Measured Productivity with Endogenous Markups and Economic Profits

Anthony Savagar [*†]

March 5th 2021

Abstract

I study the effect of dynamic firm entry, scale economies and oligopolistic competition on measured productivity and output amplification. These features cause measured productivity (Solow residual) to exceed pure technology. I decompose measured productivity into pure technology and an endogenous component that is caused by firm-level output variation interacting with increasing returns to scale. In turn, I show that firm-level output variation depends on economic profits and markups that vary in response to dynamic entry and oligopolistic competition. I estimate the pure technology series adjusted for profits and markups and show that it is less volatile and more persistent than a benchmark model, whilst still generating output amplification.

JEL: E32, D21, D43, L13, C62

Keywords: Markups, Firm Entry, Productivity, Scale Economies, Oligopolistic Competition

[*]Thanks to the following people for helpful comments: Huw Dixon, Patrick Minford, Vivien Lewis, Jo Van Biesebroeck, Akos Valentinyi, Harald Uhlig, Stephen Millard, Martin Kaae Jensen, David Collie, Jonathan Haskel, Nobu Kiyotaki, Esteban Rossi-Hansberg, Fernando Leibovici, Sungki Hong, Mehdi Sahneh, Mathan Satchi, Maarten De Ridder, Vasco Carvalho, Alex Clymo, Nick Kozeniauskas, Joaquim Oliveira Martins, Frederic Dufourt and funders: ESRC, Julian Hodge Institute and RES. I appreciate the hospitality of KU Leuven and St Louis Fed.

[†]asavagar@gmail.com

Recent empirical discussions of the US macroeconomy have focused on underlying product market structures. Specifically the behaviour of price markups, economic profits, scale economies and business dynamism (firm entry and exit).[1] In this paper, I develop a tractable dynamic general equilibrium model that endogenously determines each of these features, and I analyse the implications for productivity measurement and output amplification. This contributes to a well-established literature that in the absence of perfect competition and constant returns to scale, a Solow residual measure of productivity does not capture true shifts in an economy's production function (Hall 1989; Basu and Fernald 2001). That is, a Solow residual is a biased measure of *technology*.[2] I extend this literature by analysing the dynamic interaction of sluggish firm entry, oligopolistic competition and increasing returns to scale.

I provide empirical evidence that following an exogenous technology shock, aggregate output initially responds through incumbent firms expanding their production *before* new firm entry occurs. Therefore, the intensive margin of firm-level output responds faster than the extensive margin of new firm entry. I develop a model that replicates this observation and show that these firm-level intensive margin variations create endogenous productivity effects and amplify output. I identify a model consistent technology series which has a lower variance and higher persistence than a benchmark technology series.

Findings

I show that measured productivity exceeds pure technology when firm-level output variations interact with returns to scale. I decompose measured productivity into a pure technology component and a returns to scale component. The returns to scale component occurs because firm-level output variation interacts with firm-level increasing returns to scale. Firm-level output variation corresponds to endogenously varying profits and markups. Economic profits vary because of dynamic entry. Markups vary because of oligopolistic competition. The profit channel causes upward bias in measured TFP in the short run but tends to zero in the long run. Whereas, the markup channel is absent in the short run but grows during transition causing upward bias in measured TFP in the long run. The relative importance of the two effects depends on the speed of firm adjustment (business dynamism).

Given the analytical decomposition of measured TFP, I estimate a pure technology

[1] Decker, Haltiwanger, Jarmin, and Miranda (2018), Barkai (2020), and De Loecker, Eeckhout, and Unger (2020) present evidence of rising market power and declining business dynamism, whilst Basu (2019) and Syverson (2019) discuss the limitations.

[2] Typically, a Solow residual refers to the difference between the growth rate of aggregate output and the sum of the growth rates in factor inputs weighted by their share in costs or revenue. Throughout this paper the terms *Solow residual*, *measured productivity* and *measured TFP* refer to detrended output less share-weighted detrended inputs. *Technology* refers to a Hicks-neutral shifter in the production function.

series which I show is less volatile and more persistent than measured TFP (Solow residual). I simulate the model using the adjusted technology series and show that it generates significant output amplification. Measured productivity exceeds technology by 50% on impact of a shock and has 45% higher standard deviation than a benchmark constant markup instantaneous entry model. The full model, with endogenous markups and dynamic entry, generates 13% more variation in aggregate output than a benchmark constant markups, instantaneous entry model.

Model Description

I extend a neoclassical growth model with endogenous labour to include richer industrial organization features. There is no heterogeneity and the main model is in continuous time. The model consists of a representative household and representative firm. The household can invest in capital or firms. A dynamic arbitrage condition ensures the two assets yield the same return. The three key model features are dynamic entry, oligopolistic competition and increasing returns to scale. Dynamic entry leads to non-zero economic profits in the short run, whilst oligopolistic competition leads to endogenous markups. Endogenous markups imply that markups are a function of the number of firms. The model has both a one-off sunk entry cost and a period-by-period fixed overhead cost. Both of these costs are output denominated.

Dynamic entry refers to entry that is sluggish. The number of firms becomes a state variable similar to capital in a traditional RBC model. In my model the lag in firm entry is caused by an endogenous entry cost. Specifically, I assume that there is a convex entry adjustment cost that depends on the flow of entry due to some congestion effect or other externality. The endogenous entry cost creates an intertemporal zero-profit condition (dynamic arbitrage condition) that equates the cost of entry in an instance to the net present value of incumbency. Consequently, entry sluggishly adjusts to its long-run level as potential entrants evaluate whether to enter or wait for entry costs to fall in a future period with less congestion. In the long run, the arbitrage condition yields a free-entry steady-state condition that causes zero profit and no entry. By having no entry and economic profits in the short-run (on impact) and full adjustment with zero profits in the long run (steady state), the model captures a traditional Marshallian definition of the short run and the long run. This is also consistent with classic Chamberlinian monopolistic competition. Traditionally macroeconomic models have either instantaneous entry (Devereux, Head, and Lapham 1996; Comin and Gertler 2006; Jaimovich and Floetotto 2008) or a fixed number of firms (Hornstein 1993; Rotemberg and Woodford 1993). These outcomes are special cases of the model I develop.

Oligopolistic competition refers to the form of strategic interaction among firms. It

implies that firms are large in their industry, but small in the macroeconomy. Therefore, firms take into consideration their own-output effect on industry output when maximizing profits.[3] This differs from a standard Dixit-Stiglitz monopolistic competition framework which yields constant markups (Dixit and Stiglitz 1977). In the Dixit-Stiglitz framework firms are infinitesimal.[4] They do not consider their indirect effect on industry-level output and consequently markups are constant. However, when firms account for their indirect effect, markups depend negatively on the number of firms. Entry reduces markups because firms have a lower market share, and thus have a weaker effect on industry-level outcomes. A simple version of this model of oligopoly with Cournot competition is used in Dos Santos Ferreira and Dufourt (2006) in a DGE business cycle setting. Recently, it has been popularised by Atkeson and Burstein (2008) in a heterogeneous firm setting.

Scale economies are caused by a fixed cost and increasing marginal costs due to decreasing returns to scale in variable production. This yields firms with U-shaped average cost curves. This setup allows for increasing returns, constant returns or decreasing returns. There are increasing returns on the left-hand side of the U, whilst the fixed cost dominates the increasing marginal costs. There are constant returns at minimum average cost where marginal cost and average cost intersect. There are decreasing returns as the rising marginal cost exceeds the average cost on the right-hand side of the U. In steady-state, under imperfect competition the economy has increasing returns to scale. It operates where average cost exceeds marginal cost and price exceeds marginal cost implying there is a price markup. Under perfect competition, firms would operate at the minimum of their average cost curve at *minimum efficient scale*. At this point there are locally constant returns to scale and the endogenous productivity effects I discuss disappear.

Mechanism

To understand the mechanism I will explain the effect of a permanent, positive, technology shock. In the model a positive technology shock causes profits, entry, employment, investment, entry costs and productivity to increase, whilst markups and returns to scale decrease. I divide the effect into three stages: the short run which means on impact of the shock; transition during which entry is taking place; and the long run once steady state is achieved.

On impact the shock affects a fixed number of incumbent firms since firm entry

[3] In my model firms consider own-output effects because they compete under Cournot. It could also be solved for Bertrand where the firm considers own-price effect on industry price level, as in Etro and Colciago (2010) and Lewis and Poilly (2012).

[4] Yang and Heijdra (1993) stress that the Dixit-Stiglitz constant markup result is an approximation. Since products are defined on a finite set, own-price effects should be considered, but these indirect effects are approximately zero as the number of firms tends to infinity.

cannot respond due to adjustment costs. Consequently, incumbent firms bear the shock and raise their output to satisfy the additional demand. They adjust output through labour which is the only input that can respond instantaneously. The increase in incumbent firm size improves returns to scale and causes an endogenous increase in productivity.[5] Given markups and sluggish firm entry, the short-run rise in incumbents' output corresponds one-to-one with a short-run rise in operating profits. Thus initially measured productivity does not reflect solely the technology shock but also a returns to scale effect from expanded output which maps exactly to profits.

During transition, the rise in profits attracts entry. As entry occurs, it decreases incumbents' output and profits which reduces the scale effect. Entry reducing incumbent output is known as a *business stealing effect* (Mankiw and Whinston 1986). In addition to a business stealing effect, entry also causes a *competition effect* (Lewis and Poilly 2012). The competition effect decreases markups. Falling markups increases incumbents' output.[6] This enhances the scale effect. Hence, during transition entry has opposing effects on firm-level output and in turn measured productivity. It reduces firm size, via business stealing (profit arbitrage), which decreases the scale effect, but it also increases firm size, via competition decreasing markups, which increases the scale effect. The relative importance and persistence of the two effects depends on the speed of firm adjustment *i.e.* business dynamism.

In the long-run steady state, with a permanently higher level of firms, profits return to zero but markups are decreased to a permanently lower level. Consequently in zero profit steady-state, firms must produce more output in order to generate the revenue to cover fixed costs given the lower markup on each unit. Therefore in the long-run firms reside at a larger size leading to a long-run scale effect and productivity bias.

Firm-level returns to scale *and* firm-level intensive margin variation are *both* necessary and *jointly* they are sufficient for the endogenous productivity effect. In a benchmark model of monopolistic competition (constant markups à la Dixit Stiglitz) with fixed overhead costs and instantaneous entry (zero profits), firm output is constant. There is no intensive margin variation. Therefore, despite returns to scale and imperfect competition, there are no dynamic distortions.[7] Similarly, even if there are intensive margin variations but there are constant returns to scale, there are no dynamic distortions. This occurs in the model at the perfectly competitive equilibrium. Hence neither a benchmark monopolistic competition model nor a perfectly competi-

[5] By utilizing their returns to scale from the fixed cost, the incumbent firms get closer to the efficient level of production at minimum average cost where there are locally constant returns to scale. Hence despite benefiting from returns to scale, the level of returns to scale decreases as output expands.

[6] Lower markups raise output because firms must produce more units to generate the revenue to cover fixed costs when each unit solid has a lower price markup.

[7] Benchmark imperfect competition is a steady-state distortion. At a first-order approximation it does not affect model dynamics realtive to a perfect competition RBC model.

tive model satisfy these conditions.

Related Literature

It is well-documented that the interaction between imperfect competition, increasing returns to scale and technology shocks can explain procylical productivity (Basu 1996; Basu and Fernald 2001; Jaimovich and Floetotto 2008). I add to this literature by decomposing measured productivity into pure technology and endogenous components in the presence of dynamic entry, oligopolistic competition and increasing returns to scale.

My paper is most closely related to Devereux, Head, and Lapham (1996), Jaimovich and Floetotto (2008), Etro and Colciago (2010), and Bilbiie, Ghironi, and Melitz (2012), all of whom develop RBC models with endogenous firm entry.[8] Devereux, Head, and Lapham (1996) and Jaimovich and Floetotto (2008) develop models with increasing returns to scale and endogenous entry, but entry is static: it adjusts instantaneously due to a zero-profit condition. Whereas, Etro and Colciago (2010) and Bilbiie, Ghironi, and Melitz (2012) develop models of dynamic entry due to a time-to-build lag in firm creation, but returns to scale are constant.[9] My work emphasizes that linking *both* dynamic entry *and* increasing returns amplifies exogenous shocks and causes endogenous productivity effects. Furthermore, my setup allows me to distil pure technology and provide a tractable intertemporal decomposition of measured productivity.

Bilbiie, Ghironi, and Melitz (2012) is the seminal work on dynamic firm entry in a business cycle framework.[10] They show that dynamic entry is an important propagation mechanism of exogenous shocks that leads to quantitative improvements over the benchmark RBC model. Unlike my paper, the authors do not focus on distilling pure technology or explaining endogenous productivity movements due to returns to scale. They simulate their model with a technology shock process commonly used in other papers and based on King and Rebelo (1999). This technology shock process is estimated from a technology series acquired under the assumption of perfect competition. It is not the model-consistent technology series when there is imperfect competition and dynamic entry. The authors stress this point in their paper (Bilbiie, Ghironi, and

[8] Other papers with related model setups are Hornstein (1993), Rotemberg and Woodford (1993), Portier (1995), Ambler and Cardia (1998), Basu and Fernald (2001), Cook (2001), Kim (2004), and Chatterjee and Cooper (2014). These papers feature some combination of: dynamic or static entry; constant or endogenous markups; increasing returns or constant returns.

[9] These papers have *external returns to scale* implicit in the CES aggregator. These are typically referred to as variety effects (Ethier 1982) when entering in the consumption aggregator as in Bilbiie, Ghironi, and Melitz (2012) or returns to specialization (Devereux, Head, and Lapham 1996; Bénassy 1996) when entering in the output aggregator as in Etro and Colciago (2010).

[10] Ambler and Cardia (1998) and Cook (2001) develop early work moving from static to dynamic entry. They impose that zero-profits hold in expectation similar to Hopenhayn (1992). A shock causes profit variation as firms are predetermined but the following period entry fully adjusts to arbitrage profits.

