{E}_\lambda \left[d \log \mu_\theta \right]}{d \log w} = - \underbrace{\mathbb{E}_\lambda \left[1 - \delta_\theta \right]}_{\textit{Nominal rigidities only}} - \underbrace{\frac{\mathbb{E}_\lambda \left[\delta_\theta (1 - \rho_\theta) \right] \mathbb{E}_\lambda \left[\sigma_\theta (1 - \delta_\theta) \right]}{\mathbb{E}_\lambda \left[\left[\delta_\theta \rho_\theta + (1 - \delta_\theta) \right] \sigma_\theta \right]}}_{\textit{Real rigidities}}. \tag{31}$$

We see in Equation (30) that the response of output is composed of supply-side and demand-side effects. 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 (31) further disaggregates 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 (30), we find that when $\frac{d\log A}{d\log w} > 0$, aggregate TFP increases following an expansionary shock. This endogenous positive "supply shock" complements the effects of the positive "demand shock" on output. Unlike the demand-side effect, the supply-side effect continues to operate even when the elasticity of labor supply ζ is zero.

4.2 Flattening of the Phillips Curve

We can rearrange the output response given in Proposition 3 to get the slope of the wage Phillips curve. To get the CPI Phillips curve, we use the relationship between the CPI (P^Y),[25] the nominal wage, and average markups,

$$d\log P^Y = d\log w + d\log\bar{\mu} - d\log A = d\log w + \mathbb{E}_\lambda\left[d\log\mu_\theta\right]. \tag{32}$$

Both are presented in Proposition 4.

Proposition 4. *The wage Phillips curve is given by*

$$d\log w = (1+\gamma\zeta)\frac{1}{\left[\frac{d\log A}{d\log w} - \zeta\mathbb{E}_\lambda\left[\frac{d\log\mu_\theta}{d\log w}\right]\right]}d\log Y. \tag{33}$$

The CPI Phillips curve is given by

$$d\log P^Y = (1+\gamma\zeta)\frac{1+\mathbb{E}_\lambda\left[\frac{d\log\mu_\theta}{d\log w}\right]}{\left[\frac{d\log A}{d\log w} - \zeta\mathbb{E}_\lambda\left[\frac{d\log\mu_\theta}{d\log w}\right]\right]}d\log Y. \tag{34}$$

The non-parametric decomposition of the output response in Proposition 3 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 comparing the slopes of the Phillips curves in an economy with sticky prices alone and in an economy with incomplete pass-through:

$$\text{Flattening due to real rigidities} = \frac{\text{Phillips curve slope w/ nominal rigidities only}}{\text{Phillips curve slope w/ real rigidities}}. \tag{35}$$

Since real rigidities flatten the Phillips curve, this quantity is greater than or equal to one. If this quantity is, say, 1.5, this means that incorporating real rigidities decreases the

[25] In our setup, changes in the ideal price index and the CPI are the same to a first-order, i.e., $d\log P^Y = \mathbb{E}_\lambda\left[d\log p_\theta\right]$. See Footnote 13.

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 an economy with sticky prices and real rigidities but no misallocation to one where the allocative efficiency is allowed to change:

$$\text{Flattening due to the misallocation channel} = \frac{\text{Phillips curve slope w/ real rigidities}}{\text{Phillips curve slope w/ misallocation}}. \tag{36}$$

When $\frac{d \log A}{d \log w} > 0$, this quantity is also greater than one.

Proposition 5 presents the flattening of the CPI 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 in Appendix A Proposition 7).

Proposition 5. *Suppose $\delta_\theta = \delta$ for all firms. The flattening of the CPI Phillips curve due to real rigidities, compared to nominal rigidities alone, is*

$$\frac{\textit{Phillips curve slope w/ nominal rigidities only}}{\textit{Phillips curve slope w/ real rigidities}} = 1 + \frac{\mathbb{E}_\lambda[\sigma_\theta]\,\mathbb{E}_\lambda[1-\rho_\theta]}{\delta Cov_\lambda[\rho_\theta, \sigma_\theta] + \mathbb{E}_\lambda[\rho_\theta]\,\mathbb{E}_\lambda[\sigma_\theta]}. \tag{37}$$

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

$$\frac{\textit{Phillips curve slope w/ real rigidities}}{\textit{Phillips curve slope w/ misallocation}} = 1 + \frac{\bar{\mu}}{\zeta}\frac{\delta Cov_\lambda[\rho_\theta, \sigma_\theta]}{\delta Cov_\lambda[\rho_\theta, \sigma_\theta] + \mathbb{E}_\lambda[\sigma_\theta]}. \tag{38}$$

In Equation (37), we see that the flattening of the CPI Phillips curve due to real rigidities increases as pass-throughs diverge further from one. When firms exhibit complete pass-through (i.e. $\mathbb{E}_\lambda[\rho_\theta] = 1$), there is no real rigidities effect on the slope of the Phillips curve (the ratio of the slopes is one). The flattening due to real rigidities in (37) 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 CPI 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 (38) also shows that the flattening due to misallocation is decreasing in the Frisch elasticity ζ: when labor is infinitely elastic, an additional unit of labor can always be supplied at a fixed price, which means that resource misallocation has no effect on output. 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 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 two simple benchmark economies: one with CES preferences and the other with real rigidities but homogeneous firms.

