ONLINE APPENDIX

A Data Appendix

A.1 Data sources and treatment

We use the following data throughout the paper. Our main sample period used for estimation purposes is the one for which we can get data on all variables simultaneously, which is 1990Q2 to 2018Q4.

Business Dynamics Statistics (BDS): We use the BDS to construct data stratified by firm age. We use the 2018 release, which is available yearly from 1978 to 2018, and take the national data, split by firm age. We use data on the number of firms (and total employment in firms) of each age bin to calibrate our model. We also measure firm entry using this dataset, as the number of firms aged 0 in each year, and firm exit, as given in the dataset. We also use the data on job creation and job destruction by age, but instead calibrate our model to the quarterly JC and JD data from Davis, et al. (2012).

Davis, Faberman, and Haltiwanger (2012, DFH): We extensively use the data provided by Davis et al. (2012) which underlies the analysis in their paper. We are grateful to the authors for sharing the (updated) data which underlies the plots in their paper. This data set consists of quarterly data from 1990Q2 to 2018Q4. We use their estimates of aggregate job creation and job destruction, as well as layoffs and quits. We add their measure of "other separations" to layoffs. Their data are given as rates of total employment. We calibrate our model to match the average of these data over our sample, as well as the HP-filtered time series. In addition, we use data from this paper to form an estimate of the fraction of worker quits that firms replace by undertaking a replacement hire. We discuss this further below.

Bureau of Labor Statistics (BLS): We use monthly aggregate data from the Current Population Survey. We aggregate these data up to a quarterly frequency by taking the simple average. We use data on total employment (CE16OV) and unemployment (UNEMPLOY), in levels and seasonally adjusted. We additionally use data on the total number of people unemployed for less than five weeks (UEMPLT5) to construct the unemployed worker UE rate, following the approach of Shimer (2005).

Bureau of Economic Analysis (BEA): To construct our measure of labour productivity (output per worker) we use data on quarterly real GDP (GDPC1) from the BEA. Labour productivity is calculated as real GDP divided by total employment from the BLS data.

Compustat: Since our model assumes constant returns to scale, we do not allow for permanent productivity differences across firms, as these would lead to permanent differences in employment growth rates (rather than levels, as would be true in a model with decreasing returns to scale). Therefore, we calibrate our firm-level productivity process to the within-firm standard deviation of productivity shocks, rather than the across firm standard deviation. To compute this measure, we use data from Compustat. We use data on all US based firms in their sample, and use data only on sales and total employment. We deflate sales using the GDP deflator (GDPDEF, from the BEA) to create a measure of real sales, and then define firm-level labour productivity each year as real sales over employment. We drop firm-year observations with missing or negative sales or employment, and winsorize the data by dropping the top and bottom 1% of data by both yearly sales growth and employment growth. We take the log of labour productivity, and regress it on firm and year fixed effects, and take the residual as our measure of firm-level productivity, corrected for firm-level averages and aggregate changes. We take the standard deviation of this measure, which yields a value of 28.38%, computed from 291,703 firm-year observations.

A.2 Estimating the fraction of quits which are replaced

To estimate the amount of replacement hiring in the model, we draw information both from gross flows and from underlying firm-level data. Firstly, we note that the amount of replacement hiring is not simple to observe from aggregate flows, due to the fact that firms may do replacement hiring for two reasons: either to replace workers who quit, or to replace workers they lay off for being a bad match but where the firm wants to keep the job open. Through the lens of the model, the total hiring rate is equal to

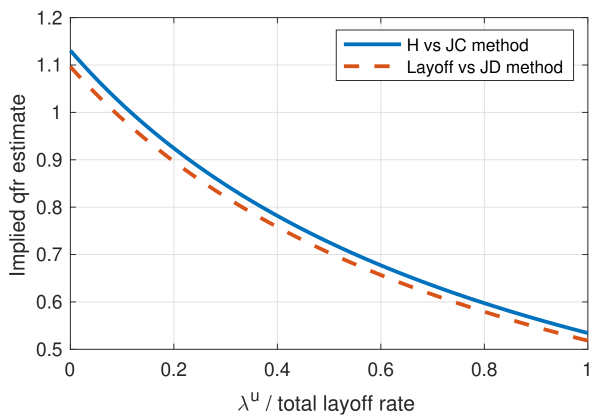
$$h_t = jc_t + qfr_t\,(q_t + \lambda^u) \tag{1}$$

where we define qfr_t as the fraction of worker quits (and layoffs due to bad worker match) which are replaced. Recall that λ^u is the rate at which workers are fired for being a bad match, but where the firm's capital remains intact so the firm has the option to hire to replace them. qfr serves as our calibration target for the amount of replacement hiring in the model. Notice that three objects in this equation are observable in the DFH dataset: hiring (h_t), job creation (jc_t), and quits (q_t). If we assumed that all layoffs were due to job destruction (and firms never fired workers with the aim of replacing them with another worker) then $\lambda^u = 0$, and estimating the degree of replacement hiring would be simple using this aggregate data alone. In this case, simply rearrange (1) to yield $qfr_t = (h_t - jc_t)/q_t$.

However, the data in DFH suggest that firms do indeed replace some of their workers who leave due to layoff, so this approach is likely not valid. In particular, in their well-known "hockeystick" plot (their Figure 7(b)) we observe that firms who have positive employment growth, and hence are expanding, still extensively use worker layoffs. In fact, our calculations below suggest that the average layoff rate for non-contracting firms appears to be around 2.73% per quarter. Given that these firms are expanding their employment on net, it is likely that they are replacing some of the workers who they have laid off, meaning that $\lambda^u > 0$. Indeed, through the lens of our model, firms which perform job creation necessarily replace worker quits.

Given that $\lambda^u > 0$ therefore seems like a more reasonable assumption, we can then return to (1) to understand the impact this has on estimated quit replacement. Expressing the relationship in steady state gives $h = jc + qfr\,(q + \lambda^u)$, where we take h, jc, and q directly from the average values in DFH's data. This gives the implied value of qfr for any assumed λ^u as $qfr = (h - jc)/(q + \lambda^u)$. Without any further information, λ^u is constrained to lie within the range 0 (in which case all layoffs are due to job destruction) and the total layoff rate in the data, l (in which case all layoffs are replaced, and not due to job destruction) [see equation (3) below]. We plot the implied value of qfr in this range in the blue line in Figure A.1 below.

Figure A.1: Estimated fraction of quits replaced vs. assumed λ^u



This procedure bounds the fraction of quits and layoffs which are replaced to be between around 50%, if all layoffs are assumed to be replaced, and 100%, if only 10% of layoffs are assumed to be replaced. Notice that for values below 10% the data implies that more than 100% of quits are replaced. Before going further, we therefore note that our chosen value for the estimation, $qfr = 80\%$, happens to lie approximately in the middle of the upper and lower bounds implied by the aggregate flow data.

One could potentially use other aggregate flow relationships, such as those between job destruction and layoffs, might help estimate the fraction of layoffs which are replaced, and hence pin down qfr. However, we found this challenging as the aggregate relationships are by definition collinear, and more information is needed. To see this, consider that job destruction is given by

$$jd_t = jd_t^l + jd_t^q = jd_t^l + (1 - qfr_t)(q_t + \lambda^u) \tag{2}$$

and layoffs by

$$l_t = jd_t^l + \lambda^u \tag{3}$$

where l_t is layoffs in the data, jd_t^l is job destruction shocks which induce layoffs, and jd_t^q is job destruction due to unreplaced quits and layoffs. Taking jc, jd, q, and l as data, (1), (2), and (3) appear to provide three equations which can solve for the three unknowns qfr_t, λ^u, and jd_t^l. However, the equations actually contain the same economic content and are collinear. Combining (2) and (3) to yield

$$qfr_t(q_t + \lambda^u) = l_t - jd_t + q_t \tag{4}$$

and rearrange (1) to yield

$$qfr_t(q_t + \lambda^u) = h_t - jc_t. \tag{5}$$

As the left hand sides of these equations are identical, the three equations together cannot be solved for a unique solution for qfr_t, λ^u, and jd_t^l. Instead, combining these two equations implies an adding-up condition which should hold in the data in theory: $jc_t - jd_t = h_t - l_t - q_t$. In practice, the adding up condition is very slightly violated, meaning that (4) and (5) provide very slightly different estimates of qfr_t for a given assumed λ^u. The estimate from (4) is given in Figure A.1 as the dashed red line, which is very similar the the previous estimate.

The discussion above shows that additional data must be included to estimate the fraction of quits which are replaced, and we investigate two approaches.

As a first approach, note that with knowledge of λ^u, the value of qfr can be calculated using the accounting relationship above. λ^u is the rate at which workers are fired for being a bad match, with the firm having the option to replace them if desired. Through the lens of the model, this can be identified as the layoff rate at expanding firms, who perform no job destruction and so any layoffs must be due to the λ^u shock. To estimate this in the data, we use the hockeystick and growth rate distribution plots in DFH.[29] We have access only to the growth rate distributions from 2006 and 2008-9 (as plotted in DFH) and so use data for the distribution and hockeysticks from 2006 to construct our estimate. Accordingly, this is data from a single year, which corresponds to an estimate for a typical non-recession year.[30] For each growth rate bin $i = -200, 199, ..., 200$, we have data on the

[29]The authors very kindly provided us with the data behind these plots, which consists of the hires, layoff, total separation, and quit rate at each growth rate bin (their Figures 6 and 8), and the (employment weighted) kernal density function of firms in each bin (their Figure 5). The data are provided on slightly different grids for each plot, and we interpolate the data onto an integer grid from -200 to 200.

[30]The results are robust to using the hockeysticks from all years (their Figure 6) integrated using the average of the 2006 and 2008-9 growth rate distributions to roughly attempt to form an estimate for all years. However, since the match of hockeysticks and growth rate distribution sample is not exact in this case, we prefer to use the data from 2006 only.

mass of employment at establishments with that growth rate (d_i) and the construct a layoff rate at that bin (l_i). We calculate the average layoff rate in all bins with non-declining employment growth as $\sum_{i=0}^{200} l_i d_i / \sum_{i=0}^{200} d_i = 2.73\%$. Under the identifying assumption that λ^u is constant across firms (as it is in the model), this implies an estimate $\lambda^u = 2.73\%$, which is 39% of the average layoff rate of 7.0% in the DFH data in 2006. Referring back to Figure A.1, we see that 39% of layoffs being potentially replaceable implies a value of qfr of approximately 80%, which is the value used in our calibration.[31] As an alternative, we also directly calculated the fraction of quits replaced from the 2006 hockeystick data for quits and hires, and found that 79.6% of quits were replaced. Specifically, we assume that expanding firm bins replace all quits and layoffs ($qfr_i = 1$). For contracting bins, we calculated the fraction of quits replaced as $qfr_i = h_i/(q_i + \lambda^u)$. Taking the $(q_i + \lambda^u)$-weighted average of qfr_i across the whole distribution yields 79.6%.