Melitz 2012, p. 321), but they also explain that an advantage of using a common technology shock process is model comparability. It is clearer to compare the model's internal propagation mechanism relative to other models that also use the same technology process, rather changing the model equations *and* the exogenous technology process. An advantage of the dynamic entry cost setup that I use – with zero-profits in steady state – is that the aggregate production function is tractable. This allows me to identify a model-consistent technology with dynamic entry, imperfect competition *and* increasing returns. My aim of gaining a tractable understanding of productivity is closer to Jaimovich and Floetotto (2008). They identify a model-consistent technology series with static entry and oligopolistic (Bertrand) competition. They show that the framework can explain 40% of variation in measured TFP and creates significant output amplification.[11] They have globally increasing returns to scale, rather than allowing for a perfect competition outcome as in my analysis with U-shaped average cost curves. In Savagar and Dixon (2020) we study a similar model to this paper but with constant markups and we relate short-run endogenous productivity to capacity utilization. Constant markups allow us to take a more tractable approach but trivializes intensive margin variations leading to a theoretical, rather than quantitative, focus.

The papers by Jaimovich and Floetotto (2008), Etro and Colciago (2010), and Bilbiie, Ghironi, and Melitz (2012) all include endogenous markups that are countercylical due to procyclical firm entry. Procylical firm entry is well-documented (Tian 2018), whereas countercylical markups are debated. Bilbiie, Ghironi, and Melitz (2012) investigate endogenous markups caused by demand-side complementarities when there are translog preferences à la Feenstra (2003). Etro and Colciago (2010) study supply-side oligopolistic competition markups, similar to my work. They show that both Cournot and Bertrand forms of strategic interaction improve RBC moment matching. Lewis and Poilly (2012) estimate the size of the competition effect under the demand-side translog approach and supply-side oligopolistic approach. They find the competition effect is stronger under the translog approach. I favour the supply-side Cournot approach because it appears more relevant for *firm* entry, whereas demand-side complementarities seem more relevant for *product* entry as in Bilbiie, Ghironi, and Melitz (2012).

I employ an endogenous entry cost model based on the partial equilibrium model of Datta and Dixon (2002) and general equilibrium, perfect competition model of Brito and Dixon (2013). The endogenous entry cost setup aids my tractable analysis in continuous-time as entry costs tend to zero in the long-run implying a zero-profit

[11] In a supplementary appendix, Jaimovich and Floetotto (2008) include dynamic entry as a robustness check. They find that it causes weaker amplification than their main result. This is a quantitative robustness exercise that does not elabourate on the mechanisms.

steady state and equivalence between cost shares and revenue shares. The importance of endogenous entry costs are widely recognised in industrial organization literature. In particular, the entry cost setup I use is similar to S. Das and S. P. Das (1997). Recently, entry congestion effects have been adopted in several macroeconomic models such as Lewis and Poilly (2012), Bergin and Lin (2012), Berentsen and Waller (2015), Poutineau and Vermandel (2015), Bergin, Feng, and Lin (2016), Gutiérrez, Jones, and Philippon (2019), and Boar and Midrigan (2020) and also in finance by Loualiche (2019). Lewis (2009) provides empirical evidence on the importance of the entry congestion channel for monetary policy transmission. Entry congestion effects are typically motivated by advertising costs or some factor required to setup a firm that is in inelastic supply. In Aloi, Dixon, and Savagar (2021) we interpret congestion as a queuing dynamic which may result from regulatory barriers to entry such as business permits. We present empirical evidence that such procedures slow firm creation.

Lastly, the paper is related to recent work by Baqaee and Farhi (2020) that provides non-parametric, static decompositions of measured productivity for economies with heterogeneity and inefficiencies. The focus of my paper is more parametric and the decompositions are dynamic. There is no role for selection or reallocation.

Outline

Section 1 presents VAR evidence on the response of firm entry and output to a technology shock. Section 2 sets up the model environment. Section 3 solves household and firm optimization problems and characterises equilibrium. Section 4 decomposes measured productivity and estimates a model-consistent technology series. Section 5 simulates the model. Section 6 concludes.

1 Empirical Evidence

The mechanism I explore in this paper rests on output *per firm* (intensive margin) adjusting in the short run and interacting with scale economies, whereas over time firm entry (extensive margin) adjusts which alleviates the short-run scale effect. In this section I provide empirical evidence that firm entry is sluggish and in aggregate it responds more slowly than output. This implies variations in firm intensive margin are present in the short-run, but dissipate in the long-run as the extensive margin of firm entry adjusts.

Figure 1 presents impulse response functions for aggregate output and the stock of firms following a one standard deviation shock to the error term of the technology regression. It shows that aggregate output responds faster than the stock of firms to a technology shock. The results follow from a VAR with four lags applied to quarterly

US data that is detrended with a fourth-order polynomial.[12] The results support the mechanism I outline in this paper. The short-run adjustment of firms is much slower and smaller in magnitude than the response of aggregate output. This implies that on average, output per firm, rises on impact of the shock. The contribution of the intensive margin on impact is more important than the extensive margin, but the extensive margin grows in importance as entry adjusts. In an economy with increasing returns to scale, my theory predicts that this dynamic will create short-run endogenous productivity movements.

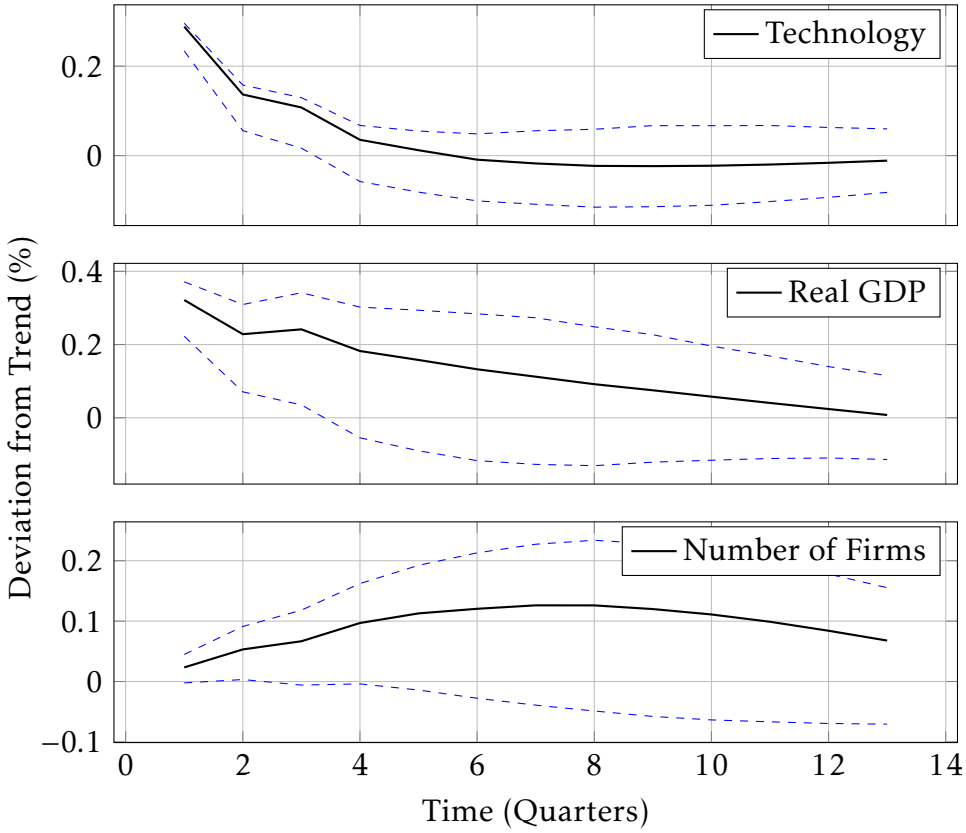


Figure 1: VAR(4) IRF to 1SD Technology Shock (90% confidence intervals)

2 Model Setup

This section describes the model economy. It presents the primitives that characterise households and firms and their objectives and constraints.

[12]The VAR results present 90% bootstrap confidence intervals (10,000 runs). A fourth-order polynomial is sufficient to make the data stationary. The technology series is constructed as $\hat{A} = \frac{\nu}{\tilde{\mu}}\left[\hat{Y} - \tilde{\mu}\tilde{s_L}(\hat{L} - \hat{K}) - \tilde{\mu}\hat{K} - (1-\tilde{\mu})\hat{N}\right]$. This adjusted technology series is consistent with the model of imperfect competition, scale economies, and dynamic entry that I develop in this paper. It is equivalent to a Solow Residual with unit markup $\tilde{\mu} = 1$ and constant returns $\nu = 1$. The parameters are calibrated as $s_L = 0.7$, $\nu = 0.95$ and $\mu = 1.3$. The VAR results are similar if we use a Solow Residual rather than our corrected technology measure.

2.1 Household

A representative household chooses consumption $\{C(t)\}_0^\infty \in \mathbb{R}$, labour supply $\{L(t)\}_0^\infty \in [0,1]$ and firm entry (firm investment) $\{E(t)\}_0^\infty \in \mathbb{R}$ to maximise lifetime utility $U : \mathbb{R}^2 \to \mathbb{R}$ subject to the evolution of state variables capital $\{K(t)\}_0^\infty \in \mathbb{R}_+$ and number of firms $\{N(t)\}_0^\infty \in [1,\infty)$.

Assumption 1. *Instantaneous utility* $u : \mathbb{R} \times [0,1] \to \mathbb{R}$ *satisfies*

1. *Strictly increasing in consumption and strictly decreasing in labour* $u_C > 0, u_L < 0$.
2. *Inada conditions on consumption and labour supply hold*

$$\lim_{L\to\infty} u_L = -\infty, \lim_{C\to\infty} u_C = 0, \lim_{L\to 0^+} u_L = 0, \lim_{C\to 0^+} u_C = +\infty.$$

3. *Twice differentiable with diminishing marginal utilities* $u_{LL}, u_{CC} \leq 0$.
4. *Additively separable* $u_{CL} = u_{LC} = 0$.[13]

The household maximizes utility subject to a budget constraint which states that expenditure on consumption, capital investment and firm investment equates to income from renting capital, wages from labour and profits from owning firms

$$C(t) + I(t) + Z(t) = rK(t) + wL(t) + \pi N(t).$$

The household takes equilibrium rental rate, wage rate and firm profits $r, w, \pi \in \mathbb{R}_+$, which are in consumption good prices, as given. Capital investment $I(t) \equiv \dot{K}(t)$ is given by the flow of capital as there is no depreciation. Firm investment $Z(t) \equiv (\zeta/2)\dot{N}^2$ is given by a quadratic entry cost that depends on the flow of entry. This represents a congestion effect in firm entry. The parameter $\zeta \in (0,\infty)$ controls the congestion effect. As $\zeta \to 0$, congestion disappears, entry adjusts instantaneously, and profits are always zero. As $\zeta \to \infty$, the model has a fixed number of firms. $\rho \in \mathbb{R}_+$ is the household discount factor.

2.2 Firm

On the production side of the economy there is a nested CES aggregator. Aggregate output Y consists of a continuum of industries $\jmath \in [0,1]$ each producing $Q_\jmath$. Industry output consists of $N \geq 1$ firms each producing $y_{\jmath\imath}$, $\imath \in 1 \ldots N$. At the aggregate level and industry level there is perfect competition, whereas at the firm level there

[13] Additive separability and diminishing marginal utility imply that u is concave in (C,L) since $u_{CC}u_{LL} - u_{CL}^2 > 0$ always holds (negative semi-definite Hessian).

is oligopolistic competition. With oligopolistic competition, a firm can affect own-price $P_{\jmath\imath}$ through a direct effect, as with monopolistic competition, but also through an indirect effect via its effect on the industry level output. The result is an endogenous demand elasticity that becomes more elastic with more firms, hence a markup that decreases in entry. Both levels of aggregation have constant returns (no variety effects).[14]

The aggregate output production function is

$$Y(t) = \left[\int_0^1 Q_\jmath(t)^{\frac{\theta_I - 1}{\theta_I}} d\jmath\right]^{\frac{\theta_I}{\theta_I - 1}}, \quad \theta_I \geq 1. \tag{1}$$

The aggregator has constant returns and $\theta_I \in [1, \infty)$ is between-industry substitutability.[15] The industry output production function is

$$Q_\jmath(t) = N(t)\left[\frac{1}{N(t)}\sum_{\imath=1}^{N} y_{\jmath\imath}(t)^{\frac{\theta_F - 1}{\theta_F}}\right]^{\frac{\theta_F}{\theta_F - 1}}, \quad \theta_F > \theta_I \geq 1 \tag{2}$$

The aggregator has constant returns (no variety effects) and $\theta_F \in (1, \infty)$ is between-firm substitutability. The assumption that $\theta_F > \theta_I$ ensures that products within an industry are more substitutable than products in different industries. Firm $\imath$ in industry $\jmath$ produces output:

$$y_{\jmath\imath}(t) := \max\{AF(k_{\jmath\imath}(t), \ell_{\jmath\imath}(t)) - \phi, 0\} \tag{3}$$

$\phi \in \mathbb{R}_+$ is an output-denominated overhead cost.[16] $A \in \mathbb{R}_+$ is a scale parameter reflecting the production technology.

Assumption 2. *The production function $F : \mathbb{R}^2_+ \supseteq (k, \ell) \to \mathbb{R}_+$ satisfies:*

(i) F is twice differentiable and strictly increasing in capital and labour: $F_k, F_\ell > 0$.

(ii) The Inada conditions hold:

$$\lim_{k\to 0} F_k(k, \ell) = \lim_{\ell \to 0} F_\ell(k, \ell) = \infty, \ \lim_{k\to\infty} F_k(k, \ell) = \lim_{\ell\to\infty} F_\ell(k, \ell) = 0, \ F(0, 0) = 0.$$

[14] Returns to scale of the aggregate production functions are not the same as returns to scale at the firm level. Hence we have constant returns in the aggregate production functions, but increasing, constant or decreasing returns at the firm level.

[15] The $\theta = 1$ should be specified as a Cobb-Douglas aggregator on a continuum.

[16] The fixed overhead parameter implies profits will be zero in steady-state despite market power. The overhead cost means that marginal costs do not measure returns to scale. We focus on the case where marginal costs are increasing, but average costs are decreasing, so there are (locally) increasing returns to scale.

(iii) F is concave which implies

$$F_{k\ell} = F_{\ell k} > 0,\ F_{kk}, F_{\ell\ell} < 0,\ F_{kk}F_{\ell\ell} - F_{k\ell}^2 > 0.$$

(iv) F is homogeneous of degree $\nu \in (0,1)$ which implies $\nu F = F_k k + F_\ell \ell$.

