Delayed Adjustment and Persistence in Macroeconomic Models

Estimated impulse responses of investment and hiring typically peak well after the impact of a shock. Standard models with adjustment costs in capital and labor do not exhibit such delayed adjustment, but we argue that it arises naturally when we relax the assumption that the production technology is separable over time. This result, which holds for both convex and nonconvex cost functions, is strong enough to match the persistence observed in the data for reasonable parameter values. We discuss some evidence for our explanation and ways to test it.

JEL codes: E24, J61, J62
Keywords: persistence, adjustment costs, organizational capital

1. INTRODUCTION

A typical impulse response function for investment estimated from aggregate data to a technology or monetary policy shock (e.g., Altig et al. 2011) has a hump shape. Investment jumps up in response to the shock, and then continues to increase before gradually falling back to zero. Most macroeconomic models since the first contributions to real business cycle theory correctly predict the sign and size of this response (King and Rebelo 1999), but have trouble explaining why there is a lag before investment peaks after a shock.

We are grateful to Begoña Domínguez, Giuseppe Fiori, Isaac Gross, Jorge Miranda Pinto, Roberto Pancrazi, Ander Perez Orive, the editor, and an anonymous referee for helpful comments and discussions, and to Marie Le Mouël for sharing the data of their task-based measure of organizational capital investment.

MARIA VUKOTIĆ and THIJS VAN RENS are at Economics Department, University of Warwick, United Kingdom, CV4 7AL Coventry (E-mails: J.M.van-Rens@warwick.ac.uk, M.Vukotic@warwick.ac.uk).

Received September 29, 2020; and accepted in revised form June 29, 2022.

Journal of Money, Credit and Banking, Vol. 0, No. 0 (January 2023) © 2023 The Authors. Journal of Money, Credit and Banking published by Wiley Periodicals LLC on behalf of Ohio State University.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
A similar puzzle arises for investments in labor input. In frictional labor market models as in Diamond (1982), Mortensen (1982), and Pissarides (1985), employment is a state variable, in which firms may invest through costly hiring. Estimates show a clear hump shape not only in the response of employment (Christiano, Eichenbaum, and Evans 1999), but also in that of job finding rate (as a measure of hiring), to technology (Canova, Lopez-Salido, and Michelacci 2010, Canova, Lopez-Salido, and Michelacci 2013), and monetary policy shocks (White 2018).

In this paper, we propose a small and plausible modification to standard models that generates the type of hump-shaped impulse responses for investment and hiring observed in the data. We relax the assumption, implicit in almost all macroeconomic models, that the production technology is intertemporally separable. In combination with standard adjustment costs in capital and labor, a nonseparable production technology gives rise to delayed adjustment: the peak of hiring and investment takes place a while after a shock has hit the economy. We show that for reasonable parameter values, the delay in adjustment is long enough to match the persistence observed in the data, and that it arises for nonconvex as well as convex adjustment cost functions.

Modern theories of investment are microfounded versions of Lucas (1967)’s “flexible accelerator” model: investment is increasing in the distance between the actual and the desired stock of capital or labor. Depending on the specifics of the model, capital adjusts gradually (with convex adjustment costs) or instantaneously (with fixed adjustment costs or irreversible investments) to its target. While intuitively attractive, these models have the counterfactual implication that investment is highest immediately after a change in demand or productivity, when the capital stock is furthest away from its target. In reality, firms slowly increase their investments, with most investment happening as much as 18 months after a shock.

We are not the first to notice that macroeconomic models do not seem to match the persistence in macroeconomic aggregates. The lack of propagation in these models is a long standing puzzle (Cogley and Nason 1995; Rotemberg and Woodford 1996), although the literature seems to have focused more on the lack of amplification, perhaps because adjustment costs provide a straightforward way to increase the autocorrelation in the model. As opposed to the early contributions on propagation, we draw a sharp distinction between the persistence in stock and flow variables, arguing that adjustment costs may explain persistence and hump-shaped responses in the stocks (capital and employment), but they cannot by themselves account for persistence in the flows (investment and hiring).

Many researchers are well aware of the “persistence puzzle” (Christiano, Eichenbaum, and Trabandt 2018), and often resort to cost-of-change adjustment costs, as in Christiano, Eichenbaum, and Evans (2005). Groth and Khan (2010)
estimate cost-of-change adjustment costs using U.S. industry-level data and show that industry-level costs are small compared to the level of costs used in the literature, even after taking aggregation bias into account. We show that the dynamics of our model with a nonseparable production function are very similar to the dynamics of models with cost-of-change adjustment costs, even though we use standard adjustment costs in the levels of investment and hiring. Previous studies have shown that hump-shaped responses of investment to shocks can also be generated in a model with sectoral heterogeneity (Fiori 2012), or with a type of time-to-build assumption where firms invest in complementary projects of uncertain duration (Lucca 2007). We will provide some suggestive evidence in favor of our mechanism, but are not able to test it against these alternative explanations, which may be at work as well.

We model nonseparabilities in the production technology by introducing an additional state variable, which we label organizational capital, which acts as a storage technology for capital and labor input. This is the simplest way to relax the extreme assumptions that all current capital and labor input immediately contributes to production, and that current capital and labor are the only inputs in production. Organizational capital is the accumulation of organizational investment, infrequent activities that are crucial to the firm in the long run, but do not immediately benefit production in the short run. The infrequent nature of these activities generates a margin of adjustment for production. When faced with higher demand or productivity, firms can temporarily expand production without investing in more capital or hiring more workers. Eventually, further depleting the stock of organizational capital becomes costly, and investment and hiring increase slowly, as they do in the data.

A good example of an organizational investment from our own production technology as academics is giving a research seminar. Giving a seminar does not immediately contribute to the production of research papers. In fact, it takes time away from directly productive activities like analyzing data or writing text. However, the comments we receive from colleagues and potential referees at the seminar do affect the quality of our paper, and may influence the direction of our work for many months or even years afterward. More generic examples of organizational investments are machine maintenance, employee training, and staff meetings to coordinate team work.

The most direct evidence for the mechanism we have in mind comes from a, now somewhat dated, survey of plant managers by Fay and Medoff (1985). In this survey, managers recalled how many workers they let go in the last recession, and were then asked how many they could have let go while still meeting demand. The difference, which, on average, amounted to 6% less workers fired than would have been feasible, was interpreted as labor hoarding. More importantly for our purposes, a follow-up question about what happened to the “hoarded” workers revealed that half of the 6% were assigned to “other work,” including (in order of importance): cleaning, painting, maintenance of equipment, equipment overhaul, and training, all of which are examples of what we would call investments in organizational capital. Unfortunately, the Fay and Medoff (1985) survey provides only a snapshot. Therefore, we also consider capacity utilization as a proxy measure for lack of organizational investments, and show that the dynamics of capacity utilization in the data (Fernald 2012) are
consistent with the predictions of our model, even though we do not target this variable in the calibration.

The interpretation of nonseparabilities in production as organizational investment relates this paper to the literature on organizational and intangible capital. A number of papers show that organizational capital and other intangible assets are important part of the productivity and stock market value of firms (Prescott and Visscher 1980, Blanchard and Kremer 1997, Brynjolfsson and Hitt 2000, Lev, Radhakrishnan, and Zhang 2009, Hall 2000b, McGrattan and Prescott 2012, Conesa and Domínguez 2013, Eisfeldt and Papanikolaou 2013). We contribute to this literature by analyzing the effect of organizational capital on business cycle dynamics. We also explore further ways to test our model using an empirical literature aiming to measure organizational capital (Atkeson and Kehoe 2005, Black and Lynch 2005, Corrado, Hulten, and Sichel 2009, Squicciarini and Mouel 2012).

The remainder of this paper is organized as follows. To set the stage, in Section 2, we first analyze a simplified business cycle model with adjustment cost in employment and use it to document the persistence puzzle for hiring. We continue working with this simple model in Section 3, but add a nonseparable production technology to show that model gives rise to delayed adjustment. Section 4 simply shows that the argument for hiring in the previous two sections goes through for investment as well. In Section 5, we add a bit more realism to the model, which now features a nonseparable production function in both labor and capital, and show that the delay in adjustment is quantitatively important and matches the persistence in hiring and investment observed in the data. Section 6 aims to provide some evidence for the mechanism by documenting the dynamics for proxies of organizational investment and analyzing the implications of the model for differences across industries. Section 7 concludes.

