The Supply-Side Effects of Monetary Policy

David R. Baqaee

University of California, Los Angeles

Emmanuel Farhi

Harvard University

Kunal Sangani

Harvard University

We propose a supply-side channel for the transmission of monetary policy. We show that when high-markup firms have lower pass-throughs than low-markup firms, then positive demand shocks, such as monetary expansions, alleviate cross-sectional misallocation by reallocating resources to high-markup firms. Consequently, positive "demand shocks" are accompanied by endogenous positive "supply shocks" that raise productivity and lower inflation. We derive a tractable, four-equation model where monetary shocks generate hump-shaped productivity responses. In our calibration, the supply-side effect amplifies the total impact of monetary shocks on output by about 70%. We provide empirical evidence validating our model's predictions using identified monetary shocks.

I. Introduction

How do demand shocks, such as monetary shocks, affect an economy's productivity? A common view is that they do not. Instead, aggregate productivity

Emmanuel Farhi tragically passed away in July 2020. He was a once-in-a-lifetime friend and collaborator, and we dedicate this paper to his memory. We are grateful to Harald Uhlig and three anonymous referees for helpful suggestions. We thank Andy Atkeson, Saki

Electronically published March 12, 2024

Journal of Political Economy, volume 132, number 4, April 2024.

https://doi.org/10.1086/727287

is determined by long-run institutional and technological forces that are orthogonal to short-run demand disturbances.

Yet, aggregate productivity, as measured by labor productivity or the Solow residual, is sensitive to demand shocks. In fact, variations in monetary and fiscal policy explain between a quarter and a half of the observed movements in aggregate total factor productivity (TFP) at business-cycle frequencies (see, e.g., Evans 1992). This empirical finding is robust across time and across countries.[1] One interpretation of this result is that aggregate productivity is mismeasured—for example, due to variable capacity utilization or external returns—resulting in a spurious relationship between measured productivity and shifts in demand.

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

The effect of monetary shocks on the allocation of resources yields a new channel through which monetary policy affects real variables, which we call the *misallocation channel*. Under conditions matching empirical patterns on firms, monetary shocks generate procyclical, hump-shaped movements in aggregate TFP. The endogenous supply shock caused by the misallocation channel complements the traditional effects of the demand shock on employment and output. Incorporating the misallocation channel heightens the response of output to demand shocks and dampens the response of prices. For example, an expansionary monetary shock boosts aggregate TFP, leading to a larger increase in output without as much inflation. Hence, the misallocation channel increases monetary nonneutrality and flattens the Phillips curve.

Bigio, Ariel Burstein, Oleg Itskhoki, Ivan Werning, Jon Vogel, and other seminar participants for helpful comments. This paper received support from the National Science Foundation (grant 1947611). This paper was edited by Harald Uhlig.

[1] The failed invariance of aggregate TFP to demand shocks is also observed by Hall (1990). Cozier and Gupta (1993), Evans and dos Santos (2002), and Kim and Lim (2004) extend the analysis to Canada, the Group of Seven countries, and South Korea.

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

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

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

To formally analyze these reallocations, we relax the CES demand system in the New Keynesian model using a nonparametric generalized Kimball (1995) demand system.[3] These preferences can accommodate variety-specific downward-sloping residual demand curves of any desired shape while remaining tractable. We couple this demand system with sticky prices using Calvo (1983) frictions.[4] Our model is flexible enough to exactly match cross-sectional and time-series estimates of the firm size distribution and firm-level pass-throughs, with realistic heterogeneity in

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

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

firms' price elasticities of demand and desired markups. We consider how TFP and output respond to an aggregate demand shock that raises nominal costs in such a model. Our comparative statics do not impose any additional parametric structure on preferences and are disciplined by measurable sufficient statistics from the distribution of firms.

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

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

While we illustrate these intuitions in a single-period model, we also extend the benchmark, infinite-horizon New Keynesian model to incorporate these supply-side effects. In the dynamic model, changes in aggregate TFP, output, inflation, and the interest rate satisfy a four-equation system.[7] Relative to the benchmark model, the Taylor rule and the Euler equation are the same but the New Keynesian Phillips curve (NKPC) is different. Our model features a flatter Phillips curve with endogenous cost-push shocks due to shifts in aggregate TFP. Those movements in aggregate TFP are pinned down by the fourth equation, which closes the system. Our model

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

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

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

is disciplined by four sufficient statistics from the firm distribution: the average markup, the average price elasticity of demand, the average desired pass-through, and the covariance of markups and desired pass-throughs.

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

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

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

[8] We follow Baqaee, Farhi, and Sangani (2021) and solve a differential equation to back out the Kimball demand system from data on firm-level sales and pass-throughs. This approach is also preferable to using an off-the-shelf functional form for preferences since it does not impose the counterfactual restrictions baked in by parametric families of preferences. We provide an explicit calibration exercise in app. G (apps. A–J are available online) showing that the most popular off-the-shelf functional form, Klenow and Willis (2016), is incapable of simultaneously matching all the relevant sufficient statistics in the data.

line with the change in the firm employment distribution in the United States from 1978 to 2018.[9]

Using identified monetary shocks, we provide empirical support for both the macro- and the micro-level predictions of our model. At the macroeconomic level, we show that aggregate productivity in the United States—as measured by labor productivity, the Solow residual, or the cost-based Solow residual—is responsive to Romer and Romer (2004) monetary shocks, in line with the findings of Evans (1992).[10] At the microeconomic level, our model ties the increase in aggregate productivity during demand-driven expansions to reallocations toward high-markup firms. Using Compustat data on public firms, we find that expansionary monetary shocks cause high-markup firms to grow relative to low-markup firms in terms of their input usage. This is because firms with high markups cut their markups relative to low-markup firms after a monetary expansion.[11] As a result, both markup dispersion and the dispersion of firm-level revenue productivity (TFPR) fall during demand-driven expansions (as documented by Kehrig 2011; Meier and Reinelt 2020). Finally, in keeping with our model's predictions, we show that productivity is more responsive to monetary shocks in industries with higher concentration (measured by the market share of top firms).

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

A strand of this literature is devoted to explaining the strength and persistence of the real effects of monetary policy shocks, which cannot be explained by nominal rigidities alone given the frequency of price adjustment. Ball and Romer (1990) introduce real rigidities, which

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

[10] Specifically, we use the Wieland and Yang (2020) extension of the Romer and Romer (2004) shocks. We do not use capacity-utilization-adjusted measures of aggregate TFP (e.g., Basu, Fernald, and Kimball 2006; Fernald 2014) in our empirical exercises. This is because the exogeneity conditions used to identify utilization-adjusted TFP—that sectoral TFP is orthogonal to oil price shocks and monetary shocks—are invalid in our model. Indeed, our core result is that sectoral TFP is endogenous to such shocks.

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

complement nominal rigidities to increase monetary nonneutrality.[12] 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, Itskhoki, and Rigobon (2010) and Gopinath and Itskhoki (2011). Our paper complements this literature by showing that incomplete pass-through, when paired with firm-level heterogeneity, results in another mechanism by which monetary policy affects output.

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

The differential cross-sectional response of firms to monetary policy links the slope of the Phillips curve in our analysis to moments of the firm distribution, such as industrial concentration. Here our study

[12] Ball and Romer (1990) have also spawned a large literature of theoretical developments on real rigidities, which characterize the conditions under which real rigidities can generate observed levels of persistence in the real effects of monetary shocks. Kimball (1995) formulates a model where real rigidities arise from nonisoelasticity of demand curves. Eichenbaum and Fisher (2004) and Dotsey and King (2005) investigate how relaxing assumptions of constant elasticities of demand interact with other frictions to generate persistence. Klenow and Willis (2016) compare the predictions of models where real rigidities are generated by a kinked demand curve vs. sticky intermediate prices. Mongey (2021) shows that real rigidities can be more powerful (and the extent of pass-through significantly diminished) under dynamic oligopolistic competition.

is complemented by Etro and Rossi (2015), Andrés and Burriel (2018), Corhay, Kung, and Schmid (2020), and Wang and Werning (2020), who also discuss mechanisms by which an increase in concentration may contribute to a decline in inflation and flattening of the Phillips curve; our work is unique among these in identifying the misallocation channel of monetary policy as a potential source for this effect.

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

Structure of the paper.—Section II introduces a simple single-period model and defines the equilibrium. Sections III and IV describe the response of aggregate TFP and output (real GDP) to a monetary shock in the single-period model. Section V generalizes the static model from the previous sections to a fully dynamic setting. Section VI contains our quantitative results, including an extension with menu costs. Section VII provides empirical evidence at both the macro and the micro level for the mechanisms described in the model. In section VIII, we summarize some extensions discussed in more detail in the appendixes, including an alternative microfoundation using oligopolistic (rather than monopolistic) competition and versions of the model with multiple sectors, multiple factors, input-output linkages, and sticky wages. Section IX concludes. All proofs are in appendix A.

II. Model Setup

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

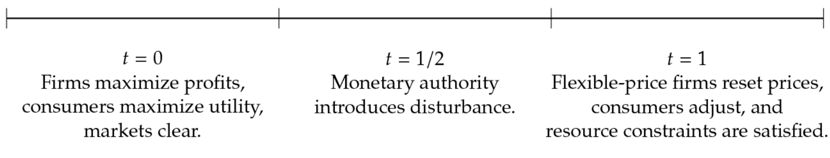

Fig. 1.—Single-period model timing.

profits, while firms with sticky prices keep prices unchanged from the initial equilibrium. Households adjust consumption and labor to maximize utility.[13] We describe the behavior of households, firms, and the monetary authority in turn.

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

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

where $1/\gamma$ represents the intertemporal elasticity of substitution and ζ represents the Frisch elasticity of labor supply. The consumption bundle Y consists of different varieties of goods indexed by $\theta \in [0, 1]$. Consumers have homothetic preferences over goods, and the consumption bundle Y is defined implicitly by[14]

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

Here y_θ represents the consumption of variety θ and Φ_θ is an increasing and concave function. CES preferences are the special case when $\Phi_\theta = \Phi$ is a power function.