CES Example.

We obtain the CES benchmark by setting $\Upsilon(x) = x^{\frac{\sigma-1}{\sigma}}$, 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 Lemma 1,

$$d \log A = 0. \tag{39}$$

Applying Proposition 4, the slope of the CPI Phillips curve is

$$d \log P^Y = \frac{1 + \gamma\zeta}{\zeta} \frac{\delta}{1 - \delta} d \log Y. \tag{40}$$

This is the traditional New Keynesian Phillips Curve.[26] Nominal rigidities, captured by the Calvo parameter $\delta < 1$, flatten the Phillips curve. As δ approaches one, prices become perfectly flexible, monetary shocks load on price rather than output, and the Phillips curve becomes vertical.

Homogeneous Firms Example.

We now relax the assumption of CES preferences, but consider a continuum of homogeneous firms: all firms have the same price-stickiness ($\delta_\theta = \delta$) 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., $\rho < 1$.[27]

Since markups are uniform, the cross-sectional allocation of resources across firms in

[26] See, for example, Galí (2015). Equation (40) can be replicated exactly from Galí (2015) pg. 63 by setting $\beta = 0$ and assuming constant returns to scale.

[27] The presence of incomplete pass-throughs has been most thoroughly documented in the international economics literature, where studies (see, for example, Gopinath et al. 2010) document incomplete pass-through of exchange-rate shocks into prices.

the initial equilibrium is efficient. Applying Lemma 1, we have

$$d \log A = 0. \tag{41}$$

Unlike the CES case, incomplete pass-throughs imply that flexible-price firms will not wholly 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 monetary shocks load to a greater degree on output. Following Proposition 4, the slope of the CPI Phillips curve is

$$d \log P^Y = \frac{1 + \gamma\zeta}{\zeta} \rho \frac{\delta}{1 - \delta} d \log Y. \tag{42}$$

The presence of real rigidities flatten the CPI Phillips curve compared to the CES case above. We measure the amount of flattening due to real rigidities as the the inverse ratio of the slope of the Phillips curve with real rigidities to the slope under CES preferences. In this case, the amount of flattening is

$$\frac{\text{Phillips curve slope w/ nominal rigidities only}}{\text{Phillips curve slope w/ real rigidities}} = \frac{1}{\rho}. \tag{43}$$

That is, the amount of flattening increases as ρ deviates further from the complete pass-through benchmark.

4.3 Discussion

The one-period model presented in the above sections helps to illustrate the key mechanisms that drive an endogenous aggregate TFP response to monetary shocks: namely, the interaction of heterogeneous markups with heterogeneity in either pass-throughs or price-stickiness.

Our model predicts that an expansionary demand shock simultaneously causes an increase in aggregate productivity and a decrease in the dispersion of firm-level revenue productivity (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. Moscarini and Postel-Vinay (2012) offer evidence that employment at large firms is more sensitive to the business cycle than small firms; in our single factor model, this is compatible with the cyclical reallocation to high-markup firms that we

describe.[28]

A key implication of our model is that it links the slope of the Phillips curve to objects of the firm distribution, such as elasticities, pass-throughs, and markups.[29] This finding means that industrial organization—and in particular industrial concentration—play 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.

Our results naturally apply to any specific parametric aggregator (e.g., Klenow and Willis (2016) preferences). More generally, the productivity effects of monetary policy that we identify also appear in models of oligopolistic competition that are populated by a discrete number of firms instead of a continuum of infinitesimal firms in monopolistic competition. As discussed above, the nested CES model of Atkeson and Burstein (2008) generates markups and pass-throughs that conform with Marshall's second law of demand, and that model yields similar implications (we show this in Appendix F). 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 the *New Keynesian Model with Misallocation*, a four-equation system that generalizes the workhorse three-equation model in Galí (2015) to account for imperfect pass-through and endogenous aggregate TFP. We provide a high-level walk-through of the derivation to highlight the key intuitions; the detailed setup is available 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 homo-

[28] We interpret these results with caution, since as noted by Fort et al. (2013), comparisons of the sensitivity of small and large firms to the business cycle also pick up confounders such as firm age. Recently, Crouzet and Mehrotra (2020) find that the sensitivity of sales to the business cycle is the same across the firm size distribution, except for the top 1% of firms who are less sensitive to business cycles. There is no significant difference in the response of sales to monetary policy shocks across the entire firm distribution. We note that reallocations in sales shares are not equivalent to reallocations in resources between firms, due to changes in markups. For instance, a net reallocation of resources to high-markup firms may be consistent with a decrease in the sales shares of those firms, provided that their realized pass-throughs are sufficiently small.