2. THE DYNAMICS OF EMPLOYMENT AND HIRING

In this section, we set up a model environment that allows us to illustrate the lack of propagation in standard business cycle models. A lack of persistence is present both in investment and in hiring, but the puzzle is more pronounced for hiring. Therefore, we focus on hiring and simplify the model by assuming the capital stock is fixed (we revisit the lack of propagation in investment in Section 4). Our starting point is a business cycle model with adjustment costs. For reasons of exposition, we keep the model as simple as possible.

2.1 A Simple Model of Employment Adjustment

Our economy produces $Y_t$ goods in each period $t$, according to a production technology that requires only labor $N_t$,

$$Y_t = A_t N_t^{1-\alpha},$$  \hspace{1cm} (1)
where $A_t$ is the state of technology, which is normalized to have mean 1, and diminishing returns to labor in production are measured by the parameter $\alpha \in (0, 1)$. We analyze the response of the model to a one-time, unexpected and permanent change in technology $A_t$. The deterministic case allows us to formally describe the model dynamics. In the quantitative analysis in Section 5, we will allow $A_t$ to follow a stochastic Markov process.

Employment $N_t$ increases or decreases through hiring $h_t > 0$ or firing $h_t < 0$ according to the following law of motion:

$$N_t = N_{t-1} + h_t,$$

where we assumed that employment does not depreciate, that is, there are no exogenous separations or quits.

We assume that both the goods market and the labor market are perfectly competitive, so that the equilibrium is efficient and we can consider the social planner’s problem. Furthermore, we assume that the utility function is linear in consumption and leisure, so that the intertemporal consumption allocation is irrelevant and the social planner’s optimization problem is equivalent to maximizing profits,

$$\max_{\{h_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \left( \frac{1}{1 + r} \right)^t \left[ Y_t - \gamma N_t - g(h_t) \right]$$

subject to (2), where $r$ is the discount rate, $\gamma$ is the disutility from working (and the wage), and $g(.)$ is an adjustment costs function. The optimal hiring policy depends on the form of this function.

2.2 Optimal Hiring Policy

We analyze two cases for the adjustment costs function. Below, we use quadratic costs, which is probably the most prevalent cost function in the literature. In Online Appendix A.2, we analyze optimal hiring policy with fixed adjustment costs. These two cases, which are in some sense opposite extremes, convey the intuition for the model dynamics under convex and nonconvex adjustment costs more generally.

As a benchmark, first consider the frictionless optimal allocation. In the absence of adjustment costs, the planner sets hiring to achieve the optimal level of employment, so that we obtain the frictionless optimal level of employment $N^*_t$ simply by maximizing (3), with $g(h_t) = 0$, over $N_t$.

$$N^*_t = \left( \frac{1 - \alpha}{\gamma} \right)^{1/\alpha} A_t.$$

In a model with adjustment costs, we can think about the frictionless optimal level as the desired level of employment. The optimal hiring policy aims to achieve a balance between bringing employment close to its desired level while keeping adjustment costs low.
We use quadratic adjustment costs, \( g(h) = \frac{1}{2} \psi h^2 \), which can be thought of as an approximation of any convex adjustment costs function. A convex cost function implies that adjustment costs get very small for small amounts of hiring, so that infinitesimal adjustment in employment is costless and therefore reversible. This provides an incentive to smooth out adjustment over time, and the optimal response to a change in technology under quadratic adjustment costs is to hire (or fire) a small number of workers in each period over a long time, as described in Lemma 1.

**Lemma 1.** With quadratic adjustment costs, the optimal hiring policy in response to a change in technology in the employment adjustment problem described by equations (3), (1), and (2) can be described by Euler equation (5) for hiring, and has the following properties:

1. Hiring (or firing) starts immediately and continues for all periods after the shock, eventually approaching zero as employment \( N_t \) approaches its desired level \( N^*_t \) as in (4).
2. Hiring monotonically declines over time as employment \( N_t \) adjusts to close the gap from its desired level \( N^*_t - N_t \).

The proof of Lemma 1 is immediate from the Euler equation for hiring (5), which is derived from a straightforward dynamic programming problem in Online Appendix A.1,

\[
h_t = \frac{\gamma}{\psi} \left( \left( \frac{N^*_t}{N_t} \right)^{\alpha} - 1 \right) + \frac{1}{1 + r} E_t h_{t+1} \approx \frac{\alpha \gamma}{\psi} \left( \frac{N^*_t - N_t}{N^*_t} \right) + \frac{1}{1 + r} E_t h_{t+1}, \quad (5)
\]

where the second equality follows from a first-order Taylor approximation. This optimal hiring policy may be compared to the optimal hiring policy under nonconvex adjustment costs, which is described in Online Appendix A.2.

### 2.3 The Dynamics of Hiring in the Model

The dynamics of hiring predicted by the model differ depending on the specification of adjustment costs. However, here, we focus on a feature of the dynamics of hiring that is common across different specifications for adjustment costs. For both quadratic and fixed adjustment costs, hiring \( h_t \) is (weakly) monotonic in the distance of employment from its desired level, and zero when that distance equals zero, see Lemma 1 above and Lemma A.1 in Online Appendix. We might label this feature of the optimal hiring policy the "flexible accelerator" property, following Clark (1917), Samuelson (1939), and Lucas (1967).

An implication of the flexible accelerator property is that hiring (or firing) is highest immediately after a shock hits the economy, when the distance between the desired and actual levels of employment is largest. This prediction seems inconsistent with the hump-shaped impulse responses that are typically estimated using structural Vector autoregressions (VARs), as discussed in the introduction.
While we derived the flexible accelerator property and its corollary that hiring peaks on impact of a shock in a very specific and simple environment, these predictions are a good deal more general than the assumptions of our model. If the cost function includes elements of both nonconvex and convex costs, that is, functions that are in between the “extremes” of fixed and quadratic costs, if employment depreciates or if shocks are mean-reverting, then it is generally not optimal to adjust to the frictionless optimal level of employment. However, in all of these cases, hiring is still monotonic in the distance between the current level and some “desired” level of employment, and these models still predict that hiring peaks immediately after a shock.

2.4 A Calibration Target for Persistence

To compare the dynamics of hiring in the model to those in the data, we would ideally want to know the response of hiring to the distance between the current and desired levels of employment. In general, it is not possible to estimate this response, because the desired level of employment $N^*_t$ is not observed. However, in the context of our simple benchmark model, we can obtain an observable proxy. Using production function (1) to eliminate technology $A_t$ from expression (4), we see that in our model, the distance of employment from its desired level is log-proportional to labor productivity.

$$\frac{N_t^*}{N_t} = \left( \frac{1 - \alpha Y_t}{\gamma N_t} \right)^{1/\alpha}. \tag{6}$$

Thus, under the assumptions of our model, we can measure the response of hiring $h_t$ to a change in technology by regressing the hiring rate on lags of labor productivity $Y_t/N_t$.

A moving-average (MA) regression of the hiring rate on labor productivity provides a simple and intuitive way to summarize the comovement of hiring with other macroeconomic aggregates and is likely to be informative about the response of hiring to shocks. The advantage of this regression over estimated impulse responses from a structural VAR is that we do not have to take a stance on the type of shocks that drive changes in the desired level of employment, which makes it a useful calibration target. An even simpler target, like the autocorrelation of the hiring rate, would not be able to distinguish between persistence in hiring due to propagation of the model and persistence that is due to persistence in the shocks that drive business cycles. As a final advantage of our calibration target for the dynamics in hiring, we note that the logic of the approach extends to other variables, and in particular that the dynamics of investment can be meaningfully summarized by an MA regression of investment on capital productivity, see Section 4.