The representative consumer maximizes utility subject to the budget constraint

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

where w represents the wage, L represents total hours, and Π represents profits. Maximization yields the inverse-demand curve for variety θ:

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

where the *price aggregator* P is given by

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

[13] We relax the single-period-ahead Calvo friction when we introduce the dynamic model in sec. V. In the infinite-horizon model, each firm changes its price at a constant hazard rate.

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

and P^Y represents the ideal price index.[15] Equation (1) shows that relative demand for a variety θ is dictated by the ratio of its price to the price aggregator P. Hence, firms compete with the rest of the market via a single price and quantity aggregator. Equation (1) also illustrates the appeal of these preferences: we can create downward-sloping demand curves of any desired shape by choosing an appropriate type-specific aggregator Φ_θ.

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

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

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

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

subject to its residual demand curve (1).

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

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

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

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

where the markup function is given by the Lerner formula,[16]

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

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

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

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

A firm of type θ has a probability δ_θ of being able to reset its price at time $t = 1$. These nominal rigidities may be heterogeneous across firm types. Flexible-price firms reset prices in $t = 1$ according to the optimal price and markup formulas above, and sticky-price firms keep their prices unchanged.

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

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

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

Monetary authority.—At time $t = 1/2$, there is an unexpected shock to the nominal wage. We interpret this shock as a disturbance introduced by the monetary authority. We could equivalently have the monetary authority choose any other nominal variable in the economy, such as the overall price level or money supply; the nominal wage is especially convenient, as it directly affects the marginal cost of every firm.[17]

We say that the shock is expansionary if the nominal wage in period 1 is higher than the one in period 0, since in this case the increase in nominal marginal cost decreases markups for firms whose prices cannot adjust, and this reduction in markups boosts labor demand and hence output.

[17] For concreteness, we interpret increases in nominal marginal cost $d \log w > 0$ to be the consequence of monetary easing. However, the basic intuition will apply to other kinds of demand shocks as well, since other shocks to aggregate demand will also raise nominal marginal costs and hence lead to productivity-increasing reallocations. In the dynamic version of the model in sec. V, changes in the nominal wage can be caused by either interest rate shocks in the Taylor rule or discount factor shocks in the Euler equation.

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

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

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

We write $\mathbb{E}$ to denote $\mathbb{E}_x$ when $x_\theta = 1$ for all θ. The operator $\mathbb{E}_x$ operates a change of measure by putting more weight on types θ with higher values of x_θ. We denote the sales share density of firm type θ by[18]

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

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

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

Log-linearization around initial equilibrium.—In what follows, we consider first-order perturbations around an initial equilibrium caused by a change in the nominal wage. For any variable X, we denote its log deviation from its initial value as $d\log X$. More formally, since all variables in this single-period model can be written as implicit functions of the wage w, we use $d\log X$ as a shorthand for $d\log X/d\log w \times \Delta\log w$, where $\Delta\log w$ represents a small change in w and the derivatives are evaluated at the initial steady state.[19]

III. Productivity Response

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

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

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

$$Y = AL.$$

Since labor is the sole factor in our model economy, A equals both aggregate TFP and aggregate labor productivity.[20]

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

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

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

LEMMA 1 (TFP and changes in markups). Following a monetary shock, the change in aggregate productivity is given by

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

Proof. By Shepard's lemma, $d \log P^{Y} = \mathbb{E}_{\lambda}[d \log \mu_{\theta}] + d \log w$. Substitute this into $d \log A = d \log P^{Y} Y - d \log P^{Y} - d \log L$ and use the fact that log changes in the labor share of income are negatively related to log changes in the average markup: $d \log(wL/P^{Y} Y) = -d \log \bar{\mu}$.[21] QED

Lemma 1 demonstrates that when the average markup rises more than individual markups on average ($d \log \bar{\mu} > \mathbb{E}_{\lambda}[d \log \mu_{\theta}]$), aggregate productivity A increases due to a composition effect toward firms with higher markups. To see this composition effect explicitly, expand the change

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

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

in the aggregate markup, $d\log\bar{\mu} = -\mathbb{E}_\lambda[(\bar{\mu}/\mu_\theta)d\log(\lambda_\theta/\mu_\theta)]$, and substitute it into (5). This yields our next lemma.

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

$$d\log A = -\mathrm{Cov}_\lambda[(\bar{\mu}/\mu_\theta), d\log\lambda_\theta/\mu_\theta] = -\mathrm{Cov}_\lambda[(\bar{\mu}/\mu_\theta), d\log \text{ Costs}_\theta], \quad (6)$$

where $\text{Costs}_\theta = wl_\theta$ are proportional to $\lambda_\theta/\mu_\theta$.

Aggregate productivity rises when changes in inputs, $d\log\text{Costs}_\theta$, negatively covary with inverse markups $1/\mu_\theta$.[22] In this case, labor is reallocated from low- to high-markup firms. Since high-markup firms are inefficiently too small relative to low-markup firms, such a reallocation boosts aggregate productivity. Lemma 2 is quite general; it continues to hold in the dynamic version of the model (sec. V) and within each sector in a version of the model with intermediate inputs and multiple sectors.[23] A corollary of lemma 2 is that if initial markups are identical, then a monetary shock has no first-order effect on TFP regardless of how markups change (i.e., regardless of $d\log\mu_\theta$).

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

$$d\log y_\theta - d\log Y = -\sigma_\theta[d\log p_\theta - d\log P]. \quad (7)$$

Firms that lower their price relative to the market-level price expand in relative terms following a monetary shock. Using the fact that $d\log y_\theta = d\log l_\theta$, we can combine (7) and (6) to get[24]

$$d\log A = (\bar{\mu}/\mathbb{E}_\lambda[\sigma_\theta])\mathrm{Cov}_\lambda[\sigma_\theta, d\log p_\theta]. \quad (8)$$

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

[23] In a multisector model, changes in the gross productivity of a sector are given by lemma 2 as long as all firms within a sector buy inputs at the same prices (see app. F).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

TABLE 1
Reallocations Due to Covariance of Desired Pass-Through ρ and Price Elasticity σ, in Response to an Expansionary Shock

Price Elasticity σ	Low Pass-Through (ρ)	High Pass-Through (ρ)
Low σ (high markup)	**Price/markup falls relative to other firms**	Price/markup rises relative to other firms
High σ (low markup)	Price/markup falls relative to other firms	**Price/markup rises relative to other firms**

where $\mathbb{E}_\lambda[\rho_\theta|\text{flex}]$ and $\text{Cov}_\lambda[\sigma_\theta, \rho_\theta|\text{flex}]$ represent the average pass-through and the covariance of price elasticities and pass-throughs for firms conditional on having flexible prices.[25] Using (12) with (11) yields the main result of this section in proposition 1.

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

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

and κ_ρ and κ_δ are nonnegative constants

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

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

A. *Mechanism I: Heterogeneous Desired Pass-Through*

If price stickiness is homogeneous across firms ($\delta_\theta = \delta$), then proposition 1 simplifies to the following.

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

$$d\log A = \kappa_\rho \text{Cov}_\lambda[\sigma_\theta, \rho_\theta]\, d\log w, \quad (\kappa_\rho \geq 0).$$

Table 1 illustrates why a positive covariance between price elasticities and pass-throughs leads to an increase in aggregate TFP following an expansionary shock ($d\log w > 0$). Firms with high pass-throughs increase their prices by more than firms with low pass-throughs. When price elasticities positively covary with pass-throughs, firms predominantly lie on

[25] That is, $\mathbb{E}_\lambda[\rho_\theta|\text{flex}] = \mathbb{E}_{\lambda\delta}[\rho_\theta]$ and $\text{Cov}_\lambda[\sigma_\theta, \rho_\theta|\text{flex}] = \text{Cov}_{\lambda\delta}[\sigma_\theta, \rho_\theta]$.

the boldface diagonal axis in table 1: the relative price of firms with initially high markups fall relative to other firms, reallocating resources toward those firms. By lemma 2, this boosts aggregate productivity.

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

DEFINITION 1. *Marshall's third law of demand* states that desired markups are increasing in quantity and desired pass-throughs are decreasing in quantity.[26] That is,

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

If Marshall's third law holds and firms face the same residual demand curve, then a monetary expansion will raise aggregate productivity. This is because large firms will have higher markups and lower pass-throughs. Marshall's third law of demand has strong empirical support (see, e.g., empirical estimates of pass-throughs by firm size from Amiti, Itskhoki, and Konings 2019) and holds in a variety of models. For example, oligopolistic competition models, such as Atkeson and Burstein (2008), satisfy Marshall's third law of demand.[27]

While Marshall's third law is sufficient to generate a positive covariance in corollary 1, it is not necessary. Markups and pass-throughs may be correlated for reasons unrelated to firm size, such as quality or nicheness (e.g., as shown empirically by Chen and Juvenal 2016; Auer, Chaney, and Sauré 2018).

B. *Mechanism II: Heterogeneous Price Stickiness*

Consider the case where pass-through is instead homogeneous but price stickiness is not.

COROLLARY 2 (Heterogeneous price rigidity). If desired pass-through is homogeneous across firms ($\rho_\theta = \rho$),[28] then

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

Table 2 illustrates why a positive covariance between markups and price stickiness causes an expansionary shock ($d\log w > 0$) to increase

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

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

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

TABLE 2
Reallocations Due to Covariance of Price Stickiness δ and Price Elasticity σ, in Response to an Expansionary Shock

Price Elasticity σ	More Sticky Firms (Low δ)	More Flexible Firms (High δ)
Low σ (high markup)	**Price/markup falls relative to other firms**	Price/markup rises relative to other firms
High σ (low markup)	Price/markup falls relative to other firms	**Price/markup rises relative to other firms**

TFP. In response to an increase in the nominal wage, firm types with more flexible prices will raise their prices relative to firms with less flexible prices. If high-markup firms are more sticky than low-markup firms, then firms predominantly lie on the boldface diagonal axis in table 2. This results in a reallocation of resources toward firms with initially high markups and away from firms with initially low markups, thereby improving allocative efficiency as per lemma 2.