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

geneous price-stickiness across firms. Households consume according to the 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,

$$d\log i_t = \phi_\pi d\log \pi_t + \phi_y d\log Y_t + v_t, \tag{44}$$

where $d\log \pi_t = d\log P_t^Y - d\log P_{t-1}^Y$ is CPI inflation in period t, and v_t is a monetary policy shock.

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

$$\max_{p_{i,t}} \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1}(1+r_{t+j})}(1-\delta_i)^k y_{i,t+k}(p_{i,t} - \frac{w_{t+k}}{A_i})\right]. \tag{45}$$

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 6.

Proposition 6. *Suppose $\delta_\theta = \delta$ for all firms. The movements of aggregate variables are described by the following four-equation system.*

Taylor Rule.

$$d\log i_t = \phi_\pi d\log \pi_t + \phi_y d\log Y_t + v_t. \tag{46}$$

Dynamic IS equation.

$$d\log Y_t = d\log Y_{t+1} - \frac{1}{\gamma}(d\log i_t - d\log \pi_{t+1}). \tag{47}$$

Misallocated NK Phillips Curve.

$$d\log \pi_t = \beta d\log \pi_{t+1} + \varphi \mathbb{E}_\lambda\left[\rho_\theta\right]\frac{1+\gamma\zeta}{\zeta} d\log Y_t - \alpha d\log A_t. \tag{48}$$

Endogenous TFP Equation.

$$d\log A_t = \frac{1}{\kappa_A} d\log A_{t-1} + \frac{\beta}{\kappa_A} d\log A_{t+1} + \frac{\varphi}{\kappa_A}\frac{1+\gamma\zeta}{\zeta}\bar{\mu}\frac{Cov_\lambda[\rho_\theta, \sigma_\theta]}{\mathbb{E}_\lambda[\rho_\theta]} d\log Y_t, \tag{49}$$

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

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 latter two equations. Consider the Misallocated New Keynesian Phillips Curve (NKPC). We note two key differences: first, in the standard New Keynesian Phillips Curve, the coefficient on $d\log Y_t$ is $\varphi\frac{1+\gamma\zeta}{\zeta}$.[30] In the Misallocated NKPC, 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 Phillips curve. Second, changes in the 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, further flattening the Phillips curve.

Finally, the path of aggregate TFP is pinned down in the fourth equation. 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.

The New Keynesian Model with Misallocation generalizes the static model presented in the above sections, as commented on by Corollary 3.

Corollary 3. *Suppose output, aggregate TFP, and the CPI 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 2 and Proposition 3.*

[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. Pass-throughs for the largest firms in our data rarely fall below 0.3, which suggest $\alpha > 0$ holds nearly always.

5.2 Solution Strategy

We present a high-level walk-through of a derivation of the Misallocated NKPC and the Endogenous TFP Equation.

We start with the firm maximization problem described in Equation (45). The optimal reset price $p^*_{i,t}$ for profit maximization satisfies

$$\mathbb{E}\left[\sum_{k=0}^{\infty} \frac{1}{\prod_{j=0}^{k-1}(1+r_{t+j})}(1-\delta_i)^k y_{i,t+k}\left[\frac{dy_{i,t+k}}{dp_{i,t}}\frac{p^*_{i,t}}{y_{i,t+k}}\frac{p^*_{i,t}-\frac{w_{t+k}}{A_i}}{p^*_{i,t}}+1\right]\right]=0. \tag{50}$$

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

In the limiting case where prices are completely flexible ($\delta_i = 1$), firm i sets its price each period as

$$d\log p^*_{i,t} = (1-\rho_i)d\log P_t + \rho_i d\log w_t, \tag{51}$$

where P_t is the price aggregator defined in Equation (5). The optimal price of a firm with a high pass-through (i.e., ρ_i close to one) moves closely with shocks to marginal cost from the nominal wage. Firms with low pass-through, on the other hand, place less weight on own marginal cost, and more weight on the expected price level in the economy.

With some manipulation, the log-linearization of Equation (50) yields,

$$d\log p^*_{i,t} = [1-\beta(1-\delta_i)]\sum_{k=0}^{\infty}\beta^k(1-\delta_i)^k\left[\rho_i d\log w_{t+k} + (1-\rho_i)d\log P_{t+k}\right]. \tag{52}$$

Compared to the case without nominal rigidities, a firm with $\delta_i < 1$ is forward looking, incorporating 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.

Next, we consider the movement of firm prices over time. We rewrite Equation (52) recursively, and for each firm type θ, as

$$d\log p^*_{\theta,t} = [1-\beta(1-\delta_\theta)]\left[\rho_\theta d\log w_t + (1-\rho_\theta)d\log P_t\right] + \beta(1-\delta_\theta)d\log p^*_{\theta,t+1}. \tag{53}$$

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,

$$\mathbb{E}\left[d\log p_{\theta,t}\right] = \delta_\theta \mathbb{E}\left[d\log p^*_{\theta,t}\right] + (1-\delta_\theta)\mathbb{E}\left[d\log p_{\theta,t-1}\right] \tag{54}$$