It is important to note that the MA regression we propose does not recover the impulse response function of hiring. Even in the context of our simple model,
labor productivity is endogenous, and we make no claim of causality in the regression. There are two reasons for this. First, while many structural shocks will affect the labor market through labor productivity (technology shocks, but also monetary policy shocks and other consumption demand shocks), other shocks will not. In particular, to the extent that the response of hiring to exogenous changes in labor supply is different from its response to other shocks, this will not be captured by our regression. Second, extensions to our simple model may break the link between the desired level of employment and labor productivity. For example, if wages strongly comove with productivity, for instance, because workers have strong bargaining power in wage negotiations, then the $\gamma$ in expression (6) will be a function of labor productivity, partially offsetting the effect of productivity on the desired level of employment. Our claim is that while our MA regression does not equal any impulse response function, it will be informative about it, and we support this claim by showing below that the regression inherits many of the properties of the response of hiring to identified structural shocks to technology and monetary policy. In particular, the estimates show a clear hump shape in the dynamics of hiring.

In the next subsection, we use our MA measure to compare the dynamics of hiring implied by the model with those in the data.

2.5 The Persistence Puzzle

Figure 1 shows the dynamics of hiring (job finding rate), as measured by an MA regression of the hiring rate $h_t$ on labor productivity $Y_t/N_t$ in the model with adjustment costs in employment and in the data.\footnote{For the hiring rate, we use the job finding probability from Shimer (2012), and labor productivity is output per worker from the BLS Labor Productivity and Costs program.} Our measure for the dynamics of hiring in the model closely mirrors the impulse response of hiring to a technology shock, and, in particular, inherits its property that hiring is largest upon impact of the shock. With quadratic adjustment costs, hiring peaks when productivity changes and then slowly reverts to zero.

In the data, hiring peaks more than 2 years after the distance between the desired and current levels of employment is largest. This is consistent with the hump-shaped impulse responses for hiring found in structural VAR models with identified technology (Canova, Lopez-Salido, and Michelacci 2010, Canova, Lopez-Salido, and Michelacci 2013) or monetary policy shocks (White 2018). The benchmark models with adjustment costs cannot replicate this property.

The failure of standard models with adjustment costs to replicate the delayed response in hiring observed in the data is what we call the persistence puzzle. In the next section, we show how the model is able to match the observed dynamics in hiring if we allow for a nonseparable production technology, and that this result holds for different specifications of the adjustment costs function.
Fig 1. Persistence: Model versus Data.

Notes: Persistence in hiring and investment in the data (black solid line with gray standard error bands) and in a model with standard separable production function with convex adjustment costs (red dashed line). The figure shows the coefficients of an MA regression of hiring $h_t$ on labor productivity (output per hour) $Y_t/N_t$ and investment $i_t$ on capital productivity $Y_t/K_t$ for the period from 1948:Q1 to 2007:Q4, and the response over a simulated sample of the model over 100,000 periods.
3. DELAYED ADJUSTMENT

In this section, we introduce the main idea of this paper: with a nonseparable production function, the model predicts that hiring peaks not when a shock hits, but several periods after. We label this property of the model delayed adjustment. Below, we first examine delayed adjustment in the context of the simple model from the previous section. In Section 5, we explore the quantitative importance of delayed hiring to match the data with an extended version of that model.

3.1 Nonseparable Production Technology

Standard production functions, like (1), are separable over time. Labor input in period $t$ contributes to production in the same period, and current-period labor is the only labor input in production. These are extreme assumptions that are unlikely to be satisfied. In reality, many tasks that workers perform do not immediately generate production, for example, cleaning, maintenance, training, or participating team meetings. Of course, these tasks are productive; surely, productivity would decrease if the office was never cleaned, machines were not maintained, workers never learned anything new, and no meetings were held to coordinate between workers. However, the effect of these tasks on production realizes in future periods and may last for a long time, so that they need to be performed only infrequently.

We model the effect of infrequent tasks on production by introducing an additional state variable, which we label organizational capital. When workers perform organizational, or infrequent, tasks, their labor does not directly enter into the production function but is used to accumulate organizational capital. Organizational capital enters into the production function and depreciates when no workers invest into it by performing organizational tasks. This gives rise to the production function,

$$
Y_t = \phi A_t (e_t N_t)^{1-\alpha} + (1 - \phi) B_t L_t^{\rho}, \tag{7}
$$

where $L_t$ is the stock of organizational capital, which evolves according to,

$$
L_t = (1 - \lambda) L_{t-1} + \tilde{\lambda} ((1 - e_t) N_t)^{1-\alpha}, \tag{8}
$$

where $e_t$ is the fraction of total labor input that is used for regular productive activities, which is a new choice variable, $\phi$ is a parameter governing the relative importance of current production versus organizational tasks, $B_t$ represents shocks to the productivity of labor in producing organizational capital and—like $A_t$— is normalized to have mean 1. As in the simple model, we will analyze the response of hiring to a one-time, unexpected and permanent change in $A_t$, keeping $B_t = 1$ fixed for most of the analysis. The parameter $\lambda$ denotes the rate of depreciation or organizational capital and $\tilde{\lambda} = [(r + \lambda)/(1 + r)\lambda^{1-\rho}]^{1/\rho}$ is a correction factor to undo the effect of $\lambda$ on the relative importance of organizational tasks versus current productive
activities. Finally, $\rho$ is a parameter measuring diminishing returns to organizational tasks in the production versus the use of organizational capital. We would expect $\rho$ to lie between $1 - \alpha$, in which case the diminishing returns to organizational capital in production are the same as for regular labor but there are no diminishing returns in the production of organizational capital, and 1, which implies diminishing returns to organizational tasks in the production of organizational capital but no diminishing returns to organizational capital in the production of output.

The production technology described by equations (7) and (8) stays as close as possible to a standard production function while allowing for intertemporal nonseparability in production. We assume that the only difference between regular productive tasks and organizational tasks is that the effect of organizational tasks on production is spread out over a long time period. How long this period is, is determined by the depreciation rate of organizational capital $\lambda$. Production function (7) reduces to (1) not only for $\phi = 1$ (no role for organizational capital in production), but also for $\lambda = 1$ ("organizational" tasks, like regular productive activities, affect production in the current period only), up to a normalization of the productivity shock. In the frictionless optimal steady state, the two types of labor enter into the production function symmetrically, and the only difference is their relative productivity $\phi A_t / (1 - \phi) B_t$, see equation (10) below. Our final assumption on the production function, and the only one that is not justified by symmetry, is additive separability, implying that output produced using regular labor is perfectly substitutable with output produced using organizational capital. This assumption is for simplicity, and we will show in Section 5 that it is not qualitatively important for our results.

The nonseparable production technology provides firms with a storage technology for labor in the form of organizational capital. This storage technology allows them to intertemporally smooth labor input and adds an additional margin of labor adjustment: by postponing organizational tasks and reallocating labor to daily productive activities, firms can temporarily increase output without increasing labor input. Below, we explore how this additional margin of adjustment changes the dynamics of hiring.

### 3.2 Optimal Hiring Policy

As before, the optimal hiring policy depends on the specification for adjustment costs, and we analyze the same two cases as in Section 2 above: quadratic costs in the text below and fixed costs in Online Appendix B.3. We show that a nonseparable

3. Notice that $\tilde{\lambda} = 1$ if $\lambda = 1$ and $\tilde{\lambda} \rightarrow \lambda$ for $r \rightarrow 0$. The reason that $\tilde{\lambda} \neq \lambda$ in general is due to the difference between the steady state and the static optimum allocation.

4. With $\lambda = 1$, production is given by $Y_t = \phi A_t (e_t N_t)^{1 - \alpha} + (1 - \phi) B_t ((1 - e_t) N_t)^{1 - \alpha}$. Since the production function is now separable over time, the fraction of workers allocated to each type of labor $e_t$ will be chosen statically to maximize production in each period, so that $e_t$ satisfies $\phi A_t e_t^{\alpha} = (1 - \phi) B_t (1 - e_t)^{1 - \alpha}$. Substituting the optimal $e_t$ into the production function gives $Y_t = \phi A_t e_t^{\alpha} N_t^{1 - \alpha} = \tilde{A}_t N_t^{1 - \alpha}$, where $\tilde{A}_t = [(\phi A_t)^{1/\alpha} + ((1 - \phi) B_t)^{1/\alpha}]^{\alpha}$. 


production technology introduces delay in the optimal response of hiring, and that this happens for both specifications for adjustment costs.