3 Equilibrium Conditions

3.1 Household Equilibrium Conditions

The household solves

$$\max_{K,C,L,N,E} \mathrm{U} := \int_0^\infty e^{-\rho t}\, u(C(t), 1 - L(t))\, \mathrm{dt}$$

$$\text{s.t.} \quad \dot{K}(t) = rK(t) + wL(t) + \pi N(t) - C(t) - \frac{\zeta}{2}E(t)^2 \tag{4}$$

$$\dot{N}(t) = E(t). \tag{5}$$

Initial conditions on the state variables are given by $K(0) = K_0$ and $N(0) = N_0$. The optimal solutions satisfy

$$w = -\frac{u_L(L(t))}{u_C(C(t))} \tag{6}$$

$$\dot{C} = -\frac{u_C(C(t))}{u_{CC}(C(t))}(r - \rho) \tag{7}$$

$$\zeta \dot{E} = r\zeta E(t) - \pi. \tag{8}$$

Therefore there are five equations in five unknowns $\{C, E, K, N, L\}$. There are two transversality conditions for the state variables

$$\lim_{t\to\infty} e^{-\rho t} u_C(C(t))K(t) = 0 \qquad \lim_{t\to\infty} e^{-\rho t} u_C(C(t))\zeta E(t)N(t) = 0.$$

Together with the two initial conditions, there are four boundary conditions corresponding to the four differential equations. Equation (6) is the intratemporal condition between consumption and labour and equation (7) is the intertemporal consumption condition. Equation (8) is a firm asset pricing equation. It states there is no arbitrage between investing in firms and capital. To see this note that $s(t) \equiv \zeta E(t)$ is the relative price of a firm (in units of marginal utility).[17] Hence the arbitrage condition

[17] Marginal utility u_C is the shadow price of consumption and $u_C \zeta E(t)$ is the shadow price of a firm. Therefore $s(t)$ is the shadow price of a firm relative to the shadow price of consumption. Hence $s(t) \equiv \zeta E(t)$ is the value of an additional firm in consumption unit terms.

states that the return to owning a firm π plus the change in firm value $\zeta \dot{E}$ equates to the return on the entry cost invested in capital $r \times \zeta E(t)$.

3.2 Firm Equilibrium Conditions

The final goods and sectoral goods producers maximize profits subject to their production functions which have constant return to scale. They operate under perfect competition so treat prices as given. The resulting inverse demands take constant elasticity form.

Final Goods Problem

Final goods producers solve:

$$\max_{Q_\jmath,\ \jmath \in [0,1]} \quad PY - \int_0^1 Q_\jmath P_\jmath \, \mathrm{d}\jmath$$

subject to (1). The first-order condition gives inverse demand for industry $\jmath$

$$P_\jmath = \left(\frac{Q_\jmath}{Y}\right)^{-\frac{1}{\theta_I}} P, \quad \forall \jmath \in [0,1]. \tag{9}$$

The corresponding price index is

$$P = \left(\int_0^1 P_\jmath^{1-\theta_I} d\jmath\right)^{\frac{1}{1-\theta_I}}.$$

Sectoral Goods Problem

Sectoral good producers solve:

$$\max_{y_{\jmath\imath},\ \imath \in (1 \ldots N)} \quad P_\jmath Q_\jmath - \sum_1^N y_{\jmath\imath} P_{\jmath\imath}$$

subject to (2). The first-order condition gives inverse demand for intermediate producer $\imath$ in sector $\jmath$ as follows

$$P_{\jmath\imath} = \left(\frac{y_{\jmath\imath}}{Q_\jmath}\right)^{-\frac{1}{\theta_F}} N^{-\frac{1}{\theta_F}} P_\jmath. \tag{10}$$

The corresponding price index is:

$$P_J = \left(\sum_{1}^{N} P_{J\imath}^{1-\theta_F} \right)^{\frac{1}{1-\theta_F}}.$$

Firm Problem

The individual firm operates under Cournot competition. When choosing output to maximize profits they recognise that this affects industry demand. The strength of this effect depends on their market share which causes an endogenous markup. The firm solves its production decision in two stages. First, the firm decides optimal input choices to produce a given output through cost minimization. Second, the firm decides optimal output to maximize profits. This illustrates that firms face non-constant marginal cost curves.

To choose inputs optimally for a given level of output, the firm solves

$$\begin{aligned} C(w, r, y_{J\imath}) \coloneqq & \min_{k_{J,\imath}, \ell_{J,\imath}} w\ell_{J,\imath} + rk_{J,\imath} \\ \text{s.t.} \quad & y_{J\imath} \leq AF(k_{J\imath}, \ell_{J\imath}) - \phi \end{aligned} \tag{11}$$

where factor prices are in final good prices $w \equiv w^{\text{nom.}}/P$ and $r \equiv r^{\text{nom.}}/P$. The optimality conditions are:

$$r = \lambda_{J\imath} AF_k(k, \ell) \qquad w = \lambda_{J\imath} AF_\ell(k, \ell) \qquad y_{J\imath} + \phi = AF(k_{J\imath}, \ell_{J\imath}).$$

where $\lambda_{J\imath}$ is the Lagrange multiplier which is equal to the marginal cost *at the optimal levels* of k, ℓ. Using the optimality conditions and Euler's homogeneous function theorem, we can represent the cost function as

$$C(w, r, y_{J\imath}) = \frac{\partial C(w, r, y_{J\imath})}{\partial y_{J\imath}} \nu(y_{J\imath} + \phi).$$

Integrating this partial differential equation gives the cost function in multiplicatively-separable form where $G(r, w, 1)$ is an arbitrary function independent of $y_{J\imath}$:

$$\mathrm{C}(r, w, y_{J\imath}) = \left(\frac{y_{J\imath} + \phi}{A} \right)^{\frac{1}{\nu}} G(r, w, 1).$$

Therefore the marginal cost is

$$\mathrm{MC} \equiv \frac{\partial C(w, r, y_{J\imath})}{\partial y_{J\imath}} = \frac{1}{\nu} \frac{\mathrm{C}(r, w, y_{J\imath})}{y_{J\imath} + \phi} = \frac{1}{\nu A^{\frac{1}{\nu}}} \left(y_{J\imath} + \phi \right)^{\frac{1}{\nu} - 1} G(r, w, 1) > 0. \tag{12}$$

The elasticity of the cost function with respect to output is $\varepsilon_{Cy} = [\nu(1+s_\phi)]^{-1}$, which we later define as returns to scale. The elasticity of the marginal cost curve with respect to output is $\varepsilon_{MCy} = (1-\nu)\varepsilon_{Cy}$.[18] This shows that the marginal cost is increasing if $\nu \in (0,1)$, constant if $\nu = 1$, and decreasing if $\nu \in (1,\infty)$.

Given the cost function the firm maximizes profits subject to inverse demand. Inverse demand follows from combining aggregate and sectoral demand functions. Therefore the firm solves

$$\begin{aligned} \max_{y_{J,\iota}} \pi_{J\iota} &= \frac{P_{J\iota}}{P} y_{J\iota} - C(w,r,y_{J\iota}) \\ \text{s.t.} \quad \frac{P_{J\iota}}{P} &= \left(\frac{y_{J\iota}}{Q_J(y_{J\iota})}\right)^{-\frac{1}{\theta_F}} \left(\frac{Q_J(y_{J\iota})}{Y}\right)^{-\frac{1}{\theta_I}} N^{-\frac{1}{\theta_F}}. \end{aligned} \tag{13}$$

Operating profits are in final good prices $\pi \equiv \pi^{\text{nom.}}/P$. The first-order condition implies a firm chooses output such that marginal revenue equals to marginal cost, which implies that a firm chooses $y_{J\iota}$ to satisfy

$$\mu_{J\iota} = \frac{\epsilon_{J\iota}}{\epsilon_{J\iota} - 1} \tag{14}$$

where $\mu_{J\iota} \in (1,\infty)$ is the markup of price over marginal cost and $\epsilon_{J\iota} \in (1,\infty)$ is the price elasticity of demand. They are defined as follows:

$$\epsilon_{J\iota} \equiv -\frac{\partial y_{J\iota}}{\partial P_{J\iota}} \frac{P_{J\iota}}{y_{J\iota}} \text{ and } \mu_{J\iota} \equiv \frac{P_{J\iota}/P}{\partial C(w,r,y_{J\iota})/\partial y_{J\iota}}.$$

The firm-level inverse demand (13) and industry-level production function (2), yield an expression for the price elasticity of demand

$$\epsilon_{J\iota} = \left[\frac{1}{\theta_F} + \left[\frac{1}{\theta_I} - \frac{1}{\theta_F}\right] \frac{\partial Q_J}{\partial y_{J\iota}} \frac{y_{J\iota}}{Q_J}\right]^{-1}, \quad \text{where } \frac{\partial Q_{J\iota}}{\partial y_{J\iota}} \frac{y_{J\iota}}{Q_{J\iota}} = \frac{y_{J\iota}^{\frac{\theta_F-1}{\theta_F}}}{\sum_{\iota=1}^{N} y_{J\iota}^{\frac{\theta_F-1}{\theta_F}}}.$$

The price elasticity depends on *exogenous* between-firm and between-industry substitutability and *endogenous* industry elasticity to own-output. The industry elasticity to own output is equal to market share.[19] Dividing the optimal input choice conditions

[18] Throughout the paper, for two variables X, Y the notation ε_{YX} represents the elasticity of Y with respect to X, that is $\varepsilon_{YX} \equiv \frac{\partial Y}{\partial X}\frac{X}{Y} = \frac{\partial \ln X}{\partial \ln Y}$.

[19] The special cases of a single-firm market and many-firm market occur when the share is 1 or 0.

by $P_{\jmath\imath}/P$ allows us to write them in terms of the markup:[20]

$$w = \frac{P_{\jmath\imath}}{P}\frac{AF_\ell(k_{\jmath\imath},\ell_{\jmath\imath})}{\mu_{\jmath\imath}}, \quad r = \frac{P_{\jmath\imath}}{P}\frac{AF_k(k_{\jmath\imath},\ell_{\jmath\imath})}{\mu_{\jmath\imath}}.$$

These conditions in conjunction with the firm output function and the markup pricing rule, which depends on aggregate and sectoral production functions, characterise equilibrium production for the firm.

3.3 Market Clearing

Factor market clearing requires the following feasibility conditions

$$K(t) = \int_0^1 \left[\sum_{\imath=1}^{N} k_{\jmath\imath}(t)\right] d\jmath \quad \text{and} \quad L(t) = \int_0^1 \left[\sum_{\imath=1}^{N} \ell_{\jmath\imath}(t)\right] d\jmath. \tag{15}$$

Goods market clearing implies output is divided among consumption, capital investment and firm investment $Y(t) = C(t) + I(t) + Z(t)$ which implies

$$Y(t) = C(t) + \dot{K} + \frac{\zeta}{2}E(t)^2. \tag{16}$$

3.4 Symmetric Equilibrium

The production function of a firm is symmetric with respect to all intermediate inputs. Therefore we study symmetric equilibria where all firms produce the same output and charge the same price. Under symmetry, intermediate variables are identical:

$$\forall(\jmath,\imath) \in [0,1]\times[1,N(t)]: \quad y_{\jmath\imath} = y,\ k_{\jmath\imath} = k,\ \ell_{\jmath\imath} = \ell.$$

The factor market clearing conditions imply

$$K = Nk, \quad L = N\ell. \tag{17}$$

Since there are constant returns at the aggregate and sectoral output level, price and output aggregation is simple.[21] The price indices show that all prices are equal under symmetry. We treat the final good as the numeraire therefore $P = P_\jmath = P_{\jmath\imath} = 1$. Aggregate and industry demand imply

$$Y = Q = Ny.$$

[20] Similarly the cost function is $C(r,w,y_{\jmath\imath}) = \frac{P_{\jmath\imath}}{P}\frac{\nu(y_{\jmath\imath}+\phi)}{\mu_{\jmath\imath}}$.

[21] This is the supply-side equivalent of there being no variety effects.

Under symmetry the price elasticity of demand is $\epsilon(N) = \left[\frac{1}{\theta_F} + \left(\frac{1}{\theta_I} - \frac{1}{\theta_F}\right)\frac{1}{N}\right]^{-1}$ and the markup becomes

$$\mu(N) = \left[1 - \frac{1}{\theta_F} - \frac{1}{N}\left(\frac{1}{\theta_I} - \frac{1}{\theta_F}\right)\right]^{-1} \tag{18}$$

where $N \geq 1$ and $\theta_F > \theta_I \geq 1$. The markup is decreasing in the number of firms because entry decreases incumbents' market share $1/N$, which increases price elasticity, which decreases price setting ability.[22] The elasticity of the markup with respect to the number of firms is

$$\varepsilon_{\mu N} = -\frac{\mu}{N}\left(\frac{1}{\theta_I} - \frac{1}{\theta_F}\right) = -\left[\left(1 - \frac{1}{\theta_F}\right)\left(\frac{1}{N}\left(\frac{1}{\theta_I} - \frac{1}{\theta_F}\right)\right)^{-1} - 1\right]^{-1} < 0$$

Firm-level output combined with aggregate output implies:

$$Y = N^{1-\nu}AF(K,L) - N\phi. \tag{19}$$

The factor market equilibrium conditions become

$$r = \frac{1}{\mu(N)}AN^{1-\nu}F_K(K,L) \tag{20}$$

$$w = \frac{1}{\mu(N)}AN^{1-\nu}F_L(K,L). \tag{21}$$

Definition 1. An equilibrium consists of paths of consumption, entry, labour, capital, firms, wages, capital rental and operating profits

$$[C(t), E(t), L(t), K(t), N(t), w(t), r(t), \pi(t)]_{t=0}^{\infty},$$

such that the representative household maximizes its utility given initial asset holdings of capital ($K(0) > 0$) and firms ($N(0) \geq 1$) and taking the time path of prices and profits $[w(t), r(t), \pi(t)]_{t=0}^{\infty}$ as given; firms maximize profits taking the time path of factor prices $[w(t), r(t)]_{t=0}^{\infty}$ as given; and factor prices and profits $[w(t), r(t), \pi(t)]_{t=0}^{\infty}$ are such that all markets clear.

Table 1 summarises the equilibrium conditions under symmetry. The model conditions form a system of ten equations in ten endogenous variables $(Y, r, w, \pi, \mu, L, C, E, K, N)$. The model reduces further to a four-dimensional dynamical system in (C, E, K, N).

[22] Note the following special cases. With homogeneous goods (perfectly substitutable goods within an industry) $\theta_F \to \infty$ and $\epsilon(N) = \theta_I N$ then the markup is $\mu(N) = \theta_I N/(\theta_I N - 1)$. A further simplification is to also assume $\theta_I = 1$ (a Cobb-Douglas aggregator across the continuum of sectors *i.e.* perfectly differentiated goods) , which implies price elasticity of demand $\epsilon(N) = N$ and markup $\mu(N) = N/(N-1)$. A second special case, is to leave θ_I and θ_F but assume a continuum of single-firm industries such that $N = 1$. In this case, the price elasticity of demand is constant $\epsilon = \theta_I$ and the markup is also constant $\mu = \theta_I/(\theta_I - 1)$. This is the monopolistic competition, or Dixit-Stiglitz, case.