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

$$\begin{aligned}\mathbb{E}[d\log p_{\theta,t} - d\log p_{\theta,t-1}] - \beta\mathbb{E}[d\log p_{\theta,t+1} - d\log p_{\theta,t}] \\ = \varphi\left[-\mathbb{E}[d\log p_\theta] + \rho_\theta d\log w_t + (1-\rho_\theta) d\log P_t\right],\end{aligned} \tag{55}$$

where

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

We are now almost finished. Equation (55) 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 (55), we can recover the movement of the CPI price index.[32] We get,

$$d\log\pi_t - \beta d\log\pi_{t+1} = \varphi\left[\mathbb{E}_\lambda\left[\rho_\theta\right](d\log w_t - d\log P_t) + (d\log P_t - d\log P^Y_t)\right]. \tag{56}$$

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 CPI $d\log P^Y_t$ —depend solely on the average markup and the distribution of markups. In particular, the following four identities allow us to express all quantities in terms of output and aggregate TFP:

$$d\log P_t - d\log P^Y_t = \bar{\mu}^{-1}\, d\log A_t \tag{57}$$

$$d\log P^Y_t - d\log w_t = \mathbb{E}_\lambda\left[d\log\mu_\theta\right] \tag{58}$$

$$d\log A_t = d\log\bar{\mu}_t - \mathbb{E}_\lambda\left[d\log\mu_{\theta,t}\right] \tag{59}$$

$$d\log Y_t = \frac{1}{1+\gamma\zeta}\left(d\log A_t - \zeta\mathbb{E}_\lambda\left[d\log\mu_{\theta,t}\right]\right) \tag{60}$$

Equations (58)-(60) have been seen before, but Equation (57) is new: it comes from log-linearizing and rearranging the expression for the price aggregator in (5).[33] Substituting

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

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

these identities into Equation (56) and rearranging yields the Misallocated NKPC in Proposition 6.

The Endogenous TFP Equation can also be derived by rearranging Equation (55). In particular, first note that variations in TFP depend on variations in firms' markups:

$$d\log A_t = d\log\bar{\mu} - \mathbb{E}_\lambda\left[d\log\mu_\theta\right] = \bar{\mu}\left(\frac{\mathbb{E}_\lambda\left[\sigma_\theta d\log\mu_{\theta,t}\right]}{\mathbb{E}_\lambda\left[\sigma_\theta\right]} - \mathbb{E}_\lambda\left[d\log\mu_\theta\right]\right). \tag{61}$$

The changes in markups can in turn be derived from (55) 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 (61).

5.3 Discussion

The model presented in Proposition 6 provides a tractable, four-equation system that can be used to calibrate economies with realistic heterogeneity in markups and pass-throughs. The system fully 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 would trigger similar reallocations across firms and hence also raise aggregate TFP, like the monetary shocks highlighted here.

Like the standard model, the New Keynesian Model with Misallocation can be used

$d\log P_t^Y + \bar{\mu}^{-1} d\log A_t$.

as a building block for more complex models. We offer one such extension to multiple sectors, multiple factors, input-output linkages, and sticky wages in Appendix D.

6 Empirical Calibration

We now turn to data to assess the empirical importance of the misallocation channel. This section is organized as follows. First, we describe our non-parametric estimation procedure. Second, we implement this procedure using empirical pass-through estimates from Amiti et al. (2019) and Belgian firm 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 $\Upsilon(\cdot)$ and tune parameters to match moments from the data. However, these parametric specifications of preferences may mute features of the data, giving rise to counterfactual properties in aggregate.[34]

Instead, we follow Baqaee and Farhi (2020a) and back out the shape of $\Upsilon(\cdot)$ from the data. To do this, note that, in the cross-section of firms, firm productivities A_θ and markups μ_θ must satisfy the following differential equations:

$$\frac{d\log\lambda_\theta}{d\theta} = \frac{\rho_\theta}{\mu_\theta - 1}\frac{d\log A_\theta}{d\theta}, \tag{62}$$

$$\frac{d\log\mu_\theta}{d\theta} = (1-\rho_\theta)\frac{d\log A_\theta}{d\theta}. \tag{63}$$

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

[34]See Footnote 7. As an example, in Appendix E, we show that Klenow and Willis (2016) preferences, commonly used in the literature, cannot simultaneously match the relationship between pass-throughs and firm size as well as average markups. Under standard calibrations of these preferences, pass-throughs vary very little as a function of firm size, and hence the covariance between markups and pass-throughs is counterfactually too low, which means that standard calibrations greatly understate the supply-side effects of a monetary shock.

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 $d \log \mu_\theta / d \log mc_\theta = \rho_\theta - 1$.

We note that these differential equations make use of a restriction implicit in the Kimball demand system: the time-series pass-through that describes a firm's price response to a cost shock is equal to the cross-sectional pass-through, which describes how firms' prices change as marginal cost changes over the productivity distribution.

Combining the two differential equations yields

$$\frac{d \log \mu_\theta}{d\theta} = (\mu_\theta - 1)\frac{1 - \rho_\theta}{\rho_\theta}\frac{d \log \lambda_\theta}{d\theta}. \tag{64}$$

Given sales shares λ_θ and pass-throughs ρ_θ, we can use this differential equation recover markups μ_θ up to a constant. We will choose the boundary condition to match 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 elasticities. The distributions of pass-throughs, markups, elasticities, and sales shares are sufficient to compute all results.