As a second approach, we consider the notion of replacement hiring in Elsby et al. (2021). They define a broad notion of replacement hiring using JOLTS data as follows. For each establishment, they consider replacement hires as the minimum of gross hires and quits in a given quarter. They then sum across establishments, and find that, by this definition, around 45% of all hires are replacement hires. Doing the same exercise on simulated data from our model finds that 40% of hires are replacement hires. As mentioned in the text, our model also generates that around 50% of firms have zero net employment change over 3 months, and since these firms also lose workers to quits, this serves as another measure of replacement hiring. Elsby et al. (2021) find this number to be around 55% and 65% in the QCEW and JOLTS data respectively. Finally, their strictest measure of replacement hiring is the total hiring at firms with zero net employment change as a fraction of total hiring. This number is 7.5% in their data, and 7.1% in our model. By all these measures our model generates a substantial amount of replacement hiring, close to the measures in the data. This provides an alternative justification for our calibrated value of $qfr = 0.8$, which delivers a sensible amount of replacement hiring by these alternative measures, and suggests that our results would be robust to instead using these measures as targets in our estimation.

A.3 Procyclicality of JD: robustness to alternative assumptions and data sources

In this section we discuss the correlations between JD, layoffs, and other variables with unemployment in more detail. In the Introduction, we argued that the overall JD rate is procyclical (positive correlation with unemployment) while the layoff rate is countercyclical. Highlighting this difference is, to our knowledge, is novel. For this reason, we provide extensive robustness showing that this

[31] If roughly 40% of layoffs are replaceable and the replacement rate is 80%, this implies that 32% of layoffs are actually replaced.

difference in cyclicality between JD and layoffs is robust. We first discuss the robustness of the result in relation to many different data treatments, still using our main DFH (Davis et al. 2012) dataset. We then show that the procyclicality of JD also holds on a different dataset, the yearly BDS dataset.

Robustness on main DFH dataset We provide an exhaustive battery of checks to show that no particular data treatment assumptions are driving our results. These are presented in the series of tables below, which we discuss in turn.

Firstly, our findings are robust to various detrending assumptions. In Table 3 we plot the correlation of job destruction (jd), layoffs (lo), job destruction net of layoffs ($jd - lo$), and job creation (jc) with the level of unemployment. We compute the correlation of unemployment at time t with the labour market flows between t and $t+1$ (as we also do in the model) so these correlations show the correlation between unemployment and how labour market flows then evolve in the next three months. Our baseline results detrend all variables using the HP-filter with parameter 10^5, following Shimer (2005), as shown in the first column. All of the flows are procyclical, apart from layoffs. The remaining columns apply different detrending methods: HP-filter with parameter 1600, Baxter-King approximate bandpass filter (frequencies 6 to 32 quarters), linear detrending, and finally the raw data without detrending. The correlations all maintain the same signs across all different detrending assumptions. The only exception is layoffs when not detrended, which now also become procyclical. Inspecting the raw data reveals that this is because layoffs have been trending downwards over time which creates a spurious correlation with unemployment, which highlights the importance of detrending data to deal with the well know slowing down of many labour market flows in recent decades.

Table 3: Cyclical correlations of JD, layoffs, and JC with unemployment

	HP 10^5	HP 1600	BK	Linear	Raw data
jd	-0.226	-0.170	-0.135	-0.228	-0.387
lo	0.143	0.156	0.211	0.164	-0.227
$jd - lo$	-0.607	-0.521	-0.700	-0.596	-0.457
jc	-0.270	-0.073	-0.068	-0.401	-0.446

Note: Each row gives the correlation of the variable with unemployment. Each column detrends the data with a different method before computing the correlations, with HP indicating the HP-filter with a given parameter, BK the Baxter-King approximate bandpass filter (frequencies 6 to 32 quarters), Linear indicating linear detrending, and Raw data the data without detrending. JC, JD, and layoff rates are from the quarterly data used by Davis, Faberman and Haltiwanger (2012), updated by these authors. Unemployment is the number of unemployed individuals using quarterly data from the Current Population Survey. The data are from 1990Q2 to 2018Q4, and we compute the correlation between unemployment at date t with the flows (e.g. for JD) between t and $t + 1$.

Secondly, we stress the important distinction between the correlation of flows with the *level* of unemployment and the *change* in unemployment. The two approaches give different results, as shown and discussed by Moscarini and Postel-Vinay (2012) and Haltiwanger et al. (2018). Our empirical and model findings are robust to this distinction, because these correlations measure different concepts, which our model helps clarify. In Table 4 we give the correlations with the change in unemployment. This shows that job destruction is positively correlated with the change in unemployment, despite being negatively correlated with the level of unemployment. This does not cast doubt on the results presented in levels, but actually highlights the different cyclical features that the level and change in unemployment capture.

To understand this further, consider that in our model the change in unemployment is by construction driven by JC and JD as

$$\dot{U}_t = JD_t - JC_t \tag{6}$$

Thus, it is very natural that JD must be positively correlated with the change in unemployment, because (6) shows that an increase in JD is literally the mechanical driver of an increase in $\dot{U}$. In our model, it is also true that JD and layoffs (JC) are positively (negatively) correlated with the change in unemployment, and hence our model is simultaneously consistent with the correlations of JD, layoffs, and JC with *both* the level and change in unemployment. Specifically, we compute log HP-filtered data as we did for our main results in Table 2. The correlation of each variable with the first difference of unemployment in the data (model) is -0.5247 (-0.6789) for JC; 0.7302 (0.4598) for JD; and 0.6611 (0.7536) for layoffs.

Table 4: Cyclical correlations of JD, layoffs, and JC with Δu

	HP 10^5	HP 1600	BK	Linear	Raw data
jd	0.759	0.726	0.817	0.742	0.488
lo	0.721	0.699	0.746	0.734	0.459
$jd - lo$	0.147	0.073	0.045	0.038	-0.005
jc	-0.494	-0.417	-0.747	-0.536	-0.056

Note: Each row gives the correlation of the variable with the change in unemployment. Each column detrends the data with a different method before computing the correlations, with HP indicating the HP-filter with a given parameter, BK the Baxter-King approximate bandpass filter (frequencies 6 to 32 quarters), Linear indicating linear detrending, and Raw data the data without detrending. JC, JD, and layoff rates are from the quarterly data used by Davis, Faberman and Haltiwanger (2012), updated by these authors. Unemployment is the number of unemployed individuals using quarterly data from the Current Population Survey. The data are from 1990Q2 to 2018Q4, and we compute the correlation between unemployment change ($u_{t+1} - u_t$) with the flows (e.g. for JD) between t and $t + 1$.

At the same time, how can JD be negatively correlated with U while being positively correlated with ΔU? This is because the two correlations capture different frequencies of the data. The $(JD, \Delta U)$ correlation captures the very short term correlation between high JD and rising U which is driven by *layoffs*. The (JD, U) correlation instead captures the longer term correlation between JD and the level of unemployment, which is driven by *unreplaced quits*. Essentially, JD is made up by a fast moving component (layoffs) and by a slow moving component (unreplaced quits) with different correlations with unemployment, and the $(JD, \Delta U)$ and (JD, U) correlations have opposite signs because each picks up a different component respectively.

Thirdly, the correlations are robust to whether data are expressed in levels or logs. In Table 5 we repeat the correlations this time taking the logs of both the flows and the unemployment level. Since $jd - lo$ sometimes takes negative values it cannot be logged, and so is excluded from this exercise. The correlations retain the same signs and properties as the data in levels in Table 3.

Table 5: Cyclical correlations of JD, layoffs, and JC with unemployment (all logged)

	HP 10^5	HP 1600	BK	Linear	Raw data
$\log jd$	-0.254	-0.113	-0.058	-0.262	-0.389
$\log lo$	0.174	0.230	0.301	0.175	-0.211
$-$	0	0	0	0	0
$\log jc$	-0.305	-0.123	-0.159	-0.433	-0.447

Note: Each row gives the correlation of the log of variable with log unemployment. Each column detrends the data with a different method before computing the correlations, with HP indicating the HP-filter with a given parameter, BK the Baxter-King approximate bandpass filter (frequencies 6 to 32 quarters), Linear indicating linear detrending, and Raw data the data without detrending. JC, JD, and layoff rates are from the quarterly data used by Davis, Faberman and Haltiwanger (2012), updated by these authors. Unemployment is the number of unemployed individuals using quarterly data from the Current Population Survey. The data are from 1990Q2 to 2018Q4, and we compute the correlation between unemployment at date t with the flows (e.g. for JD) between t and $t + 1$.

Fourthly, an important data issue is how to treat "Other Separations" in the DFH dataset. Separations in their dataset are split into layoffs, quits, and other separations. Since other separations are classified neither as layoffs or quits, which make up the only two types of separation in our model, we must deal with this extra category somehow. In practice, this does not affect the results because other separations is a relatively small category. In our main dataset, we add other separations to layoffs and hence treat them as layoffs rather than quits. In Table 6 we instead add other separations to quits, and show that the correlations all retain the same signs as in the main specification in Table 3.

Table 6: Cyclical correlations with unemployment (alternative treatment of other separations)

	HP 10^5	HP 1600	BK	Linear	Raw data
jd	-0.226	-0.170	-0.135	-0.228	-0.387
lo	0.205	0.186	0.229	0.240	-0.184
$jd - lo$	-0.637	-0.527	-0.651	-0.652	-0.604
jc	-0.270	-0.073	-0.068	-0.401	-0.446

Note: Each row gives the correlation of the variable with unemployment. Each column detrends the data with a different method before computing the correlations, with HP indicating the HP-filter with a given parameter, BK the Baxter-King approximate bandpass filter (frequencies 6 to 32 quarters), Linear indicating linear detrending, and Raw data the data without detrending. JC, JD, and layoff rates are from the quarterly data used by Davis, Faberman and Haltiwanger (2012), updated by these authors. Unemployment is the number of unemployed individuals using quarterly data from the Current Population Survey. The data are from 1990Q2 to 2018Q4, and we compute the correlation between unemployment at date t with the flows (e.g. for JD) between t and $t + 1$.

Finally, we consider the robustness our our results to an alternative subsample of the data. Inspecting the time series for JD in Figure 2, we see that the period following the Great Recession exhibits the clearest negative correlation between JD and unemployment, and so one might worry that the post-Great Recession era is driving our results. It could be that the negative correlation is a feature novel to the Great Recession, which featured a very pronounced collapse in job-to-job quits. To investigate this, in Table 7 we repeat our correlations but excluding the post-Great Recession period, computing the correlations only up to 2010Q1. As can be seen in the table, all correlations maintain their same signs on this smaller subsample.

Overall, these various exercises indicate that the correlations we document are a robust feature of the data, and are not specific to any specific detrending and sampling assumptions.