There are four boundary conditions given by the two initial conditions on the state variables and the two transversality conditions on the state variables. Capital and number of the firms (K, N) are the state variables. Consumption and entry (C, E) are the jump variables. Labour is defined implicitly as $L(C, K, N)$ through labour market equilibrium.

Household	
Labour Supply	$w = -\frac{u_L(L)}{u_C(C)}$
Consumption Euler	$\dot{C} = -\frac{u_C(C)}{u_{CC}(C)}(r - \rho)$
Entry Arbitrage	$\zeta \dot{E} = r\zeta E - \pi$
Budget Constraint	$\dot{K} = rK + wL + \pi N - C - \frac{\zeta}{2}E^2$
Entry Rate	$\dot{N} = E$
Firm	
Markup (Cournot)	$\mu = \left[1 - \frac{1}{\theta_F} - \frac{1}{N}\left(\frac{1}{\theta_I} - \frac{1}{\theta_F}\right)\right]^{-1}$
Labour Demand	$w = \frac{1}{\mu} A N^{1-\nu} F_L(K, L)$
Capital Rental	$r = \frac{1}{\mu} A N^{1-\nu} F_K(K, L)$
Aggregate Output	$Y = AN^{1-\nu} F(K, L) - N\phi$
Market Clearing	
Aggregate Accounting	$Y = C + \dot{K} + \frac{\zeta}{2}E^2$

Table 1: Equilibrium Conditions

3.4.1 Reduced-form Optimal Profit Expression

A crucial reduced-form relationship from the equilibrium conditions relates profits, output, markups and fixed costs. Combining the equilibrium conditions for the household budget constraint and goods market clearing condition yields an aggregate operating profits condition

$$N\pi = Y - wL - rK.$$

Then if we substitute in equilibrium factor prices, we get a reduced-form optimal profit expression:

$$N\pi = \left(1 - \frac{\nu}{\mu(N)}\right)(Y + N\phi) - N\phi. \tag{22}$$

We can also express this in intensive margin form which links firm-level output, profits and markups:

$$\pi = \left(1 - \frac{\nu}{\mu}\right)(y + \phi) - \phi. \tag{23}$$

3.5 Steady-State Equilibrium

In steady-state equilibrium $\dot{C} = \dot{E} = \dot{K} = \dot{N} = 0$ at steady-state levels $\tilde{C}, \tilde{E}, \tilde{K}, \tilde{N}$, where tildes denote variables at steady state.[23] In steady state the interest rate equals the discount factor $\tilde{r}(\tilde{C}, \tilde{K}, \tilde{N}) = \rho$; profits are zero $\tilde{\pi}(\tilde{C}, \tilde{K}, \tilde{N}) = 0$; output equals consumption $\tilde{Y}(\tilde{C}, \tilde{K}, \tilde{N}) = \tilde{C}$ and entry is zero $\tilde{E} = 0$. The zero-profit outcome follows from the entry arbitrage equation and entry rate definition. It implies that in steady state costs and revenues are equal. Furthermore it allows us to understand firm-level output behaviour.

Proposition 1. *Steady-state output per firm is increasing in the number of firms and the fixed-cost share $\tilde{s}_\phi \equiv \phi/\tilde{y}$ is decreasing in the number of firms:*

$$\tilde{y} = \frac{\nu\phi}{\mu(\tilde{N}) - \nu}, \qquad \tilde{s}_\phi = \frac{\mu(\tilde{N})}{\nu} - 1.$$

The result follows by noting that $\pi = \left(1 - \frac{\nu}{\mu}\right)(y + \phi) - \phi$, then imposing zero profits and rearranging. The result arises because output depends on the markup which is decreasing in the number of firms due to oligopolistic competition. The intuition is that entry decreases the markup so a firm increases output in order to generate enough revenue to cover fixed costs in zero-profit steady state.

3.6 Labour Market

We can understand labour market behaviour by equating labour supply with demand, then log-linearizing, which gives

$$\hat{L}(t) = \left(\varepsilon_{u_L L} - \varepsilon_{F_L L}\right)^{-1} \left[\hat{A} + \varepsilon_{u_C C}\hat{C}(t) + \varepsilon_{F_L K}\hat{K}(t) - (\varepsilon_{F_L K} + \varepsilon_{F_L L} + \varepsilon_{\mu N})\hat{N}(t)\right] \tag{24}$$

where hat notation represents deviation from steady state $\hat{X} \equiv \dot{X}/\tilde{X}$.[24] Given the assumptions on the production function and utility function, labour responds positively to technology, negatively to consumption, positively to capital and positively to number of firms. The number of firms coefficient is positive because $\nu < 1$ and by Euler's homogeneous function theorem $\varepsilon_{F_L K} + \varepsilon_{F_L L} = \nu - 1$.[25] The elasticity of the marginal utility of consumption to further consumption is negative ($\varepsilon_{u_C C} < 0$) due to dimishing marginal utility.

[23] The conditions are nonlinear, and steady-state may not be well-defined.

[24] In our quantitative model specification: $\varepsilon_{u_L L} = \eta$, $\varepsilon_{F_L L} = \beta - 1$, $\varepsilon_{F_L K} = \alpha$, $\varepsilon_{u_C C} = -\sigma$ and $\varepsilon_{\mu N}$ is a nonlinear function of $\tilde{N}, \theta_I, \theta_N$ where $\tilde{N}$ in an intractable function of model parameters.

[25] The pre-multiplying reciprocal is strictly positive which ensures labour market equilibrium existence. It represents the difference between the gradients of the postively-sloped labour-supply curve $\varepsilon_{wL}^{\text{Supply}} = \varepsilon_{u_L L}$ and negatively-sloped labour-demand curve $\varepsilon_{wL}^{\text{Demand}} = \varepsilon_{F_L L}$.

3.7 Scale Economies

Scale economies are defined as the inverse elasticity of costs to output, which is equal to the average cost to marginal cost ratio:

$$\text{RTS} \equiv \varepsilon_{CY}^{-1} = [C_Y(Y/C)]^{-1} = \text{AC}/\text{MC}.$$

There are increasing returns if $\text{AC}/\text{MC} \in (1,\infty)$, constant returns if $\text{AC}/\text{MC} = 1$ and decreasing returns if $\text{AC}/\text{MC} \in (0,1)$. Our model setup has two expressions for returns to scale. The first is

$$\text{RTS} = \nu(1+s_\phi), \quad \text{where } s_\phi = \frac{\phi}{y}.$$

This measure of scale economies is based on technical parameters of the production function. It follows from dividing the optimal cost function, $\text{C} = \nu\,\text{MC}(y+\phi)$, by $y\,\text{MC}$ thus giving AC/MC. The second expression for returns to scale is

$$\text{RTS} = \mu(1-s_\pi), \quad \text{where } s_\pi = \frac{\pi}{Py}.$$

This measure of scale economies is based on behavioural factors relating to the decisions of profit-maximizing firms. It follows from the profit expression $\pi = (P-\text{AC})y = \left(1-\frac{\text{AC}}{\text{MC}}\frac{\text{MC}}{P}\right)Py = \left(1-\frac{\text{RTS}}{\mu}\right)Py$. The two expressions for returns to scale are linked via the reduced-form profit expression:

$$\text{RTS} = \nu\left(1+s_\phi\right) = \mu\left(1-s_\pi\right) \gtreqless 1. \tag{25}$$

This relates profit share, fixed cost share and the markup.

If we (log-)linearize the two expressions we get

$$\begin{aligned}
\hat{\text{RTS}} &= \hat{\mu} - \hat{s}_\pi \\
\hat{\text{RTS}} &= \left(\frac{\tilde{s}_\phi}{1+\tilde{s}_\phi}\right)\hat{s}_\phi = \left(1-\frac{\nu}{\tilde{\mu}}\right)\hat{s}_\phi, \quad \text{where } \hat{s}_\phi = -\hat{y}
\end{aligned}$$

where we define $\hat{s}_\pi \equiv \dot{s}_\pi$.[26] The second expression emphasizes that returns to scale are countercyclical and will have a lower variance than the fixed cost share variance. The first expression shows that returns to scale are countercylical due to a positive relationship with countercyclical markups and a negative relationship with procyclical profits.

[26] The notation is inconsistent with our usual notation that $\hat{X} \equiv \dot{X}/\tilde{X}$ because the profit-share are zero in steady state.

3.7.1 Scale Economies in Steady State

Proposition 2. *In steady state the markup and returns to scale are equivalent. Therefore there are increasing or constant returns in steady state and steady-state returns to scale are increasing in the number of firms.*

$$\tilde{\text{RTS}} = \nu(1 + \tilde{s}_\phi) = \mu(\tilde{N}) \geq 1$$

3.7.2 Scale Economies in Related Literature

In Etro and Colciago (2010) and Bilbiie, Ghironi, and Melitz (2012) there are no overhead costs $\phi = 0$ and marginal costs are constant $\nu = 1$, therefore RTS $= \nu(1 + s_\phi) = 1$ or, by the equilibrium profit condition $\pi = \left(1 - \frac{1}{\mu}\right)y$, then RTS $= \mu(1 - s_\pi) = 1$.[27] Even though dynamic entry creates the same short-run intensive margin (output per firm) and long-run extensive margin (output across all firms) effects as in this paper, the firm size variations do not interact with scale economies, therefore there are no endogenous productivity effects through this mechanism. In Jaimovich and Floetotto (2008) the production function has fixed costs $\phi > 0$ and a flat marginal cost $\nu = 1$, with $s_\pi = 0$ due to instantaneous entry.[28] Since there are always zero profits, returns to scale are equivalent to the fixed cost and markup RTS $= 1 + s_\phi = \mu$. Since there are zero profits N is determined instantaneously. Consequently output and the fixed cost share are determined instantaneously following a permanent shock. The transition dynamics of returns to scale are equivalent to the markup. There is no strengthening effect from profits.

4 Measured Productivity

The reduced-form optimal profit condition (22) gives the number of firms as a function of aggregate output, which in turn can be used to remove the aggregate fixed cost component from the aggregate output expression (19). Consequently, under equilibrium conditions, we can write aggregate output in Cobb-Douglas form:

$$Y = \left(\frac{A}{\pi + \phi}\right)^{\frac{1}{\nu}} \left(1 - \frac{\nu}{\mu}\right)^{\frac{1}{\nu}-1} \left(\pi + \frac{\nu}{\mu}\phi\right) F(K, L)^{\frac{1}{\nu}}. \tag{26}$$

[27] In BGM the profit share (gross of entry costs) is always positive and $\mu > 1$. They do not have an overhead cost which in our model offsets profit, instead they have a fixed entry cost that equates to profit in the long run. In our model entry costs are zero in the long run.

[28] The production function parameters imply a flat marginal cost and globally downward sloping average cost. Hence there are globally increasing returns and equilibrium only exists with imperfect competition $\mu > 1$. There is no Walrasian benchmark at *minimum efficient scale*.

The full derivation is in the appendix. This expression shows that aggregate output is a function of factor inputs and an efficiency term that I refer to as measured TFP.

Definition 2 (Measured TFP)**.** Measured TFP is the fraction of aggregate output that is not explained by factor inputs in capital and labour where the denominator is adjusted to be homogeneous of degree 1:

$$\text{TFP} \equiv \frac{Y}{F(K,L)^{\frac{1}{\nu}}}.$$

I use the term 'measured' to recognise that this is the TFP one would acquire from a traditional Solow Residual measurement exercise.

Proposition 3. *Measured TFP is equivalent to a Solow Residual. That is, measured TFP equals to aggregate output less share-weighted factor inputs:*

$$\hat{\text{TFP}} = \hat{Y} - \tilde{s}_K \hat{K} - \tilde{s}_L \hat{L},$$

where hat notation represents deviation from trend ($\hat{X} \equiv \dot{X}/\tilde{X}$) and steady-state factor shares are given by $\tilde{s}_K \equiv rK/PY$ and $\tilde{s}_L \equiv wL/PY$. Steady-state factor shares in revenue and costs are equal due to zero-profit steady state.

To show that measured TFP is equivalent to a Solow Residual, log-linearize $\text{TFP} = Y/F(K,L)^{1/\nu}$ around steady-state to give

$$\hat{\text{TFP}} = \hat{Y} - \frac{1}{\nu}\left(\tilde{\varepsilon}_{FK}\hat{K} + \tilde{\varepsilon}_{FL}\hat{L}\right).$$

In zero-profit steady state, output elasticities are given by $\tilde{\varepsilon}_{FK} = \nu\tilde{s}_K$ and $\tilde{\varepsilon}_{FL} = \nu\tilde{s}_L$ (see appendix). Hence,

$$\hat{\text{TFP}} = \hat{Y} - \tilde{s}_K \hat{K} - \tilde{s}_L \hat{L}.$$

4.1 Output Elasticities

Factor shares are related to production function elasticites as follows

$$s_L = \varepsilon_{FL}\frac{(1+s_\phi)}{\mu} = \varepsilon_{FL}\frac{(1-s_\pi)}{\nu} \text{ and } s_K = \varepsilon_{FK}\frac{(1+s_\phi)}{\mu} = \varepsilon_{FK}\frac{(1-s_\pi)}{\nu}$$

where shares are in revenue terms. Full derivations are in the appendix. When evaluated at zero-profit steady state revenue (R) and cost (C) shares, $s^R/(1-s_\pi) = s^C$, are equal and output elasticity to an input is equal to the product of ν and total expenditure of the input in revenue or cost. Notice that this is the popular ratio-estimator of the markup. That is, the markup is the ratio of output elasticity $\varepsilon_{YL} = \varepsilon_{FL}(1+s_\phi)$ to the input share s_L, thus $\mu = \varepsilon_{YL}/s_L$.

4.2 Measured TFP in Steady State

In steady-state measured TFP depends endogenously on the number of firms via the markup

$$T\tilde{F}P = \nu \left[\frac{A}{\mu(\tilde{N})} \left(\frac{\mu(\tilde{N}) - \nu}{\phi} \right)^{1-\nu} \right]^{\frac{1}{\nu}}$$

Since there are increasing returns in steady state and output per firm is increasing in entry, TFP is increasing in entry due to increasing returns.[29]

Proposition 4. *Steady-state measured TFP is increasing in the number of firms.*

4.3 Measured TFP Decompositions

Equation (26) shows that measured TFP is not equivalent to technology as the TFP term contains fixed parameters ν, ϕ and endogenous variables π, μ. In this section, I present three decompositions of measured TFP that illustrate the channels through which measured TFP and underlying technology differ over time following a permanent technology shock.