The planner still maximizes the expected net present value of profits, as in (3), but she now has an additional choice variable $e_t$, the fraction of labor to allocate to regular productive versus organizational tasks.

$$\max_{[e_t, h_t]} \sum_{t=0}^{\infty} \left( \frac{1}{1 + r} \right)^t [Y_t - \gamma N_t - g(h_t)]. \quad (9)$$

where $Y_t$ is given by production function (7), subject to the laws of motion for employment (2) and organizational capital (8).

It is again useful as a benchmark to solve for the frictionless allocation. Setting $g(h_t) = 0$, maximizing (9) over $e_t$ and $N_t$, and assuming that the organizational capital was in steady state before technology changed, we find that the frictionless optimal level of employment $N^*_t = N^*$ is constant over time and given by

$$N^* = \left( \frac{(1 - \alpha)\phi A}{\gamma} \right)^{1/\alpha} + \left( \frac{(1 - \alpha)(1 - \phi)B}{\gamma} \right)^{1/\alpha} \quad (10)$$

with $e^* N^* = ((1 - \alpha)\phi A/\gamma)^{1/\alpha}$, workers are allocated to regular productive activities and the remaining $(1 - e^*) N^*$ working on organizational tasks, see Online Appendix B.1 for the derivation.

With convex adjustment costs, $g(h_t) = \frac{1}{2} \psi h^2$, and a standard separable production function, hiring jumps in response to a change in technology, and then slowly and monotonically declines to zero as employment approaches its frictionless optimal level, see Lemma 1. With a nonseparable production technology, hiring still only jumps on impact of the shock, and eventually declines to zero as employment approaches the frictionless optimal. However, the decline in hiring need not be monotonic. For a relevant range of parameter values, hiring first increases after the shock before starting to decrease and declining to zero, as described in Proposition 1. Thus, peak hiring happens after a delay.

**Proposition 1.** With quadratic adjustment costs, the optimal hiring policy in response to a change in technology in the employment adjustment problem described by equations (3) and (2) and a nonseparable production technology described by (7) and (8) has the following properties.

1. Hiring (or firing) starts immediately and continues for all periods after the shock, eventually approaching zero as employment $N_t$ approaches its desired level $N^*_t$ as in (10).
2. Hiring (firing) is delayed, in the sense that after its initial jump it first increases, then peaks, and finally declines over time, if the discount rate $r > 0$, adjustment costs $\psi > 0$, the relative importance of organizational capital in production $1 - \phi \in [0, 1]$, are all sufficiently high, and the depreciation rate of organizational
capital $\lambda \in [0, 1]$ and diminishing returns in organizational capital $\rho \in [1 - \alpha, 1]$ are sufficiently low.

3. If the parameter conditions for delayed adjustment are satisfied, then the amount and length of delay (the difference between peak hiring and initial hiring) increases with the discount rate $r > 0$, adjustment costs $\psi > 0$, and the relative importance of organizational capital $1 - \phi \in [0, 1]$, and decreases with the depreciation rate of organizational capital $\lambda \in [0, 1]$.

The proof of Proposition 1 is a straightforward application of dynamical systems, and is implemented numerically, see Online Appendix B.2. The intuition for the result can be seen from the Euler equation for hiring,

$$h_t = \frac{\gamma}{\psi} \left( \frac{(e^* N^*)^\alpha}{e_t N_t} - 1 \right) + \frac{1}{1 + r} h_{t+1} \simeq \frac{\alpha \gamma}{\psi} \left( \frac{e^* - e_t}{e^*} + \frac{N^* - N_t}{N^*} \right) + \frac{1}{1 + r} E_t h_{t+1},$$  \hspace{1cm} (11)

which may be compared with the Euler equation (5) for the model with separable production technology. With a nonseparable production technology, the urgency of hiring or firing is no longer summarized by the deviation of employment from its desired level $N^* - N_t$, but depends also on the fraction of labor that is optimally allocated to current production $e^* - e_t$, which, in turn, depends on the state of organization in the firm. If the organizational capital stock was at its optimal steady state level before an unexpected increase in technology, then it still is after the shock hits. Therefore, it is initially costless for firms to disinvest in organization, reallocating workers from organizational to current productive tasks. This reduces the incentive for hiring. Over time, however, organizational capital gets depleted, which negatively affects production and profits. When this happens, more workers are allocated to organizational tasks again, and firms need to hire more workers to achieve the desired level of output.

The length of the delay increases with adjustment costs $\psi$, and with the discount rate, which increases the incentive for delaying the adjustment costs. The length of delay decreases with the depreciation rate of organizational capital $\lambda$, which determines how fast underinvestment in organizational tasks leads to a decline in the organizational capital stock. This last parameter will be important as a lever to match the data, see Section 5 below. These results are qualitatively similar for nonconvex adjustment costs, see Online Appendix B.3.

3.3 Persistence in Hiring

The optimal hiring policy with a nonseparable production technology is summarized in Figure 2, which shows the response of hiring to an increase in technology. The figure shows this response both for quadratic adjustment cost, and for fixed adjustment costs as analyzed in Online Appendix B.3. While the hiring policy looks quite different depending on the specification of adjustment costs, the optimal policy
in both cases involves delay, in the sense that either all or most hiring takes place later than the impact of the shock. Delayed adjustment in the model with nonconvex adjustment costs is a nonlinear effect, which depends on the size of the shock. With convex adjustment costs, whether there is delayed adjustment depends on parameters. In this model, we can linearize the equilibrium conditions without qualitatively affecting the dynamics, which greatly facilitates incorporating nonseparabilities into larger scale macroeconomic models.

The intuition for why delayed adjustment may be optimal in our model is straightforward. The nonseparable production technology, in particular the slow-moving organizational capital stock, acts as a storage (or smoothing) technology for labor. This storage technology provides firms with an intensive margin for labor adjustment: firms may postpone organizational tasks and reallocate workers to current productive activities, effectively borrowing labor from the future. Using this intensive margin immediately increases production and therefore profits. And the intensive margin is initially costless, because the organizational capital stock is slow to depreciate, and remains at its current level even when organizational investment drops. Eventually, however, the borrowed labor needs to be paid back. The (slow) depreciation of the organizational capital stock negatively affects production and profits, and this cost increases over time, so that the firm is forced to allocate more workers to organizational tasks again. Having exhausted the intensive margin of adjustment, firms must then turn to the extensive margin and hire more workers.
The type of delay predicted by our model is endogenous, in the sense that delayed adjustment is optimal in response to a single shock, even if no further shocks hit the economy. This makes it different from the delay discussed, for example, in Dixit and Pindyck (1994) in the context of investment, and used, for example, in Bachmann (2012) in the context of employment, which we might call exogenous delay. In a model with nonconvex adjustment costs, but with a separable production technology, as in these studies, a firm may choose not to respond to a shock while it is “waiting for new information” (Dixit and Pindyck 1994, p.9), that is, to take a “wait and see” approach (Schreft, Singh, and Hodgson 2005). However, new information in this context means new shocks. If those new shocks are such as to further increase the benefits of adjustment, then the firm may decide to adjust after a “delay.” However, if the new shocks revert the effect of the first shock, then adjustment may never happen. In our model with a nonseparable production technology, on the other hand, delayed adjustment will happen in response to some shocks, but it does not happen immediately. The distinction is important, because estimated impulse responses from a VAR, if correctly identified, describe the response of the economy to a single shock. Therefore, only a model with endogenously delayed adjustment can explain the hump shapes in the estimated response of hiring.

We will turn to the quantitative implications of our model in Section 5 below, but first take a brief detour into the dynamics of capital and investment in the next section.

4. THE DYNAMICS OF CAPITAL AND INVESTMENT

The impulse response of capital investment to technology and monetary policy shocks, like that of hiring, shows a clear hump shape (Altig et al. 2011). For this reason, the Dynamic stochastic general equilibrium (DSGE) literature often assumes so-called “cost-of-change adjustment costs,” that is, costs that are quadratic in the change in investment $i$ rather than its level, $g(i, i_{t-1}) = \frac{1}{2}\psi (i / i_{t-1})^2$ (Christiano, Eichenbaum, and Evans 2005, Christiano, Eichenbaum, and Trabandt 2018). In this section, we argue that standard “level” adjustment costs, $g(i) = \frac{1}{2}\psi i^2$, in combination with a nonseparable production technology, give rise to very similar dynamics as cost-of-change adjustment costs with a standard separable production technology. We show that the persistence puzzle we documented for hiring in Section 2.5 holds for investment as well, and argue that since the model is symmetric in capital and labor, our results for the dynamics of hiring with a nonseparable production function apply equally to investment in a model with fixed labor.

