Slow recoveries, worker heterogeneity, and the zero lower bound

Federico Ravenna*and Carl E. Walsh†

First draft: May 2013
This draft: March 2014

Abstract

We show that a new Keynesian model with search and matching labor market frictions, worker heterogeneity, and a zero lower bound on the nominal interest rate is able to account qualitatively for several characteristics of the Great Recession and the sluggish recovery of the U.S. job market. With worker heterogeneity, a slow recovery lowers the average productivity of the pool of unemployed workers as less productive workers experience a higher inflow into unemployment and lower outflow from unemployment. This compositional effect lowers the expected surplus for firms of creating new jobs. Compared to a model with homogeneous workers, worker heterogeneity in a persistent downturn results in: (1) a much larger increase in unemployment; (2) a similar output recovery but a delayed and much slower recovery of unemployment (a jobless recovery); (3) little downward wage pressure despite considerable slack in the labor market; (4) a fall in measured match efficiency and a long-lived shift in the Beveridge curve. These results are obtained despite the assumption that wages are fully flexible. We show that the employment consequences of the zero lower bound on the nominal interest rate are small if workers are homogeneous but are large in the presence of worker heterogeneity.

*HEC Montreal, Institute of Applied Economics. Email: federico.ravenna@hec.ca.
†Department of Economics, University of California, Santa Cruz. Email: walshc@ucsc.edu.

1 Introduction

The behavior of U.S. unemployment since the trough of the Great Recession in June 2009 has posed a puzzle for economists and a dilemma for policy makers. While differential sectorial shocks have been suggested as one explanation for the slow recovery of employment, others have argued that such structural factors are less important; instead, the U.S. experienced a significant negative aggregate shock that produced employment losses across the entire economy (Lazear and Spletzer, 2012). Implicit in this latter view is the notion that the U.S. experience simply reflects the larger than normal negative shock associated with the Great Recession. However, it is now recognized that the cyclical behavior of the U.S. labor market has changed over the past 25 years. From the late 1940s through the 1980s, U.S. recoveries were relatively short, characterized by quick recoveries with GDP growing by about 10 percent in the first two years after the trough of the business cycle. In contrast, in the three most recent recoveries, GDP grew on average only 5.4 percent during the first two years of the recovery, and the two years following the trough of the Great Recession saw real GDP grow by only 2.2 percent. In contrast to pre-1990 business cycles, employment has taken much longer to rebound during slow recoveries. During the last three recoveries, for example, the growth in nonfarm payroll employment has lagged output growth by several quarters (Bachmann, 2011).

The contrasting behavior of nonfarm business output and employment over the last four recessions is shown in Figure 1. Each panel shows (log) output and employment of the nonfarm business sector, normalized to equal zero in the quarter identified by the NBER as the cyclical peak.[1] Each figure begins 8 quarters before the cyclical peak. The severity of the Great Recession stands out clearly, as does the weak employment growth during recent recoveries, particularly after the 2001 recession and the Great Recession. The Great Recession also saw an unprecedented increase in layoffs, with job destruction accounting for 52% of the increase in unemployment (Petrosky-Nadeau, 2012)

General equilibrium models of the business cycle with search frictions in the labor market and flexible wages are well known to generate counterfactually low volatility of employment relative to output (Shimer, 2005). These models also imply employment adjusts quickly. Thus, it has proven difficult to use these models to understand fully the Great Recession. In this paper, we investigate the role worker heterogeneity can play in generating output recoveries characterized by weak employment growth. We build

[1] The 1980 and 1982 recessions are treated as a single cyclical episode for the purposes of the figure.

on a simple model of worker heterogeneity previously developed in Ravenna and Walsh (2012b).

Workers differ along many dimensions. Some, such as educational level, specific job skills or experience, age, and gender may be easily observable. Other differences are difficult for prospective employers to observe, and Mincer-wage regressions that condition on observable characteristics of workers exhibit large unexplained residual variation in wages across workers (see Lemieux, 2006, and Hornstein et al., 2011). Other aspects of labor market outcomes are also difficult to explain based on observable worker characteristics. For example, Dickens and Triest (2012) estimate a multinormal logit model of involuntary separation transition probabilities using the 2008 wave of the SIPP. Controlling for age, education, race, and gender, their estimated equation has an R-squared of 0.129, suggesting heterogeneity of worker experiences within groups classified based on standard observable characteristics is important. It is this ex-ante unobservable heterogeneity that is the focus of our model.

We assume workers are of two types: high average productivity and low average productivity. While an unemployed worker's type is unobserved ex ante, a firm can observe the productivity type of its existing employees. We also assume a firm that is hiring engages in a process of interviewing, or screening, during which the firm is able to observe the productivity of a job applicant. The aggregate separation rate consists of exogenous and endogenous components. Firms employ an (optimal) cutoff productivity strategy; any job applicant whose productivity exceeds the cutoff is hired; any existing worker whose productivity is below the cutoff is fired. This cutoff productivity is endogenous, so the aggregate job separation rate is endogenous. To maintain a simple model structure, we assume the productivity of high productivity workers is nonstochastic and always exceeds the critical cutoff value. These workers, therefore, are subject only to the exogenous separation hazard. In contrast, we assume workers of low average productivity experience idiosyncratic fluctuations in productivity. If the worker's total productivity is too low, the worker endogenously separates if she was employed and is screened out through the interview process if she was unemployed and obtained an interview.

The presence of workers with different productivity generates a composition effect – a negative, exogenous aggregate productivity shock leads to a rise in endogenous separations and skews the composition of the employment pool towards workers with higher level of match-specific productivity. However, in standard models with endogenous separation (e.g., den Haan, Ramey, and Watson 1997) all unemployed workers are ex-ante

identical, and changes in the flow of job losers do not affect the expected productivity of job-searchers. With worker heterogeneity, the composition of the pools of employed and unemployed workers varies endogenously, depending on the changing composition of flows into and out of unemployment over the business cycle. The rise in unemployment disproportionately affects the low productivity workers, and, consequently, causes a decline in the average quality of the pool of unemployed workers through two channels. First, the job separation rate for low productivity workers rises and second, the job finding rate for such workers falls as more are screened out by firms during the recruitment process.