6.2 Data and Parameter Values

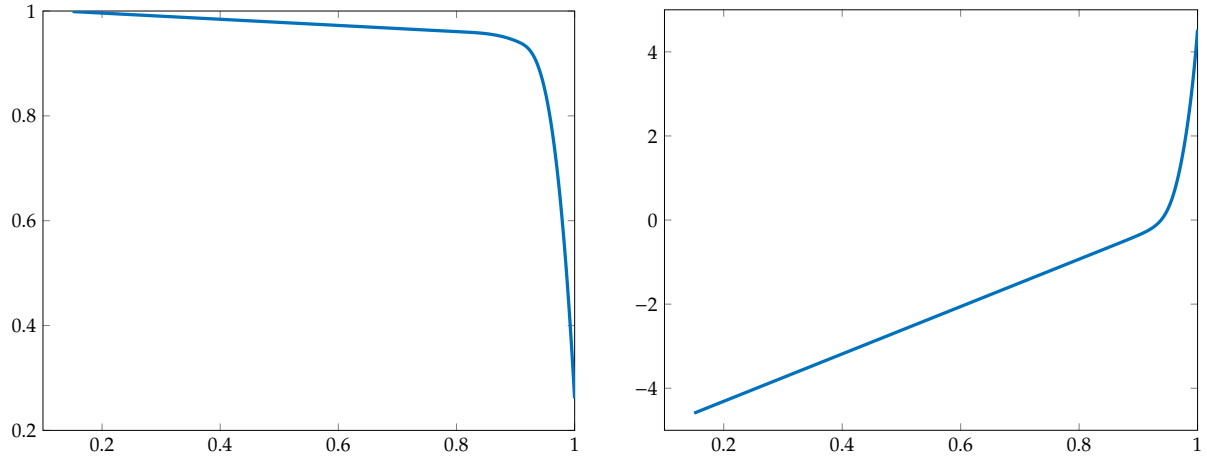
To implement this procedure, we use pass-throughs by firm size from Belgium estimated by Amiti et al. (2019), paired with data on the firm sales distribution from Belgium. Amiti et al. (2019) use annual firm-product level data (Prodcom) collected by Statistics Belgium from 1995-2007. Using exchange rate shocks as an instrument for changes in marginal cost, they are able to identify partial equilibrium pass-throughs by firm size for their sample of manufacturing firms.

We merge their estimates of pass-throughs ρ (as a function of size) with the sales distribution λ for the entire Belgian manufacturing sector.[35] Prodcom does not include very small firms, so we assume when extrapolating the Amiti et al. (2019) estimates to the universe of Belgian manufacturing firms that the pass-through for small firms not sampled by Prodcom is equal to one. This is consistent with Amiti et al. (2019), who estimate that the average pass-through for the smallest 75% of firms in Prodcom is 0.97.

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

[35] This procedure is also used in Baqaee and Farhi (2020a). We replicate the procedure here, and provide additional details in Appendix G.

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 $\log \lambda_\theta$

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 $\delta_\theta = \delta = 0.5$.

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

[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%.

Table 1: Parameters for empirical calibration.

Parameter	Estimate	Source
Statics calibration:		
$\bar{\mu}$	1.15	In line with micro-estimates from Konings et al. (2005), Amiti et al. (2019) and macro-estimates from Gutiérrez and Philippon (2017), Barkai (2020)
γ	1	Balanced growth preferences
ζ	0.2	Estimates from Chetty et al. (2011), Martinez et al. (2018), Sigurdsson (2019)
δ	0.5	At an average price duration of one year (see, e.g., Taylor 1999, Nakamura and Steinsson 2008), this implies a period length of 6 months
Dynamics calibration:		
δ	0.25	Period length of one quarter, at an average price duration of one year
ϕ_y	0.5 / 4	Galí (2015)
ϕ_π	1.5	Galí (2015)
β	0.99	Corresponding to 4% annual interest rate
ϵ_0	25bp	Galí (2015) (100bp annualized)
ρ_v	0.7	Strong persistence

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 5). 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 CPI 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 change as we vary the Frisch elasticity, the degree of industrial concentration, the average

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

markup, and the level of price-stickiness.

The Frisch elasticity.