The model in Sections 2 and 3 assumed that production requires only labor. To focus on the dynamics of investment, we can make the opposite extreme assumption that production requires only capital $K$. Then, the simple production function (1) would be replaced by $Y_t = A_t K_\alpha$, where $\alpha \in (0, 1)$ is the capital share, whereas if we allow for intertemporal nonseparability, production function (7) becomes

$$Y_t = \phi A_t (u_t K_t) + (1 - \phi) B_t L_t$$

(12)
and
\[ L_t = (1 - \lambda)L_{t-1} + \tilde{\lambda}((1 - u_t)K_t)^\alpha, \]

(13)

where all parameters have the same interpretation as before and \( u_t \) is the fraction of the capital stock that is used for current production, with \( 1 - u_t \) of capital being used for investing in organizational or intangible capital. Perhaps, the easiest way to interpret \( 1 - u_t \) is as the fraction of machines that are shut down for maintenance. The planner may adjust the capital stock by investing or disinvesting in it, \( K_t = K_{t-1} + i_t \).

This model for capital adjustment is almost completely symmetric to the model for labor adjustment, with the capital share \( \alpha \) playing the role of the labor share \( 1 - \alpha \) in that model, except that the way investment affects profits is slightly different from the way hiring does. Maintaining the same assumptions of linear utility and competitive markets as before, the planner maximizes the expected net present value of profits,

\[ \max_{\{i_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \left( \frac{1}{1 + r} \right)^t [Y_t - i_t - g(i_t)]. \]

(14)

The difference with the labor adjustment model is that investment in capital lowers profits, whereas the stock of labor reduces profits (or utility) in objective function (3). The Euler equation for investment in this model is given by

\[ i_t = \frac{r}{(1 + r)^\psi} \left( \frac{u^* K^*}{u_t K_t} \right)^{1-\alpha} - 1 + \frac{1}{1 + r} i_{t+1}, \]

(15)

where \( u_t = u^* = 1 \) if \( \phi = 1 \) or \( \lambda = 1 \), see Online Appendix C for the derivation. Comparing this equation to Euler equation for hiring (11), it is clear that the model for capital adjustment model is symmetric to the one for labor adjustment under a parameter restriction on the value of leisure in that model, \( \gamma = r/(1 + r) \).

The symmetry between the models for capital and labor adjustment allows us to extend our results for hiring dynamics to investment.

The desired capital stock is log-proportional to capital productivity for \( \phi = 1 \),

\[ \frac{K^*_t}{K_t} = \left( \frac{\alpha(1 + r) Y_t}{r K_t} \right)^{1/(1-\alpha)}, \]

(16)

see Online Appendix C and compare to (6). Therefore, a regression of log investment on an MA for log capital productivity is a meaningful summary of the persistence in investment, see the discussion in Section 2.4. The bottom panel in Figure 1 shows the dynamics of investment in the data (private nonresidential fixed investment, net of consumption of fixed capital, from the NIPA), as well as in the model with a standard separable production function (\( \phi = 1 \)). The figure documents a persistence puzzle for investment, which is very similar as the puzzle for hiring that we documented earlier, although less severe. In the model, investment monotonically declines after the initial
impact of the shock, whereas in the data, the response is hump-shaped and peaks only after 5–8 quarters (compared to 8–12 quarters for hiring).

A nonseparable production function brings the dynamics of investment closer to the data.

**Proposition 2.** *The optimal investment policy in response to a change in technology in the capital adjustment problem described by equation (14) and a nonseparable production technology described by (12) and (13) exhibits delayed adjustment, both for fixed adjustment costs and for quadratic adjustment costs, as described in Proposition 5 (in Online Appendix B.3), and Proposition 1 replacing hiring with investment and employment with capital.*

Qualitatively, a nonseparable production technology can explain the persistence puzzle in hiring as well as in investment. Whether this mechanism is sufficient to match the data is a quantitative question, to which we now turn. Since production in macroeconomic models usually requires both labor and capital, there is an additional quantitative question whether the model can match the persistence in both hiring and investment for the same parameter values.

5. **PERSISTENCE IN MACROECONOMIC MODELS**

We showed that in a simple model with nonseparable production technology, the optimal policy for hiring and investment involves delayed adjustment. In this section, we argue that this delay is quantitatively relevant, in the sense that it brings the model dynamics substantially closer to the data. As a by-product, we also show that the result goes through if there are adjustment costs in both capital and labor and if we extend the model in other dimensions to make it more similar to the type of DSGE models typically used in the literature. The main quantitative exercise is to calibrate the model parameters to the literature as much as possible, and then to evaluate whether there exist values for the free parameters describing the nonseparability in production, which generate persistence in hiring and investment as observed in the data. In Section 6, we think about whether the parameters we need are reasonable, and whether we can find additional evidence to test our story.

5.1 **Quantitative Analysis**

We use the following production technology for output using labor and capital,

\[
Y_t = \left[ \phi(A_t(u_t,K_t)^{\alpha}(e_t,N_t)^{1-\alpha})^{\frac{\alpha-1}{\alpha}} + (1 - \phi)(B_tL_t^\rho)^{\frac{\alpha-1}{\alpha}} \right]^{\frac{\alpha}{\alpha-1}},
\]

(17)

where organizational capital \(L_t\) evolves according to

\[
L_t = (1 - \lambda)L_{t-1} + \tilde{\lambda}((1 - u_t)K_t)^{\frac{\alpha}{\rho}}((1 - e_t)N_t)^{\frac{1-\alpha}{\rho}},
\]

(18)
which is a straightforward extension of (7) and (12). There are only two new parameters: \( \sigma \), the elasticity of substitution between current production and organizational capital, and \( \theta \), which measures decreasing returns to scale in production. We need decreasing returns, because with constant returns to scale and perfect competition, the size of the firm is indeterminate and the model (with linear utility) does not have a steady state. We assume that technology \( A_t \) is stochastic and follows an exogenous Markov process, and keep the technology for organizational capital production \( B_t \) fixed at unity for most of the analysis.

There are adjustment costs in both labor and capital. Consistent with the literature, we let the adjustment cost functions be quadratic in the relative adjustment, that is, in hiring and investment as a fraction of employment and capital, respectively, so that

\[
g_N(h_t) = \frac{1}{2} \psi_N(h_t/N_t)^2 \quad \text{and} \quad g_K(i_t) = \frac{1}{2} \psi_K(i_t/K_t)^2.
\]

Furthermore, we assume that the stocks of both employment and capital depreciate,

\[
N_t = (1 - \delta_N)N_{t-1} + H_t = N_{t-1} + h_t,
\]

\[
K_t = (1 - \delta_K)K_{t-1} + I_t = K_{t-1} + i_t,
\]

where \( \delta_N \) is the separation rate and \( \delta_K \) is the depreciation rate for physical capital. Notice that our timing assumptions imply that depreciated employment and capital may be reinstalled within the period, so that \( \delta_N \) and \( \delta_K \) are gross depreciation rates. Also, note that we assume that adjustment costs depend on hiring \( h_t \) and investment \( i_t \) net of replacement hiring/investment, rather than on total hiring \( H_t \) and investment \( I_t \).

We maintain the assumption that utility is linear, but we allow for a preference shock as a stand-in for all nontechnology shocks. Thus, we assume that per-period welfare is given by

\[
Z_t = [Y_t - I_t - g_N(h_t) - g_K(i_t)] - \gamma N_t,
\]

where \( Z_t \) is stochastic and follows an exogenous Markov process that is independent of \( A_t \). This second shock brings the correlation matrix of the model variables closer in line with data by breaking the almost-perfect comovement between variables in a one-shock model, and is also needed to help a simple real business cycle (RBC)-type model like ours to match the relative volatility of labor market variables, hiring and employment. The equilibrium conditions for the quantitative model are listed in Online Appendix D.