Table 7: Cyclical correlations of JD, layoffs, and JC with unemployment (1990Q2 to 2010Q1)

	HP 10^5	HP 1600	BK	Linear	Raw data
jd	-0.075	-0.198	-0.082	-0.005	-0.167
lo	0.278	0.189	0.421	0.290	0.035
$jd - lo$	-0.613	-0.619	-0.750	-0.514	-0.518
jc	-0.346	-0.076	-0.320	-0.524	-0.434

Note: Each row gives the correlation of the variable with unemployment. Each column detrends the data with a different method before computing the correlations, with HP indicating the HP-filter with a given parameter, BK the Baxter-King approximate bandpass filter (frequencies 6 to 32 quarters), Linear indicating linear detrending, and Raw data the data without detrending. JC, JD, and layoff rates are from the quarterly data used by Davis, Faberman and Haltiwanger (2012), updated by these authors. Unemployment is the number of unemployed individuals using quarterly data from the Current Population Survey. The data are from 1990Q2 to 2010Q1, and we compute the correlation between unemployment at date t with the flows (e.g. for JD) between t and $t+1$.

Robustness to different dataset: BDS To further check the robustness of the correlation between JD and unemployment, we show that it holds on a different dataset. We turn to the BDS dataset, which is the yearly dataset we used for our firm age distribution calculations. This dataset features three differences from the DFH data. Firstly, it has a longer sample period, from and we use data from 1978 to 2018. Secondly, since it is yearly it measures the yearly JD and JC flows, which may in principle be different from the quarterly flows in DFH. Finally, it does not measure layoffs, and so we can only check the correlations of JD and JC with unemployment. Nonetheless, we identify the same patterns on this dataset.

We compute the correlations of the aggregate JD and JC rates with unemployment, where we calculate unemployment as the average unemployment within each year from the BLS data. We log and HP-filter all data with a HP-filter parameter of 100. Given that the dataset is yearly, it is very important to consider timing, and we correlate unemployment in year t with the JC and JD flows between year t and $t+1$, to capture the correlation of unemployment with the rates going forward, as we do in the model. We find that both JC and JD are procyclical, just as in the quarterly DFH data: their correlations with unemployment in the yearly BDS data are -0.1035 and -0.1952 respectively. Hence, the procyclicality of JD appears to be a robust feature of both the quarterly and yearly data.

B Model Appendix

B.1 Proofs

Proof of Lemma 1: For convenience we suppress reference to Ω. It is immediate that $q(i) = \lambda_u + \lambda_1$ for $i \leq i^h$. Consider now $i > i^h$ where equation (16) in the text implies equilibrium quit rate

$$q(i) = \lambda_u + \frac{\lambda_1}{\lambda}\int_i^1 \frac{h(j)[1-U]G'(j)}{\alpha + (1-\alpha)G(j)}dj.$$

Now $\lambda_1 = \phi\lambda_0$ and $\lambda = \lambda_0 U + \lambda_1(1-U)$ implies $\lambda_1/\lambda = \phi/[U + \phi(1-U)]$ while $\alpha \equiv \lambda_0 U/\lambda = U/[U + \phi(1-U)]$. Substituting out λ_1/λ and α in the above yields

$$q(i) = \lambda_u + \int_i^1 \frac{h(j)G'(j)}{\frac{U}{\phi(1-U)} + G(j)}dj.$$

Because $h(j) = JC(j) + q(j)$ for $j \geq i > i^h$ we also have

$$h'(j) = JC'(j) - \frac{h(j)G'(j)}{\frac{U}{\phi(1-U)} + G(j)}.$$

Now define $Z(j) = \frac{U}{\phi(1-U)} + G(j)$ and so $Z'(j) = G'(j)$. Integration by parts establishes:

$$\int_i^1 Z'(j)h(j)dj = [Z(j)h(j)]_i^1 - \int_i^1 Z(j)\left[JC'(j) - \frac{h(j)G'(j)}{\frac{U}{\phi(1-U)} + G(j)}\right]dj$$

and simplifying yields:

$$Z(1)h(1) - Z(i)h(i) = \int_i^1 Z(j)JC'(j)dj.$$

Integrating by parts then yields:

$$Z(1)h(1) - Z(i)h(i) = Z(1)JC(1) - Z(i)JC(i) - \int_i^1 Z'(j)JC(j)dj.$$

Substituting out $h(1) = \lambda_u + JC(1)$, $h(i) = JC(i) + q(i)$ implies:

$$\begin{aligned} Z(i)q(i) &= Z(1)\lambda_u + \int_i^1 Z'(j)JC(j)dj \\ &= Z(i)\lambda_u + \int_i^1 Z'(j)[\lambda_u + JC(j)]dj. \end{aligned}$$

Using $Z(i) = \frac{U}{\phi(1-U)} + G(i)$ and $Z'(j) = G'(j)$ then establishes the Lemma.

Proof of Proposition 2: Note these wage strategies are firm size invariant and have the job ladder property. By construction, beliefs B1-B3 are consistent with these equilibrium wage strategies, Bayes Rule and the restriction to monotone beliefs. Because these wage strategies have the job ladder property, the worker quit strategies described in equation (15) in the main text are optimal with the distribution of wage offers $F(.)$ given by equation (16) in the main text. All that remains is to show is these wage strategies are indeed optimal. It is easy to verify that firms $i < i^h$ (who have $v < c_0$) strictly prefer to post wage $w' = w_{\min}$ [paying a higher wage reduces the employee quit rate but the firm's gain by doing so $v < c_0$ and this wage deviation is profit reducing]. Conversely each hiring firm is actually indifferent to posting any wage $w' \in [w_{\min}, \overline{w}]$ while posting wage $w' > \overline{w}$ is strictly profit decreasing [because quit rates cannot fall further]. Hence each firm's wage strategy is indeed privately optimal which completes the proof of Proposition 2.

B.2 Entry with more than one worker

In this section we show how to extend the model in the main text to allow a new firm to start with more than one employee, but $n_0 \in \{1, 2, .., N_0 + 1\}$ employees. This version of the model is the one used in the calibration. The reason for this extension is that in the data we observe that the majority of new firms have more than one employee in their first year. This extension allows the model to be closer to the data in this dimension as well for allowing start-up size to be potentially sensitive to the aggregate state Ω_t, while keeping the rest of its formulation as in the main text.

In particular, assume there is a unit measure of entrepreneurs who independently seek business ventures. At rate μ_0 an entrepreneur identifies a possible business venture whose investment cost $c^E \geq 0$ is considered an independent random draw from cost distribution $H^E(.)$. If the entrepreneur chooses not to invest, the venture is lost with no recall. If the entrepreneur invests, a start-up is created with $n_u \in \mathbb{N}^+$ employees drawn randomly from the pool of unemployed workers. Its productivity $i \sim U[0, 1]$ is then revealed at which point we refer to the start-up as a new firm. The new firm thus has n_u employees, is in state i, and subsequently pays wages and expands/contracts like all other existing firms. Each new firm also has N_0 immediate potential expansion job opportunities where $N_0 \in \mathbb{N}^+$ is exogenous. The job creation process is the same as for existing firms: associated with each potential expansion is an independent cost draw $c^{JC} \sim H^{JC}$ and recruitment cost c_0 to hire a worker. If the new firm invests its initial size n_0 increases by 1. If the new firm does not invest, the expansion opportunity is lost with no recall. Hence each new firm begins life with initial employment $n_0 = n_u + \widetilde{n}_i$ where hires $\widetilde{n}_i$ are a binomially distributed random variable with N_0 independent trials and an [endogenous] probability of investment which depends on (i, Ω).

Note that by allowing new firms to start with $n_0 > 1$ we need to amend the definition of equilibrium by adding: $P^E(\Omega)$ is the probability an entrepreneur invests in a start-up in state Ω and, given realised i, its starting size $n_0 = n_u + \widetilde{n}(i, \Omega)$ maximises expected profit. We now turn to derive $P^E(\Omega)$.

Firm Optimality [New Firms] Suppose in state Ω, an entrepreneur creates a new firm with revealed productivity $i \sim U[0,1]$. If $i < i^c$ the entrepreneur closes the firm [because $v(i,\Omega) < 0$]. If $i \in [i^c, i^h)$ the firm survives but the entrepreneur does not invest in new jobs [because $v(i,\Omega) < c_0$] and so initial firm size $n_0 = n_u$. For $i \geq i^h$, the entrepreneur invests in each expansion opportunity if and only if realised $c^{JC} \leq v(i,\Omega) - c_0$. Thus start-up employment $n_0 = n_u + \widetilde{n}(i,\Omega)$ where $\widetilde{n}(i,\Omega)$ is a binomially distributed random variable with expected value $N_0 H^{JC}(v(i,\Omega) - c_0)$. The expected value of a start-up is therefore

$$\Pi^{SU}(\Omega) = \int_{i^c}^{1} \left\{ n_u v(i,\Omega) + N_0 \int_0^{v(i,\Omega)-c_0} [v(i,\Omega) - c_0 - c'] dH^{JC}(c') \right\} di.$$

Hence given the investment opportunity, the entrepreneur proceeds with a new start-up when $c^E \leq \Pi^{SU}(\Omega)$ and so $P^E(\Omega) = H^E(\Pi^{SU}(\Omega))$. Initial firm size [in expectation] is then $En_0 = n_u + N_0 H^{JC}(v(i,\Omega) - c_0)$, noting that $H^{JC}(v(i,\Omega) - c_0) = 0$ for all $i < i^h$. Further, since for $i \geq i^h$ it holds that $H^{JC}(v(i,\Omega) - c_0) = jc(i,\Omega)/\mu_1$, then for these firms $En_0 = n_u + \frac{N_0}{\mu_1} jc(i,\Omega)$.

Job Creation and Aggregate Hires This extension also requires us to amend the expression for total hires in the main text. In particular, to calculate the total job creation flow in this case, by the definition of G, $[1-U]G'(i)jc(i,\Omega)$ describes the total job creation flow from all existing $i \geq i^h$ firms. Similarly the uniform distribution implies gross job creation flows (excluding the initial unemployed workers) at new $i \geq i^h$ firms is $\mu_0 P^E(\Omega) N_0 H^{JC}(v(i,\Omega) - c_0)$. Adding both flows together yields the total job creation flow (excluding the initial unemployed workers) by firm productivity

$$JC(i,\Omega) = \left\{ [1-U]G'(i) + \frac{\mu_0}{\mu_1} N_0 P^E(\Omega) \right\} jc(i,\Omega). \tag{7}$$

Finally, the modification to job creation flows implies that the expression for total hiring flows in the text is modified, since startups now hire more than one worker. In particular, equation (4) in the main text is replaced with

$$H(i,\Omega) = \begin{cases} 0 & \text{if } i < i^h \\ [1-U]G'(i)q(i,\Omega) + JC(i,\Omega) & \text{if } i \geq i^h, \end{cases} \tag{8}$$

where $JC(i,\Omega)$ now includes the hiring flow from incumbent firms and the hiring flow (excluding the first “free” hires from unemployment) of entrant firms. This modification has an important economic implication, since productive startup firms may now hire by poaching workers from other firms.