The first decomposition is simplest and most intuitive. It shows that measured TFP differs from underlying technology because firm-output variation (intensive margin) interacts with scale economies. The second and third decompositions go further in analyzing the components of intensive margin variation.

In all three graphical illustrations the solid, black, horizontal line represents a permanent technology shock and the solid, blue, downward-sloping line represents measured TFP given the permanent technology shift. These two lines are equivalent to eachother in all three figures. The dashed lines represent the components of measured TFP which differ across each figure. The dashed lines sum to give the measured TFP line.

4.3.1 Measured TFP Intensive Margin Decomposition

The most intuitive expression for TFP follows from linearizing TFP expressed as TFP $= y/(y+\phi)^{1/\nu}$ which yields

$$\hat{\text{TFP}}(t) = \frac{1}{\nu}\hat{A}(t) + \left(1 - \frac{1}{\tilde{\mu}}\right)\hat{y}(t). \tag{27}$$

The result shows that variations in TFP are due to underlying technology and a second term encompassing the interaction of firm output variations with the markup, which

[29] Under perfect competition ($\nu < 1$ and $\mu = 1$) steady-state firm output is fixed at its minimum efficient scale (minimum AC) and TFP is maximized.

equals RTS in steady state.[30] If $\nu < 1$, TFP overestimates technology regardless of the second term. The second term is zero with either perfect competition ($\tilde{\mu} = 1$) or constant firm output ($\hat{y} = 0$). Under perfect competition, firms operate at their minimum average cost where there are unit markups and constant returns ($\tilde{\text{RTS}} = \tilde{\mu} = 1$) so output variations do not affect TFP.[31] Under constant firm output, scale economies are irrelevant as scale is fixed. Therefore, returns to scale and a markup do not necessarily cause TFP to incorrectly measure technology. For example, with monpolistic competition (constant markups) and instantaneous entry (zero profits) firm-level output is always fixed so the second term in the decomposition disappears. Furthermore, if $\nu = 1$ then TFP and technology are equivalent.

Figure 2 decomposes measured TFP transition into technology and intensive margin components following a permanent technology shock. The horizontal dashed line shows that $1/\nu$ amplifies the technology shock directly. The decreasing dashed line shows the adjustment in firm-level output.

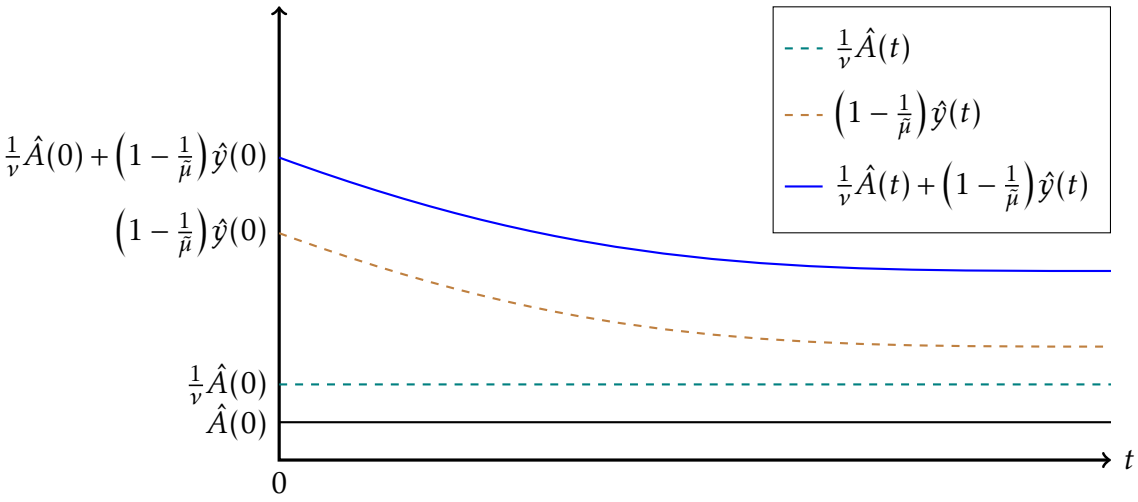


Figure 2: Measured TFP Decomposition, Firm Output

The variance and autocovariance of measured TFP fluctuations are

$$\text{var}(\hat{\text{TFP}}_t) = \left(\frac{1}{\nu}\right)^2 \text{var}(\hat{A}_t) + \left(1-\frac{1}{\tilde{\mu}}\right)^2 \text{var}(\hat{y}_t) + 2\left(\frac{1}{\nu}\right)\left(1-\frac{1}{\tilde{\mu}}\right)\text{cov}(\hat{A}_t, \hat{y}_t)$$

$$\text{cov}\left(\hat{\text{TFP}}_t, \hat{\text{TFP}}_{t-1}\right) =$$

$$\left(\frac{1}{\nu}\right)^2 \text{cov}\left(\hat{A}_t, \hat{A}_{t-1}\right) + \left(\frac{1}{\nu}\right)\left(1-\frac{1}{\tilde{\mu}}\right)\left[\text{cov}\left(\hat{A}_t, \hat{y}_{t-1}\right) + \text{cov}\left(\hat{y}_t, \hat{A}_{t-1}\right)\right] + \left(1-\frac{1}{\tilde{\mu}}\right)^2 \text{cov}\left(\hat{y}_t, \hat{y}_{t-1}\right).$$

Given firm-level output and technology co-vary positively contemporaneously, the variance of $\hat{\text{TFP}}$ is greater than the variance of technology $\hat{A}$. If firm-level output and

[30] The intensive margin coefficient is the price-cost margin (Lerner Index).

[31] For perfect competition to exist $\nu < 1$ is necessary, so the $\hat{A}$ coefficient continues to amplify technology fluctuations.

technology co-vary positively with a lead and lag, the autocovariance of measured TFP exceeds the autocovariance of technology.

Special Cases: Perfect Competition and Monopolistic Competition

The cases of perfect competition ($\nu < 1$ and $\mu = 1$) and monopolistic competition ($\nu \leq 1$ and $\mu > 1$) with zero profits (instantaneous entry) give the same variance and autocovariance for measured TFP and technology:

$$\text{var}(\hat{\text{TFP}}) = \left(\frac{1}{\nu}\right)^2 \text{var}(\hat{A}) \quad \text{and} \quad \text{cov}\left(\hat{\text{TFP}}_t, \hat{\text{TFP}}_{t-1}\right) = \left(\frac{1}{\nu}\right)^2 \text{cov}\left(\hat{A}_t, \hat{A}_{t-1}\right).$$

Put another way, monopolistic competition with instantaneous entry has the same dynamics as perfect competition (at the first order), and there is no distortion between technology and TFP with $\nu = 1$. In the case of perfect competition, $\tilde{\mu} = 1$ so the output variation coefficient is zero. In the case of monopolistic competition with instantaneous free entry there is no variation in output $\hat{y} = 0$ to interact with the positive scale economies.[32] Therefore, in both cases the second term in the linearized TFP expression is zero. This is the term that captures the interaction of intensive margin variations and scale economies. The variance and covariance results imply that the autocorrelation of measured TFP and technology are equal regardless of ν.[33] Since the coefficient of an AR(1) model gives the autocorrelation, our numerical results show the invariance of autocorrelation to different values of ν.

4.3.2 Measured TFP Profit-Markup Decomposition

Given the first basic decomposition into technology and firm output variation, we can now investigate further the drivers of firm output variation in the next two decompositions. Linearizing the reduced-form profit expression shows that output variations occur due to markup and profit variations:[34]

$$\hat{y}(t) = \tilde{\mu}\left(\frac{1}{\nu\phi}\hat{\pi}(t) - \frac{1}{\tilde{\mu} - \nu}\hat{\mu}(t)\right).$$

[32] Fixed output occurs because $\hat{\mu} = 0$ and $\hat{\pi} = 0$. We could represent thee variance and autocovarance in $\hat{A}_t, \hat{\mu}_t, \hat{\pi}_t$ terms rather than $\hat{A}_t, \hat{y}_t$. The cost is many more covariates.

[33] To see this plug the variance and covariance of $\hat{\text{TFP}}$ into the autocorrelation formula: $\text{cov}(\hat{\text{TFP}}_t, \hat{\text{TFP}}_{t-1}) / \sqrt{\text{var}(\hat{\text{TFP}}_t)\,\text{var}(\hat{\text{TFP}}_{t-1})}$.

[34] Since profits are zero in steady state $\tilde{\pi} = 0$, our notation for linearized profits is $\hat{\pi} = \dot{\pi}$, rather than $\hat{y} = \dot{y}/\tilde{y}$ and $\hat{\mu} = \dot{\mu}/\tilde{\mu}$ (and all other hat variables).

From this we can see that endogenous variation in markups and profit raise the variance of output. Their covariance further reinforces the effect.[35]

$$\text{var}(\hat{y}) = \left(\frac{\tilde{\mu}}{\nu\phi}\right)^2 \text{var}(\hat{\pi}) + \left(1 - \frac{\nu}{\tilde{\mu}}\right)^{-2} \text{var}(\hat{\mu}) - \frac{\tilde{\mu}}{\nu\phi}\left(1 - \frac{\nu}{\tilde{\mu}}\right)^{-1} \text{cov}(\hat{\pi}, \hat{\mu}).$$

Using our output expression in the basic decomposition, we can express TFP fluctuations as consisting of technology, markup and profit fluctuations:

$$\hat{\text{TFP}}(t) = \frac{1}{\nu}\hat{A}(t) - \left(\frac{\tilde{\mu} - 1}{\tilde{\mu} - \nu}\right)\hat{\mu}(t) + \left(\frac{\tilde{\mu} - 1}{\nu\phi}\right)\hat{\pi}(t) \tag{28}$$

Figure 3 decomposes measured TFP transition into technology, markup and profit components following a permanent technology shock. The horizontal dashed line shows that $1/\nu$ amplifies the technology shock directly. The decreasing dashed line shows an instantaneous increase in profits that decreases to zero in the long run. The increasing dashed line shows the markup effect that begins at zero but increases as markups decrease.

Dynamic firm entry causes the profit effect in the short run combined with oligopolistic competition causes the delay in the markup effect in the long run. The positive technology shock causes a gradual increase in the number of firms to a permanently higher steady-state level. In the short run, the absence of entry means profits increase, free from arbitrage, and there is no markup effect because markups only respond to entry. In the long run, the entry of firms to arbitrage profits removes the profit effect and causes a long-run markup effect as markups are permanently lower. If firm adjustment is fast, the profit effect disappears quickly and the markup effect quickly dominates.

[35] The covariance of profits and markups is negative because as profits jump postively on impact markups do not deviate from trend, but as firms enter to bring profits back to trend, the markup deviates negatively away from trend.

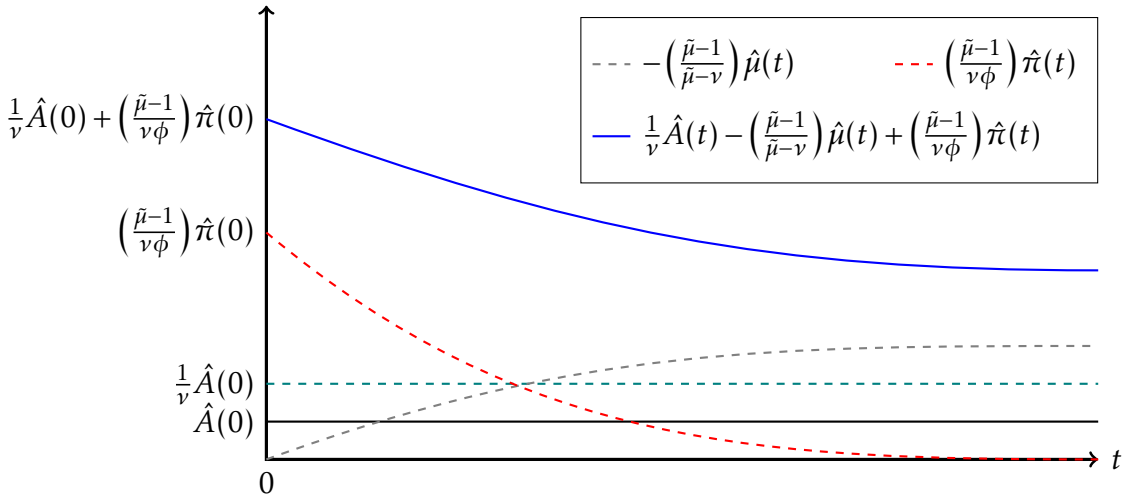


Figure 3: Measured TFP Decomposition, Markups and Profits

Special Cases: The profit-markup decomposition is useful to pinpoint the two novel mechanisms in the model. If there were instantaneous entry and therefore zero profits, firm size is fixed each period as the number of firms is determined instantaneously. In this case the measured TFP curve is horizontal at its current long-run asymptote, and consists of the direct technology component and markup component which is now flat rather than rising. There is no profit component. Similarly, if we also assume markups are fixed, then the markups component disappears too and measured TFP is equal to the existing dashed horizontal line. Lastly, if there were constant markups but dynamic entry (non-zero profits), the measured TFP curve would be composed of the downward-sloping profit effect and direct teachnology effect, which it would converge on in the long-run as profits disappeared.