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

Although we allow for the possibility that price stickiness varies systematically with firm type, we do not pursue this mechanism further and point interested readers to Meier and Reinelt (2020). When we quantify the model, we assume that there is no variation in price stickiness and instead focus on heterogeneity in desired pass-through only. This is because whereas there is robust empirical support for Marshall's third law of demand, the covariance of price stickiness with markups is less well documented.[29]

IV. Output Response and the Phillips Curve

In the above section, we showed that aggregate TFP can respond to monetary shocks. In this section, we show how monetary shocks are transmitted to output, taking into account the endogenous response of aggregate productivity. We show that the change in output can be decomposed into three channels: (1) nominal rigidities (as in a CES economy with sticky prices), (2) real rigidities due to imperfect pass-through (which arise from strategic complementarities in pricing à la Kimball 1995), and (3) the *misallocation channel*, which is due to the endogenous response of aggregate TFP.

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

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

A. *Output Response*

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

PROPOSITION 2 (Output response). Following a shock to the nominal wage $d\log w$, the response of output is

$$d\log Y = \underbrace{\frac{1}{1+\gamma\zeta}\, d\log A}_{\text{supply-side effect}} + \underbrace{\frac{\zeta}{1+\gamma\zeta}\,\mathbb{E}_\lambda[-d\log\mu_\theta]}_{\text{demand-side effect}}, \tag{15}$$

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

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

Equation (15) breaks down the response of output into a supply-side and demand-side effect. The demand-side effect of an expansionary shock arises from the average reduction in markups, which increases labor demand (and employment). The supply-side effect is due to changes in aggregate TFP and arises from changes in the economy's allocative efficiency.

Equation (16) further decomposes the demand-side effect into the effect of sticky prices and the effect of real rigidities. The first is the standard New Keynesian channel: nominal rigidities prevent sticky-price firms from responding to the shock. As a result, markups fall for a fraction $\mathbb{E}_\lambda[1-\delta_\theta]$ of firms. This reduction in the markups of sticky-price firms boosts labor demand, employment, and ultimately output.

This sticky-price effect in (16) is amplified by real rigidities, which arise from imperfect pass-through. When pass-through is incomplete, flexible-price firms increase prices less than one-for-one with the marginal cost shock. As a result, the markups of flexible-price firms also fall. Together, the reduction in the markups of both sticky-price and flexible-price firms increase labor demand, which spurs employment and output.

The supply-side effect, alternatively, is concerned with the efficiency with which labor is used. Returning to (15), we find that when aggregate TFP increases following an expansionary shock $(d\log A/d\log w > 0)$, the endogenous positive "supply shock" complements the effects of the positive "demand shock" on output. We term this channel the *misallocation channel.*

Interestingly, whereas the demand-side effect is increasing in the size of the elasticity of labor supply ζ, the supply-side effect is decreasing in ζ. In fact, the supply-side effect is strongest when labor is inelastically supplied ($\zeta = 0$). Alternatively, as the Frisch elasticity of labor supply approaches infinity, the supply-side effect becomes irrelevant for output. This is because reallocations to high-markup firms, which boost productivity, also have a negative effect on labor demand. When the Frisch is infinite, the positive reallocation benefits are exactly canceled out by reductions in employment, which contracts due to the expansion of high-markup firms.

B. *The Misallocation Channel and the Phillips Curve*

We now construct the Phillips curve—the relationship between the output gap and inflation generated by a demand shock—in the model and show that the misallocation channel flattens its slope.[30] We derive the slope of the wage Phillips curve by rearranging the output response in proposition 2. To get the price Phillips curve, we use the relationship between the CPI P^Y, the nominal wage, and average markups,

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

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

PROPOSITION 3 (Wage and price Phillips curves). Let $d \log A / d \log w$ and $d \log \mu_\theta / d \log w$ respectively denote the total derivatives of $\log A$ and $\log \mu_\theta$ with respect to the exogenous nominal wage $\log w$. The wage Phillips curve is given by

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

The price Phillips curve is given by

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

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

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

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

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

PROPOSITION 4 (Flattening of the Phillips curve). Suppose that $\delta_\theta = \delta$ for all firms. The flattening of the price Phillips curve due to real rigidities, compared with nominal rigidities alone, is

$$1 + \frac{\mathbb{E}_\lambda[\sigma_\theta]\mathbb{E}_\lambda[1-\rho_\theta]}{\delta \mathrm{Cov}_\lambda[\rho_\theta, \sigma_\theta] + \mathbb{E}_\lambda[\rho_\theta]\mathbb{E}_\lambda[\sigma_\theta]}. \tag{17}$$

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

$$1 + \frac{\bar{\mu}}{\zeta}\frac{\delta \mathrm{Cov}_\lambda[\rho_\theta, \sigma_\theta]}{\delta \mathrm{Cov}_\lambda[\rho_\theta, \sigma_\theta] + \mathbb{E}_\lambda[\sigma_\theta]}. \tag{18}$$

In equation (17), we see that the flattening of the Phillips curve due to real rigidities increases as average pass-throughs fall (as in Kimball 1995). The flattening due to real rigidities in (17) is also decreasing in price flexibility δ. As price flexibility increases, the price aggregator moves more closely with shocks to marginal cost; hence, the "pricing-to-market" effect from incomplete pass-throughs is less powerful.

The flattening of the Phillips curve due to the misallocation channel depends positively on the covariance of pass-throughs and elasticities ($\mathrm{Cov}_\lambda[\rho_\theta, \sigma_\theta]$). The misallocation channel also flattens the Phillips curve more when the Frisch elasticity ζ is low, since the supply-side effect is stronger when labor is inelastically supplied. Finally, since the expansion of high-markup firms relative to low-markup firms occurs only for flexible-price firms, the misallocation channel is relatively more important when prices are more flexible.

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

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

Since desired markups are uniform, the initial allocation of the economy is efficient and the misallocation channel is absent. Applying proposition 3, the slope of the price Phillips curve is

$$d\log P^Y = \frac{1+\gamma\zeta}{\zeta}\frac{\delta}{1-\delta}\,d\log Y. \tag{19}$$

This is the traditional NKPC.[31] Nominal rigidities, captured by the Calvo parameter $\delta < 1$, flatten the Phillips curve. As δ approaches one, prices become perfectly flexible and the Phillips curve becomes vertical.

Representative firm example.—We now relax the assumption of CES preferences but consider an economy with a representative firm: all firms have the same price stickiness ($\delta_\theta = \delta$), the same residual demand curve $\Phi'_\theta = \Phi'$, and the same productivity ($A_\theta = 1$). The homogeneous firms in this economy have identical markups, $\mu_\theta = \mu$, and pass-throughs, $\rho_\theta = \rho$. By deviating from CES, however, we allow firms' desired pass-throughs to be incomplete ($\rho < 1$).

Since markups are uniform, the cross-sectional allocation of resources across firms in the initial equilibrium is still efficient. Hence, as in the CES example, the misallocation channel is still absent. Unlike the CES case, however, incomplete desired pass-through implies that flexible-price firms will not fully adjust prices to reflect increases in marginal cost from a monetary shock. As noted by Kimball (1995), compared with the CES economy, prices in this economy are slower to respond, and hence the slope of the price Phillips curve is flatter:

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

In particular, proposition 4 implies that the amount of flattening due to the real rigidities channel is $1/\rho$.

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

[31] See, e.g., Galí (2015). Equation (19) can be replicated exactly from Galí (2015, 63) by setting $\beta = 0$ and assuming constant returns to scale.

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

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

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

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

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

C. *Discussion*

Before moving on to the dynamic version of the model, we discuss some of implications and extensions of the results in this section. First, unlike the standard model, our model links the slope of the Phillips curve to the industrial organization of the economy, via statistics such as the covariance of pass-throughs and price elasticities. This means that industrial concentration plays a role in shaping the Phillips curve. We consider this effect quantitatively in section VI, where we illustrate the effect of increasing industrial concentration on the Phillips curve slope.

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

V. The New Keynesian Model with Misallocation

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

A. *Four-Equation Dynamic Model*

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

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

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

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

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

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

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

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

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

As in Galí (2015), we log-linearize all variables around the no-inflation steady state. Macroeconomic aggregates such as output Y and aggregate productivity A are endogenous outcomes that depend on the path of shocks; for parsimony, we simply write log-linearized variables $d \log Y$ and $d \log A$, with the understanding that these endogenous variables are functions of the entire path of monetary and discount factor shocks.[32]

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

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

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

$$d\log i_t = \phi_\pi d\log \pi_t + \phi_y d\log Y_t + v_t, \qquad \text{(Taylor rule)}$$

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

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

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

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

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

Whereas the Taylor rule and Euler equation are the same as in the three-equation model, the last two equations are different. Start by considering the amended Phillips curve. We note two key differences: first, in the standard NKPC, the coefficient on $d\log Y_t$ is $\varphi((1+\gamma\zeta)/\zeta)$.[33] In proposition 5, this coefficient is multiplied by the average pass-through $\mathbb{E}_\lambda[\rho_\theta]$. As in the static version of the model, imperfect pass-through moderates the response of prices to nominal shocks and hence flattens the NKPC. More importantly, changes in aggregate TFP enter the Phillips curve as endogenous, negative cost-push shocks, given by $\alpha d\log A_t$.[34] This means that procyclical movements in aggregate TFP further dampen the response of inflation to an expansionary shock.

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

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

The final equation in proposition 5 pins down the path of aggregate TFP. When markups covary negatively with pass-throughs, output booms, $d\log Y_t > 0$, driven by either monetary shocks or discount factor shocks, are concomitant with improvements in aggregate productivity. Furthermore, unlike the standard New Keynesian model, which consists of only forward-looking terms, the movement of aggregate TFP depends on a backward-looking term. As a result, the augmented four-equation model may generate endogenous hump-shaped impulse responses to monetary shocks.