We find that allowing for worker heterogeneity increases the volatility of employment relative to output by 600%. More interestingly, our model can generate a number of correlations that standard search and matching models with nominal rigidities find difficult to explain. For example, an aggregate demand shock that reduces output and labor demand increases separations and generates a substantial endogenous fall in productivity among unemployed workers, even if measured labor productivity is barely affected. The productivity dynamics is explained by the fact that unemployment falls disprortionately on low productivity workers. This composition effect also implies that average wages may not fall in a downturn even though we assume nominal wages are flexible and set through Nash bargaining.

Worker heterogeneity also generates endogenous fluctuations in the measured efficiency of the matching function. During a recession, the cutoff productivity level determining job finding rates and separation rates rises, increasing separations and decreasing hiring as more low productivity worker are screened out in the interview process. The empirical Beveridge curve shifts out since soon after the trough of the recession firms increase the posting of vacancies. Firms though are willing to fill vacancies only with relatively high productivity workers, while the increase in unemployment, mostly accounted for by low-productivity workers, is persistent, as job losers are screened out more aggressively.

We also show that in the model, slow output recoveries translate into very sluggish employment recoveries. Consequently, relative to a model without worker heterogeneity, the speed at which the employment gap is closed is much slower than the rate at which output recovers. The recovery is also jobless: employment does not start increasing until several quarters after the beginning of the recovery in output. Finally, slow recoveries have a higher employment sacrifice ratio: shocks that imply slower output recovery lead to a large increase in the cumulative employment gap relative to the cumulative output gap.

The effects of worker heterogeneity, causing high unemployment to significantly and adversely affect the average skill-quality of the pool of unemployed workers, are larger in more persistent recessions. A more persistent fall in aggregate demand leads to a more than proportional fall in the implied productivity among unemployed workers, and slow output recoveries end up being jobless recoveries. Thus, the composition effect may play a larger role in periods like the Great Recession, when a large negative demand shock combined with limitations on monetary policy contribute to a large and very persistent downturn.

If long recessions and slow recoveries generate high and persistent unemployment, what accounts for the slow recovery from the Great Recession? We explore the role of monetary policy constrained by the zero lower bound (ZLB) as one explanation for a slow recovery. Figure 2 shows nonfarm output and the real federal funds rate, defined as the nominal funds rate in quarter t minus the realized inflation rate in $t+1$. Inflation is measured by the 4-quarter change in the log of the index for personal consumption expenditures. A striking feature of the figure is the different behavior of the real interest rate during the four recessions. Real rates fell during the 1980 recession, reflecting in part the sharp fall in interest rates in reaction to the credit controls imposed by the Carter administration. The rise in the real rate in 1981 was an explicit Federal Reserve policy action under Chairman Volcker designed to bring down inflation, and monetary policy is generally viewed as responsible for the subsequent 1981 recession. In contrast, the real rate rose sharply in late 2008 even as the Federal Reserve pushed the funds rate to zero in an attempt to expand the economy. And while the real rate has remained lower since early 2009 than it was at the start of the Great Recession, it has fallen less than in the much milder 2001 recession, despite the greater severity of the Great Recession.

We use our framework to study the behavior of the economy when hit by a combination of shocks replicating the Great Recession. We model the fall in output as the combination of two shocks: a persistent negative demand shock and a fall in households' discount rate. As a consequence of these shocks, the nominal interest rate is pushed to ZLB. We find that the zero lower bound in our model does not play a major role in accounting for persistently high unemployment. Absent time-varying worker heterogeneity, the combination of a persistent fall in demand and limitations to monetary policy due to the ZLB cannot explain either the magnitude of the fall in employment or its slow subsequent recovery. In contrast, with heterogeneous workers, our model predicts that unemployment would recovery slowly even in the absence of any constraint on the ability

of the monetary authority to cut nominal interest rates. Imposing the ZLB does amplify the implied TFP decline among the unemployed, leading to a large and sustained increase in unemployment.

The model also implies that, despite the persistence of an unemployment gap, wages rebound quickly, closing 90% of the gap from the previous peak after only 6 quarters, while employment has closed less than half of the gap. Thus, high unemployment continues even though wages are flexible. In other words, our model does not attribute high unemployment to wage rigidity. Finally, throughout the recession and the recovery in output, labor productivity is virtually unchanged. These outcomes are qualitatively consistent with the data from the Great Recession.

The remainder of this paper is organized as follows. The next section provides a brief review of related literature. Our model is presented in section 3. Since the basic model we employ is developed in Ravenna and Walsh (2012b), our description here is kept to a minimum. Section 4 discusses the distortions present in the model that cause the competitive market equilibrium to differ from the social planners allocation. The role of productivity heterogeneity and the composition effect is investigated in a calibrated version of the model in section 5. The role persistence plays is examined in section 5.4. Section 6 discussions the effects of monetary policy on the labor market, while the impact of the zero lower bound and alternative policy rules for monetary policy are studied in section 6.1. Conclusions are discussed in the final section.