4.3.3 Measured TFP Factor-Input Decomposition

If we log-linearize firm output directly, we can see it varies through technology, capital, labour and number of firms:

$$\hat{y}(t) = \tilde{\mu}\left(\frac{\hat{A}}{\nu} - \hat{N}(t) + \tilde{s}_K\hat{K}(t) + \tilde{s}_L\hat{L}(t)\right).$$

Hence, TFP fluctuations consist of four components:

$$\hat{\text{TFP}}(t) = \frac{\tilde{\mu}}{\nu}\hat{A}(t) + (\tilde{\mu}-1)\left(\tilde{s}_K\hat{K}(t) + \tilde{s}_L\hat{L}(t) - \hat{N}(t)\right).$$

Figure 4 emphasizes the important role of labour increasing on impact. This drives the short-run profit response in the second decomposition and the short-run output response in the first decomposition. On impact of a technology shock, the state variables are fixed $\hat{K}(0) = \hat{N}(0) = 0$ but labour $\hat{L}(0)$ jumps. Equation (24) shows that the

labour jump is driven by a the shock to $\hat{A}$ and a movement of $\hat{C}$. Labour increases on impact if the direct technology effect offsets the negative effect of a rise in consumption.[36]

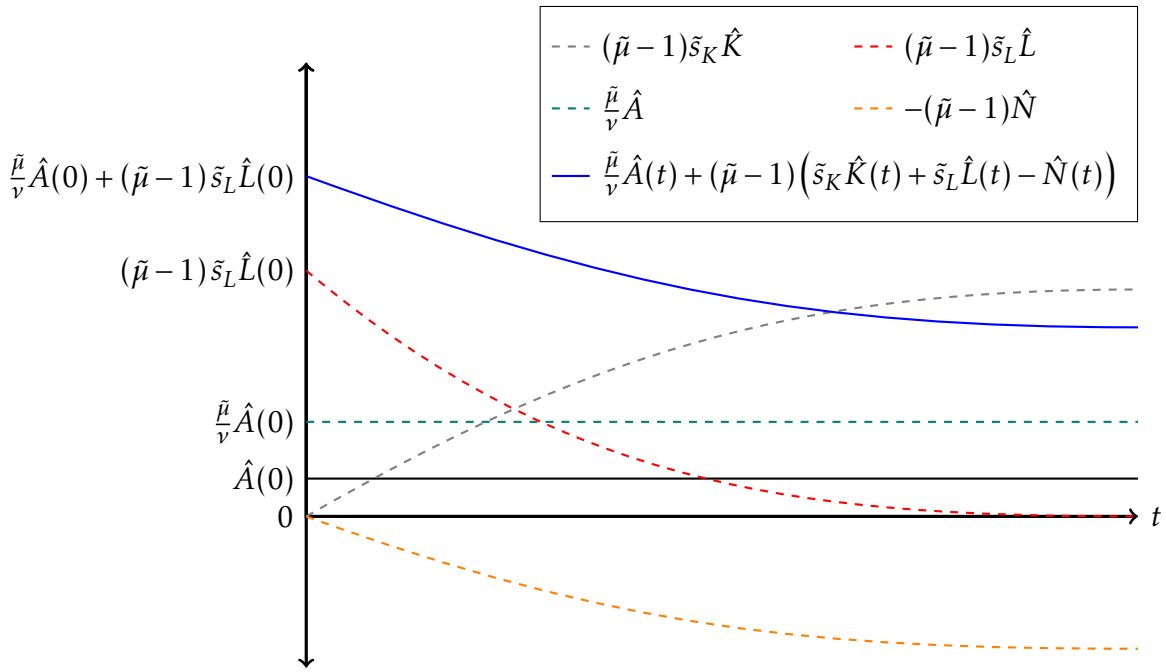


Figure 4: Measured TFP Decomposition, Factor Inputs

4.4 Identifying Technology

We can derive a model-consistent technology series by log-linearizing the aggregate production function around steady-state and then re-arranging. Log-linearized aggregate output is

$$\hat{Y} = \frac{\tilde{\mu}}{\nu}\hat{A} + \tilde{\mu}\tilde{s}_K\hat{K} + \tilde{\mu}\tilde{s}_L\hat{L} + (1-\tilde{\mu})\hat{N}.$$

We can remove the capital share using the zero-profit steady-state result $\tilde{s}_K + \tilde{s}_L = 1$.[37] Then rearranging for technology gives

$$\hat{A} = \frac{\nu}{\tilde{\mu}}\left[\hat{Y} - \tilde{\mu}\tilde{s}_L(\hat{L}-\hat{K}) - \tilde{\mu}\hat{K} - (1-\tilde{\mu})\hat{N}\right].$$

This derivation of technology is the correct measure for other models with imperfect competition and non-constant marginal cost.[38] The measure of technology is equiva-

[36] The illustration shows $\hat{L}$ converging to zero. This is a special case that occurs with log consumption utility such that $\varepsilon_{u_C C} = -1$ in (24). Income and substitution effects cancel out such that a change in technology has no effect on long-run labour.

[37] Not removing $\tilde{s}_K$ gives $\hat{A} = \frac{\nu}{\tilde{\mu}}\left[\hat{Y} - \tilde{\mu}\tilde{s}_L\hat{L} - \tilde{\mu}\tilde{s}_k\hat{K} - (1-\tilde{\mu})\hat{N}\right]$.

[38] It is invariant to the presence of dynamic entry or endogenous markups. Though the endogenous markup means the markup term μ is evaluated at its steady-state value which depends on the steady-state number of firms, rather than being a constant parameter. Also, the presence of a fixed costs ϕ does

lent to measured TFP (Solow Residual) when $\nu = \tilde{\mu} = 1$: $\hat{\text{TFP}} = \hat{Y} - \tilde{s}_L\hat{L} - \tilde{s}_K\hat{K}$. However, this case is not permissible in our model due to the fixed cost term.[39]

4.4.1 Data and Calibration for Technology

To acquire our technology series we need detrended time series for capital $\hat{K}$, labour $\hat{L}$ and number of firms $\hat{N}$. We also need to assign values to the steady-state labour share $\tilde{s}_L$, steady-state markup $\tilde{\mu}$, and the degree of returns to variable inputs ν.[40]

Technology Series Data

I use data for the US from 1992 Q3 - 2018 Q4 at a quarterly frequency. For aggregate output, I use real, seasonally-adjusted, GDP in 2012 prices from the Bureau of Economic Analysis (BEA). For labour, I use quarterly hours of all persons in the nonfarm business sector from the Bureau of Labour Statistic (BLS). For capital, I use annual capital services for the private non-farm business sector (excluding government enterprises) from the BLS. I linearly interpolate capital services to get quarterly data. For number of firms, I use quarterly number of establishments from the BLS Quarterly Census of Employment and Wages (QCEW). I remove population trends by converting data to per capita terms using BLS population level data.

The empirical counterpart to the continuous-time, log-linearized, variables in deviation from steady state $\hat{X} = \dot{X}/\tilde{X}$ is the logged data less the trend in the logged data, all scaled by 100 to give percentage units. The resulting series $\hat{X} = [\ln X_t - p(\ln X_t)] \times 100$, where $p(\cdot)$ is the trend, are approximate percentage deviations from trend.

Technology Series Calibrated Parameters

I assume the labour share is $\tilde{s}_L = 0.7$ which is a common approximation for the US labour share (Jaimovich and Floetotto 2008). The parameter ν is more unusual. It represents the slope of the marginal cost curve or the sum of output elasticities (net of the fixed cost) $\nu = \varepsilon_{FK} + \varepsilon_{F\ell}$. I follow Atkeson and Kehoe (2005) who have a similar model setup. They use $\nu = 0.95$ based on surveying production function estimation literature. This calibration is adopted by Restuccia and Rogerson (2008) and a large literature that follows them. The estimate is close to recent estimates by Ruzic and Ho (2019).[41] The markup $\mu(\tilde{N})$ is not a fixed parameter. It is determined endogenously

not affect the technology expression explicitly. The fixed cost is subsumed in the factor shares, which are shares in output Y which deducts fixed costs.

[39] The reduced-form equilibrium condition $\nu(1 + \tilde{s}_\phi) = \tilde{\mu}(1 - \tilde{s}_\pi)$ cannot hold with zero profits, unit markup and flat marginal cost, unless fixed costs are zero.

[40] Alternatively, we could use $\frac{\tilde{\mu}}{\nu} = 1 + \tilde{s}_\phi$ where $\tilde{s}_\phi = \phi/\tilde{y}$, if we preferred to specify the steady-state fixed-cost share. In either case, we must specify two out of the three terms: $\tilde{s}_\phi, \tilde{\mu}, \nu$.

[41] Atkeson and Kehoe (2005) specify $\nu/\mu = 0.85$ implying a markup of $\mu = 1.11$. The 0.85 ratio is adopted by Restuccia and Rogerson (2008) and subsequent papers. In these papers, there are no fixed

as a function of steady-state number of firms, which is a nonlinear function of model parameters, and it is directly affected by between-industry θ_I and between-firm θ_F product substitutability. I calibrate the markup to $\tilde{\mu} = 1.3$ for the technology process and I use this as a steady-state target in my simulations. This value is consistent with recent estimates for the US by Hall (2018) and is consistent with Jaimovich and Floetotto (2008). I study the effect of a high markup $\tilde{\mu} = 1.5$ and low markup $\tilde{\mu} = 1.1$ on the technology properties below.

4.4.2 Technology Series Properties

In Table 2 and Figure 5a, we show how different values of steady-state markup $\tilde{\mu}$ affect the unconditional variance of our detrended technology series. A higher value of $\tilde{\mu}$ raises the variance. The second column in Table 2 represents measured TFP (Solow Residual) since our adjusted TFP measure equals the Solow Residual when $\nu = \mu = 1$, and figure 5 benchmarks against this value.[42] The results are consistent with our decomposition that for a given variation in technology there is a larger variation in measured TFP. The second and third rows report the properties of the adjusted technology series from fitting the following AR(1) model to the derived series:

$$\hat{A}_t = \rho_{\hat{A}} \hat{A}_{t-1} + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \sigma_\epsilon^2\right).$$

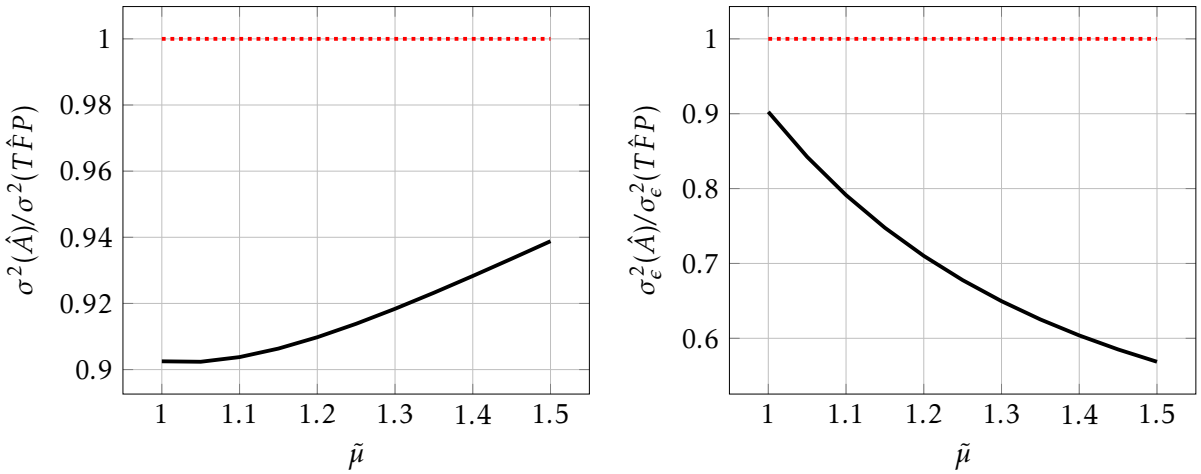
The results show that the variance of the AR(1) error term is decreasing in the markup.

ν	1	0.95			
$\tilde{\mu}$	1	1	1.1	1.3	1.5
$\sigma^2(\hat{A})$	2.5711	2.3204	2.3237	2.3612	2.4138
AR(1) Estimates					
$\rho_{\hat{A}}$	0.9755	0.9755	0.9793	0.9839	0.9863
σ_ϵ^2	0.2305	0.2080	0.1824	0.1497	0.1311

Table 2: Technology Series Properties (quadratic detrend and $s_L = 0.7$)

costs, so by $\nu/\mu = (1 - s_\pi)/(1 + s_\phi)$, the ratio implies 15% profit share, whereas in our work fixed costs offset the profits in steady state, and the ratio implies a fixed cost share of 18%. This is similar to direct estimates of the overhead cost share in De Loecker, Eeckhout, and Unger (2020). They measure the overhead cost share in total costs not total revenue, but in our setup in steady-state profits are zero so revenue and cost shares are equivalent. They measure overhead by an accounting cost called *S,G&A* (Selling, General & Administrative).

[42] Holding $\tilde{\mu} = 1$, we see that the decrease in the technology series variance as ν decreases from 1 to 0.95 is consistent with theory $\mathrm{var}(\hat{A}) = \nu^2 \mathrm{var}(\hat{SR})$, that is $2.3204 = 0.95^2 \times 2.5711$. We can also see the invariance of the autocorrelation term as ν changes from 1 to 0.95.



(a) Relative Unconditional Variance

(b) Relative Innovation Variance

Figure 5: Adjusted Technology Properties Relative to Measured TFP

5 Simulations

I simulate a discrete-time version of the model using the technology series we identified in the previous section.

5.1 Model Calibration

To acquire technology I assigned values to the labour share, markup and marginal cost slope $\tilde{s}_L, \tilde{\mu}, \nu$. The labour share and markup are endogenously determined in the model so we shall use them as targets when calibrating other parameters. I assume an isoelastic utility function and a Cobb-Douglas production function that is homogeneous of degree $\nu \equiv \alpha + \beta$. The α and β parameters are output elasticities to capital and labour:

$$U(C,L) = \frac{C^{1-\sigma}}{1-\sigma} - \xi \frac{L^{1+\eta}}{1+\eta} \quad \text{and} \quad F(k,\ell) = k^\alpha \ell^\beta.$$

The parameter η is the inverse Frisch elasticity which I set the same as Bilbiie, Ghironi, and Melitz (2012), Etro and Colciago (2010), and King and Rebelo (1999), all of whom use the same utility function. I also follow this literature for the discrete-time discount factor which corresponds to quarterly time periods. I set the between-sector and between-firm substitutability according to Jaimovich and Floetotto (2008), which is also close to Etro and Colciago (2010) and other papers with oligopolistic competition (Atkeson and Burstein 2008; Lewis and Poilly 2012).[43] I choose ξ

[43] Generally authors set the between-industry substitutability to be close to 1, implying perfectly differentiated industries, and the between-firm substitutability to be large, implying homogeneous firms. Hence there are homogeneous firms within unique industries.

such that steady-state labour is one-third $\tilde{L} = 1/3$. The more noteworthy parameters for our analysis are those that relate to scale economies α, β, ϕ and firm adjustment ζ. To derive a series for technology we calibrated the steady-state labour share $\tilde{s}_L = 0.7$, and marginal cost slope (sum of output elasticities) to $\nu = 0.95$. These figures imply values for the production function elasticities α and β in zero-profit steady-state. Specifically, the factor market equilibrium conditions imply that in steady-state $\beta = \tilde{\varepsilon}_{FL} = \tilde{s}_L \nu = 0.7 \times 0.95 = 0.665$ and $\alpha = \tilde{\varepsilon}_{FK} = (1 - \tilde{s}_L)\nu = 0.3 \times 0.95 = 0.285$.[44] The entry congestion parameter ζ controls the speed of adjustment of firms – it is not present in steady state – therefore I set it such that the simulated series of firms has the same persistence as a series of firms from US data.[45] I choose the fixed cost ϕ to give a steady-state markup of 1.3. The fixed cost control $\tilde{N}$ and in turn the markup. I set the persistence and innovation variance of the technology process according to the results in Table 2 when $\tilde{\mu} = 1.3$.