The remainder of the description of the model and its expressions remain unchanged.

B.3 Finite state space

Suppose finite productivity states so that firm productivity $p = p^{is}$ with $i \in \{1, 2, ..., I\}$. A very important property of the model is that the aggregate state now reduces to a finite vector, a result which holds even in the extended case with endogenous worker reservation wages $R = R(\Omega)$ and $\phi < 1$. This property does not arise in the standard Burdett and Mortensen (1998) framework, e.g. Coles (2001), Moscarini and Postel Vinay (2013) and Audoly (2020), and only holds in Coles and Mortensen (2016) for the special case $\phi = 1$.

The reason for the finite state space result is simple but subtle and reflects the replacement hiring process. Define N_i as the measure of workers employed in firms in state i, the employment vector $\underline{N} = (N_1, ..., N_I)$ where adding up implies unemployment $U = 1 - \sum_{i=1}^{I} N_i$. We now show the aggregate state reduces to vector $\Omega = (s, \underline{N})$ with corresponding vector of firm values $\underline{v}(\Omega) = \{v_i(\Omega)\}_{i=1}^{I}$, job creation rates $\{jc_i(\Omega)\}_{i=1}^{I}$ and so on as previously determined in the main text.

The first step is to extend the notation because firms in the same state i post different wages; i.e. there is equilibrium wage dispersion within each state i. The cleanest approach is to assume firms select wage strategies as follows: i) On start-up, a firm is allocated a wage rank $\chi \sim U[0, 1)$. In the stationary equilibrium, firm (i, χ, n, Ω) posts wage with rank χ in the firm i wage distribution. ii) On receiving a firm specific productivity shock with updated productivity i' the firm also updates to a new wage rank $\chi' \sim U[0, 1)$. Because all χ-wage strategies yield equal value, such wage selection is consistent with equilibrium. We choose this wage selection process because it guarantees first order stochastic dominance in wages, and so a worker will always quit to a higher wage offer.

To match to the notation in the main text, consider the following partition of line $[0, 1]$ into a grid $\{x_0, x_1, .., x_I\}$ where $x_0 = 0$, $x_i = x_{i-1} + \gamma_{0i}$ and $x_I = 1$. A firm in state $i \in \{1, 2, ..., I\}$ with wage rank $\chi \in [0, 1)$ is correspondingly defined as being in state $x \in [0, 1]$ where $x = x_{i-1} + \chi[x_i - x_{i-1}]$. Each start-up is then equivalently defined as having initial state $x \sim U[0, 1]$, where $p^s(x) = p^{is}$ for $x \in [x_{i-1}, x_i) \subset [0, 1]$. The only material difference to the continue productivity case is that firm productivity $p^s(x)$ is increasing in $x \in [0, 1]$ but not strictly increasing. The underlying wage structure described in equation (14) in the main text, however, continues to apply and equation (19) in the main text describes the equilibrium values $v_i(\Omega)$.

So why is the state space finite? The critical property is that despite there being wage and quit rate dispersion across firms $x \in [x_{i-1}, x_i)$ within a productivity level, all such firms have identical expected employment dynamics. Why? Because all firms with $i \geq i^h$ immediately replace any worker who quits and so their expected employment dynamics are independent of χ. Additionally, all firms with $i < i^h$ post the same wage, $w_{\min}$ and so their expected employment dynamics are also independent of χ. Now recall that $G(x)$ describes the distribution of employment across firms $x \in [0,1]$. By definition of the partition above, firm $x = x_i$ has productivity $i+1$ and rank 0 and so $G(x_i) = \sum_{j=1}^{i} N_i/(1-U)$. Because firm size is orthogonal to rank we then have

$$G(x) = \frac{\sum_{j=1}^{i-1} N_j + \frac{x-x_{i-1}}{x_i-x_{i-1}} N_i}{1-U} \text{ for all } x \in [x_{i-1}, x_i),$$

is continuous but $\underline{N}$ is a sufficient statistic for $G(.)$. Thus with finite productivity states the previous analysis goes through with the added simplification that the aggregate space is a finite vector $\Omega = (s, \underline{N})$.

B.4 Finite productivity model summary

We now briefly summarise the equations of the finite productivity model as we use it in our quantitative work which include minor additions to the model made relative to the continuous productivity model. Our calibrated model features $i^c(\Omega) = 1$ at all times, and we present the equations for this case of the model. Recall from the previous section that we specialise to a finite number of productivities $i = 1, ..., I$, where within each productivity level firms additionally separate into different wage ranks $\chi \in [0,1]$. We then define the overall wage rank across all firms as $x \in [0,1]$.

Firm HJB and policy functions: All firms with the same productivity p_i achieve the same value v_i, regardless of their wage rank. With aggregate shocks the HJB includes the aggregate state $\Omega = (s, N_1, ..., N_I)$. If $i^c(\Omega) = 1$ at all times, the N_i evolve continuously over time. In this case, the HJB can be written

$$\begin{aligned}(r+\delta_F)v_i(\Omega) = a_s p_i - c_f - w_{\min} - (\lambda_1(\Omega) + \lambda_u)\min[v_i(\Omega), c_0] + \mu_1 E_c \max[v_i(\Omega) - [c_0 + c], 0] \\ - \delta_D E_c \min[v_i(\Omega), c] + \alpha_\gamma \sum_j \gamma_{ij}(v_j(\Omega) - v_i(\Omega)) \\ + \alpha_a \sum_{s'} \gamma_{s,s'}(v_i(s', \underline{N}) - v_i(\Omega)) + \sum_{j=1}^{I} \frac{\partial v_i(\Omega)}{\partial N_j} \dot{N}_j(\Omega). \quad (9)\end{aligned}$$

Notice that we extend the model relative to the main text by introducing a flow cost of capital maintenance, c_f. This is a cost which must be paid each period to maintain each existing unit of capital. The introduction of c_f does not change the economics of the model, but it is useful as it allows us to more easily partition the productivity bins into those which do and do not replace quits (see below for more details), and is also used as the subsidy in our counterfactual calibration with a higher value of c_0. The only aggregate "price" which affects firm value is the scalar $\lambda_1(\Omega)$. The expectations over JC and JD cost draws have closed form solutions under the assumed distributions. The hiring threshold $i^h(\Omega)$ is defined as the first i for which $v_i(\Omega) > c_0$.

In the finite productivity model, most policy functions – and in particular those which relate to net employment dynamics – depend only on i and not the wage rank. Specifically, the job creation rate per employee is $jc_i(\Omega) = \mu_1 H^{JC}(v_i(\Omega) - c_0)$. The job destruction rate is $jd_i(\Omega) = \delta_D[1 - H^{JD}(v_i(\Omega))]$ for firms with $i > i^h(\Omega)$ and $jd_i(\Omega) = \delta_D[1-H^{JD}(v_i(\Omega))]+\lambda_1(\Omega)+\lambda_u$ otherwise. Entrants who draw productivity i have average initial employment $\bar{n}_{0,i}(\Omega) = n_u + \frac{N_0}{\mu_1} jc_i(\Omega)$. Here, $n_u = 2$ is the number of initial workers a firm can draw from unemployment for free upon startup. We found it helpful to set $n_u = 2$ rather than $n_u = 1$ to ensure that even unproductive ($i = 1$) entrants start with more than one employee. This slows down the firm exit process for unproductive entrants, who otherwise exit as soon as their one and only initial employee is poached or fired. We define $\hat{n}_{0,i}(\Omega) = \frac{N_0}{\mu_1} jc_i(\Omega)$ as average entrant size excluding the initial n_u free hires from unemployment.

Evolution of employment distribution: The total mass of employment at each productivity bin evolves according to:

$$\dot{N}_i(\Omega) = \mu_0 P^E(\Omega)\gamma_{0i}\bar{n}_{0,i}(\Omega) + N_i \left[jc_i(\Omega) - jd_i(\Omega) - \delta_F - \alpha_\gamma \sum_{j \neq i} \gamma_{ij} \right] + \alpha_\gamma \sum_{j \neq i} \gamma_{ji} N_j. \tag{10}$$

The first term on the right hand side is the inflow of employment from firm entry. The term in square brackets gives net job creation accounting for job creation and destruction including the firm exit shock. The terms proceeded by α_γ gives the transition of firms across productivity bins. Total unemployment is $U = 1 - \sum_i N_i$. The distribution across firm wage ranks can then be calculated from our closed form solution:

$$G(x, \Omega) = \frac{\sum_{j=1}^{i-1} N_j + \frac{x - x_{i-1}}{x_i - x_{i-1}} N_i}{1 - U} \text{ for all } x \in [x_{i-1}, x_i). \tag{11}$$

Recall that we define the boundaries $x_0 = 0$, $x_i = x_{i-1} + \gamma_{0i}$ and $x_I = 1$. A firm in state $i \in \{1, 2, ..., I\}$ with wage rank $\chi \in [0, 1)$ is correspondingly defined as being in state $x \in [0, 1]$ where $x = x_{i-1} + \chi[x_i - x_{i-1}]$.

Quits and hires across the x distribution: To close the model, we need to calculate the offer arrival rate $\lambda_1(\Omega)$. To do this, we must solve the quit rates across the wage rank distribution. As in the continuous productivity model, the quit rate for incumbent firms at any wage rank x is given by equation (8) in the main text, which we rewrite in our x notation as

$$q(x, \Omega) = \begin{cases} \lambda_u + \lambda_1(\Omega) & \text{if } x \in [0, x^h(\Omega)) \\ \lambda_u + \frac{\phi \int_x^1 \{JC(y,.) + \lambda_u[1-U]G'(y)\} dy}{U + \phi[1-U]G(y)} & \text{if } x \geq x^h(\Omega) \end{cases} \tag{12}$$

where $x^h(\Omega) = x_{i^h(\Omega)-1}$ corresponds to the lowest ranked hiring firm, who has $i = i^h(\Omega)$ and $\chi = 0$. The total job creation flow at each x is $JC(x, \Omega) = \left\{ [1 - U]G'(x) + \frac{\mu_0}{\mu_1} N_0 P^E(\Omega) \right\} jc_i(\Omega)$. The closed form solution for $G(x)$ similarly defines a closed form solution for $G'(x)$, which is well defined except at the x_i boundaries where $G(x)$ is non-differentiable. Performing the integration in (12) yields a closed form solution for $q(x,\Omega)$ for any $x \geq x^h(\Omega)$:

$$q(x,\Omega) = \lambda_u + \frac{(1-\chi)\left[(jc_i+\lambda_u)N_i + \mu_0 P^E(\Omega)\gamma_{0,i}\hat{n}_{0,i}\right] + \sum_{j=i+1}^{I}\left[(jc_j+\lambda_u)N_j + \mu_0 P^E(\Omega)\gamma_{0,j}\hat{n}_{0,j}\right]}{U/\phi + \sum_{j=1}^{i-1} N_j + \chi N_i} \tag{13}$$

This equation uses the reverse mapping $\chi(x)$ to find the χ associated with the current x from $\chi = \frac{x-x_{i-1}}{\gamma_{0i}}$, and similarly the productivity level $i(x)$ associated with the current x. The hiring rate for incumbent firms can then be simply computed as $h(x,\Omega) = jc_i(\Omega) + \mathbf{1}(i \geq i^h(\Omega))q(x,\Omega)$. Finally, $\lambda_1(\Omega)$ is just $q(x,\Omega)$ evaluated at $x = x^h(\Omega)$ and then subtracting λ_u, which gives

$$\lambda_1(\Omega) = \frac{\sum_{j=i^h(\Omega)}^{I}\left[(jc_j(\Omega)+\lambda_u)N_j + \mu_0 P^E(\Omega)\gamma_{0,j}\hat{n}_{0,j}(\Omega)\right]}{U/\phi + \sum_{j=1}^{i^h(\Omega)-1} N_j} \tag{14}$$

This closes the model, and provides sufficient information to simulate the model keeping track only of the finite employment vector $\underline{N}$. In particular, the model can be solved and simulated using only the HJB (9), the evolution of total employment by productivity bin (10), and the closed-form solution for the job offer arrival rate (14).