2 Related Literature

Several explanations have been proposed to account for the changes in the behavior of labor market variables over the US business cycle over the last three recessions. Berger (2012) finds that lower unionization and the lifting of firing restrictions after the 1980s has allowed firms to fire workers more selectively, leading to acyclical labor productivity and jobless recoveries. Groshen (2011) documents that in the last three recessions, temporary layoffs accounted for about 10% of the total increase in unemployment, as opposed to a range between 30% and 55% for the previous four recessions. Galí and Van Rens (2010) show that a decline in labor market frictions in a DSGE model with wage rigidities can explain the change in the cyclicality of labor productivity and the increase in the volatility of unemployment. The interaction between long-term job polarization across routine and non-routine employment and the business cycle, has been suggested as an explanation

for jobless recoveries (Jaimovic and Siu, 2012). The recent fall of unemployment in routine occupation, which has been faster then the simultaneous fall in unemployment in non-routine occupations, casts doubt on the role of job polarization as a cause of slow employment recovery after the Great Recession (Albanesi and Sahin, 2013). Petrosky-Nadeau (2012) builds a DSGE model where a credit crunch leads to the destruction of the least productive jobs, resulting in persistent unemployment and stagnant or rising wages, similarly to our results.

Worker and match heterogeneity play a key role in several models in the search and matching literature and in models with job-to-job transitions (e.g., Guerrieri 2007, Nagypal 2007, Nagypal and Mortensen 2007, Krause and Lubik 2010 and Tasci 2007). Bils, Chang and Kim (2009) and Mueller (2011) study the implications of skill heterogeneity for wages and labor market flows over the business cycle, but they assume segmented labor markets and only consider productivity shocks.

Rogerson and Pries (2005) allow for persistent job-specific productivity variation, and firms screen workers based on limited information on their productivity. As the match productivity is revealed over time, separations take place. They assume the average productivity of unemployed workers is not state-dependent, and the authors focus on steady state results rather than on the dynamics of labor market variables over the business cycle. In a model with heterogeneous skills and exogenous separation rates, Pries (2008) shows that the composition effect has a large impact on the cyclical value of vacancies and thus on the behavior of employment flows. Pries sets the relative covariance of separation rates for high and low productivity workers exogenously.

While the framework we propose is closely related to this previous work and relies on a similar mechanism in affecting incentives to post vacancies, we provide a framework with nominal rigidities that allows aggregate demand and the role of monetary policy to be analyzed. In addition, the average productivity of unemployed workers is state-dependent in our model and we focus on the dynamics of labor market variables over the business cycle. The relative covariance of separation rates for high and low efficiency workers is also endogenous in our model and can vary depending on the nature of the shock processes.

Our modeling framework is related to several contribution in the literature on labor and nominal frictions. We include nominal rigidities in a model with unemployment, as do Blanchard and Galí (2007, 2010), Gertler, Sala and Trigari (2008), Gertler and Trigari (2009), Ravenna and Walsh (2008, 2011a, 2011b), Walsh (2003, 2005), and Galí

(2011). However, these contributions with the exception of Walsh (2003, 2005) assume an exogenous separation rate, and all these previous papers assume homogenous workers. Our model includes endogenous separations, as in Den Haan, Ramey, and Watson (2000). Contrary to their model, we assume a portion of the match-productivity is worker-specific rather than match-specific.

In our model, slow recoveries in output will endogenously result in jobless recoveries, contrary to the pattern after a V-shaped recessions. All that is needed is that there be enough time variation in heterogeneity across workers' efficiency levels. Several of the changes in the characteristics of US labor markets since the 1980s are consistent with our hypothesis (a fall in search frictions, an increase in dismissals relative to temporary layoffs, or a lifting of firing restrictions), as well as any shock that requires reallocation of the workforce across sectors with loss of industry-specific human capital.

In Ravenna and Walsh (2012), we showed that composition effects play a larger role in economies with smaller gross labor flows, such as many European economies, suggesting that in normal times, the composition effect is relatively less important in accounting for U.S. unemployment dynamics. We demonstrate that even when labor flows are calibrated to match U.S. evidence, the composition effect plays a more significant role in the face of very large demand shocks. We then examine the consequences of large demand shocks when monetary policy is constrained by the zero lower bound on nominal interest rates.

3 The model of productivity heterogeneity[2]

The model consists of households, wholesale and retail firms, and a monetary policy authority. The representative household purchases consumption goods, holds bonds, and supplies labor to wholesale firms. The labor market is characterized by search and matching frictions. Wholesale firms produce a homogeneous good that is sold in a competitive market to retain firms. Retail firms transform the wholesale good into differentiated final goods which are sold to households for consumption and to wholesale firms to use in posting job vacancies.

[2] This section draws from Ravenna and Walsh (2013).

3.1 Households

The household consists of a continuum of members, a fraction $\bar{\gamma}$ of whom are of low (l) efficiency. The remaining $1-\bar{\gamma}$ are of high (h) efficiency. Workers of each type are either employed or searching for jobs. We follow the literature in assuming households pool consumption. Households are also the owners of all firms in the economy.

Firms must interview applicants to determine the worker's efficiency type. The aggregate number of interviews per period is determined through random matching as in standard matching models of the labor market, and all job seekers have identical interview-finding probability, regardless of skill level. At the interview, the job applicant's skill type is revealed. We assume the productivity of an h worker is high enough that it guarantees a positive surplus in all states.[3] Thus, if the skill level is revealed in the interview to be h, the worker is hired and produces with probability equal to one.

Regardless of whether employed or unemployed, each low-efficiency worker i receives a new idiosyncratic stochastic productivity level $a_{i,t}$ each period. We assume $a_{i,t}$ is serially uncorrelated and drawn from a distribution with support (0 1]. While productivity is randomly drawn in each period for a low-efficiency worker, the worker's efficiency-type, h or l, is permanently assigned.[4] Only low-efficiency unemployment workers with $a_{i,t} > \bar{a}_t$ will be hired, where $\bar{a}_t$ is an endogenously determined threshold level of productivity that will depend on an aggregate productivity shock and on the markup of retail over wholesale prices. In the absence of direct hiring and firing costs, $\bar{a}_t$ will also be the cut off value for determining whether an existing employed low-efficiency worker is retained by the firm. That is, from the perspective of the firm, the decision to retain or fire an existing low-efficiency worker with productivity $a_{i,t}$ is the same as the decision to screen out or hire a newly interviewed low-efficiency worker with productivity $a_{i,t}$.