Capital Elast.	α	0.285
Labour Elast.	β	0.665
Fixed Cost	ϕ	0.070
Entry Congestion	ζ	1.500
Risk Aversion	σ	1.000
Discount Factor	ρ	0.990
Labour Weight	ξ	2.760
Labour Elast.	η	0.250
Industry (inter) Subs.	θ_I	1.001
Firm (intra) Subs.	θ_F	19.600
Tech. Persist.	$\rho_{\hat{A}}$	0.979
Tech. Innov. Var.	σ^2_ϵ	0.1497

Table 3: Parameter Values

5.2 Model Results

The impulse response functions from our simulated model reflect the dynamics of our VAR analysis. Following a one standard deviation shock to technology, there is an immediate strong response in aggregate output, but the adjustment of firms begins small and increases over time. Consequently there is a short-run jump in the intensive margin activity of incumbent firms $\hat{y}$ which interacts with returns to scale

[44] With $\nu = 1$ the output elasticities would equal to factor shares.

[45] The first-order autocorrelation of HP-filtered, quarterly, data on the number of firms in the US economy 1992Q3–2018Q4 is 0.956. This is calculated using BLS Business Employment Dynamics time series data on the number of establishments.

causing an additional endogenous productivity effect on top of the technology shock. The measured TFP response exceeds underlying technology by 50% on impact. The increase in incumbent firm size on impact means firms utilise their fixed costs and returns to scale. Consequently the fixed-cost share and returns to scale decrease on impact closer to constant returns to scale. Subsequently as the initial intensive margin expansion declines the fixed cost share and returns to scale measures return to their higher values.

As we showed in our qualitative decompositions, we can decompose the expansion in firm-level output into markup and profit variation or factor-input variation. Initially the firm-level output expansion corresponds to a proportional jump in operating profits because firm entry is yet to adjust, so all the expansion is accrued to incumbents as profits. But over time firms enter stealing profits and decreasing markups. Notably the profit effect disappears by 20 quarters which coincides with the markup effect reaching its maximum. The markup decreases slowly in response to competition from new entry. This buoys firm-level output leading to the persistent endogenous productivity effect. The fall in markups causes the resilience in incumbents' output because they must produce more in order to cover their fixed costs of production, given they now have a smaller markup on each unit they sell.

Entry creates opposing effects on the firm intensive margin and consequently the endogenous productivity effect. On the one hand, entry leads to *business stealing* which diminishes operating profits, output per firm and consequently productivity. On the other hand, entry leads to a competition effect which decreases markups, raising output per firm and strengthening the endogenous productivity effect.

The final three panels (middle-right and bottom two) interpret the output expansions in terms of factor inputs. The middle-right demonstrates that on impact – when there is no firm entry – expansion in the intensive margin and expansion in aggregate output are equivalent: all aggregate output expansion is due to incumbents raising their output. Subsequently as firms enter the role of the firm-level intensive margin expansion diminishes and the differences between the $\hat{Y}$ and $\hat{y}$ curves shows the growing importance of new entrants' output contribution to the aggregate. The same logic applies to labour, which is the only factor of production that is able to adjust on impact and therefore on impact causes all the output expansion which is not caused directly by the $\hat{A}$ shock. Then as entry adjusts the emerging gap between $\hat{\ell}$ and $\hat{L}$ represents the growing importance of new firms in the composition of aggregate hours. Rather than working more hours at an incumbent firm as they do initially, over time workers setup new firms and allocate hours to these.

The bottom two panels show the components of aggregate output and firm-level (intensive margin) output (excluding $\hat{A}$ component). They emphasize that all of the short-run effect occurs through hours adjusting since both capital and firms are fixed

state variables on impact. This also explains the equal response of aggregate output and firm level output on impact.

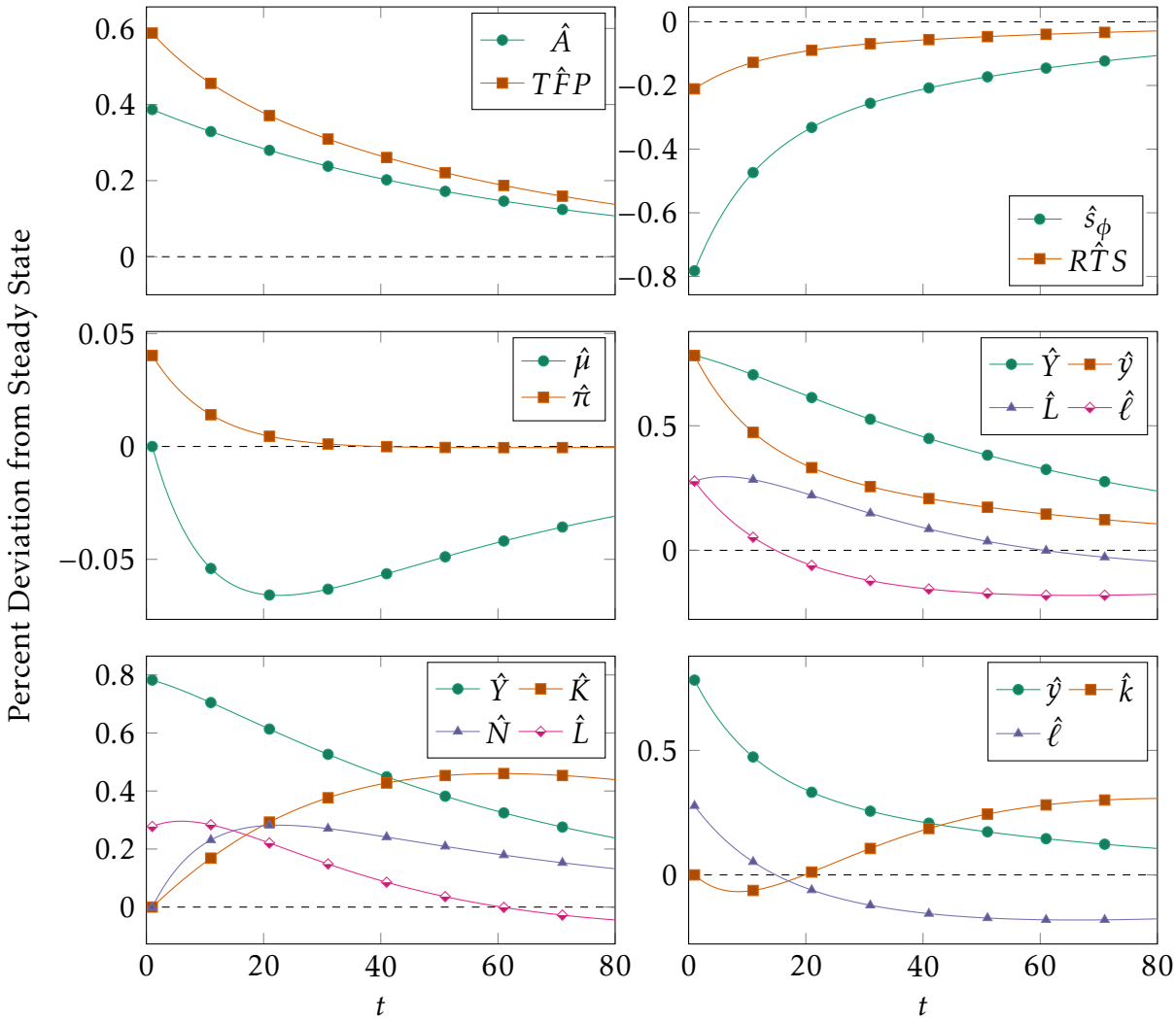


Figure 6: IRF Dynamic Entry Oligopolistic Competition

5.3 Model Comparison

I compare four versions of the model that isolate each key model mechanism in turn. The full model has oligopolistic competition and dynamic entry. I compare this to the model with oligopolistic competition and instantaneous entry and the model with monopolistic competition and dynamic entry. The benchmark model has neither model feature: it has monopolistic competition and instantaneous entry. As I have explained, the dynamics of a benchmark monopolistic competition model are the same as an RBC model. The only distortion in such models are static and occur in steady state. Therefore the IRFs for the static entry monopolistic competition model are identical to the perfect competition RBC model. The models are all simulated with the same cali-

bration of persistence and variance for the technology process. The technology series we have identified is correct for all models regardless of whether there is dynamic or static entry or monopolistic or oligopolistic competition. The calibration from Table 3 is the same for all models, except in the monopolistic competition case the markup is constant $\mu = \theta_I/(\theta_I - 1)$ and I calibrate it to equal 1.3 by setting $\theta_I = 4.33$. This is the endogenously determined level of the markup with oligopolistic competition. With static entry there are no entry adjustment costs therefore $\zeta = 0$. The steady-state outcomes are identical across all four models. This is because we set the markup to be the same and the dynamic entry feature has not effect on steady state.

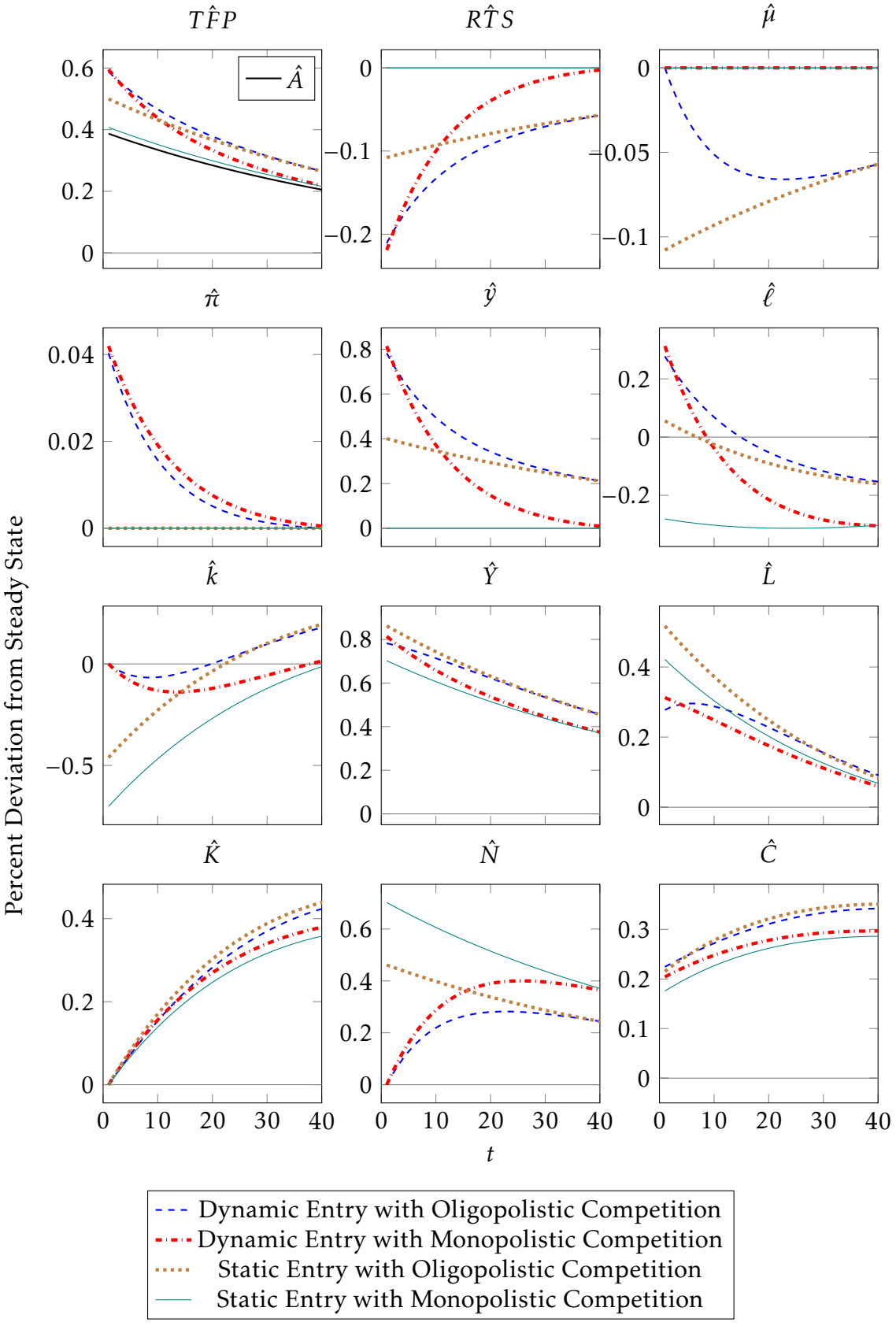


Figure 7: IRF Model Comparison

The TF̂P and $\hat{y}$ panels show that the addition of dynamic entry to the baseline

model causes a more pronounced short-run effect than the addition of oligopolistic competition. After approximately 10 quarters, the oligopolistic competition effect exceeds the dynamic entry effect leading to a less volatile but more persistent effect on $\hat{y}$. This can be seen by comparing the IRFs for the two intermediate cases: dynamic entry with monopolistic competition and static entry with oligopolistic competition.

	Static Entry		Dynamic Entry	
	Monop.	Oligop.	Monop.	Oligop.
$\sigma(Y)$	0.9394	1.1602	1.0870	1.0590
σ(TFP)	0.5300	0.6545	0.7767	0.7685

Table 4: Model Moments Comparison

Table 4 shows that dynamic entry causes a large increase in measured TFP volatility relative to model variants with static entry, but once dynamic entry is assumed the form of strategic interaction plays little role. The increase in the volatility of measured TFP relative to the benchmark static entry monopolistic competition model is 45%. In terms of aggregate output fluctuations, the full model generates 13% more variation in output than the benchmark model. However, the greatest amplification of aggregate output occurs in the model with static entry and oligopolistic competition. This is consistent with the IRFs which show a strong response of aggregate labour and in turn aggregate output for this variant of the model.