C Numerical Methods Appendix

Steady state: For given parameter values, solving the core equations of the model in steady state reduces to solving for a vector of I values v_i and employment stocks N_i, as well as the arrival rate λ_1. This is a simple problem to solve using the steady state versions of (9), (10), and (14). Intuitively, one can guess a value of λ_1, solve the HJB for the values v_i, use the implied policy functions to calculate the employment stocks N_i, and use these to update your guess for λ_1. In practice, we solve the model in steady state at the same time as calibrating our parameters, which involves calculating other statistics, which we detail further below.

Calculating average quit and hiring rates involves integrating over the wage rank distribution, x. To do this, we build an uniform grid over $\chi \in [0,1]$ with 1,000 nodes. This is then combined with the x_i to build a grid over x with $I \times 1,000$ nodes. We calculate all integrals on this grid using trapezoid integration.

To calculate the firm age and size distribution we solve for the densities of firms on grids for age and size. This is thus reminiscent of the non-stochastic simulation approach of Young (2010), or the methods of Achdou et al. (2021). To calculate the size distribution, we build a grid over firm sizes. Recalling that the number of employees in a given firm is an integer, we define a size grid as integer values from 0 to 20,000 employees. We solve for the mass of firms at each size s and productivity i. We use the firm dynamics processes (job creation, destruction, entry, exit, and so on) to construct flow rates across these joint size-productivity bins, which we use to build a matrix of transitions. We can then solve for the steady-state density of the number of firms at each size-productivity bin by inverting this matrix.

To calculate the age distribution we follow a similar process. However, since age is a continuous number, we discretise the age grid on a uniform grid from age 0 to age 26 years old (since this is the maximum age bin recorded in the BDS data) with 1,000 nodes. We solve for the mass of firms and employment at each age-productivity bin. Finally, to compute the mass of active firms at each age, we actually need to compute the joint age-size distribution, since we define firms with 0 employees as having exited. To do this, we must solve for the joint age-size-productivity distribution, which we do using the same methods. Given the high dimension of this object, we solve for this distribution using a reduced firm size grid from 0 to 1,000 employees, and confirm that raising this maximum has no impact on the moments for which this distribution is used.

To compute the growth rate density and hockeystick plots (Figure 6 in the main text) we simulate a panel of firms for one quarter, with their initial states drawn from the steady-state size-productivity distribution. We simulate a panel of one million firms and calculate net employment growth and gross flows from the beginning to the end of the quarter.

Business cycle: For given parameter values, solving the core equations of the model over the business cycle reduces to solving for the functions $v_i(\Omega)$, $\dot{N}_i(\Omega)$, and $\lambda_1(\Omega)$ over the state space $\Omega = (s, N_1, ..., N_I)$, using the equations (9), (10), and (14). In terms of approximation, it actually suffices to approximate only the function $v_i(\Omega)$, as the values of $\dot{N}_i(\Omega)$, and $\lambda_1(\Omega)$ can then be calculated exactly using (10) and (14) at any grid point or point in a simulation.

We approximate $v_i(\Omega)$ using second order polynomials in $N_1, ..., N_I$, with different coefficients for each of the discrete productivity level pairs i, s. In particular, first adjust the value function notation to $v_i(\Omega) = v_{i,s}(N_1, ..., N_I)$ to acknowledge that aggregate productivity s is also a discrete state. Then for each i, s we approximate $v_{i,s}(N_1, ..., N_I)$ as

$$v_{i,s}(N_1, ..., N_I) \simeq h^0_{i,s} + \sum_{j=1}^{I} \left(h^1_{i,s,j} N_j + h^2_{i,s,j} N_j^2 \right) \tag{15}$$

where $h^0_{i,s}$ $h^1_{i,s,j}$ and $h^2_{i,s,j}$ are scalar coefficients to be estimated. $h^0_{i,s}$ is the intercept, and $h^1_{i,s,j}$ and $h^2_{i,s,j}$ capture the first and second order effect on the value of firms with state i of changing total employment of firms with productivity j . Notice that we exclude cross terms in the second-order approximation, since these are known to typically be unstable given that the $N_1, ..., N_I$ tend to be highly correlated.

For each i, s this approximation uses $1 + 2 \times I = 1 + 2 \times 5 = 11$ coefficients. Since we use $I = 5$ idiosyncratic and $S = 3$ aggregate productivity nodes, this gives $5 \times 3 \times 11 = 165$ coefficients to estimate.

To solve the HJB, we need to know both the level of value and its derivative with respect to the aggregate employment bins. The derivatives are easy to compute given our approximation, as

$$\frac{\partial v_{i,s}(N_1, ..., N_I)}{\partial N_j} = h^1_{i,s,j} + 2h^2_{i,s,j} N_j \tag{16}$$

In order to solve the full business-cycle version of the model, we use the following procedure:

1. We need a grid of values for $(N_1, ..., N_I)$ to approximate our second-order polynomial on. To generate this, we use a Sobol set, which generates values of the $(N_1, ..., N_I)$ vector which are roughly equally spaced between a minimum and maximum value for each N_i. We generate 50 of such vectors, denoting the values of $(N_1, ..., N_I)$ at each candidate z as $\underline{N}_z = (N_{1,z}, ..., N_{I,z})$ for $z = 1, ..., 50$. Note that the aggregate state at any grid point is now denoted s, z, where s corresponds to aggregate productivity and z to the vector of $\bar{N}$ values.

2. Generate initial guesses for the parameters $h^0_{i,s}$ $h^1_{i,s,j}$ and $h^2_{i,s,j}$. Generate an initial guess for $\lambda_1(\Omega) = \lambda_{1,s,z}$ at each aggregate state. Generate an initial guess for $\dot{N}_i(\Omega) = \dot{N}_{i,s,z}$ at each aggregate state.

3. Given these guesses, solve the value function (9) for values $v_{i,s}(\Omega) = v_{i,s,z}$ at each idiosyncratic productivity node and aggregate state node. In solving (9), replace $\lambda_1(\Omega)$ with the current guess $\lambda_{1,s,z}$, and the drift term, $\sum_{j=1}^{I} \frac{\partial v_i(\Omega)}{\partial N_j} \dot{N}_j(\Omega)$, using i) the current guesses for the value function derivative implied by $h^1_{i,s,j}$ and $h^2_{i,s,j}$ and ii) the current guess for the drifts $\dot{N}_{i,s,z}$.

4. Using the new values of $v_{i,s,z}$, perform OLS regressions on (15) to update the parameters $h^0_{i,s}$ $h^1_{i,s,j}$ and $h^2_{i,s,j}$ with dampening.

5. Using the new values of $v_{i,s,z}$, calculate the new policy functions for job creation and destruction. Use these to update the drifts $\dot{N}_{i,s,z}$ using (10) and the offer arrival rates $\lambda_{1,s,z}$ using (14), both with dampening.

6. Return to step 3 and iterate to convergence.

As a measure of the accuracy of our second-order approximation, the R^2 of the regressions used to fit the polynomial is 99% on average across the $I \times S = 15$ regressions. This R^2 is a measure of the error between the predicted value from the second-order polynomial and the exactly computed value from the HJB of the $v_{i,s,z}$ on the nodes where the HJB is evaluated.

With our approximated policy function parameters $h^0_{i,s}$, $h^1_{i,s,j}$, and $h^2_{i,s,j}$ in hand, we can simulate the aggregate model, calculating all other objects exactly using the true nonlinear equations of the model. When estimating the model, we simulate using a one month aggregate time step $\Delta t = 1$, but finer grids do not affect the results. For most data comparisons, we aggregate up to quarterly data via averaging and HP-filter the model data as in the data.

Post estimation, we also simulate our impulse response to a typical recession using the same procedure. We additionally simulate the age and size distributions over time, by first solving for the aggregate dynamics, and then extending our age and size distribution codes to allow for aggregate dynamics.

Overview of estimation procedure: We pre-set six parameters, and our estimation procedure then chooses 23 parameters to minimize the distance to a large number of moments. To speed up the estimation, we split the estimation into two layers: an inner loop and an outer loop. Conditional on outer loop parameter values, in the inner loop we solve for the values of 11 parameters to exactly hit 11 moments. Intuitively, each of these 11 parameters has a tight link to a particular moment, which we are able to exploit to quickly solve for the value of these parameters. In the outer loop, we use the remaining 12 parameters to minimize the average distance to a set of 18 moments using a global minimization routine, every time repeating the inner loop procedure.

A key step in speeding up our estimation is that we calibrate many parameters in the non-stochastic steady state of the model (i.e. a version of the model without aggregate shocks) as is standard in heterogeneous agent modelling. In brief, the procedure operates as follows:

1. Guess values for the 12 outer loop parameters.

2. Given the current guess for the outer loop parameters, use the inner loop to exactly solve for the 11 inner loop parameters, to exactly hit the inner loop moments. These moments are calculated in the non-stochastic steady state of the model.

3. Given the values of the inner and outer loop parameters, now solve the full model (out of steady state), and simulate to construct aggregate time series.

4. Calculate the moments used in the outer loop. The moments related to the firm age distribution are calculated from the non-stochastic steady state, and the moments related to the business cycle are calculated from the business cycle simulation.

5. Calculate the distance measure of the outer loop moments to the moments in the data. Update the outer loop parameters using the global minimization routine, and return to step 2. Repeat until the global minimization routine completes.

For our global minimization routine in the outer loop, we program a simplification of the "TikTak" algorithm of Arnoud et al. (2019). Specifically, we draw 12,000 initial guesses of the outer loop parameters from a Sobol set, and calculate the outer loop moments at each guess. We then choose the five best performing guesses and run a local optimizer (pattern search) at each to find the local minima, and choose the lowest error among these as our final estimate.