In addition to these idiosyncratic shocks, we also allow for efficiency-biased aggregate productivity shocks z_t^h and z_t^l given by

$$z_t^j = z_t \phi_t^j;\ j = h, l$$

[3] This assumption is for simplicity as it will imply that endogenous separations and interviews that do not lead to hires only involve low skilled workers.

[4] We could assume match productivity is also random for high skill workers. If the support of the distribution is such that high-skill workers productivity for the least productive match is sufficiently higher then low-skill workers productivity for the least productive match, the basic results of our model would be unchanged.

where z is a common aggregate productivity shock and ϕ_t^j is a efficiency-specific shock. Thus, total productivity of a low-efficiency worker i is $z_t^l a_{i,t}$.while that of all high-efficiency workers is z_t^h.

We neglect labor force participation decisions and normalize the total workforce to equal one:[5]

$$L^l + L^h = L = 1,$$

where L^j denotes the labor force of type j, $j = h, l$. Let $\bar{\gamma} = L^l/L$ be the (fixed) fraction of the total labor force that is of low efficiency.

Let $N_t = N_t^l + N_t^h$ be total employment, where N^j be the number of type j workers who are employed, and let

$$\xi_t \equiv \frac{N_t^l}{N_t^l + N_t^h} = \frac{N_t^l}{N_t}$$

be the fraction of employed workers of low efficiency. The representative household maximizes

$$\mathrm{E}_t \sum_{i=0}^{\infty} \beta^i \delta_t^i \left\{ D_t \frac{\mathcal{C}_{t+i}^{1-\sigma}}{1-\sigma} - \left[v(h_{t+i}^h)(1-\xi_{t+i})N_{t+i} + \xi_{t+i} N_{t+i} \int_{\bar{a}_t}^1 v(h_{i,t+i}^l) f(a) da \right] \right\}, \quad (1)$$

where $\sigma > 0$ is the coefficient of relative risk aversion, δ_t^i is a discount rate shock, D_t is an aggregate preference shock, $\mathcal{C}_t$ is the sum of a market-purchased composite consumption good C_t and home produced consumption by unemployed workers $C_t^u = (1 - N_t)w^u$. In (1), the term

$$v(h_{t+i}^h)(1-\xi_{t+i})N_{t+i} + \xi_{t+i} N_{t+i} \int_{\bar{a}_t}^1 v(h_{i,t+i}^l) f(a) da$$

is the disutility to the household of having N_t members working, where hours worked depends on type and the idiosyncratic productivity shocks. We assume $v(h_{t+i}) = \ell h_{t+i}^{1+\chi}/(1+\chi)$.

Market consumption C_t is a Dixit-Stiglitz composite good consisting of the differentiated products produced by retail firms and is defined as

$$C_t = \left[\int_0^1 c_{kt}^{\frac{\theta-1}{\theta}} dk \right]^{\frac{\theta}{\theta-1}} \qquad \theta > 0.$$

[5] Erceg and Levin (2013) argue that it is important to model labor force participation and not just focus on the unemployment rate to understand the Great Recession.

Given prices p_{kt} for the final goods, this preference specification implies the household's demand for good k is

$$c_{kt} = \left(\frac{p_{kt}}{P_t}\right)^{-\theta} C_t, \tag{2}$$

where the aggregate retail price index P_t is defined as

$$P_t = \left[\int_0^1 p_{kt}^{1-\theta} dj\right]^{\frac{1}{1-\theta}}.$$

If i_t is the nominal rate of interest, the representative household's first order conditions imply the following must hold in equilibrium:

$$\lambda_t = \beta(1+i_t)\mathrm{E}_t\left(\frac{P_t}{P_{t+1}}\right)\lambda_{t+1}, \tag{3}$$

where λ_t denotes the total marginal utility of consumption at time t.

3.2 Labor flows, wholesale firms, wages and vacancies

The timing of labor market activity is as follows. The stock of producing matches (filled jobs) in period t is N_t of which $1-\xi_t$ are quality h and ξ_t are quality l. At the start of each period, there is an exogenous separation probability, denoted by ρ^x, that affects all employed workers, regardless of efficiency level. Workers who are not in a match at the start of the period, or who do not survive the exogenous separation hazard, are unemployed and seek new interviews. There are

$$S_t = 1 - (1-\rho^x)\, N_{t-1}$$

such job seekers. We define the end-of-period number of unemployed workers as

$$U_t = 1 - N_t.$$

The two measures of unemployment can differ as some job seekers find employment (and produce) during the period.[6]

[6] In search models based on a monthly period of observation, it is more common to assume workers hired in period t do not produce until period $t+1$. In this case, the number of job seekers in period t plus the number of employed workers adds to the total work force. Because we base our model on a quarterly frequency, we allow for some workers seeking jobs to find jobs and produce within the same period.

After exogenous separation occurs, all aggregate shocks realizations are observed. This allows wholesale firms to determine $\bar{a}_t$, the cutoff point for low-efficiency productive that will determine hiring and retention.[7] The time t idiosyncratic productivity shocks $a_{j,t}$ associated with employed low-efficiency workers and low-efficiency workers who are interviewed are observed. If $F(.)$ is the cumulative distribution function for $a_{i,t}$, then a fraction $1-F(\bar{a}_t)$ of type l workers receive productivity levels $a_{i,t} > \bar{a}_t$. With probability $\rho_t^n \equiv F(\bar{a}_t)$ a low-efficiency worker's productivity draw will be less than $\bar{a}_t$. An existing employee whose current idiosyncratic productivity is $a_{i,t} < \bar{a}_t$ is fired; an unemployed low-efficiency worker with $a_{i,t} < \bar{a}_t$ is not hired.