6 Conclusion

The paper investigates the effect of firm entry on measured productivity over the business cycle. I consider that entry is non-instantaneous leading to temporary profits and entry affects the price markups that incumbents charge. Together with increasing returns to scale, these mechanisms can explain short-run procylical productivity and long-run persistence following a technology shock. The theory explains that productivity is exacerbated on impact, since firms cannot adjust immediately so incumbents gain profits and expand output, and in the long run underlying productivity persists as firm entry strengthens competition.

References

Adjemian, Stéphane, Houtan Bastani, Michel Juillard, Fréderic Karamé, Junior Maih, Ferhat Mihoubi, George Perendia, Johannes Pfeifer, Marco Ratto, and Sébastien Villemot (2011). *Dynare: Reference Manual Version 4*. Dynare Working Papers 1. CEPREMAP.

Aloi, Marta, Huw Dixon, and Anthony Savagar (2021). "Labor Responses, Regulation, and Business Churn". In: *Journal of Money, Credit and Banking* 53.1, pp. 119–156. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jmcb.12694.

Ambler, Steve and Emanuela Cardia (1998). "The Cyclical Behaviour of Wages and Profits under Imperfect Competition". In: *Canadian Journal of Economics* 31.1, pp. 148–164.

Atkeson, Andrew and Ariel Burstein (2008). "Pricing-to-market, trade costs, and international relative prices". In: *The American Economic Review* 98.5, pp. 1998–2031.

Atkeson, Andrew and Patrick J Kehoe (2005). "Modeling and measuring organization capital". In: *Journal of Political Economy* 113.5, pp. 1026–1053.

Baqaee, David Rezza and Emmanuel Farhi (2020). "Productivity and misallocation in general equilibrium". In: *The Quarterly Journal of Economics* 135.1, pp. 105–163.

Barkai, Simcha (2020). "Declining Labor and Capital Shares". In: *The Journal of Finance* 75.5, pp. 2421–2463.

Basu, Susanto (1996). "Procyclical productivity: increasing returns or cyclical utilization?" In: *The Quarterly Journal of Economics* 111.3, pp. 719–751.

— (2019). "Are price-cost markups rising in the united states? a discussion of the evidence". In: *Journal of Economic Perspectives* 33.3, pp. 3–22.

Basu, Susanto and John Fernald (2001). "Why Is Productivity Procyclical? Why Do We Care?" In: *New Developments in Productivity Analysis*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 225–302.

Bénassy, Jean-Pascal (1996). "Monopolistic competition, increasing returns to specialization and output persistence". In: *Economics Letters* 52.2, pp. 187–191.

Berentsen, Aleksander and Christopher Waller (2015). "Optimal Stabilization Policy with Search Externalities". In: *Macroeconomic Dynamics* 19 (03), pp. 669–700.

Bergin, Paul R., Ling Feng, and Ching-Yi Lin (2016). "Firm entry and financial shocks". In: *The Economic Journal*.

Bergin, Paul R. and Ching-Yi Lin (2012). "The dynamic effects of a currency union on trade". In: *Journal of International Economics* 87.2, pp. 191–204.

Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz (2012). "Endogenous Entry, Product Variety, and Business Cycles". In: *Journal of Political Economy* 120.2, pp. 304–345.

Boar, Corina and Virgiliu Midrigan (2020). *Efficient Redistribution*. Working Paper 27622. National Bureau of Economic Research.

Brito, Paulo and Huw Dixon (2013). "Fiscal policy, entry and capital accumulation: Hump-shaped responses". In: *Journal of Economic Dynamics and Control* 37.10, pp. 2123–2155.

Chatterjee, Satyajit and Russell Cooper (2014). "Entry And Exit, Product Variety, And The Business Cycle". In: *Economic Inquiry* 52.4, pp. 1466–1484.

Comin, Diego and Mark Gertler (2006). "Medium-term business cycles". In: *American Economic Review* 96.3, pp. 523–551.

Cook, David (2001). "Time to enter and business cycles". In: *Journal of Economic Dynamics and Control* 25.8, pp. 1241–1261.

Das, Sanghamitra and Satya P Das (1997). "Dynamics of entry and exit of firms in the presence of entry adjustment costs". In: *International Journal of Industrial Organization* 15.2, pp. 217–241.

Datta, Bipasa and Huw Dixon (2002). "Technological Change, Entry, and Stock-Market Dynamics: An Analysis of Transition in a Monopolistic Industry". In: *American Economic Review P&P* 92.2, pp. 231–235.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (2020). "The rise of market power and the macroeconomic implications". In: *The Quarterly Journal of Economics* 135.2, pp. 561–644.

Decker, Ryan A, John Haltiwanger, Ron S Jarmin, and Javier Miranda (2018). *Changing Business Dynamism and Productivity: Shocks vs. Responsiveness*. Working Paper 24236. National Bureau of Economic Research.

Devereux, Michael B, Allen C Head, and Beverly J Lapham (1996). "Aggregate fluctuations with increasing returns to specialization and scale". In: *Journal of economic dynamics and control* 20.4, pp. 627–656.

Dixit, Avinash K. and Joseph E. Stiglitz (1977). "Monopolistic Competition and Optimum Product Diversity". English. In: *The American Economic Review* 67.3, pp. 297–308.

Dos Santos Ferreira, Rodolphe and Frédéric Dufourt (2006). "Free entry and business cycles under the influence of animal spirits". In: *Journal of Monetary Economics* 53.2, pp. 311–328.

Ethier, Wilfred J (1982). "National and international returns to scale in the modern theory of international trade". In: *The American Economic Review* 72.3, pp. 389–405.

Etro, Federico and Andrea Colciago (2010). "Endogenous Market Structures and the Business Cycle*". In: *The Economic Journal* 120.549, pp. 1201–1233.

Feenstra, Robert C (2003). "A homothetic utility function for monopolistic competition models, without constant price elasticity". In: *Economics Letters* 78.1, pp. 79–86.

Gutiérrez, Germán, Callum Jones, and Thomas Philippon (2019). *Entry costs and the macroeconomy*. Tech. rep. National Bureau of Economic Research.

Hall, Robert E (1989). *Invariance properties of Solow's productivity residual*. Tech. rep. National Bureau of Economic Research.

— (2018). *New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy*. Working Paper 24574. National Bureau of Economic Research.

Hopenhayn, Hugo A. (1992). "Entry, Exit, and Firm Dynamics in Long Run Equilibrium". In: *Econometrica* 60.5, pp. 1127–1150.

Hornstein, Andreas (1993). "Monopolistic competition, increasing returns to scale, and the importance of productivity shocks". In: *Journal of Monetary Economics* 31.3, pp. 299–316.

Jaimovich, Nir and Max Floetotto (2008). "Firm dynamics, markup variations, and the business cycle". In: *Journal of Monetary Economics* 55.7, pp. 1238–1252.

Kim, Jinill (2004). "What determines aggregate returns to scale?" In: *Journal of Economic Dynamics and Control* 28.8, pp. 1577–1594.

King, Robert G. and Sergio T. Rebelo (1999). "Resuscitating real business cycles". In: *Handbook of Macroeconomics*. Ed. by J. B. Taylor and M. Woodford. Vol. 1. Handbook of Macroeconomics. Elsevier. Chap. 14, pp. 927–1007.

Lewis, Vivien (2009). "Business Cycle Evidence On Firm Entry". In: *Macroeconomic Dynamics* 13 (5), pp. 605–624.

Lewis, Vivien and Céline Poilly (2012). "Firm entry, markups and the monetary transmission mechanism". In: *Journal of Monetary Economics* 59.7, pp. 670–685.

Loualiche, Erik (2019). "Asset Pricing with Entry and Imperfect Competition". In: *Working Paper*.

Mankiw, N Gregory and Michael D Whinston (1986). "Free entry and social inefficiency". In: *The RAND Journal of Economics*, pp. 48–58.

Pfaff, Bernhard (2008). "VAR, SVAR and SVEC Models: Implementation Within R Package vars". In: *Journal of Statistical Software* 27.4.

Portier, Franck (1995). "Business formation and cyclical markups in the French business cycle". In: *Annales d'Economie et de Statistique*, pp. 411–440.

Poutineau, Jean-Christophe and Gauthier Vermandel (2015). "Financial frictions and the extensive margin of activity". In: *Research in Economics* 69.4, pp. 525–554.

Restuccia, Diego and Richard Rogerson (2008). "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments". In: *Review of Economic Dynamics* 11.4, pp. 707–720.

Rotemberg, Julio and Michael Woodford (1993). *Dynamic General Equilibrium Models with Imperfectly Competitive Product Markets*. Working Paper 4502. National Bureau of Economic Research.

Ruzic, Dimitrije and Sui-Jade Ho (2019). "Returns to Scale, Productivity Measurement, and Trends in U.S. Manufacturing Misallocation". In: *Working Paper*.

Savagar, Anthony and Huw Dixon (2020). "Firm entry, excess capacity and endogenous productivity". In: *European Economic Review* 121.

Syverson, Chad (2019). "Macroeconomics and market power: Context, implications, and open questions". In: *Journal of Economic Perspectives* 33.3, pp. 23–43.

Tian, Can (2018). "Firm-level entry and exit dynamics over the business cycles". In: *European Economic Review* 102, pp. 298–326.

Yang, Xiaokai and Ben J Heijdra (1993). "Monopolistic Competition and Optimum Product Diversity: Comment". In: *American Economic Review* 83.1, pp. 295–301.

A Output Elasticities and Factor Shares

Take factor market equilibrium $w = \frac{AN^{1-\nu}}{\mu} F_L$ and multiply by $\frac{L}{AN^{1-\nu}F} = \frac{L}{Y+N\phi}$ to give $\frac{wL}{Y+N\phi} = \frac{\varepsilon_{FL}}{\mu}$, then noting our definition for factor shares and fixed-cost share, gives $s_L \frac{1}{1+s_\phi} = \frac{\varepsilon_{FL}}{\mu}$ then rearrange for production function elasticities. The process is symmetric for wage and rental markets. Hence

$$\varepsilon_{FL} = \frac{\mu}{(1+s_\phi)} s_L, \quad \varepsilon_{FK} = \frac{\mu}{(1+s_\phi)} s_K,$$

Additionally from the profit condition under firms optimizing choices

$$1 - s_\pi = \frac{\nu}{\mu}(1+s_\phi) \quad (= s_L + s_K)$$

Therefore

$$\varepsilon_{FL} = \frac{\nu}{(1-s_\pi)} s_L, \quad \varepsilon_{FK} = \frac{\nu}{(1-s_\pi)} s_L.$$

B Reduced-form Aggregate Production Function

We can obtain reduced-form aggregate output by removing the number of firms N from the aggregate production function using the maximized profit equation. The aggregate production function is

$$Y = N^{1-\nu} AF(K,L) - N\phi.$$

The aggregate budget constraint under optimal factor prices is:

$$N\pi = \left(1 - \frac{\nu}{\mu}\right)(Y + N\phi) - N\phi.$$

Therefore the number of firms is

$$N = \left(1 - \frac{\nu}{\mu}\right)\left(\pi + \frac{\nu}{\mu}\phi\right)^{-1} Y.$$

Substituting N into aggregate output and re-arranging for Y yields

$$Y = \left(\frac{A}{\pi+\phi}\right)^{\frac{1}{\nu}} \left(1 - \frac{\nu}{\mu}\right)^{\frac{1}{\nu}-1} \left(\pi + \frac{\nu}{\mu}\phi\right) F(K,L)^{\frac{1}{\nu}}$$

where we define *measured TFP* as $\text{TFP} \equiv \left(\frac{A}{\pi+\phi}\right)^{\frac{1}{\nu}} \left(1 - \frac{\nu}{\mu}\right)^{\frac{1}{\nu}-1} \left(\pi + \frac{\nu}{\mu}\phi\right)$.

C Identifying Technology

If we log-linearize the expression for aggregate output we get

$$\ln Y = \ln[AN^{1-\nu}F(K,L) - N\phi]$$

$$\dot{Y} = \frac{1}{Y}\left[\frac{Y+N\phi}{A}\dot{A} + \frac{Y+N\phi}{N}(1-\nu)\dot{N} + \frac{Y+N\phi}{F(K,L)}(F_K\dot{K} + F_L\dot{L}) - \phi\dot{N}\right]$$

$$\hat{Y} = (1+s_\phi)\hat{A} + (1+s_\phi)\varepsilon_{FK}\hat{K} + (1+s_\phi)\varepsilon_{FL}\hat{L} + (1-\nu(1+s_\phi))\hat{N}$$

where $s_\phi \equiv \frac{N\phi}{Y}$, $\varepsilon_{FK} = F_K\frac{K}{F}$ and $\varepsilon_{FL} = F_L\frac{L}{F}$. Hence replacing elasticities in terms of factor shares, log-linearized output can be written:

$$\hat{Y} = \frac{\mu(1-s_\pi)}{\nu}\hat{A} + \mu s_K\hat{K} + \mu s_L\hat{L} + (1-\mu(1-s_\pi))\hat{N}$$

where $s_\pi \equiv \frac{\pi N}{Y}$, $s_K \equiv \frac{rK}{Y}$ and $s_L \equiv \frac{wL}{Y}$. Evaluating coefficients at steady state where $\tilde{s}_\pi = 0$ gives

$$\hat{Y} = \frac{\tilde{\mu}}{\nu}\hat{A} + \tilde{\mu}\tilde{s}_K\hat{K} + \tilde{\mu}\tilde{s}_L\hat{L} + (1-\tilde{\mu})\hat{N}.$$

Furthermore, in steady state $\tilde{s}_L + \tilde{s}_K = 1$ which allows us to express in terms of labour shares

$$\hat{Y} = \frac{\tilde{\mu}}{\nu}\hat{A} + \tilde{\mu}\tilde{s}_L(\hat{L}-\hat{K}) + \tilde{\mu}\hat{K} + (1-\tilde{\mu})\hat{N}$$

Either of these expressions can be rearranged to provide a true technology series $\hat{A}$.

$$\hat{A} = \frac{\nu}{\tilde{\mu}}\left[\hat{Y} - \tilde{\mu}\tilde{s}_K\hat{K} - \tilde{\mu}\tilde{s}_L\hat{L} - (1-\tilde{\mu})\hat{N}\right]$$

$$\hat{A} = \frac{\nu}{\tilde{\mu}}\left[\hat{Y} - \tilde{\mu}\tilde{s}_L(\hat{L}-\hat{K}) - \tilde{\mu}\hat{K} - (1-\tilde{\mu})\hat{N}\right]$$