The calibration of the model is summarized in Table 1. For most of the parameters, we choose values that are commonly used in the literature. In this spirit, we calibrate the quarterly discount rate \( r \) to 3% to match the average return on equity, the capital share \( \alpha \) to 0.33, \( \theta = 0.89 \) to match the markup of 12.5% of a monopolistically competitive firm (Galí 2015, p. 67), choose a depreciation rate for capital \( \delta_K \) of 2.5% (King and Rebelo 1999), a (gross) separation rate \( \delta_N \) of 20% per quarter (Galí and van Rens 2020), and calibrate the marginal rate of substitution between consumption and leisure \( \gamma \) to match the average employment population ratio \( \bar{N} = 0.7 \). We set the autocorrelation of \( A_t \) to match the corresponding parameter for total factor productivity in the data, normalize the standard deviation of innovations in \( A_t \) to 1%, and
set the stochastic process for $Z_t$ equal to that for $A_t$, loosely based on the estimates in (Smets and Wouters 2007, Table 1B), showing that the autocorrelations and standard deviations of nontechnology shocks are in the same order of magnitude as those of technology shocks.

The calibration of the production and adjustment technologies is crucial for this paper. For the adjustment costs in capital, we set $\psi_K = 4.34$ to match Jermann (1998)’s estimate of the elasticity of the investment rate $i_t/K_t$ to Tobin’s $q$ ($1/\psi_K = 0.23$). Since there are few direct estimates of adjustment costs in employment, the literature often sets this parameter to match a volatility target. We follow this practice and set $\psi_N$ to match the amplitude of the response of hiring. The parameters $\phi$, $\rho$, $\sigma$, and $\lambda$, which describe the nonseparability in the production technology, are specific to our model, and consequently, there is no guidance in the literature on how to calibrate these parameters. In our quantitative exercise, we set these parameters, together with the adjustment costs $\psi_N$, to match both the amplitude and the dynamics of the hiring response. We then evaluate the performance of the model in two ways. First, in Table 2, we illustrate the effect of a nonseparable production technology in terms of a standard set of business cycle statistics. Second, we treat the response of investment as an overidentifying restriction and evaluating the model fit of the response of investment for the same parameter values that we calibrated to the response of hiring, see Section 5.2 below. We also explore how the parameters of the production function affect the predictions of the model for the dynamics of hiring and investment.

As most models with adjustment costs, ours predicts a volatility of investment relative to output that is lower than in the data, and a volatility of consumption that is higher. This problem is aggravated by the introduction of a nonseparable production technology, but only very slightly. The volatility of employment is slightly lower than in the data in the model with separable production, and slightly higher in the
TABLE 2  
BUSINESS CYCLE STATISTICS

<table>
<thead>
<tr>
<th></th>
<th>Separable production</th>
<th>Org. capital (N)</th>
<th>Org. capital (K)</th>
<th>Org. capital (N, K)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data Adjustment costs in:</td>
<td>Adjustment costs in:</td>
<td>Adjustment costs in:</td>
<td>Adjustment costs in:</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>K</td>
<td>N, K</td>
<td>N, K</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sigma_1 / \sigma_Y )</td>
<td>2.69</td>
<td>16.8</td>
<td>0.90</td>
<td>1.27</td>
</tr>
<tr>
<td>( \sigma_N / \sigma_Y )</td>
<td>1.09</td>
<td>0.96</td>
<td>1.24</td>
<td>1.06</td>
</tr>
<tr>
<td>( \sigma_Y / \sigma_K )</td>
<td>1.28</td>
<td>0.09</td>
<td>0.98</td>
<td>0.94</td>
</tr>
<tr>
<td>( \sigma_N / \sigma_Y )</td>
<td>0.59</td>
<td>0.30</td>
<td>0.37</td>
<td>0.49</td>
</tr>
<tr>
<td>( \sigma_Y / \sigma_N )</td>
<td>0.75</td>
<td>2.59</td>
<td>1.01</td>
<td>0.97</td>
</tr>
<tr>
<td>( \text{corr}(I, I_{t-1}) )</td>
<td>0.87</td>
<td>0.24</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td>( \text{corr}(N, N_{t-1}) )</td>
<td>0.89</td>
<td>0.95</td>
<td>0.72</td>
<td>0.93</td>
</tr>
<tr>
<td>( \text{corr}(Y/Y_{t-1}, K_{t-1}) )</td>
<td>0.85</td>
<td>0.71</td>
<td>0.71</td>
<td>0.85</td>
</tr>
<tr>
<td>( \text{corr}(Y/K_{t-1}, Y_{t-1}/K_{t-1}) )</td>
<td>0.69</td>
<td>0.70</td>
<td>0.71</td>
<td>0.74</td>
</tr>
<tr>
<td>( \text{corr}(C/C_{t-1}) )</td>
<td>0.80</td>
<td>0.32</td>
<td>0.72</td>
<td>0.87</td>
</tr>
</tbody>
</table>

model with nonseparable production, but close to the data in both models. The relative volatility of both labor and capital productivity is higher in the model with a nonseparable production technology, bringing the volatility of these series closer to that observed in the data. Nonseparable production increases the persistence of investment and employment, bringing the autocorrelation coefficient in investment closer to the data, but taking the autocorrelation in employment (a bit) further from its observed equivalent. The downside of this is that the autocorrelation coefficients of capital and labor productivity decline to levels lower than observed in the data, in the case of capital productivity substantially so. The autocorrelation in consumption moves away from the data as well. Overall, while nonseparability in the production technology does not seem to be an unambiguous improvement in the performance of the model in matching a standard set of business cycle statistics, it does not clearly deteriorate the ability of the model to match the data either. Of course, model fit is a much broader concept than is reflected in these statistics, and we now turn to the predictions of the model for the dynamics of hiring and investment, as measured by our measure of persistence introduced in Section 2.4.

5.2 The Dynamics of Hiring and Investment

Figure 3 shows the model impulse responses of hiring and investment for the model with a separable and a nonseparable production technology. The three lines in this figure correspond to increasing levels of importance of organizational capital in production: a separable production function, the calibrated baseline, and a version of the model with a slower depreciation of organizational capital. The model with a nonseparable production function replicates the hump-shaped impulse responses for hiring and investment, and the length of the delay increases with a lower depreciation rate of organizational capital. Thus, the results we proved in Sections 3 and 4 hold for
a more general model, in which production requires both capital and labor, with a standard calibration for the parameters.

Next, we ask the question whether we can find parameters for the production technology so that the dynamics of hiring and investment match the persistence observed in the data. As a summary measure of the dynamics of hiring and investment, we use MA regressions, over a 24-quarter horizon, of these variables on labor and capital productivity, respectively, as introduced in Sections 2.5 and 4. We find parameter values that minimize the distance between estimated empirical and theoretical response functions of hiring over a 10-quarter horizon. The mean empirical response is the black solid line in the top panel of Figure 1. For each combination of parameters, we simulate the model 1,000 times and find the mean theoretical response. We loop over possible combinations of parameters until we find the combination that minimizes the distance, that is, the weighted sum of squared differences between the mean of the model and empirical estimates, where the weighting function is a diagonal matrix with sample variances of model estimates along the diagonal. We truncate the responses at lag 10 because the hump-shaped response of the series is most pronounced over this horizon. We display the resulting theoretical responses together with empirical responses over a 24-lag horizon. The results of this exercise are presented in Figure 4, and the values for $\varphi$, $\rho$, $\sigma$, and $\lambda$ we used for these figures are reported in Table 1.

It is clear from Figure 4 that the model has no trouble replicating the persistence in hiring observed in the data, and it also gets close to matching the responses of investment with the same parameter values, even though we did not target this response in the calibration. Hiring responds little initially and peaks after just over 2 years in the data, whereas investment jumps more on impact but peaks around the same time as hiring. The calibrated model matches the response of hiring almost perfectly. The model also gets close to matching the amount of delay (0.43 vs. 0.21 in the data) and the length of the delay (8 quarters in the model vs. 5 in the data), even though we did not target this response in the calibration.
Fig 4. Persistence in Hiring and Investment: Models versus Data.