Wholesale firms post vacancies V_t. The number of vacancies, together with the number of job seekers, determine the number of interviews I_t via a standard CRS matching function:

$$I_t = \psi V_t^{1-a} S_t^a, \qquad 0 < \alpha < 1, \, \psi > 0, \tag{4}$$

The probability a job seeker gets an interview is $k_t^w \equiv I_t/S_t = \psi\theta_t^{1-a}$ where $\theta_t \equiv V_t/S_t$. Firms interview $k_t^f V_t$ workers in the aggregate, where $k_t^f = \psi\theta_t^{-a}$ is the probability a given vacancy receives an applicant to interview. Because of worker heterogeneity, the probabilities of being interviewed and being hired will differ by the worker's efficiency level. The job finding probability is identical to the interview rate for high-efficiency workers, $k_t^w = \psi\theta_t^{1-\alpha}$, while it is lower, and equal to

$$k_t^{w,l} = k_t^w \left(1 - \rho_t^n\right) < k_t^w$$

for low-efficiency workers.

Let S^j be the number of type j workers who are seeking jobs and $S_t = S_t^h + S_t^l$. Then the probability a worker drawn from the pool of unemployed job seekers is low efficiency is

$$\gamma_t \equiv \frac{S_t^l}{S_t^l + S_t^h} = \frac{S_t^l}{S_t}.$$

The overall job finding probability can be defined as

$$\gamma_t k_t^{w,l} + (1-\gamma_t) k_t^w = \left(1 - \gamma_t \rho_t^n\right) k_t^w.$$

With heterogeneous workers, a job opening that would be filled and lead to production

[7] We show below that $\bar{a}_t$ is the same for all firms.

if a high-efficiency applicant is interviewed may go unfilled if a low-efficiency worker is interviewed. New hires H_t are given by the number of interviewees who are of high efficiency, all of whom are hired, plus the number of interviewees who are of low efficiency times the fraction of these with productivity levels that exceed $\bar{a}_t$:

$$H_t = (1-\gamma_t)k_t^w S_t + (1-\rho_t^n)\,\gamma_t k_t^w S_t = (1-\gamma_t\rho_t^n)\,k_t^w S_t. \tag{5}$$

Note that fewer workers are hired than are interviewed: $H_t < k_t^w S_t$. The probability a randomly selected unemployed worker is screened out in the interview process (i.e., actually gets interviewed with a firm, is of low efficiency but has a $a_{i,t} < \bar{a}_t$ and so is not hired) is $\gamma_t\rho_t^n$. Screening implies new hires depend on the endogenous average quality of the pool of unemployed workers γ_t and the aggregate productivity and markup which we show below affects ρ_t^n.

Low-efficiency workers employed in existing matches that survived the exogenous separation hazard also receive a new productivity shock and are retained if and only if $a_{i,t} > \bar{a}_t$. Thus, actual employment in period t is equal to

$$\begin{aligned} N_t &= (1-\rho^x)\left[\left(1-\xi_{t-1}\right) + \xi_{t-1}(1-\rho_t^n)\right] N_{t-1} + H_t \\ &= (1-\rho^x)\left(1-\xi_{t-1}\rho_t^n\right) N_{t-1} + H_t \end{aligned}$$

The total separation rate is $(1-\rho^x)\left(1-\xi_{t-1}\rho_t^n\right)$, and the share of low efficiency employed workers evolves according to

$$\xi_t = (1-\rho_t^n)\left[\frac{(1-\rho^x)\xi_{t-1}N_{t-1} + \gamma_t k_t^w S_t}{N_t}\right]. \tag{6}$$

Job seekers at t who are of quality l equal the total number of low-efficiency workers minus the number of matches of quality l that survive the exogenous separation hazard. Hence,

$$\gamma_t = \frac{L^l - (1-\rho^x)\xi_{t-1}N_{t-1}}{S_t}. \tag{7}$$

In deriving (6) and (7) we assume workers who suffer exogenous separations can search within the same period, while those who experience endogenous separation, which occurs after shocks are realized during the period, cannot search until the following period.[8]

[8] Combining eqs. (6) and (7), it can be seen that job seekers at t who are of quality l arise from three sources: low-skilled workers who were searching for jobs in $t-1$ and failed to be hired; those employed

Since $a_{i,t}$ is *i.i.d.*, the model does not generate any endogenous distribution of efficiency-related productivity (each l worker may be more or less productive in every period), and an l worker can become less productive even if already in a match. But the share of low-efficiency workers in the unemployment pool, γ_t, is endogenous, so the efficiency-weighted productivity of both the workforce and the pool of unemployed changes over time. In particular, a burst of separations raises the average productivity of surviving matches and lowers the average efficiency level of the pool of unemployed job seekers.

Wholesale firms post vacancies V_t, interview and screen applicants, make hiring and retention decisions, and produce a homogenous output sold in a competitive market at price P_t^w. Define $\mu_t = P_t/P_t^w$ as the retail-price markup. Thus, expressed in terms of final retail goods, the current surplus of a firm-worker match involving a high-efficiency worker is

$$s_t^h = \left(\frac{z_t^h h_t^h}{\mu_t}\right) - \frac{v(h_t^h)}{\lambda_t} - w_t^{h,u} + q_t^h, \tag{8}$$

where h_t^h denote hours worked by an employed high-efficiency worker (so a high-efficiency worker produces $z_t^h h_t^h$ of the wholesale good), $v(h_t^h)$ is the disutility of hours worked, λ_t is the marginal utility of consumption, $w_t^{u,h}$ is the value of an unmatched high-efficiency worker's outside opportunity, and q_t^h is the continuation value of a match with a high-efficiency worker. All type h workers will work the same hours since they have the same productivity, and h_t^h is chosen optimally to maximize the match surplus. This implies

$$v_h'(h_t^h) = \left(\frac{z_t^h}{\mu_t}\right)\lambda_t \tag{9}$$

Let $h_{i,t}^l$ be the hours worked by an employed low-efficiency worker i; $h_{i,t}^l$ will depend on the worker's idiosyncratic productivity realization. The surplus of a match involving a low-efficiency worker is

$$s_{i,t}^l = \left(\frac{a_{i,t} z_t^l h_{i,t}^l}{\mu_t}\right) - \frac{v(h_{i,t}^l)}{\lambda_t} - w_t^{l,u} + q_t^l, \tag{10}$$