We expect the misallocation channel to increase in importance as the Frisch elasticity decreases. Recall the intuition from Proposition 5: the change in output is given by

$$d \log Y = d \log A + d \log L, \tag{65}$$

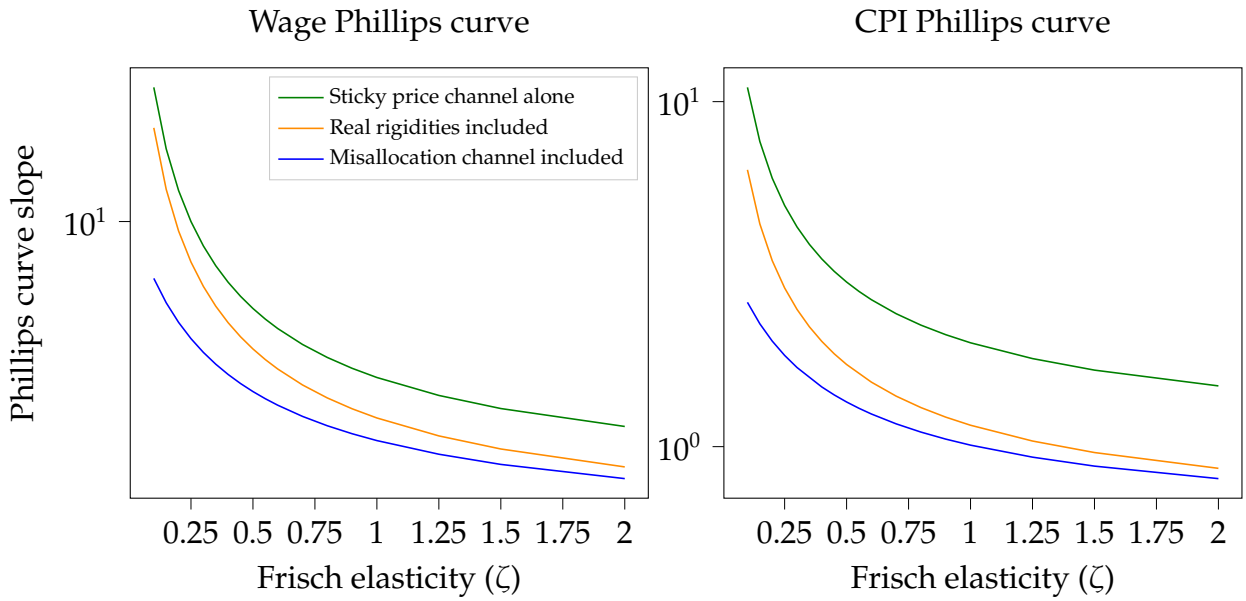
so changes in $d \log A$ dictate movements in $d \log Y$ when labor is inelastic.

We vary the Frisch elasticity and show the results in Figure 3. 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.

Incorporating the misallocation channel allows us to generate flatter Phillips curves at lower levels of the Frisch elasticity. At $\zeta = 2$, the flattening of the Phillips curve due to the misallocation channel is 7%; decreasing the Frisch elasticity to $\zeta = 0.1$ increases the flattening due to the misallocation channel dramatically to 142%. 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

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



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]

$$\frac{y_\theta}{A_\theta} \sim \text{Truncated Pareto}(\xi, H). \tag{66}$$

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 $\Upsilon(\cdot)$ estimated from the Belgian data, we pin down the distributions of markups, sales shares, and pass-throughs.

We present the slope of the Phillips curve as we vary the employment distribution parameter ξ. 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 decrease ξ.

To illustrate the role of concentration, we consider an increase in the Gini coefficient of firm employment, which measures the distribution's degree of inequality. Our model predicts that increasing the Gini coefficient from 0.80 to 0.85 flattens the CPI 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 G.2. 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 CPI Phillips curve by 26%.

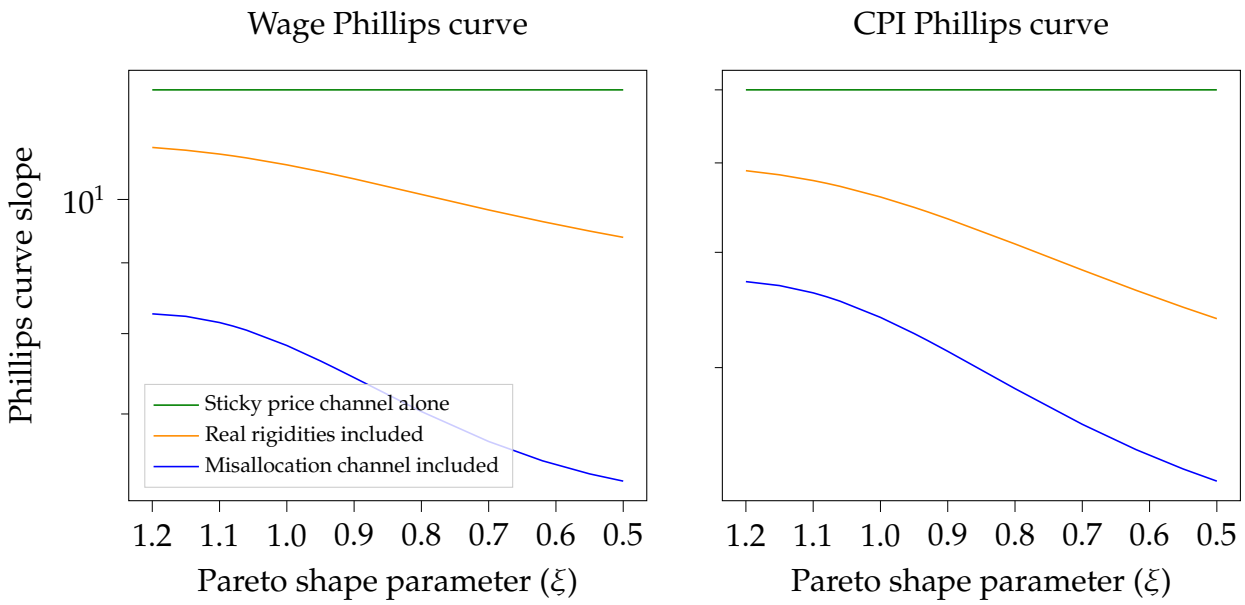
Other parameters.