Proposition 5 also generalizes the static model presented in sections II–IV as shown by the following corollary.

Corollary 3 (Static model as special case). Suppose that output, aggregate TFP, and the price level are in steady state at $t = 0$. When the discount factor $\beta = 0$, the effect of shocks on impact are the same as the static results from propositions 1 and 2.

B. *Proof Sketch*

Before calibrating the model, we provide a high-level walk-through of the derivation for proposition 5 to highlight the key intuitions; the detailed derivation is in appendix A. The derivation of the Euler equation is standard, so we focus instead on the Phillips curve and the TFP equations. Start with the firm maximization problem described in equation (20). The optimal reset price $p_{i,t}^{\text{flex}}$ for profit maximization satisfies

$$\mathbb{E}\left[\sum_{k=0}^{\infty}\frac{1}{\prod_{j=0}^{k-1}(1+r_{t+j})}(1-\delta_i)^k y_{i,t+k}\left[\frac{dy_{i,t+k}}{dp_{i,t}}\frac{p_{i,t}^{\text{flex}}}{y_{i,t+k}}\frac{p_{i,t}^{\text{flex}}-(w_{t+k}/A_i)}{p_{i,t}^{\text{flex}}}+1\right]\right]=0. \quad (21)$$

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

With some manipulation, the log-linearization of equation (21) yields

$$d\log p_{i,t}^{\text{flex}} = [1-\beta(1-\delta_i)]\sum_{k=0}^{\infty}\beta^k(1-\delta_i)^k[\rho_i d\log w_{t+k} + (1-\rho_i)d\log P_{t+k}]. \quad (22)$$

When prices are fully flexible, this simplifies to the static optimality condition in (9). Compared with the case without nominal rigidities, a firm with sticky prices is forward looking and incorporates expected future prices and marginal costs into its reset price today. Just as in the completely flexible benchmark, firms with high pass-throughs are more responsive to expected changes in their own marginal costs, while firms with low pass-throughs are more responsive to expected changes in the economy's price aggregator.

Rewrite equation (22) recursively, and for each firm type θ, as

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

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

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

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

$$\begin{aligned}&\mathbb{E}[d\log p_{\theta,t} - d\log p_{\theta,t-1}] - \beta\mathbb{E}[d\log p_{\theta,t+1} - d\log p_{\theta,t}]\\ &\quad = \varphi[-\mathbb{E}[d\log p_{\theta,t}] + \rho_\theta d\log w_t + (1-\rho_\theta)\, d\log P_t],\end{aligned} \tag{23}$$

where $\varphi = \delta/(1-\delta)(1-\beta(1-\delta))$. Since equation (23) pins down type θ firms' average price over time, we can recover the movements of aggregate variables, such as the CPI, aggregate TFP, and output, by manipulating this expression and averaging over firm types.

For instance, by taking the sales-weighted expectation of both sides in equation (23), we recover the movement of the CPI:[35]

$$\begin{aligned} d\log\pi_t - \beta d\log\pi_{t+1} = \varphi\big[\,&\mathbb{E}_\lambda[\rho_\theta](d\log w_t - d\log P_t)\\ &+ (d\log P_t - d\log P^Y_t)\big].\end{aligned} \tag{24}$$

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 $d\log P^Y_t$—can be reexpressed in more familiar terms using the following identities:

$$d\log P_t - d\log P^Y_t = \bar{\mu}^{-1} d\log A_t, \tag{25}$$

$$d\log P^Y_t - d\log w_t = \frac{1}{\zeta}[d\log A_t - (1+\gamma\zeta)\, d\log Y_t]. \tag{26}$$

Equation (25) can be derived by log-linearizing and rearranging the expression for the price aggregator in (2),[36] and (26) comes from rearranging (15) for the average change in markups. Substituting these identities into (24) yields the Phillips curve in proposition 5.

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

[36] Using the fact that $d\log P = \mathbb{E}_{\lambda\sigma}[d\log p_\theta]$, we get $\bar{\mu}(d\log P - d\log P^Y) = \bar{\mu}(\mathbb{E}_{\lambda\sigma}[d\log p_\theta] - \mathbb{E}_\lambda[d\log p_\theta]) = d\log A$ from (27).

Movements in TFP also come from rearranging (23). From (5), we have

$$d\log A_t = d\log\bar{\mu} - \mathbb{E}_\lambda[d\log\mu_\theta] = \bar{\mu}(\mathbb{E}_{\lambda\sigma}[d\log\mu_{\theta,t}] - \mathbb{E}_\lambda[d\log\mu_{\theta,t}]). \quad (27)$$

The changes in markups can in turn be derived from (23) 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 (27).

VI. Quantitative Results

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

A. Nonparametric Calibration Procedure

It may be tempting to use an off-the-shelf functional form for Φ and tune parameters to match moments from the data. However, there is no guarantee that parametric specifications of preferences are able to match the relevant features of the data required for generating correct aggregate properties.[37] Instead, we follow Baqaee, Farhi, and Sangani (2021) and back out the shape of the Kimball aggregator nonparametrically from the data. We summarize this approach below.

Assume that Φ_θ takes the form

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

Hence, firms differ in their productivities A_θ and taste shifters B_θ. Allowing for taste shifters is important since, in practice, two firms that charge

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

the same price in the data can have very different sales, and taste shifters allow us to accommodate this possibility.

We order firms by their size and let $\theta \in [0,1]$ be firm θ's quantile in the size distribution. Baqaee, Farhi, and Sangani (2021) show that, in the cross section, markups and sales must satisfy the following differential equation:[38]

$$\frac{d\log\mu_\theta}{d\theta} = (\mu_\theta - 1)\frac{1-\rho_\theta}{\rho_\theta}\frac{d\log\lambda_\theta}{d\theta}. \qquad (28)$$

Given data on sales shares λ_θ and pass-throughs ρ_θ, we can use this differential equation to solve for markups μ_θ up to a boundary condition. We choose the boundary condition to target a given value of the (harmonic) sales-weighted average markup, $\bar{\mu}$. We then use $\sigma_\theta = 1/(1-1/\mu_\theta)$ to recover price elasticities. The distributions of pass-throughs, markups, price elasticities, and sales shares are the sufficient statistics we need to calibrate the model.[39]

B. Data and Parameter Values

We follow the implementation in Baqaee, Farhi, and Sangani (2021) (and refer interested readers to app. A of that paper for details). To calibrate the model, we need data on pass-throughs ρ_θ and the sales density λ_θ. For pass-throughs, we use estimates of (partial equilibrium) pass-throughs by firm size for manufacturing firms in Belgium from Amiti, Itskhoki, and Konings (2019).[40] We interpolate between their point estimates with smooth splines and assume that pass-throughs go to one for the smallest firms (they find that the average pass-through for the smallest 75% of firms is already 0.97). Figure 2 shows the pass-through ρ_θ and log sales

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

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

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

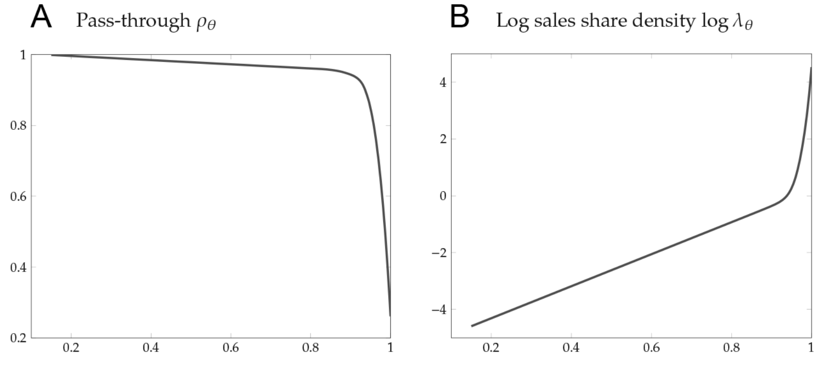

FIG. 2.—Pass-through ρ_θ and sales share density log λ_θ for Belgian manufacturing firms ordered by type θ.

share density log λ_θ as a function of θ. Pass-throughs are strictly decreasing with firm size, which means that Marshall's third law holds.

To compute the distribution of markups and elasticities from this data using equation (28), we must take a stance on the average markup. We assume that the average markup $\bar{\mu} = \mathbb{E}_\lambda[\mu_\theta^{-1}]^{-1} = 1.15$, in line with estimates from micro-data.[41]

To calibrate the rest of the model, we use standard values from the literature. We set the Frisch elasticity $\zeta = 0.2$, in line with recent estimates (see, e.g., Chetty et al. 2011; Martinez, Saez, and Siegenthaler 2018; Sigurdsson 2019), and set the intertemporal elasticity of substitution $\gamma = 1$. We consider a time period of one-quarter and set the Calvo parameter $\delta_\theta = \delta = 0.5$ according to an average price duration of about 6 months (Nakamura and Steinsson 2008). We specify the coefficients on the Taylor rule, ϕ_π and ϕ_y, to match the calibration in Galí (2015). For the dynamic model, we set the discount factor $\beta = 0.99$, corresponding to a 4% annual interest rate. We assume that monetary disturbances follow a first-order autoregressive model process $v_t = \rho_v v_{t-1} + \epsilon_t$; set

[41] The resulting markup function μ_θ is shown in fig. G.1 (figs. B.1–B.10, C.1–C.3, D.1–D.4, E.1, E.2, F.1, G.1–G.3, H.1, J.1, and J.2 are available online). The markup distribution we recover is consistent with direct estimates from the literature. Konings, Cayseele, and Warzynski (2005) use micro evidence to estimate price-cost margins in Bulgaria and Romania and find that average price-cost margins range between 5% and 20% for nearly all sectors. In the working paper version of Amiti, Itskhoki, and Konings (2019), they report that small firms in their calibration have a markup of around 14% and large firms have markups of around 30%. These micro estimated average markups are also broadly in line with macro estimates from Gutiérrez and Philippon (2017) and Barkai (2020), who estimate average markups on the order of 10%–20%. Edmond, Midrigan, and Xu (2018) also choose $\bar{\mu} = 1.15$.