C.1 Further details of estimation and parameterization

As we impose a relatively small number of productivity states, we use parameter choices to impose the key behaviour (perform JC or not, perform JD or not, replacing quits or not) of each state, rather than letting the estimation decide. The estimation is allowed to affect behaviour within each node (for example the level of JC in the node, if it is positive) and the probabilities that nodes are drawn.

Additional flow cost of capital maintenance: In the estimation we impose that $i^h = 3$ in steady state, which requires that $v_2 < c_0 < v_3$. This could be imposed using a penalty function approach, which penalises the SMD error whenever either of these inequalities is violated. We take a simpler approach, which speeds up the estimation (by allowing more parameters to be chosen in the inner loop) at the cost of introducing one new parameter, the flow cost of capital maintenance. Specifically, we firstly impose $p_2 < p_3$ in the estimation, which ensures that $v_2 < v_3$. Secondly, we then choose c_f in the inner loop to ensure that $c_0 = 0.5(v_2 + v_3)$, which guarantees that $v_2 < c_0 < v_3$. Intuitively, we thus introduce the flow cost of capital to shift the average value to ensure that c_0 lies exactly in the middle of v_2 and v_3 in steady state. We force this parameter to be small by imposing penalties if it falls below 0 or above 0.1. The estimated value of this cost is small, at 0.078, which is a small fraction of average productivity (which is one) and small compared to the hiring cost c_0. This flow cost c_f is also used as the flow subsidy (allowing $c_f < 0$) in our counterfactual calibration with a higher value of c_0.

We also impose that i^h should not move over the business cycle, so that $i^h(\Omega) = 3$ almost always during a simulation. We do this by computing the fraction of periods where $i^h(\Omega) \neq 3$ during our business cycle simulation and adding this as a penalty in the SMD function if it exceeds 1% of the simulation. At the estimated parameters, $i^h(\Omega) \neq 3$ only 0.77% of the time during our long business cycle simulation, and $i^h(\Omega) = 3$ at all times in our typical recession impulse response function plots.

Placement of cost function support parameters: We place the lower support of H^{JC} and upper support of H^{JD} so that (in steady state) i) firms with $i = 1, 2, 3$ perform JD in response to the H^{JD} shock, but never JC in response to the H^{JC} shock, and ii) firms with $i = 4, 5$ perform JC but not JD. To do this requires that i) $v_3 < \bar{c}^{JD} < v_4$, so that firms with $i = 4$ have high enough value to survive any draw of c^{JD}, and ii) $v_3 < \underline{c}^{JC} + c_0 < v_4$, so that firms with $i = 3$ have low enough value so that the minimum cost of performing JC is too high. In order to not introduce additional degrees of freedom into the model, we simply set $\bar{c}^{JD} = 0.5(v_3 + v_4)$ and $\underline{c}^{JC} = 0.5(v_3 + v_4) - c_0$, which we can impose simply at every iteration of the inner loop.

Definitions of aggregate flows: The formulas below give the aggregate flows (per unit of time) used in Table 2. These can be expressed as rates by dividing by an appropriate denominator, and are constructed to be comparable with their data counterparts.

Note that we treat the λ_u shock as layoffs into unemployment, not quits into unemployment. This means that in the model, the EU flow and layoffs are identical, as all EU moves are involuntary. Similarly, this means that all quits are job-to-job quits, so the EE flow and quit flow are identical.

Total JC flow: $JC_t = \sum_{i=1}^{I} \left[\mu_0 P^E(\Omega)\gamma_{0i}\bar{n}_{0,i}(\Omega) + N_{i,t} jc_i(\Omega)\right]$
Total JD flow: $JD_t = \sum_{i=1}^{I} \left[jd_i(\Omega) + \delta_F\right] N_{i,t}$
Total EU flow: $EU_t = \sum_{i=1}^{I} \left[\delta_D[1 - H^{JD}(v_i(\Omega))] + \delta_F + \lambda_u\right] N_{i,t}$.
Total UE flow: $UE_t = \lambda_0(\Omega)U_t + n_u\mu_0 P^E(\Omega)$
Total EE flow: $EE_t = \int_0^1 (q(x,\Omega) - \lambda_u) dG(x,\Omega) \times N_t$
Total hiring flow: $H_t = UE_t + EE_t$
Firm entry flow: $EN_t = \mu_0 P^E(\Omega)$
Firm exit flow: $EX_t = \delta_F M_t + \sum_{i=1}^{I} jd_i(\Omega) m_{i,t}^1$, where M_t is the total mass of firms with at least one employee, and $m_{i,t}^1$ is the mass of firms with productivity i and exactly one employee.

JC and JD definition in the model vs. the data: The definitions of JC and JD in the model are built to correspond as closely as possible to their notions in the data. However, practical computational limitations mean that their definitions are not identical. In particular, in the data JC is defined as the sum of employment increases across firms which saw an increase in employment between two dates, and JD is defined as the sum of employment falls at firms which saw a decrease in employment (see, e.g., DFH). Computing these measures exactly therefore would require simulating a panel of firms over the business cycle, which slows down the estimation and simulation of the model.

Instead, we are careful to segment our model so that firms in states $i = 1, 2, 3$ have declining employment at any instant of time, and firms in states $i = 4, 5$ have increasing employment at any instant in time. Abstracting for the moment from entry and exit, JC can therefore be measured as the employment change at state $i = 4, 5$ firms, which corresponds exactly to the job creation flow $\sum_{i=1}^{5} jc_i(\Omega)N_{i,t}$ in the model, since $jc_i(\Omega) = 0$ for $i = 1, 2, 3$. Similarly, JD can be measured as the employment change at state $i = 1, 2, 3$ firms, which corresponds to $\sum_{i=1}^{5} jd_i(\Omega)N_{i,t}$ since $jd_i(\Omega) = 0$ for $i = 4, 5$. This is complicated slightly by two factors. Firstly, firms may switch from being in state $i = 1, 2, 3$ to $i = 4, 5$ within the same quarter, which is the period over which JC and JD are measured in the DFH data. This means that measured and theoretical measures may differ, as a firm might receive a JC and JD shock in the same quarter. However, this is a rare occurrence, as our productivity shocks occur on average only once per quarter. Secondly, we must account for entry and exit. Entry is simple, as all entrants create jobs and therefore can simply be added to the job creation flow. Exit complicates the analysis somewhat, as even firms with $i = 4, 5$ might receive the δ_F exit shock. However, this shock is calibrated to be very rare so this does not matter much in practice.

In order to check the applicability of our JC and JD measures, we simulate a large panel of firms after estimating the model. We focus on the steady state, and simulate the firms for one quarter of data, with the firms drawn from the ergodic productivity and size distribution. We compute the job destruction rate exactly as is done on the data, and find a quarterly rate of 6.2%, while the theoretical rate as calculated by $jd = \sum_{i=1}^{I} [jd_i + \delta_F] N_i/N$ exactly equals the targeted value of 7.0%. While not identical, this difference of roughly 10% is in line with the average error of the other moments in the outer loop of our SMD routine.

Comparison of firm-level autocorrelation to data: Our parameter values generate an autocorrelation of 0.21 for yearly productivity in a year-averaged simulated firm-level productivity series, or 0.53 for quarterly productivity (within mature firms). Elsby et al. (2017) discuss empirical estimates of the persistence of idiosyncratic productivity, and find a wide range of values. Our estimate lies within this range. Specifically, Cooper, Haltiwanger, and Willis (2015) imply a quarterly autocorrelation of 0.4, which is below our value, while Abraham and White (2008) imply 0.68 which is above our value. While our productivity process has relatively low persistence, our constant returns to scale structure means that current productivity controls the growth rate of employment, not the level. Hence, even temporary productivity shocks will generate permanent effects on a firm's employment.

Further details of the Inner loop: We present below a list of the 11 parameters (plus the additional parameter c_f) chosen in the inner loop, and how the moment used to choose the parameter is calculated. The inner loop is terminated when the error in all moments is below 10^{-8}. All moments are computed in the non-stochastic steady state of the model.

The parameter p_4 is chosen to normalise aggregate labor productivity to one. The parameter p_3 is chosen to generate a standard deviation of idiosyncratic productivity of 30%. The parameter γ_2 is chosen to match that 80% of quits (and replaceable layoffs) are replaced, calculated as $qfr = \int_{x^h}^{1} q(x)dG(x) / \int_0^1 q(x)dG(x)$.

Parameters $\underline{c}_{JC}$, $\bar{c}_{JD}$, and c_f are set as discussed above. The firm entry flow μ_0 is set to hit the average size of firms, measured as N/M, where M is the mass of firms with at least one employee. Parameters μ_1, δ_D, λ_u, and ϕ are set to match the theoretically computed JC, JD, layoff, and EE quit rates. Finally, $w_{\min}$ is set to match the labor share, defined as $LS = Ew \times N/Y$, where $Ew = \int_0^1 w(x)dG(x)$.

Further details of the Outer loop: The 12 parameters chosen in the outer loop fall into two broad categories: those relating to the firm age distribution $(p_1, p_2, p_5, \gamma_{11}, \gamma_{55}, \gamma_{01}, \gamma_{05}, N_0)$ and those relating to business cycle moments $(\xi_e, \xi_{JC}, \xi_{JD}, c_0)$. The 12 age distribution moments are computed using the non-stochastic simulation of the age distribution in steady state. The six business cycle moments are computed by simulating the model for 1,000 years. The simulated data are then aggregated to a quarterly frequency and HP-filtered, as in the data. We apply a simple diagonal weighting to the moments, with the total weight given to business cycle moments slightly overweighted to ensure the model performs well on both business cycle and steady state dimensions.

We perform several parameter swaps in the outer loop estimation, searching over some hyper-parameters instead as these ensure that the estimation searches over sensible regions of the parameter space. Specifically, instead of searching over values of p_1, p_2, and p_5, we search over steady state values of JD_1, JD_2, and JC_5. Instead of searching over values of N_0, we search over values of the average number of employees of firms at the moment of entry.

The estimation finds that the productivity grid is non-monotone, as $p_1 > p_2$ and $p_5 < p_4$. Nonetheless, firm *values* remain monotone, with $v_i < v_{i+1}$ for all i, which is sufficient for the job ladder to be directed monotonically by i, and hence for our notion of equilibrium to remain well defined. The disconnect between the ordering of productivities and values occurs simply because the entrant states $i = 1, 5$ are more persistent than the mature states.

Simulated Minimum Distance (SMD) results The weighted average error of model moments to data moments in the outer loop SMD estimation is 7.5% across 18 moments. The average error for the six business cycle moments is 8.0%, where the quarterly data and model moments can be seen in Table 2, and the yearly standard deviation of firm entry is 0.0688 in the model and 0.0701 in the data. The average error for the 12 firm age moments is 6.4%, where the data and model moments can be seen in Figure 3, and the firm exit rate is 9.6% in the model and 8.3% in the data. The errors for the 11 moments in the inner loop are zero by construction.