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 G. We briefly summarize the results: Increasing the average markup $\bar{\mu}$ has no effect on the flattening due to real rigidities,

[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] Since the Gini coefficient rises monotonically as we decrease ξ, we associate the effect of increasing the Gini coefficient with the effect of decreasing ξ to generate the desired change in the Gini coefficient. 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: Decomposition of Phillips curve slope, varying ξ, which governs the distribution of employment.



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.

Increasing the price flexibility δ increases the strength of the misallocation channel, and decreases the flattening of the CPI Phillips curve due to real rigidities, for the reasons explained after the statement of Proposition 5.

6.4 Results from Dynamic Model

We calibrate the dynamic model using the pass-through and sales data from Belgium. 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 Lemma 1. In contrast, when firms have heterogeneous markups, the dispersion in TFPR across firm types increases by around 30% 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 deflation following the contractionary monetary shock and deepens the immediate response of output to the shock.

[39]For comparison, Kehrig (2011) finds that TFPR dispersion increases about 10% during recessions and

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.

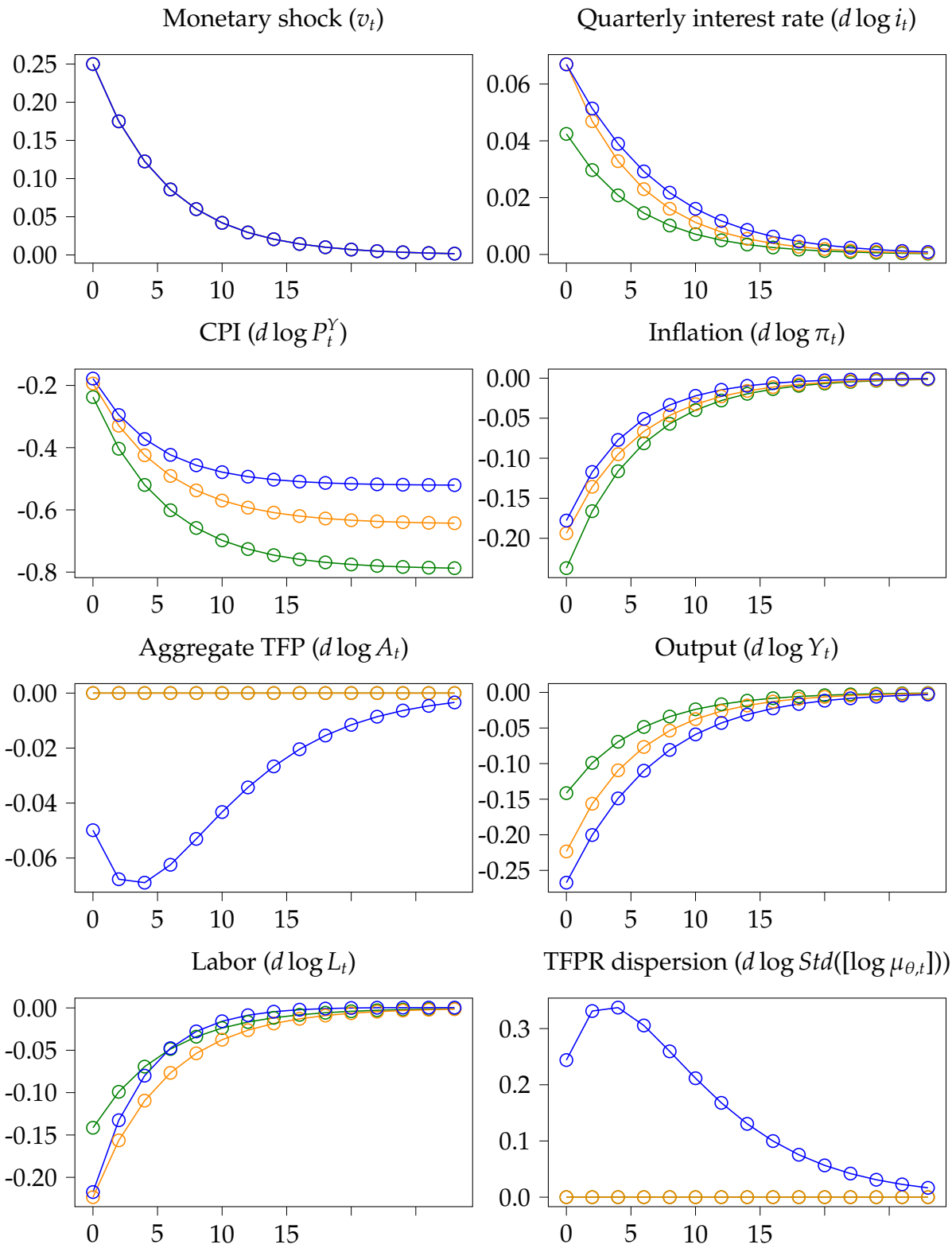


Figure 6: Impulse response function of output following a monetary policy shock.