TABLE 3
Calibrated Parameter Values for the Static and Dynamic Versions of the Model

Parameter	Description	Value
	Static Model	
$\bar{\mu}$	Aggregate markup	1.15
$1/\gamma$	Intertemporal elasticity of substitution	1
ζ	Frisch elasticity	.2
δ	Calvo friction	.5
	Additional Parameters for Dynamic Model	
ϕ_y	Output gap coefficient	.5 / 4
ϕ_π	Inflation coefficient	1.5
β	Discount factor	.99
ρ_v	Shock persistence	.7

$\rho_v = 0.7$, indicating strong persistence to the interest rate shock; and set the size of the initial interest rate shock to 25 basis points (bp). These parameter values are listed in table 3.

C. *Results from Static Model*

Table 4 reports the estimated flattening of the Phillips curve due to real rigidities and the misallocation channel in the static model (as given by proposition 4). We find that the misallocation channel is quantitatively important: compared with the real rigidities channel, which flattens the wage Phillips curve by 27% and the price Phillips curve by 73%, the misallocation channel flattens both Phillips curves by 71%.

To highlight the key forces at play in this calibration, we consider how these estimates change as we vary the Frisch elasticity and the degree of industrial concentration.[42]

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

The introduction of the misallocation channel—and its increased strength at low Frisch elasticities—may help explain the discrepancy between micro evidence on the Frisch elasticity and that required to explain

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

TABLE 4
Flattening of the Phillips Curve Due to Real Rigidities and the Misallocation Channel

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

the slope of the Phillips curve in traditional models. Evidence accumulated from quasi-experimental studies suggests that the labor supply elasticity is on the order of 0.1–0.4. To match the slope of the Phillips curve that the model with real rigidities and misallocation predicts at $\zeta = 0.2$, the model with nominal rigidities alone would require $\zeta \approx 1$. Incorporating the misallocation channel allows us to generate more monetary nonneutrality at lower levels of the Frisch elasticity.

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.

To illustrate the role of industrial concentration, we consider counterfactual firm distributions. To do so, we use a beta distribution for firm productivities, A_θ.[43] We choose the shape parameters of the beta distribution, $a = 0.14$ and $b = 15.7$, to match the Gini coefficient of firm employment in the Belgian data and the slope of the price Phillips curve in our baseline calibration.

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

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

[43] We choose the beta distribution since, as a bounded distribution, it allows us to remain within the range of productivities for which we have estimated the Kimball aggregator.

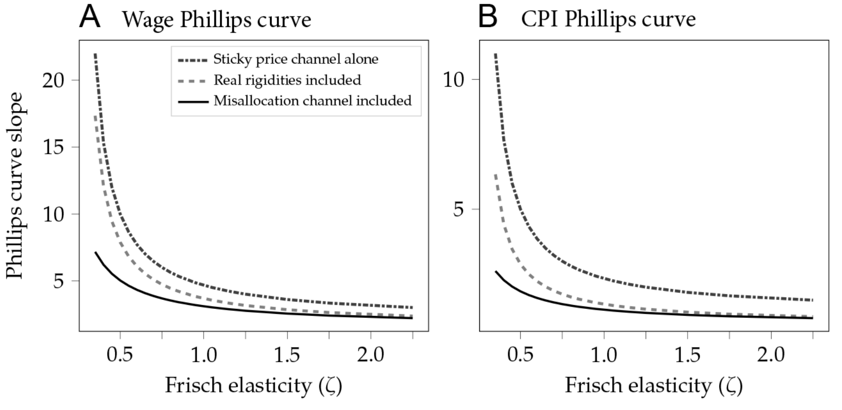

FIG. 3.—Decomposition of Phillips curve slope, varying the Frisch elasticity ζ.

with the increase in the Gini coefficient in firm employment from 1978 to 2018 in the United States (measured using the Census Business Dynamics Statistics; see app. J). Increasing the Gini coefficient from 0.72 to 0.86 (the increase in the Gini coefficient in the retail sector over the same period) flattens the price Phillips curve by 41%.

D. *Results from Dynamic Model*

Figure 5 shows the impulse response functions of aggregate variables following a persistent, 25-bp (100 bp annualized) shock to the interest rate in the dynamic model. We compare the benchmark heterogeneous firm model with a homogeneous firm model, which has real rigidities but no

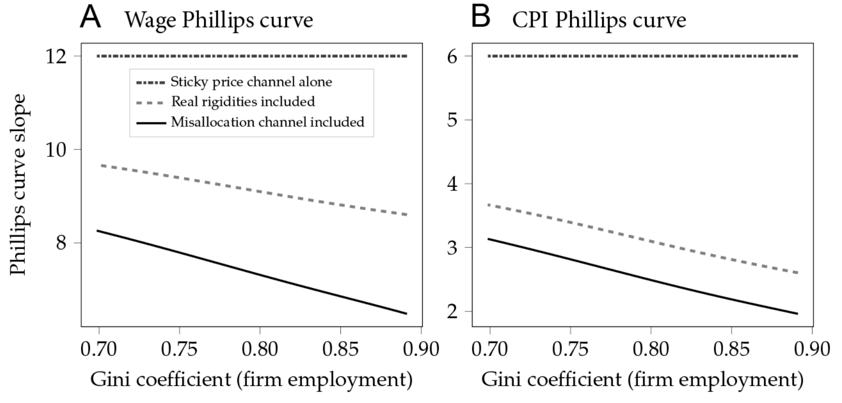

FIG. 4.—Slope of the Phillips curve and its decomposition as a function of the Gini coefficient of the employment distribution.

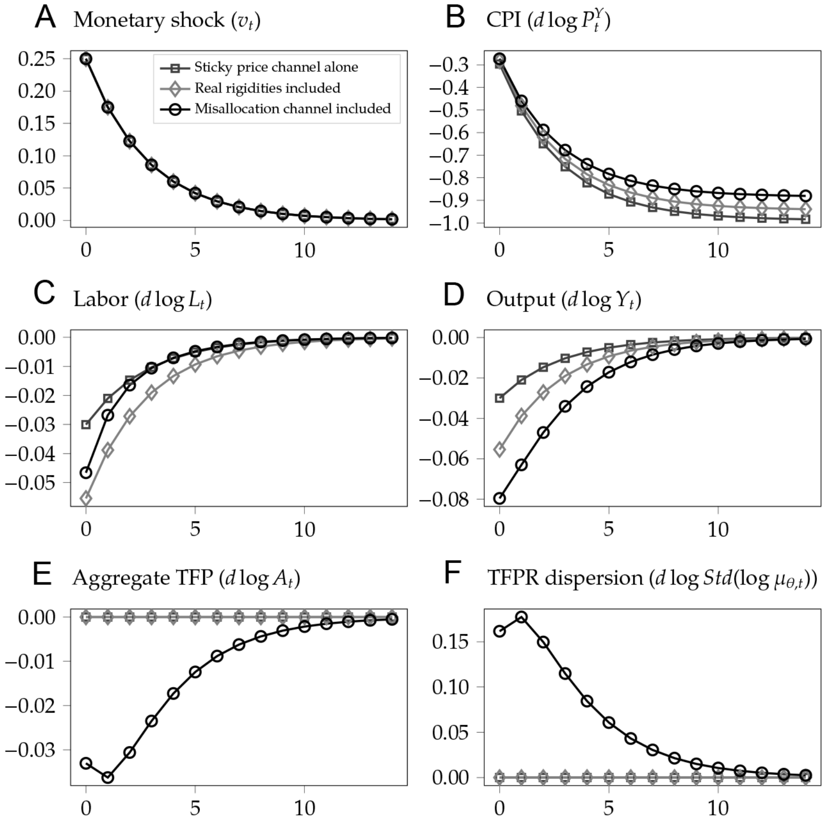

FIG. 5.—Impulse response functions following a 25-bp monetary shock.

misallocation channel, and a CES model, which has neither real rigidities nor the misallocation channel. As mentioned earlier, discount factor shocks are isomorphic to monetary shocks in our model, so the results can equally be taken to be the response of the model to discount factor shocks.[44]

In the CES and homogeneous firms case, aggregate TFP does not react to the monetary shock, as implied by lemma 2. In contrast, when firms have heterogeneous markups, aggregate TFP falls in response to the contractionary shock. The fall in aggregate TFP dampens the extent of disinflation caused by the monetary contraction and deepens the immediate response of output to the shock. The reduction in aggregate

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

TABLE 5
EFFECT OF MONETARY POLICY SHOCK ON OUTPUT

Model	Output Effect at $t = 0$	Half-Life	Cumulative Output Impact
CES	−.030	1.95	−.10
Homogeneous firms	−.055	1.95	−.18
Heterogeneous firms	−.080	2.56	−.31

TFP coincides with an increase in the cross-sectional dispersion of firm-level TFPR since high-markup firms are raising their markups relative to low-markup firms.[45] The magnitude of the increase in TFPR dispersion is broadly consistent with Kehrig (2011), who finds that TFPR dispersion increases about 10% during a typical recession and increased over 20% from 2007 to the trough of the recession in 2009.

We quantify how the misallocation channel affects real output in table 5. The contraction in output in the full model is about 45% deeper on impact than in the homogeneous firm model. The persistence of the shock's effect on real output also increases: while the CES and homogeneous firm models feature a constant half-life of just under two-quarters, the misallocation channel increases the half-life of the shock by about 30% to about 2.6 quarters.[46] In full, the misallocation channel increases the cumulative impact on output of the monetary shock by around 70%.

Figure 6 shows the covariance between firms' inverse markups and their change in markups (fig. 6*A*) and change in total input costs (fig. 6*B*). Following lemma 2, the contractionary monetary shock reallocates inputs to low-markup firms, generating the fall in TFP. This is a directly testable prediction of the model that we return to in section VII.