Identification All parameters in the inner loop are adjusted to match one assigned moment to the data, and hence we know these parameters are identified by the moments because individually adjusting their values successfully sets the errors in those moments to zero.

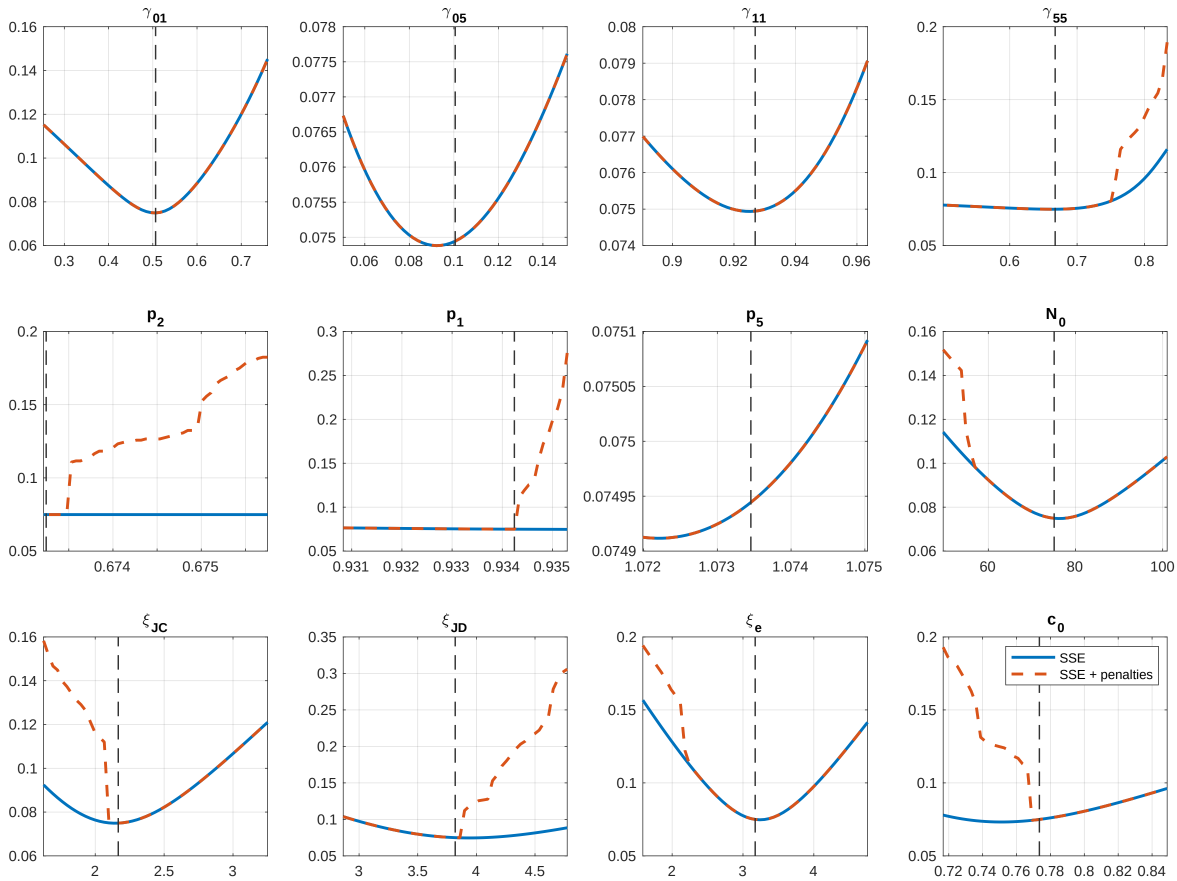
To check the identification of parameters in the outer loops we perform three experiments. First, we vary the values of our outer loop parameters one by one in a grid of values around their estimated values. The other parameters are held at their estimated values. We compute the sum of squared errors from the SMD at the new values, and plot them in Figure C.1. Each panel gives the effect of varying one parameter on the (square root of) the sum of squared errors from the outer loop. The blue line gives the sum of squared errors, and the red led the sum of squared errors plus the penalties imposed in the estimation for parameter values which lead to the model violating certain required conditions. As can be seen in the figure, the parameters are well identified with most sitting near the bottom of "U" shapes in the sum of squared error. The penalty function penalises the estimation if either: (i) the condition $v_2(\Omega) < c_0 < v_3(\Omega)$ is violated more than 1% of the time during a simulation. This would imply that the degree of replacement hiring experienced extreme counterfactual changes as entire groups of firms started or stopped replacement hiring, or (ii) matching $v_2 < c_0 < v_3$ in steady state required a value of c_f outside of the allowed small range. Some parameters are also partly identified because changing them would violate these penalties. This stresses that matching the degree of replacement hiring in the data is not something that our model can do "for free" and that it imposes constraints on the estimation. We highlight that the fact that the model can match the moments that it does while matching the degree of replacement hiring over the business cycle is not automatic, and is a success of the model.

Secondly, we verify that the parameters that we informally associate with affecting the steady state moments (firm age distribution) versus the business cycle moments actually affect these moments more than the others. In Figure C.2 we split the sum of squared errors into the part coming from firm age moment errors and the part from business cycle errors. We plot these, subtracting the true estimated errors so that the values of each line is zero at the estimated parameter values. In blue we plot the steady state moment errors, and in red the business cycle errors. The parameters associated more with the steady state are in the top two rows, and the business cycle in the bottom row. We can see that the parameters in the top two rows affect the steady state moments (blue line) much more than the business cycle moments (red line), while the opposite is true for the parameters in the bottom row. Hence, our intuitions about the role of these parameters appears to be justified.[32]

Finally, in our estimation we informally associated certain parameters with certain business cycle moments. We argued that ξ_{JC}, ξ_{JD}, and ξ_e controlled the standard deviations of job creation, job destruction, and entry respectively. In the top row of Figure C.3 we show that varying these parameters monotonically affects the associated moment in the expected way, with higher values increasing the associated elasticity and hence standard deviation. We argued that raising c_0 lowered the autocorrelation, by increasing the feedback from unemployment into firm value and hence job creation and layoff policies. In the bottom row we show how varying c_0 affects the autocorrelation of JC, JD, and layoffs

[32] We actually perform a few parameter swaps in the estimation which helps keep the parameter choices in sensible ranges. For example, rather than directly searching over p_1, we search over values of jd_1, which are used to back out the required value of p_1. This explains why adjusting the parameter values in the bottom row has exactly zero effect on the steady state moments, since these parameter swaps hold the realised firm age distribution constant by construction.

Figure C.1: Identification test: effect on sum of squared errors of varying each parameter

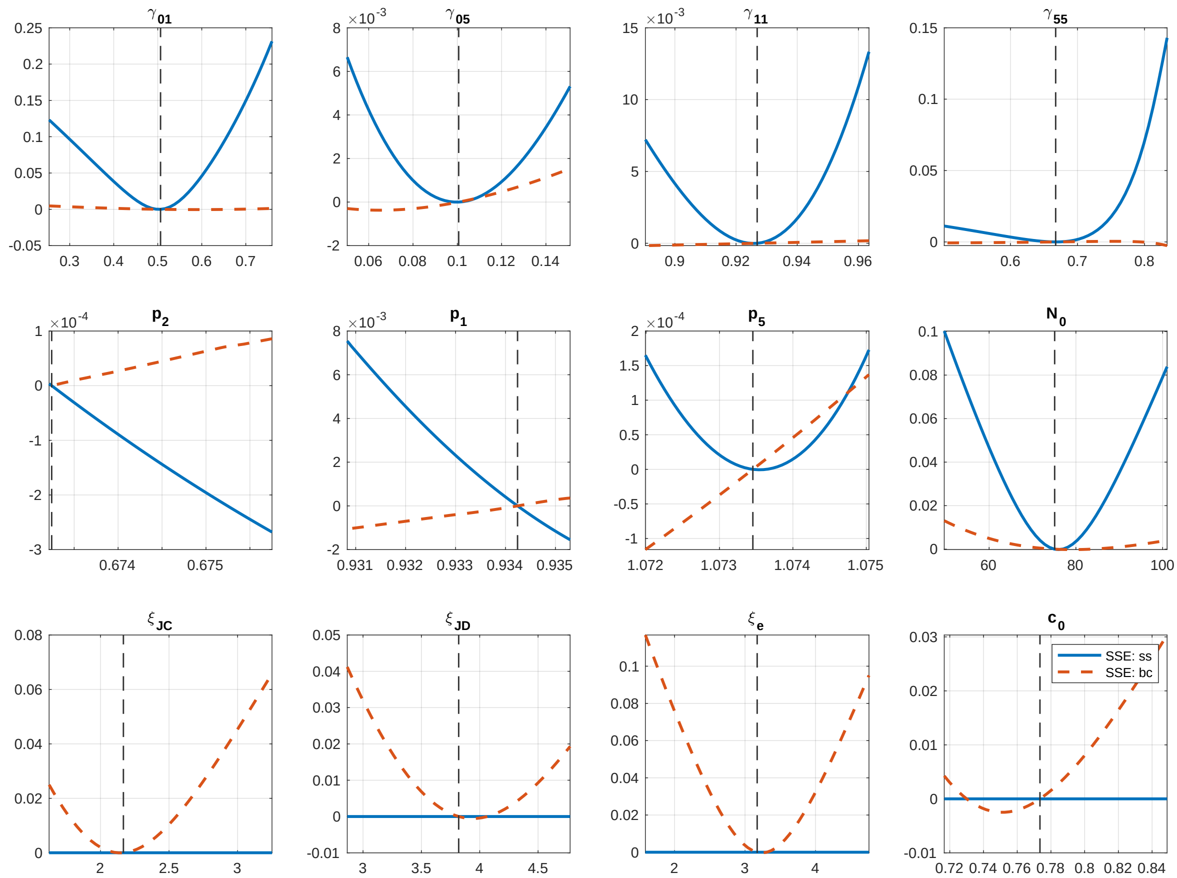


Note: Effect on sum of squared errors of varying one parameter at a time around their estimated value. Blue line gives the sum of squared errors and red line also include the penalty functions from the estimation. The vertical dashed line indicates the estimated parameter value.

respectively. This shows that, as expected, raising c_0 lowers the autocorrelation of JC and layoffs. In isolation, raising c_0 actually raises the autocorrelation of JD, contrary to expectations. This is because total JD includes both layoffs and JD from unreplaced quits, and raising c_0 also affects the dynamics of unreplaced quits by making the overall volatility of unemployment lower. However, when recalibrating ξ_{JC}, ξ_{JD}, and ξ_e to maintain the same volatilities as in the baseline estimation, raising c_0 also lowers the autocorrelation of total JD (see the counterfactual experiment in the "The importance of c_0" section of the main text).