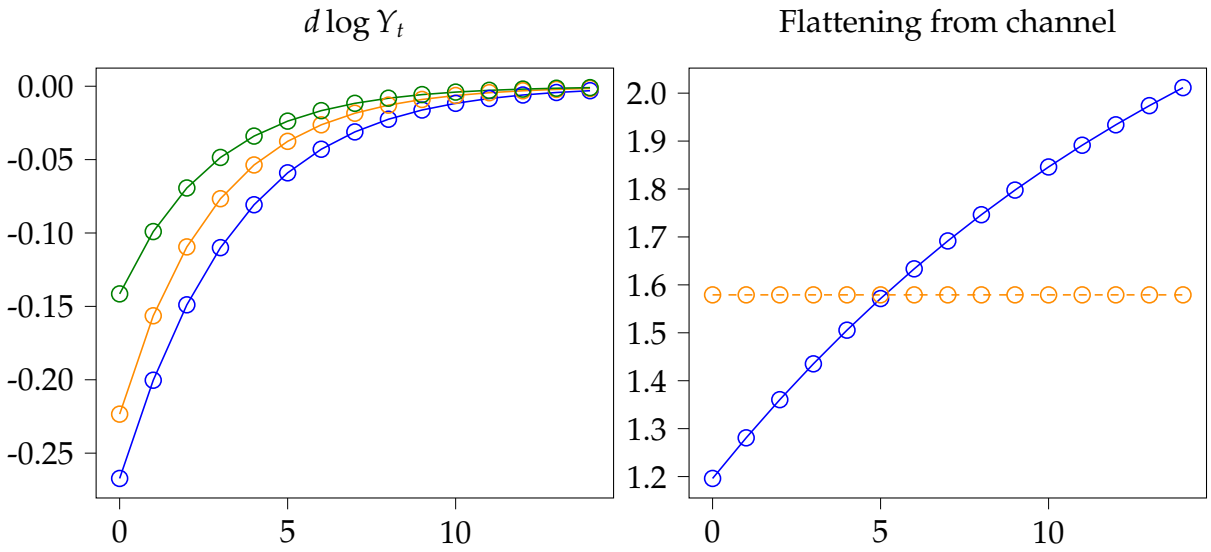


Figure 6 zooms in on the output response following the shock and computes the flattening due to real rigidities and the misallocation channel over time. We find that the misallocation channel deepens the contraction in output by about 20% on impact. The role of the misallocation rises over time, increasing the persistence of the shock's effect on output.

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 monetary shocks. For example, in Appendix C, 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.

increased over 20% from 2007 to the trough of the recession in 2009.

[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. The cumulative output impact is calculated as in Alvarez et al. (2016).

Model	Output effect at $t = 0$	Half life	Cumulative output impact
CES	-0.141	1.95	-0.472
With real rigidities	-0.223 (+58%)	1.95 (+0%)	-0.745 (+58%)
Full model	-0.267 (+20%)	2.39 (+23%)	-1.020 (+37%)

7 Extensions

Before concluding, we describe some extensions that are developed in the appendix.

Multiple sectors, multiple factors, input-output linkages, and sticky wages.

In the main text of the paper, we model an economy with a single sector and a single factor, labor. In Appendix D, we allow the economy to have a general network production structure, with multiple sectors and multiple factors. As an example, in Appendix D.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.

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 E. 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, 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 F, the misallocation channel remains quantitatively important in the oligopoly calibration.

Additional calibration results.

In the calibration section of the paper, we have provided comparative statics of our calibration results with respect to the Frisch elasticity and the degree of industrial concentration. In Appendix G, we additionally 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 C.

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 the enriched model features supply-side effects of monetary shocks on aggregate productivity and output.

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—such as the efficient cross-sectional allocation of resources in the steady state—mute important channels by which monetary policy, or demand more generally, affects the economy. In this paper, we abstract away from entry and exit and do not consider optimal monetary policy. We are pursuing these extensions in ongoing work.

References

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

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

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

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

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

Auer, R. A. 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.

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.

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.

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.

Crouzet, N. and N. R. Mehrotra (2020). Small and large firms over the business cycle. *American Economic Review 110*(11), 3549–3601.

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.

Fort, T. C., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2013). How firms response to business cycles: the role of firm age and firm size. *IMF Economic Review 61*(3), 520–559.

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.

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

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.

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

Melitz, M. J. (2018). Competitive effects of trade: Theory and measurement. *Review of World Economics 154*(1), 1–13.

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.