We provide additional calibration results in appendix D. In particular, we report the change in sales shares for firms at different percentiles of the size distribution. The sales shares of small firms are about as volatile as aggregate output, whereas the sales shares of the largest firms are less volatile. In appendix E, we show that results are quantitatively similar when monetary policy is implemented via changes in money supply (with a cash-in-advance constraint) rather than an interest rate rule. All in all, our results suggest that the misallocation channel is as powerful as the real rigidities channel in affecting the transmission of monetary policy.

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

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

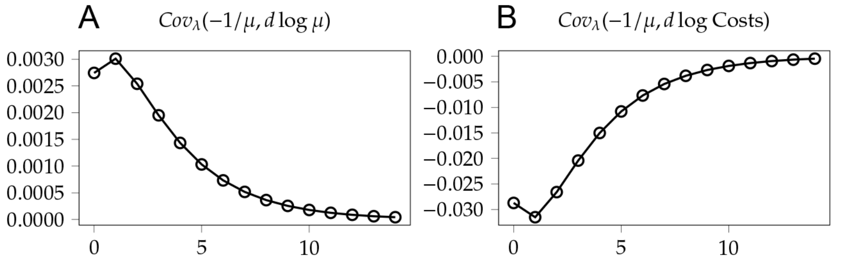

FIG. 6.—Covariance of firms' inverse markups with changes in markups and costs following a 25-bp monetary shock. The contractionary shock leads high-markup firms to increase their markups relative to low-markup firms (*A*), causing a reallocation of resources away from high-markup firms (*B*).

E. Menu Cost Calibration

The price rigidities we have explored so far take the form of Calvo frictions. A natural question is whether the effects we identify would also arise under a different model of nominal rigidities. In appendix C, we nonlinearly solve and provide impulse response functions for a quantitative model with menu costs instead.[47] We first calibrate the model under CES preferences and then replace those preferences with the Kimball demand system estimated in the Belgian data. In response to a money supply shock, the Kimball calibration generates a procyclical TFP response that increases the effect of the shock on output. Similar to our baseline results, roughly half of the movement of output on impact is due to the supply-side effect. Accordingly, the response of output on impact is more than twice as large in the Kimball calibration relative to the CES calibration.

As in the Calvo model, aggregate TFP rises in response to monetary expansions because high-markup firms have lower realized pass-through than low-markup firms. However, unlike the Calvo model, the differences in realized pass-throughs comes from the extensive (rather than the intensive) margin of price changes. Table 6 shows the intensive and extensive margins of price adjustment for firms in the Calvo and menu cost

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

TABLE 6
EXTENSIVE AND INTENSIVE MARGINS OF PRICE ADJUSTMENT IN CALVO AND MENU COST MODEL FOR 1 YEAR AFTER MONEY SUPPLY SHOCK

	Calvo Model		Menu Cost Model	
Quintile of Initial Size	Share of Firms with Price Change	Average Size of Price Change	Share of Firms with Price Change	Average Size of Price Change
1	.938	.0359	.921	.0374
2	.938	.0358	.841	.0408
3	.938	.0357	.766	.0446
4	.938	.0356	.719	.0458
5	.938	.0345	.676	.0495

NOTE.—Response to a 4-bp money supply shock in both models. The share of firms with price change reports the fraction of firms with at least one price change within 1 year of the initial shock. The average size of price change is the average magnitude of the first price change by firms in each quintile.

models in response to a similar-sized money supply shock.[48] In the Calvo model, we assume that all firms have the same degree of price stickiness δ_θ, so that all differences in realized pass-through come from intensive margin differences in the degree to which firms adjust their prices. Alternatively, in the menu cost model, high-markup firms endogenously choose to keep their prices unchanged for longer due to lower desired pass-through. As a result, large firms are less likely to change their prices in the first year after the shock. However, conditional on changing their price, large firms make slightly larger adjustments. This is because of a selection effect where large firms that choose to adjust their prices are those that have been buffeted by large idiosyncratic shocks. Lower realized pass-through of large firms—due to differences in the extensive margin of price adjustment in the menu cost model—generates the misallocation channel.[49]

Berger and Vavra (2019) find a positive correlation between (reduced-form, general equilibrium) exchange rate pass-through and dispersion in price changes in the time series. They attribute this to the intensive margin—variation in pass-through conditional on a price change—rather than the extensive margin—variation in the frequency of price

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

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

adjustment. Our baseline Calvo model is able to better match this pattern in the sense that increases in dispersion of price changes are due to differences in desired pass-through across firms, rather than variations in the probability of price adjustment caused by a monetary shock.

VII. Empirical Evidence

In this section, we provide empirical evidence in support of the reallocation mechanism described in this paper. We first present macro-level evidence on the response of aggregate TFP to identified monetary shocks. We then show at the micro level that contractionary monetary shocks lead high-markup firms to increase their markups relative to low-markup firms, leading to a deleterious reallocation of inputs across firms. Finally, we provide evidence that the contraction in productivity following monetary tightening is greater in more concentrated industries, as in figure 4.

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

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

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

For monetary shocks, we use an extended version of the Romer and Romer (2004) monetary shock series constructed by Wieland and Yang (2020) for 1969–2007. We use three different measures of aggregate productivity—labor productivity, the Solow residual, and the cost-based Solow residual (see Hall 1990).[50] We do not use utilization-adjusted TFP (e.g., Basu, Fernald, and Kimball 2006; Fernald 2014). This is because these series are identified using the assumption that sectoral productivity is orthogonal to monetary shocks, and this exogeneity condition fails in our model.

Figure 7 plots the estimated coefficients b_0^h for horizons up to 16 quarters. Following a contractionary shock, there is a significant contraction in aggregate productivity and output. The magnitude of the decline in aggregate productivity is more than half of the effect on output. This movement in aggregate productivity relative to output is moderately larger than that predicted by our model, which suggests that allocative

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

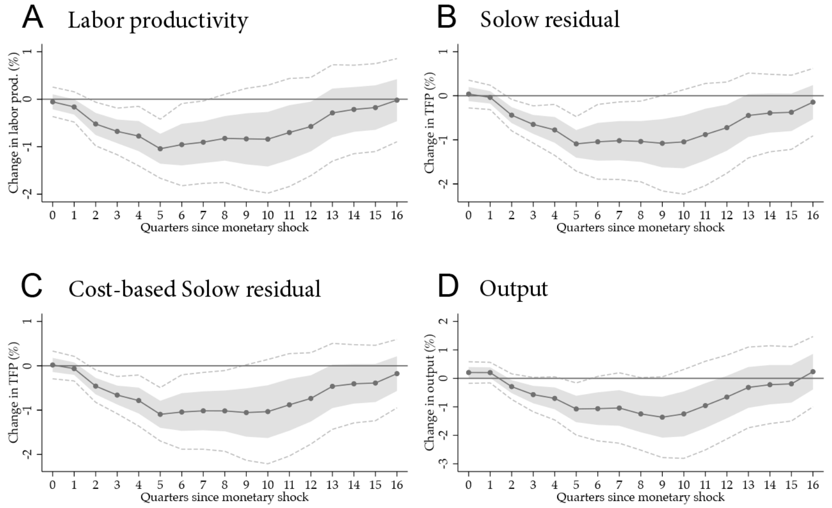

FIG. 7.—Local projection of a contractionary Romer and Romer (2004) shock (using extension by Wieland and Yang 2020) on aggregate productivity and output. Shaded region indicates Newey-West standard errors. Dashed lines indicate 95% confidence intervals. Sample covers 1969–2007.

effects explain part but perhaps not all of the aggregate productivity response.[51]

Micro-level evidence.—In our model, aggregate TFP responds to monetary shocks due to systematic reallocations among firms with different markups. We now turn to micro-level evidence on these reallocations. To do so, we use estimates of markups for publicly traded firms in Compustat. Of course, this exercise must be interpreted with caution since measuring markups accurately at high frequency is challenging and Compustat is not a representative sample of all US producers. Nevertheless, our empirical results are supportive of the basic mechanism underlying the misallocation channel.

We study the response of firm-level markup changes and input reallocations across firms to identified exogenous monetary shocks.[52] For

[51] In app. sec. B.2, we also show that labor productivity, the Solow residual, and the cost-based Solow residual are unconditionally procyclical over the period 1948–2020. The dynamic calibration in sec. VI predicts that a 1% change in output due to a monetary shock is accompanied by a 0.4% change in aggregate productivity. In fig. 7, our point estimates suggest that a 1% change in output due to a monetary shock is accompanied by a 0.7% change in aggregate productivity. Thus, the relative size of the productivity response in our model is roughly half of that in the data.

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

our baseline estimate of firm markups, we follow the user cost approach of Gutiérrez (2017) and Gutiérrez and Philippon (2017). That is, we estimate each firm's capital stock and user cost of capital. To estimate the user cost of capital, we use industry-specific depreciation rates and industry-level risk premia. We estimate profits by subtracting total estimated costs from total revenues, and we back out the markup by assuming that firms have constant returns to scale. Appendix B describes the data sources and assumptions underlying our markup estimation procedure in more detail.