For the parameters we informally associated with the moments of the firm age distribution, we did not informally associate each individual parameter with a specific moment of the firm age distribution. Instead, we envisaged choosing the parameters as a whole to match the whole distribution. Hence we do not repeat the exercises of Figure C.3 for the firm age distribution, but note that these parameters do indeed affect the firm age distribution and are well identified, as we showed in our discussion of Figure C.2.

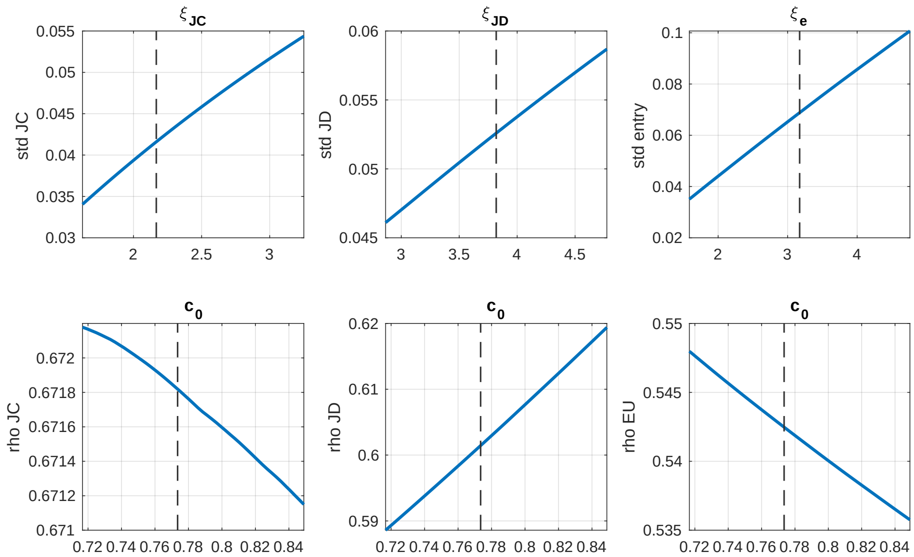
Figure C.2: Identification test: steady state versus business cycle errors



Note: Effect on sum of squared errors of varying one parameter at a time around their estimated value. Blue line gives the sum of squared errors for the firm age distribution moments, and red line for the business cycle moments. Both lines subtract the value of the moments at the estimated parameters, so are normalised to zero at the estimated parameter value. The vertical dashed line indicates the estimated parameter value.

Counterfactual calibration with higher c_0 In Section 7.3 we discussed a counterfactual calibration where we raised c_0 by 50%. This counterfactual is constructed as follows. Firstly, we fix c_0 at 50% higher than its estimated value. Secondly, we adjust the outer-loop parameters ξ_{JC}, ξ_{JD}, and ξ_e to hold the simulated standard deviations of JC, JD, and entry at the same values from our estimated model. We do not adjust the remaining outer-loop parameters and hyper-parameters from their original values, which holds the firm age distribution exactly at the original estimated distribution. Finally, given these outer-loop parameters, we re-run the inner loop, which adjusts the inner loop parameters to the values needed to continue exactly hitting the original inner loop moments. With this new calibration, we repeat our original business cycle simulation in order to compute the new moments discussed in the text.

Figure C.3: Identification test: business cycle moments



Note: Effect on the associated moment(s) of changing each parameter. This exercise is performed for each parameter we informally associated with business cycle moments. The vertical dashed line indicates the estimated parameter value.

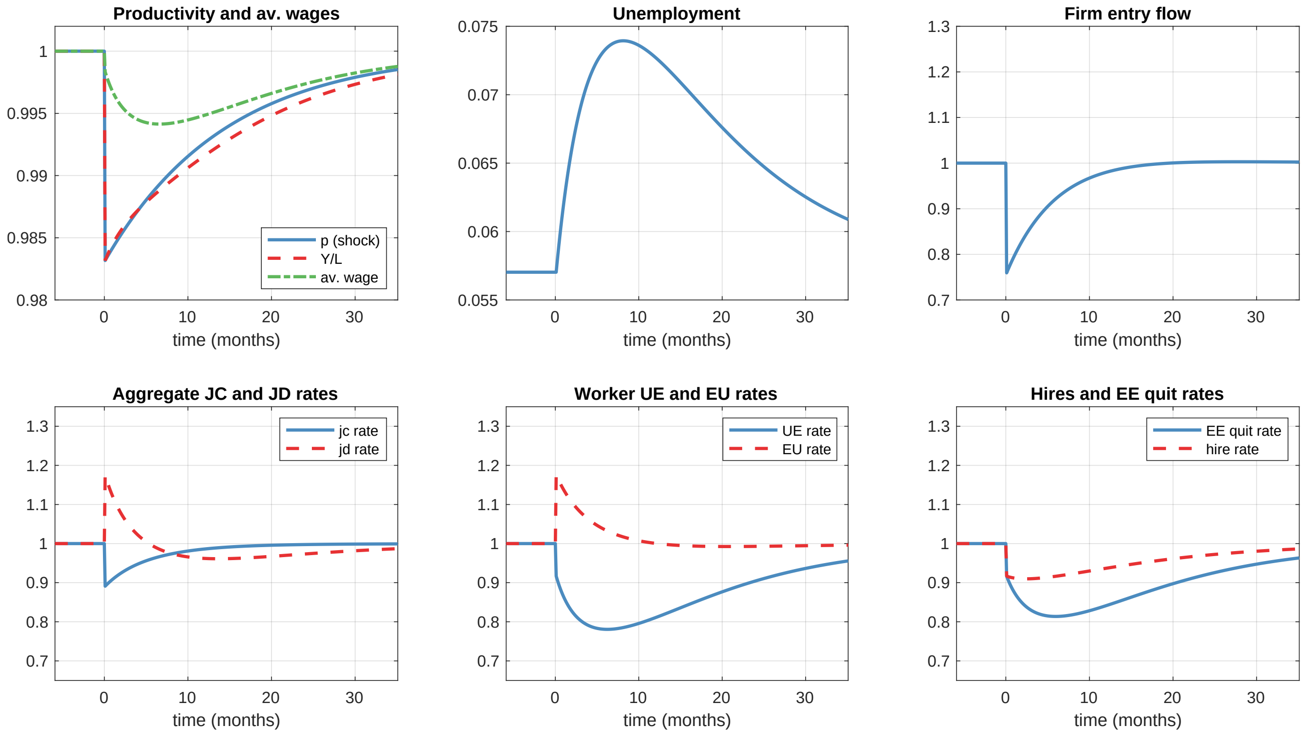
D Additional Tables and Figures

Table 8: Equilibrium policies and values in steady state

	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
v_i	0.7242	0.7244	0.8226	2.6328	2.6561
p_i	0.9342	0.6732	0.7127	1.3085	1.0735
jc_i	0	0	0	0.0446	0.0458
jd_i	0.0995	0.0995	0.0294	0	0
njc_i	-0.0996	-0.0996	-0.0295	0.0445	0.0457
N_i/N	0.0031	0.1208	0.3841	0.4848	0.0072
M_i/M	0.0498	0.1423	0.3968	0.4035	0.0076

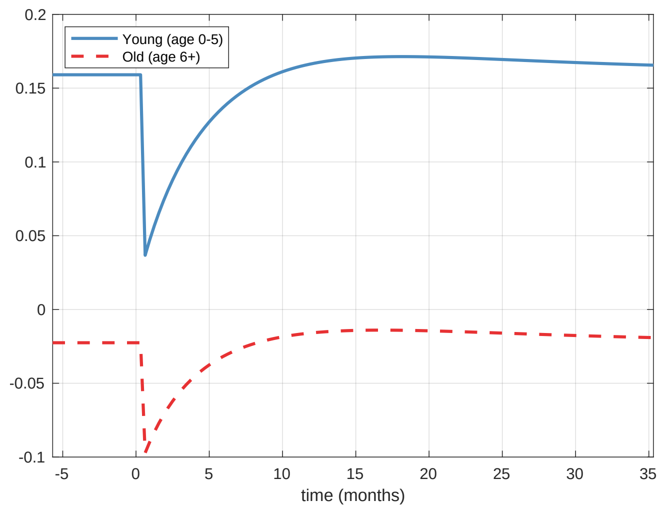
Note: Table summarizes the value and policy functions in steady state across productivity levels $i = 1, ..., I$. v_i is firm value, and p_i productivity. Value is monotonically increasing across states. jc_i and jd_i are job creation and destruction rates per employee for incumbent firms, excluding the δ_F exit shock. njc_i is net job creation: $njc_i = jc_i - jd_i - \delta_F$. The final two rows give the fraction of employment and active firms (with at least one employee) repectively at each i in the ergodic distribution.

Figure D.1: Impulse Response Function - Cyclical behaviour of key aggregates



Note: Figure plots additional aggregates from our typical recession impulse response function. See Section 7.1 for further details of the experiment.

Figure D.2: Model Impulse Response: Net JC by age



Note: Figure plots Net JC rates by firm age for our typical recession experiment. See Section 7.1 for further details of the experiment. JC and JD flows are yearly, and computed as $1 - e^{-12r}$, where r are the average theoretical monthly rates within each bin.

References

[1] Achdou, Y, J. Han, J.-M. Lasry, P.-L. Lions, B. Moll. 2021. "Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach," *The Review of Economic Studies*, 2021.

[2] Arnoud, A, F. Guvenen, T. Kleineberg. 2019. "Benchmarking Global Optimizers," *mimeo*.

[3] Audoly, R. 2020. "Firm Dynamics and Random Search over the Business Cycle," Mimeo.

[4] Burdett, K. and D. Mortensen. 1998. "Wage Differentials, Employer size and Unemployment," *International Economic Review*, 39: 257-273.

[5] Coles, M. G. and D. Mortensen. 2016. "Equilibrium Labor Turnover, Firm Growth, and Unemployment," *Econometrica*, 84: 347-363.

[6] Coles, M. G. 2001. "Equilibrium Wage Dispersion, Firm Size and Growth," *Review of Economic Dynamics*, 4(1): 159-187.

[7] Davis, S. J., R. J. Faberman and J. Haltiwanger. 2012. "Labor market flows in the cross section and over time," *Journal of Monetary Economics*, 59(1): 1-18.

[8] Elsby, M., R. Michaels and D. Ratner. 2017. "Vacancy Chains". University of Edinburgh, Department of Economics, mimeo.

[9] Moscarini, G. and F. Postel-Vinay. 2013. "Stochastic Search Equilibrium," *Review of Economic Studies*, 80(4): 1545-1581.

[10] Shimer, R. 2005. "The Cyclical Behavior of Equilibrium Unemployment and Vacancies," *American Economic Review*, 95(1): 25-49.

[11] Young, E. R. 2010. "Solving the incomplete markets model with aggregate uncertainty using the Krusell-Smith algorithm and non-stochastic simulations," *Journal of Economic Dynamics and Control*, 34(1): 36-41.