If a low-efficiency worker's productivity is too low, the surplus will be negative, leading to endogenous separation (or screening in the case of an interviewed job seeker). From (10),

in $t-2$ who survived the exogenous separation hazard but were endogenously terminated; and those employed in $t-1$ but who suffer the exogenous hazard at the start of period t.

the cutoff value of worker productivity at which the surplus produced by a low-efficiency worker equals zero is

$$\bar{a}_t = \frac{\mu_t \left(w_t^{u,l} + \frac{v(\hat{h}_t^l)}{\lambda_t} - q_t^l \right)}{z_t^l \hat{h}_t^l}, \tag{11}$$

where $\hat{h}_t^l$ maximizes the surplus and satisfies

$$v_h'(\hat{h}_t^l) \equiv \frac{\partial v(\hat{h}_t^l)}{\partial \hat{h}_t^l} = \left(\frac{\bar{a}_t z_t^l}{\mu_t} \right) \lambda_t, \; a_{i,t} \geq \bar{a}_t. \tag{12}$$

That is, hours $\hat{h}_t^l$ maximizes the joint surplus in a match with a low-efficiency worker of productivity $\bar{a}_t$.

Because the idiosyncratic productivity shocks are assumed to be serially uncorrelated, q_t^j depends on the efficiency-type of the worker in a match but is the same for all matches of the same efficiency-type. The continuation values are therefore given by

$$q_t^h = \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[(1-\rho^x) s_{t+1}^h + w_{t+1}^{u,h} \right]. \tag{13}$$

and

$$\begin{aligned} q_t^l &= \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[(1-\rho^x)(1-\rho_{t+1}^n)(s_{i,t+1}^l | a_{i,t+1} > \overline{a}_{i,t+1}) + w_{t+1}^{u,l} \right] \\ &= \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[(1-\rho^x) \int_{\bar{a}_{t+1}}^1 s_{i,t+1}^l f(a_i) da_i + w_{t+1}^{u,l} \right]. \end{aligned} \tag{14}$$

We assume wages are determined by Nash bargaining with the worker receiving a constant share η of the match surplus. Then the value of unemployment is equal to w^u plus the expected probability of being employed and receiving the surplus share ηs_{t+1}^j plus the expected value of remaining unemployed. For a high-efficiency worker this is

$$w_t^{h,u} = w^{h,u} + \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left(k_{t+1}^w \eta s_{t+1}^h + w_{t+1}^{h,u} \right), \tag{15}$$

while for a low-efficiency worker it is

$$\begin{aligned} w_t^{u,l} &= w^{h,u} + \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[k_{t+1}^w \eta (1 - \rho_{t+1}^n) \mathrm{E}_t (s_{i,t+1}^l | a_{i,t} > \overline{a}_{i,t}) + w_{t+1}^{l,u} \right] \\ &= w^{l,u} + \beta \mathrm{E}_t \left(\frac{\lambda_{t+1}}{\lambda_t} \right) \left[k_{t+1}^w \eta \int_{\bar{a}_t+1}^1 s_{i,t+1}^l f(a) da + w_{t+1}^{l,u} \right]. \end{aligned} \tag{16}$$

Matches of low-efficiency workers separate endogenously if $a_{i,t} < \bar{a}_t$. As claimed previously, $\bar{a}_t$ is the same for all firm considering the retention or hire of a low–efficiency worker. The probability of endogenous separation for a low-efficiency worker/firm match is also the probability a low-efficiency worker who receives an interview is not hired. If aggregate productivity falls or if the retail price markup μ_t increases, $\bar{a}_t$ will rise, lowering the fraction of low-efficiency unemployed that receive job offers and increasing the endogenous separation rate of already employed low efficiency workers. Low efficiency workers become a larger fraction of the unemployed pool, since the probability of separation is always higher than for high efficiency workers. Also, after a positive aggregate shock (even *i.i.d.*) the average duration of unemployment increases, as the low efficiency workers lose jobs faster and have a harder time finding new employment since they are more likely to be screened out during the interview process.

Wholesale firms post vacancies after observing aggregate variables, so their decisions are conditional on $\bar{a}_t$. If κ is the cost of posting a vacancy, expressed in terms of final goods, and firms receive a share $1 - \eta$ of the surplus from a match, the job posting condition is

$$k_t^f (1 - \eta) \left[(1 - \gamma_t) s_t^h + \gamma_t \int_{\bar{a}_t}^1 s_{i,t}^l f(a_i) da_i \right] = \kappa, \tag{17}$$

since with probability $(1 - \gamma_t)$ the firm interviews (and hires) a high-efficiency worker and with probability γ_t it interviews a low-efficiency worker. This condition can also be expressed as

$$k_t^f (1 - \eta) \left[s_t^h - \gamma_t \left(s_t^h - \int_{\bar{a}_t}^1 s_{i,t}^l f(a_i) da_i \right) \right] = \kappa.$$

Since the surplus from a high efficiency worker is greater than that from an employed low efficiency worker, a fall in the quality of the unemployment pool (a rise in γ_t) reduces the incentive to post vacancies.