We then estimate the following local projections:

$$\begin{aligned} \mathrm{Cov}_\lambda(-1/\mu_t, \Delta \log \mu_{t\to t+h}) &= a^h + \sum_{k=0}^{4} b_k^h \cdot \mathrm{MonetaryShock}_{t-k} \\ &\quad + \sum_{k=1}^{4} c_k^h \cdot \mathrm{Cov}_\lambda(-1/\mu_t, \Delta \log \mu_{t-k\to t}) + \epsilon_t^h, \\ \mathrm{Cov}_\lambda(-1/\mu_t, \Delta \log \mathrm{Costs}_{t\to t+h}) &= \tilde{a}^h + \sum_{k=0}^{4} \tilde{b}_k^h \cdot \mathrm{MonetaryShock}_{t-k} \\ &\quad + \sum_{k=1}^{4} \tilde{c}_k^h \cdot \mathrm{Cov}_\lambda(-1/\mu_t, \Delta \log \mathrm{Costs}_{t-k\to t}) \\ &\quad + \epsilon_t^h, \end{aligned}$$

where $\mathrm{Cov}_\lambda(-1/\mu_t, \Delta \log \mu_{t\to t+h})$ represents the sales-weighted covariance between inverse markups at time t and the change in markups from time t to time $t+h$, $\mathrm{Cov}_\lambda(-1/\mu_t, \Delta \log \mathrm{Costs}_{t\to t+h})$ represents the sales-weighted covariance between inverse markups at t and the change in total costs, and $\mathrm{MonetaryShock}_t$ represents the (extended) Romer and Romer (2004) shock in quarter t.[53] This is a direct test of the model, as in lemma 2. Figure 6 shows that in our calibrated model, a contractionary shock leads to relative increases in the markups of high-markup firms ($\mathrm{Cov}_\lambda(-1/\mu, \Delta \log \mu) > 0$) and a reallocation of resources toward low-markup firms ($\mathrm{Cov}_\lambda(-1/\mu, \Delta \log \mathrm{Costs}) < 0$).[54]

Figure 8 shows estimates of b_0^h and $\tilde{b}_0^h$ following a monetary shock. As figure 8*A* shows, a contractionary shock leads high-markup firms to increase their markups relative to low-markup firms; the result, in figure 8*B*, is a reallocation of resources away from high-markup firms and toward low-markup firms. In figure 8*C* and 8*D*, we estimate a panel version of the above specifications across three-digit North American

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

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

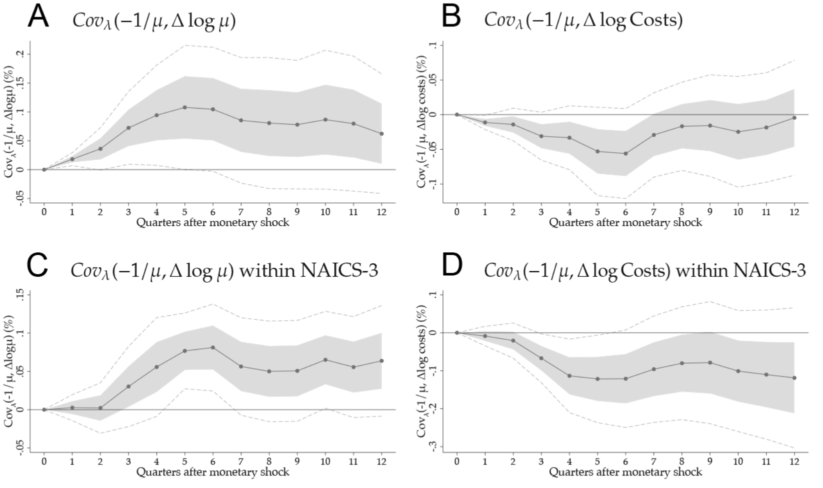

FIG. 8.—Local projection of contractionary Romer and Romer (2004) shock (using extension by Wieland and Yang 2020) on $\text{Cov}_\lambda(-1/\mu, \Delta \log \mu)$ and $\text{Cov}_\lambda(-1/\mu, \Delta \log \text{Costs})$. Shaded region indicates Newey-West standard errors (*A*, *B*) and Driscoll-Kraay standard errors (*C*, *D*). Dashed lines indicate 95% confidence intervals.

Industry Classification System (NAICS) industries with industry fixed effects.[55] Both the direction and the magnitude of the impulse responses are similar, suggesting that within-sector reallocations play an important role.

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

In appendix B, we show that our results are robust to using firm accounting profits to measure markups (fig. B.2) and to including intangible capital when estimating user cost markups (fig. B.3). Our results are also robust to using monetary shocks identified using high-frequency methods by Gorodnichenko and Weber (2016; fig. B.9).

Cross-sector evidence.—Figure 4 suggests that industrial concentration may play a role in how productivity responds to monetary shocks. All things being equal, higher industrial concentration is likely to be accompanied

[55] See app. B for the estimating equations for the industry-level specifications.

TABLE 7
Differential Response of Industry Multifactor Productivity to Monetary Shocks in Concentrated Manufacturing Industries

	Δ log MultifactorProductivity$_{i,t}$		
	(1)	(2)	(3)
Top eight firms Share$_i$ × MonetaryShock$_t$	−.0185**		
	(.00906)		
Top 20 firms Share$_i$ × MonetaryShock$_t$		−.0183**	
		(.00762)	
Top 50 firms Share$_i$ × MonetaryShock$_t$			−.0176**
			(.00699)
Industry fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Observations	1,634	1,634	1,634

Note.—Sales shares of the top eight, 20, and 50 firms in each four-digit NAICS industry are from the 2002 Economic Census for Manufacturing. Monetary shocks are from the extension of the Romer and Romer (2004) shock series by Wieland and Yang (2020).
** Significant at the 5% level.

by greater heterogeneity in pass-through and hence a greater response of productivity to monetary shocks.

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

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

where i represents the four-digit NAICS industry, t represents the year, δ_i represents industry fixed effects, and α_t represents year fixed effects.[56] The coefficient of interest is β, which indicates whether multifactor productivity in concentrated industries is differentially responsive to the monetary shock. Our calibration suggests that a contractionary monetary shock leads to a greater reduction in multifactor productivity in concentrated industries, and hence $\beta < 0$.

Table 7 shows the estimated coefficient β using three measures of industrial concentration—the sales share of the industry's top eight, 20, and 50 firms in the 2002 Economic Census for Manufacturing—and using the extended Romer and Romer (2004) shock series. For all three measures, we observe that the estimated $\beta < 0$, which suggests that the

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

productivity effects of a monetary shock are more pronounced in concentrated industries.

In tables B.2 and B.3 (tables B.1–B.3, E.1, and H.1 are available online), we show that these results are robust to using concentration data from the 2007 Economic Census and to using monetary shocks identified from high-frequency data by Gorodnichenko and Weber (2016).

VIII. Extensions

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

Multiple sectors, multiple factors, input-output linkages, and sticky wages.— The model we use in the main text of this paper is deliberately stylized for clarity. It has only one sector and only one factor of production. This means that it is missing some ingredients that are quantitatively important for how output responds to monetary shocks but that are unrelated to the mechanism this paper studies.[57] In appendix F, we show how to extend the model to have a general production network structure, with multiple sectors and multiple factors. As an example, in appendix section F.1 we consider an economy with two factors (labor and capital), a firm sector, and a "labor union" sector that generates sticky wages. The intuition underlying the supply-side effects of a monetary shock are unchanged in this extension compared with the model presented in the main text, and we find that the misallocation channel remains similar in magnitude.

Variation in markups and pass-throughs unrelated to size.—In our calibrations, we assume that markups and pass-throughs at the initial equilibrium vary only as a function of firm size. While markups and pass-throughs do vary as a function of firm size (e.g., see Amiti, Itskhoki, and Konings 2019; Burstein, Carvalho, and Grassi 2020), in practice, markups and pass-throughs also vary for reasons unrelated to size, such as firm-specific shifters in demand curves, quality differences, or markup dispersion inherited from previous periods. In appendix I, we show how our baseline results change if there is variation in markups and pass-throughs unrelated to size. We show that the supply-side effects of monetary policy are strengthened if the excess variation in markups is negatively correlated with the excess variation in pass-throughs and weakened if this correlation is positive. When excess variation in markups and pass-throughs is orthogonal, then the presence of the noise does not affect the strength of supply-side effects of monetary policy relative to our benchmark calibration.

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

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

Klenow and Willis (2016) calibration.—In the main text, we caution against using off-the-shelf functional forms for preferences. We illustrate this by calibrating our model with the commonly used Klenow and Willis (2016) specification in appendix G. We show that to match the observed relationship between pass-through and firm size (see fig. 2), large firms must have markups that are on the order of 10,000%. Under standard calibrations, which do not produce astronomically large markups for large firms, the implied pass-through function does not vary much as a function of firm size. Therefore, standard calibrations of these preferences fail to capture the cross-sectional covariance between pass-throughs and markups and hence imply counterfactually small supply-side effects.

IX. Conclusion

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

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

While we focus on heterogeneous markups in product markets, it is possible that similar distortions could exist in input markets. Specifically, if firms have heterogeneous and variable monopsony power in the labor market, then TFP would increase if firms with relatively high markdowns reduce their markdowns following an expansionary shock. Finally, our analysis is purely positive, and we leave the normative implications for optimal policy for future work.

Data Availability

Code replicating the tables and figures in this article can be found in the Harvard Dataverse, https://doi.org/10.7910/DVN/FEECAL (Baqaee, Farhi, and Sangani 2023).

References

Amiti, M., O. Itskhoki, and J. Konings. 2014. "Importers, Exporters, and Exchange Rate Disconnect." *A.E.R.* 104 (7): 1942–78.

———. 2019. "International Shocks, Variable Markups, and Domestic Prices." *Rev. Econ. Studies* 86 (6): 2356–402.

Andrés, J., and P. Burriel. 2018. "Inflation and Optimal Monetary Policy in a Model with Firm Heterogeneity and Bertrand Competition." *European Econ. Rev.* 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 Econ. J. Macroeconomics* 11 (3): 67–110.

Atkeson, A., and A. Burstein. 2008. "Pricing-to-Market, Trade Costs, and International Relative Prices." *A.E.R.* 98 (5): 1998–2031.

Auer, R. A., T. Chaney, and P. Sauré. 2018. "Quality Pricing-to-Market." *J. Internat. Econ.* 110:87–102.

Ball, L., and D. Romer. 1990. "Real Rigidities and the Non-neutrality of Money." *Rev. Econ. Studies* 57 (2): 183–203.

Baqaee, D. R., and E. Farhi. 2017. "Productivity and Misallocation in General Equilibrium." Working Paper no. 24007, NBER, Cambridge, MA.

———. 2020. "Productivity and Misallocation in General Equilibrium." *Q.J.E.* 135 (1): 105–63.

Baqaee, D. R., E. Farhi, and K. Sangani. 2021. "The Darwinian Returns to Scale." Working Paper no. 27139, NBER, Cambridge, MA.