Notes: Persistence in hiring and investment in the data (black solid line with gray standard error bands), in our baseline model (red line with diamonds), and in the model without organizational capital (blue line with stars). The figure shows the coefficients of an MA regression of hiring \( H_t \) on labor productivity (output per worker) \( Y_t/N_t \) and investment \( I_t \) on capital productivity \( Y_t/K_t \) for the period from 1948:Q1 to 2007:Q4, and the response over a simulated sample of the model over 100,000 periods.
### TABLE 3

**Robustness Analysis**

<table>
<thead>
<tr>
<th></th>
<th>Hiring</th>
<th>Investment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Amount delay</td>
<td>Length delay</td>
</tr>
<tr>
<td>Data (1948:Q1–2007:Q4)</td>
<td>0.86</td>
<td>8</td>
</tr>
<tr>
<td>— Pre-84 (1948:Q1–1984:Q4)</td>
<td>0.83</td>
<td>8</td>
</tr>
<tr>
<td>— Post-84 (1985:Q1–2007:Q4)</td>
<td>1.65</td>
<td>11</td>
</tr>
<tr>
<td>$\phi = 1$ (no OC, $e = u = 1$)</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>$\phi = 0.545$, $\lambda = 0.032$ (baseline)</td>
<td>0.53</td>
<td>9</td>
</tr>
<tr>
<td>$\phi = 0.545$, $\lambda = 0.027$ (OC depreciates less)</td>
<td>0.96</td>
<td>9</td>
</tr>
<tr>
<td>$\psi_N = 40$ (higher adjustment costs)</td>
<td>0.30</td>
<td>6</td>
</tr>
<tr>
<td>$\psi_N = 12$ (baseline)</td>
<td>0.53</td>
<td>9</td>
</tr>
<tr>
<td>$\sigma = 3$ (OC more complementary)</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>$\sigma = 3.65$ (baseline)</td>
<td>0.53</td>
<td>9</td>
</tr>
<tr>
<td>$\sigma = 4$ (OC more substitutable)</td>
<td>0.85</td>
<td>9</td>
</tr>
<tr>
<td>$B_t = A_t$, $\sigma = 3$ (OC more complementary)</td>
<td>0.20</td>
<td>12</td>
</tr>
<tr>
<td>$B_t = A_t$, $\sigma = 3.65$</td>
<td>0.50</td>
<td>12</td>
</tr>
<tr>
<td>$B_t = A_t$, $\sigma = 4$ (OC more substitutable)</td>
<td>0.70</td>
<td>11</td>
</tr>
<tr>
<td>$\alpha_L = 0$ (OC requires only labor)</td>
<td>0.34</td>
<td>10</td>
</tr>
<tr>
<td>$\alpha_L = 0.33$ (baseline)</td>
<td>0.53</td>
<td>9</td>
</tr>
<tr>
<td>$\alpha_L = 0.4$ (OC requires less labor)</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>AC over gross hiring/investment</td>
<td>0.51</td>
<td>8</td>
</tr>
</tbody>
</table>

*Note:* The amount of delay is measured as peak minus impact hiring or investment as a fraction of peak hiring/investment. The length of delay is the difference between the time of peak hiring/investment and the time of impact measured in quarters.

The share of organizational capital we need to assume to match the hiring dynamics observed in the data is $1 - \bar{u} = 1 - \bar{\epsilon} = 21\%$, corresponding to $\phi = 0.545$, see Table 1. Organizational capital is slightly less persistent than physical capital, with a depreciation rate of 3.2%. In Section 6, we discuss some evidence on whether these are reasonable parameter values and try to find ways to test our story.

### 5.3 Robustness

The nonstandard element in our model is the production technology, and this is where we focus our robustness analysis. We start with varying the parameters $\phi$, which measures the importance of organizational capital in production, and $\lambda$, its depreciation rate, which have the expected effect on the results. These and all other results discussed in this subsection are reported in Table 3.

We then consider the elasticity of substitution between current production and organizational capital $\sigma$, which is qualitatively important for the predictions of the model. For an increase in productivity to have a positive effect on the fraction of workers $e_t$ allocated to current production, $\sigma$ needs to be sufficiently greater than one. The reason is that for smaller values of $\sigma$, an increase in technology $A_t$ affects the productivity of organizational capital production just as much as or even more than that of current productive activities. If we set $B_t = A_t$, then varying $\sigma$ leaves the impulse responses virtually unaltered, suggesting that it is not the degree of substitutability that is important, but the effect of changes in $A_t$ on the relative productivity
of current production over organizational investments. Thus, we need to think of a boom as a period of high relative productivity of current production. Organizational investments are no more productive in a boom than in a recession. Then, because capital and labor are overall more valuable, firms will substitute organizational investments for productive inputs in a boom and vice versa in a recession.

Next, we turn to the symmetry in the production function between labor and capital. In the simple model with only labor in Section 3, it was relatively straightforward to justify our nonseparable production technology in equations (7) and (8) as the smallest possible deviation from a standard separable production function as in (1). But in the full model with both labor and capital, as in (17) and (18), further assumptions were required, importantly the assumption that the capital and labor shares in current productive activities are the same as in organizational capital production. Relaxing this assumption, we replace equation (18) with

\[ L_t = (1 - \lambda) L_{t-1} + \tilde{\lambda}((1 - u_t) K_t)^{\frac{\rho}{\alpha}} ((1 - e_t) N_t)^{1 - \frac{\rho}{\alpha}}. \] (21)

We start from the extreme case that organizational capital production requires only labor, \( \alpha_L = 0 \), and gradually increase the capital share in organizational capital production. As we may have expected, the model is able to generate delay in investment only if it requires capital.

Finally, we explore the robustness of our results if we assume adjustment costs over gross rather than net hiring, \( H_t = h_t + \delta_N N_{t-1} \), and investment, \( I_t = i_t + \delta_K K_{t-1} \). In this case, the model is still able to generate delayed adjustment, see Table 3, and the amount of delay is only slightly lower.

6. EVIDENCE AND IMPLICATIONS

We showed that an otherwise standard macroeconomic model with a nonseparable production technology can match the persistence in hiring and investment observed in the data, because the nonseparability creates an additional margin of adjustment that firms may use to increase factor inputs into current production by postponing other types of activities. The most direct evidence for this mechanism comes from a somewhat dated, survey by Fay and Medoff (1985). In this survey, plant managers were asked how many workers they had been forced to let go in the previous recession, and how many they could have fired while still meeting production requirements. The results showed that there was labor hoarding in the amount of 6% of workers, who were not needed during the recession but who had nevertheless not been laid off. Importantly, for this paper, the survey then asked managers to indicate how they employed these “extra” workers. The answers indicated that these workers were assigned to “other work,” including (in order of frequency) cleaning, painting, maintenance of equipment, equipment overhaul, or sent on training. In the context of our model, these types of “other work” can all be considered organizational tasks, because they do not
affect production immediately, but are likely to improve productivity in the longer run. Our calibrated model predicts in response to a negative 1% shock an increase in $e_t$ of 2%-points (a 9.9% increase from 21% in steady state to 23%). A typical recession is a shock of about three times that size, so that the prediction of our model lines up well with Fay and Medoff (1985)’s estimate of 6% of workers.

We model nonseparabilities production as an additional state variable, which we label organizational capital, because there is evidence that organization is important in economics: for the existence of firms (Prescott and Visscher 1980), to explain the large drop in output in the transition from a planned to a market economy (Blanchard and Kremer 1997), for understanding the link between information technology and skill-biased technological change (Brynjolfsson and Hitt 2000), for stock market value (Lev, Radhakrishnan, and Zhang 2009, Hall 2000a), and for asset returns (Eisfeldt and Papanikolaou 2013), and it is often meaningful to think of organization as a stock of “capital” that positively affects productivity (Hall 2000b). Organizational or intangible capital has also been shown to be important for measured productivity and business cycles (McGrattan and Prescott 2010, McGrattan and Prescott 2012, McGrattan 2017), for optimal taxation (Conesa and Domínguez 2013, Conesa and Domínguez 2018), and for the rise in the relative volatility of labor market variables (Mitra 2019). This offers further opportunities for testing the model, building on a literature trying to measure organizational capital.