Output of wholesale goods is obtained by aggregating over the output produced by employed high-efficiency workers and the output produced by employed low-efficiency

workers (i.e., those with idiosyncratic productivity levels greater than $\bar{a}_t$):

$$\begin{aligned} Q_t &= z_t^l N_t^l \left[\frac{\int_{\bar{a}_t}^1 a_{i,t} h_{i,t}^l dF(a_i)}{1 - F(\overline{a}_t)} \right] + z_t^h h_t^h N_t^h \\ &= \left\{ \phi_t^l \xi_t \left[\frac{\int_{\bar{a}_t}^1 a_{i,t} h_{i,t}^l dF(a_i)}{1 - F(\overline{a}_t)} \right] + (1 - \xi_t) \phi_t^h h_t^h \right\} z_t N_t. \end{aligned} \tag{18}$$

3.3 Retail firms

There are a continuum of retail firms, indexed by j, who purchase the wholesale good and convert it into differentiated final goods that are sold to households and wholesale firms. Retail firms maximize profits subject to a CRS technology for converting wholesale goods into final goods, the demand functions (2), and a restriction on the frequency with which they can adjust their price. Each period a firm can adjust its price with probability $1-\omega$. The real marginal cost for retail firms is the price of the wholesale good relative to the price of final output. This is just the inverse of the markup of retail over wholesale goods:

$$\mu_t \equiv \frac{P_t}{P_t^w}$$

This retail price markup is the driving force for inflation.

3.4 Monetary policy

We assume that the monetary authority in this economy implements monetary policy through a simple Taylor-type instrument rule, constrained by the requirement that the nominal rate be non-negative. The specific rule we incorporate takes the form

$$\ln(1 + i_t) = \max \left[0, \ -\ln \beta + \chi_i \ln(1 + i_{t-1}) + (1 - \chi_i) \left[\omega_\pi \pi_t + \omega_y \left(\ln Y_t - \ln \overline{Y} \right) \right] + \varepsilon_t \right]. \tag{19}$$

where ε_t is an i.i.d. policy shock. As a baseline policy we assume an inflation-targeting policy, setting $\omega_\pi = 1.5$, $\omega_y = 0$, $\chi_i = 0$.

3.5 Market clearing

Goods market clearing requires that household consumption of market produced goods equals the output of the retail sector minus final goods purchased by wholesale firms to

cover the costs of posting job vacancies. Hence, goods market equilibrium takes the form

$$Y_t = \Delta_t \left(C_t + \kappa V_t \right), \tag{20}$$

where $\Delta_t \geq 0$ is a measure of relative price dispersion.

4 Social efficiency and labor market distortions

The existence of time-varying worker heterogeneity implies that even when Nash-bargaining between workers and firms satisfies the Hosios conditions, the efficient allocation may not be attained by a market allocation. When firms separate or screen out at the interview phase low-efficiency workers, they also jointly decide the average efficiency level of the pool of searching workers from which all new matches are formed. This externality, which is not internalized by profit-maximizing firms, is not eliminated when the Hosios condition is met. We prove that, for a given path of consumption and job-finding probability, separations in the competitive equilibrium are inefficiently high. Since this translates into a higher share of low-efficiency workers among unemployed workers, lowering the expected benefit to posting vacancies, it will result in a higher unemployment rate.

In a basic new Keynesian model with search and matching frictions but a homogeneous labor force, Ravenna and Walsh (2012a) identify four distortions that arise in the competitive equilibrium with Nash bargaining over wages. The first two are standard in new Keynesian models: a non-zero steady-state markup under monopolistic competition generates a level of output that is inefficiently low, and price (or wage) rigidity generates inefficient fluctuations in the markup. The second distortion in Ravenna and Walsh (2012a), also present in our model, arises because the markup affects the optimal choice of hours and so fluctuations in the markup distorts hours from their efficient level. Finally, if the Hosios condition is not satisfied, the vacancy posting condition in the competitive equilibrium is distorted relative to the social planner's allocation. These four distortions are eliminated if (a) the Hosios condition holds; (b) the markup is constant (i.e., prices are stable); and (c) a subsidy to firms is used to raise steady-state output to its efficient level.

In the presence of labor heterogeneity, an additional distortion arises that is related to the compositional effect heterogeneity generates. Let $\bar{\mu}_t^j$ be the surplus value of an employed worker of type j in the planner's allocation. Recall that s_t^j denotes the joint

worker-firm surplus for a match involving a type j worker in the competitive equilibrium. Because type h workers are more productive,

$$s_t^h \geq s_t^l \text{ and } \bar{\mu}_t^h \geq \bar{\mu}_t^l.$$

Consider first the value of type h workers. In the appendix, we show that

$$s_t^h = \left[\left(\frac{\chi}{1+\chi}\right)\frac{z_t^h h_t^h}{\mu_t} - w^{h,u}\right] + (1-\rho^x)\beta \mathrm{E}_t\left(\frac{\lambda_{t+1}}{\lambda_t}\right)\left(1-\eta k_{t+1}^w\right) s_{t+1}^h$$

where η is the share of the surplus going to the worker and use has been made of the fact $v(h_t^h) = \ell h_t^{1+\chi}/(1+\chi)$ and h_t^h satisfies

$$\frac{v'(h_t^h)}{\lambda_t} = \frac{z_t^h}{\mu_t}.$$

These imply

$$\frac{v(h_t^h)}{\lambda_t} = \frac{z_t^h h_t}{\mu_t\left(1+\chi\right)}$$

In the efficient allocation, it holds that

$$\begin{aligned}\bar{\mu}_t^h &= \left[\left(\frac{\chi}{1+\chi}\right) z_t^h h_t^h - w^{h,u}\right] + \beta\left(1-\rho^x\right)\mathrm{E}_t\frac{\lambda_{t+1}}{\lambda_t}\left(1-\alpha k_{t+1}^w\right)\bar{\mu}_{t+1}^h \\ &\quad -\beta\left(1-\rho^x\right)\chi\left(1-\alpha\right)\mathrm{E}_t\frac{\lambda_{t+1}}{\lambda_t}\gamma_{t+1}\theta_{t+1}^{1-\alpha}X_{t+1},\end{aligned}$$