———. 2023. "Replication Data for: 'The Supply-Side Effects of Monetary Policy.'" Harvard Dataverse. https://doi.org/10.7910/DVN/FEECAL.

Barkai, S. 2020. "Declining Labor and Capital Shares." *J. Finance* 75 (5): 2421–63.

Basu, S., J. G. Fernald, and M. S. Kimball. 2006. "Are Technology Improvements Contractionary?" *A.E.R.* 96 (5): 1418–48.

Benigno, G., and L. Fornaro. 2018. "Stagnation Traps." *Rev. Econ. Studies* 85 (3): 1425–70.

Benkard, C. L., A. Yurukoglu, and A. L. Zhang. 2021. "Concentration in Product Markets." Working Paper no. 28745, NBER, Cambridge, MA.

Berger, D., and J. Vavra. 2019. "Shocks versus Responsiveness: What Drives Time-Varying Dispersion?" *J.P.E.* 127 (5): 2104–42.

Bianchi, F., H. Kung, and G. Morales. 2019. "Growth, Slowdowns, and Recoveries." *J. Monetary Econ.* 101:47–63.

Bils, M., and P. J. Klenow. 2004. "Some Evidence on the Importance of Sticky Prices." *J.P.E.* 112 (5): 947–85.

Burstein, A., V. M. Carvalho, and B. Grassi. 2020. "Bottom-Up Markup Fluctuations." Working Paper no. 27958, NBER, Cambridge, MA.

Calvo, G. A. 1983. "Staggered Prices in a Utility-Maximizing Framework." *J. Monetary Econ.* 12 (3): 383–98.

Castro, N. 2019. "The Importance of Production Networks and Sectoral Heterogeneity for Monetary Policy." Working paper.

Chen, N., and L. Juvenal. 2016. "Quality, Trade, and Exchange Rate Pass-through." *J. Internat. Econ.* 100:61–80.

Chetty, R., A. Guren, D. Manoli, and A. Weber. 2011. "Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins." *A.E.R.* 101 (3): 471–75.

Comin, D., and M. Gertler. 2006. "Medium-Term Business Cycles." *A.E.R.* 96 (3): 523–51.

Corhay, A., H. Kung, and L. Schmid. 2020. "Q: Risk, Rents, or Growth?" Working paper.

Cozier, B., and R. Gupta. 1993. "Is Productivity Exogenous over the Business Cycle? Some Canadian Evidence on the Solow Residual." Working Paper no. 93-13, Bank Canada, Ottawa, ON.

Cravino, J. 2017. "Exchange Rates, Aggregate Productivity and the Currency of Invoicing of International Trade." Working paper.

Del Negro, M., M. Lenza, G. E. Primiceri, and A. Tambalotti. 2020. "What's Up with the Phillips Curve?" Working Paper no. 27003, NBER, Cambridge, MA.

Dotsey, M., and R. G. King. 2005. "Implications of State-Dependent Pricing for Dynamic Macroeconomic Models." *J. Monetary Econ.* 52 (1): 213–42.

Edmond, C., V. Midrigan, and D. Y. Xu. 2018. "How Costly Are Markups?" Working Paper no. 24800, NBER, Cambridge, MA.

Eichenbaum, M., and J. Fisher. 2004. "Evaluating the Calvo Model of Sticky Prices." Working Paper no. 10617, NBER, Cambridge, MA.

Etro, F., and L. Rossi. 2015. "New-Keynesian Phillips Curve with Bertrand Competition and Endogenous Entry." *J. Econ. Dynamics and Control* 51:318–40.

Evans, C. L. 1992. "Productivity Shocks and Real Business Cycles." *J. Monetary Econ.* 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 Econ. J.* 1 (1): 47–70.

Fernald, J. 2014. "A Quarterly, Utilization-Adjusted Series on Total Factor Productivity." Tech. report, Fed. Reserve Bank San Francisco.

Foster, L., J. Haltiwanger, and C. Syverson. 2008. "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *A.E.R.* 98 (1): 394–425.

Fratto, C., and H. Uhlig. 2014. "Accounting for Post-crisis Inflation and Employment: A Retro Analysis." Working Paper no. 20707, NBER, Cambridge, MA.

Galí, J. 2015. *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications.* Princeton, NJ: Princeton Univ. Press.

Goldberg, P., and R. Hellerstein. 2011. "How Rigid Are Producer Prices?" Tech. Report no. 407, Fed. Reserve Bank New York.

Gopinath, G., and O. Itskhoki. 2011. "In Search of Real Rigidities." *NBER Macroeconomics Ann.* 25:261–310.

Gopinath, G., O. Itskhoki, and R. Rigobon. 2010. "Currency Choice and Exchange Rate Pass-Through." *A.E.R.* 100 (1): 304–36.

Gorodnichenko, Y., and M. Weber. 2016. "Are Sticky Prices Costly? Evidence from the Stock Market." *A.E.R.* 106 (1): 165–99.

Gutiérrez, G. 2017. "Investigating Global Labor and Profit Shares." Working paper.

Gutiérrez, G., and T. Philippon. 2017. "Declining Competition and Investment in the U.S." Working Paper no. 23583, NBER, Cambridge, MA.

Hall, R. E. 1990. "Invariance Properties of Solow's Productivity Residual." In *Growth, Productivity, Unemployment: Essays to Celebrate Bob Solow's Birthday,* edited by P. Diamond, 71–112. Cambridge, MA: MIT Press.

Hazell, J., J. Herreno, E. Nakamura, and J. Steinsson. 2020. "The Slope of the Phillips Curve: Evidence from US States." Working Paper no. 28005, NBER, Cambridge, MA.

Hooper, P., F. S. Mishkin, and A. Sufi. 2020. "Prospects for Inflation in a High Pressure Economy: Is the Phillips Curve Dead or Is It Just Hibernating?" *Res. Econ.* 74 (1): 26–62.

Hsieh, C.-T., and P. J. Klenow. 2009. "Misallocation and Manufacturing TFP in China and India." *Q.J.E.* 124 (4): 1403–48.

Jaimovich, N., S. Rebelo, and A. Wong. 2019. "Trading Down and the Business Cycle." *J. Monetary Econ.* 102:96–121.

Jordà, Ò. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *A.E.R.* 95 (1): 161–82.

Kehrig, M. 2011. "The Cyclicality of Productivity Dispersion." Tech. Report no. CES-WP-11-15, US Census Bur. Center Econ. Studies, Washington, DC.

Kim, S., and H. Lim. 2004. "Does the Solow Residual for Korea Reflect Pure Technology Shocks?" Tech. report, Seoul, Korea.

Kimball, M. S. 1995. "The Quantitative Analytics of the Basic Neomonetarist Model." *J. 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): 443–72.

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." *Rev. Econ. and Statis.* 87 (1): 124–34.

Krugman, P. R. 1998. "It's Baaack: Japan's Slump and the Return of the Liquidity Trap." *Brookings Papers Econ. Activity* (2): 137–205.

La'O, J., and A. Tahbaz-Salehi. 2022. "Optimal Monetary Policy in Production Networks." *Econometrica* 90 (3): 1295–336.

Martinez, I. Z., E. Saez, and M. Siegenthaler. 2018. "Intertemporal Labor Supply Substitution? Evidence from the Swiss Income Tax Holidays." Working Paper no. 24634, NBER, Cambridge, MA.

Matsuyama, K., and P. Ushchev. 2017. "Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems." Working Paper no. 17-109, Northwestern Global Poverty Res. Lab, Evanston, IL.

———. 2022. "Selection and Sorting of Heterogeneous Firms through Competitive Pressures." CEPR Discussion Paper no. DP17092, Centre Econ. Policy Res., London.

McLeay, M., and S. Tenreyro. 2020. "Optimal Inflation and the Identification of the Phillips Curve." *NBER Macroeconomics Ann.* 34:199–255.

Meier, M., and T. Reinelt. 2020. "Monetary Policy, Markup Dispersion, and Aggregate TFP." ECB Working Paper no. 2427, European Central Bank, Frankfurt, Germany.

Melitz, M. J. 2018. "Competitive Effects of Trade: Theory and Measurement." *Rev. World Econ.* 154 (1): 1–13.

Mongey, S. 2021. "Market Structure and Monetary Non-neutrality." Working Paper no. 29233, NBER, Cambridge, MA.

Nakamura, E., and J. Steinsson. 2008. "Five Facts about Prices: A Reevaluation of Menu Cost Models." *Q.J.E.* 123 (4): 1415–64.

Pasten, E., R. Schoenle, and M. Weber. 2020. "The Propagation of Monetary Policy Shocks in a Heterogeneous Production Economy." *J. Monetary Econ.* 116:1–22.

Restuccia, D., and R. Rogerson. 2008. "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments." *Rev. Econ. Dynamics* 11 (4): 707–20.

Romer, C. D., and D. H. Romer. 2004. "A New Measure of Monetary Shocks: Derivation and Implications." *A.E.R.* 94 (4): 1055–84.

Rossi-Hansberg, E., P.-D. Sarte, and N. Trachter. 2021. "Diverging Trends in National and Local Concentration." *NBER Macroeconomics Ann.* 35:115–50.

Rubbo, E. 2020. "Networks, Phillips Curves and Monetary Policy." Manuscript.

Sigurdsson, J. 2019. "Labor Supply Responses and Adjustment Frictions: A Tax-Free Year in Iceland." Working paper.

Smets, F., and R. Wouters. 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *A.E.R.* 97 (3): 586–606.

Smith, D., and S. Ocampo. 2021. "The Evolution of US Retail Concentration." Working paper.

Taylor, J. B. 1980. "Aggregate Dynamics and Staggered Contracts." *J.P.E.* 88 (1): 1–23.

Wang, O., and I. Werning. 2020. "Dynamic Oligopoly and Price Stickiness." Working Paper no. 27536, NBER, Cambridge, MA.

Wieland, J. F., and M.-J. Yang. 2020. "Financial Dampening." *J. Money, Credit and Banking* 52 (1): 79–113.