An ideal test of our explanation would compare the predictions of our model for the dynamics of (investments in) organizational capital directly to the data. This requires good estimates of organizational capital at sufficiently high frequency and over a sufficiently long time period. Since such an idea test is not feasible due to lack of data, we try to build our case based on a compendium of indirect evidence. In Subsection 6.1, we discuss some of the attempts to measure organizational capital, and show that the estimated share of organization in production is roughly in line with what our model needs to match the data on persistence in hiring and investment. Section 6.2 uses capacity utilization as an observable proxy for allocation of labor and capital to production versus organization, and shows that its dynamics are consistent with the dynamics predicted by our model. Finally, in Subsection 6.3, we attempt a test of our mechanism by checking whether differences in organizational capital line up with differences in persistence across industries.

6.1 Organizational Capital Share in Production

Although measuring organizational capital is far from straightforward (Lev, Radhakrishnan, and Zhang 2009), the literature has made a number of strong attempts. Atkeson and Kehoe (2005) estimate a structural model based on Prescott and Visscher (1980) and find that 8% of output is due to intangibles. Hall (2000a) uses a weight of 9% for e-capital in production and find that accumulation of e-capital contributed 15% to productivity growth over the 1990–98 period. Black and Lynch (2005) argue that employer-provided training is an important component of organizational investments and more easily measured, and find that 30% of output growth
is due to “workplace practices,” mostly training. The sources-of-growth analysis by Corrado, Hulten, and Sichel (2009) considers investments in IT and training, but also research and development (R&D) and advertising and find an income share of 15% due to intangibles in 2000–03, with growth in the share of intangibles contributing 27% to growth in labor productivity from 1995 to 2003. Finally, Squicciarini and Mouel (2012) develop a measure of organizational investments in organization by using O*Net to identify occupations, in which workers perform tasks that are classified as organizational: organizing, planning, and prioritizing work; building teams, matching employees to tasks, and providing training; supervising and coordinating activities; and communicating across and within groups. They find that over 20% of employees work primarily on organizational tasks and double the estimates used in Corrado, Hulten, and Sichel (2009).

There seems to be broad consensus in this empirical literature that the share of organizational capital in output is somewhere between 8% and 20%, and that accumulation of organizational capital accounts for a much larger contribution to growth in output and productivity. We find that to match the observed persistence in hiring and investment with our model, we need to assume that 21% of labor and capital are being used for organizational tasks in steady state, well in line with these estimates. Since we did not target the organizational capital share, nor any of the series that are used to estimate it, but instead calibrated it to the response of hiring and investment, we interpret this as evidence in favor of our model.

6.2 The Dynamics of Factor Input Allocation

Our model has strong predictions for the dynamics of factor input allocation. Real-locating workers and capital services from current production to organization acts as an intensive margin of adjustment that makes it possible for firms to adjust labor and capital. As a consequence, we would expect factor allocation to respond immediately when a shock hits the firm, and the response of \( e_t \) and \( u_t \) should not show a hump shape. To test this prediction, we need an empirical counterpart of \( e_t \) or \( u_t \).

There is little direct evidence on the allocation of workers or capital within a firm, beyond the one-time survey by Fay and Medoff (1985). Empirical measures of organizational investment are of limited use as well, because they are available at best an annual frequency and for relatively short periods. We argue that capacity utilization is a good measure for \( e_t \) and \( u_t \), as it measures changes in (current) output that cannot be explained by changes in factor inputs. Basu, Fernald, and Kimball (2006) argue that changes in hours per worker are a good proxy for changes in both labor effort and capital utilization, and Fernald (2012) provides a long quarterly time series for capacity utilization based on this idea, which we use to test the predictions of our model for the dynamics of \( e_t \) and \( u_t \).

5. Alternative proxies we considered are effort (Shea 1990) and skill acquisition (DeJong and Ingram 2001, Dellas and Sakellaris 2003). While the cyclical of these measures is consistent with our model as well, the data are annual, which makes it difficult to estimate the dynamics precisely.
In Figure 5, we show the result of the same MA regression on productivity for capacity utilization as we showed for hiring and investment in Figure 1. The response of utilization to changes in the economy is immediate, without evidence for a delayed response as for hiring or investment, consistent with the predictions of our model. This is further evidence in favor of the mechanism proposed in this paper.

6.3 Cross-Industry Evidence

The response of sectoral investment to macroeconomic shocks is hump-shaped, just as in aggregate data (Zorn 2016). This finding implies that the delayed response of investment in aggregate data is not due to a composition effect but to a mechanism that operates within-industries. Therefore, we can use the variation in the response of hiring (and investment) across industries to provide some further evidence for the mechanism we propose in this paper.

We explore to what extent the response of hiring and investment to shocks across industries is correlated with various measures of organizational capital intensity. Our model predicts that adjustment of employment and capital in industries with a higher share of organizational capital should exhibit more delay. A range of measures of organizational or intangible capital intensity is available for the United States at the
industry level, although at a relatively high level of aggregation: data on information capital intensity as suggested by Brynjolfsson, Hitt, and Yang (2002) and provided by the Bureau of Labor Statistics (2019a); data on intangible capital, organizational capital, and training intensity constructed using the perpetual inventory method from a broad range of investments, including things that are usually treated as intermediate costs in the NIPA, from INTAN-Invest (Carol Corrado 2016); a task-based measure of organizational investments produced by Squicciarini and Mouel (2012); data on e-capital from Hall (2000a); and data on employer-provided training as suggested by Black and Lynch (2005) as a measure for organizational capital and provided by the Bureau of Labor Statistics (2019b). We match these data to measures of delay in hiring and investment calculated from the US KLEMS (Bureau of Labor Statistics 2019a), see Online Appendix E for a more detailed description of the data and the measures for delay and organizational capital intensity.

The correlations between delay in hiring and organizational capital intensity we find tend to be positive, ranging from 0.7 for the percentage of workers who received formal training provided by their employer over the past year to zero for e-capital intensity, see Online Appendix E. Unfortunately, the number of industries at which the measures of organizational capital intensity are provided is too low (between 8 and 28) to estimate these correlations with any reasonable degree of certainty. We conclude that the cross-industry evidence is at least not inconsistent with the explanation for delayed adjustment proposed in this paper.

7. CONCLUSIONS

We offered an explanation for the hump-shaped impulse responses in hiring and investment in U.S. data that rely on nonseparabilities in production in combination with standard adjustment costs in labor and capital. A nonseparable production technology means that firms can intertemporally substitute labor and capital, allowing them to adjust factor inputs without the need for hiring and investment or firing and disinvestment. In combination with adjustment costs in labor and capital, this new intensive margin of adjustment generates an incentive to postpone hiring and investment in response to a shock, a feature of the model which we labeled delayed adjustment. Delayed adjustment in our model is endogenous, that is, adjustment eventually happens in response to a single shock and does not require a specific sequence of shocks, nor does it depend on the specific type of adjustment costs (nonconvex or convex). We discussed some evidence that the organizational capital share in production the model needs to match the persistence in hiring and investment observed in the data is consistent with empirical estimates of organizational and intangible capital.

Compared to the early literature on propagation (Cogley and Nason 1995, Rotemberg and Woodford 1996), we draw a sharp distinction between the persistence in stock and flow variables, arguing that although adjustment costs may explain persistence and hump-shaped responses in the stocks (capital and employment),
they cannot by themselves account for persistence in the flows (investment and hiring). This is the same observation that led Christiano (2011) to dismiss adjustment costs in capital as a “failed approach.” Following Christiano, Eichenbaum, and Evans (2005), the literature has addressed the problem by assuming adjustment costs in the change in investment rather than in capital, that is, \( g(i_t, i_{t-1}) = \frac{1}{2} \psi(i_t / i_{t-1})^2 \) instead of \( g(i_t) = \frac{1}{2} \psi i_t^2 \), see, for example, Christiano, Eichenbaum, and Trabandt (2018). We show that with a reasonably calibrated nonseparable production technology, a model with standard adjustment costs generates impulse responses that are very similar to a model with cost-of-change adjustment costs.

LITERATURE CITED


Corrado, Carol, Cecilia Jona-Lasinio, Massimiliano Iommi, and Jonathan E. Haskel. (2016) “Intangible Investment in the EU and US before and since the Great Recession and Its


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1