where the new variable X_t in the planner's allocation is defined as

$$X_t \equiv \bar{\mu}_t^h - \left[1-F(\bar{a}_t)\right]\bar{\mu}_t^l \geq 0$$

and h_t^h satisfies

$$\frac{v'(h_t^h)}{\lambda_t} = z_t^h.$$

Suppose the markup distortion has been eliminated and the Hosios condition holds. Then $\mu_t = 1$ and $\eta = \alpha$. In this case, for a given path of consumption and a given level of the job-finding probability, the difference between the private and efficient level of surplus

is given by

$$s_t^h - \bar{\mu}_t^h = \mathrm{E}_t Z_{t+1} + \beta\left(1-\rho^x\right)\mathrm{E}_t \frac{\lambda_{t+1}}{\lambda_t}\left(1-\alpha k_{t+1}^w\right)\left(s_{t+1}^h - \bar{\mu}_{t+1}^h\right) \tag{21}$$

where

$$\mathrm{E}_t Z_{t+1} = \beta\left(1-\alpha\right)\left(1-\rho^x\right)\mathrm{E}_t \frac{\lambda_{t+1}}{\lambda_t}\gamma k_{t+1}^w X_{t+1} \geq 0.$$

Eq. (21) shows that $s_t^h - \bar{\mu}_t^h$ can be expressed as the present discounted value of expected current and future values of Z_{t+i}, which is nonnegative. Thus, $s_t^h - \bar{\mu}_t^h \geq 0$ and matches involving type h workers generate an inefficiently *high* surplus in the competitive equilibrium relative to the first-best allocation.

For type l workers, the opposite condition will hold. For an employed type l worker i,

$$\begin{aligned} s_{i,t}^l - \bar{\mu}_{i,t}^l &= \beta\left(1-\rho^x\right)\mathrm{E}_t \frac{\lambda_{t+1}}{\lambda_t}\left(1-\alpha k_{t+1}^w\right)\left[1-F(\bar{a}_{t+1})\right]\left(s_{t+1}^l - \bar{\mu}_{t+1}^l\right) \\ &\quad -\beta\left(1-\alpha\right)\left(1-\rho^x\right)\mathrm{E}_t \frac{\lambda_{t+1}}{\lambda_t}\left(1-\gamma_{t+1}\right)k_{t+1}^w X_{t+1} \leq 0. \end{aligned}$$

Type l workers generate a *smaller* surplus in the competitive equilibrium than in the efficient allocation.

These differences in valuations translate into differences in the optimal cutoff productivity level that determines the endogenous separations rate and the fraction of type l unemployed workers who survive the screening process and secure jobs. Specifically, both $s_{i,t}^l$ and $\bar{\mu}_{i,t}^l$ are increasing functions of the idiosyncratic productivity of worker i. Since private matches involving type l workers end whenever $a_{i,t} < \bar{a}_t$, with $\bar{a}_t$ defined such that $s_t^l(\bar{a}_t) = 0$, it follows that

$$0 = s_t^l\left(\bar{a}_t\right) \leq \bar{\mu}_{i,t}^l\left(\bar{a}_t\right).$$

Therefore,

$$\bar{a}_t \geq \bar{a}_t^{eff},$$

where $\bar{a}^{eff}$ is the cutoff productivity level in the efficient allocation (i.e., such that $\bar{\mu}_t^l(\bar{a}_t^{eff}) = 0$). Thus, some type l workers who would remain employed in the efficient allocation, conditional on a given level of consumption growth $\frac{\lambda_{t+1}}{\lambda_t}$ and job finding probability k_t^w experience endogenous separation and become unemployed in the competitive

equilibrium. Similarly, some unemployed type l worker who obtain interviews but are screened out in the competitive equilibrium would be hired in the efficient equilibrium. Ceteris paribus, endogenous separations are too high in the competitive equilibrium implying also that average unemployment duration is inefficiently long.

5 Workers Heterogeneity and Slow Recoveries

This section discusses the role the composition effect plays in generating slow employment recoveries in response to a demand shock D_t. We ignore the ZLB constraint until section 6. To study the impact of worker heterogeneity in affecting the economy's response to an aggregate demand shock, we employ a parameterized version of the model.

5.1 Inspecting the Mechanism

The key variable for firms' hiring decisions is total factor productivity (TFP) among unemployed workers. In our model, time-varying worker heterogeneity amplifies fluctuations in TFP among unemployed workers, relative to among employed workers.

Recessions are times when there is a proportionally larger inflow into unemployment of marginal workers with low average productivity. For this to lead to a worsening of TFP among unemployed, two conditions need to hold. We first require that the share of job losers among unemployed be countercyclical. This is the case in our model, since flows between the labor force participation state and the out of the labor force state are precluded by assumption, and because job losers have a lower job-finding probability, lowering the outflow rate. In US data, instead, a large part of the flow into employment and unemployment comes from new entrants or re-entrants in the labor force, and selection bias can make the composition of these flows state-dependent. However, fig. 3 shows that once we account for temporary layoffs, the share of job losers among the unemployed is indeed strongly countercyclical, consistent with our assumption that the increase in unemployment is correlated with the increase in separations.

We then require that an increase in separations brings lower-productivity workers into the pool of unemployed. In an economy where all separations were endogenous, a fall in labor demand would lead to the firing of relatively *higher* productivity workers. At the other extreme, in an economy where all separations were exogenous, worker-specific productivity would be random, and independent of the business cycle. This would be the

case, for example, in an economy where job losers resulted from the random closing of plants employing workers of all productivity levels. In our model, we have a sufficiently low number of high-productivity workers separating exogenously, and a sufficiently high difference in TFP across high and low productivity workers, that the increase in the share of low productivity workers among job losers results in an inflow into unemployment of workers of lower average TFP.

When the composition effect is at work, it has several important implications. First, it lowers the incentive of firms to post job vacancies by reducing the expected productivity of a new match. In standard search and matching models, a rise in unemployment increases the job filling rate and so increases the expected value of posting a vacancy; the rise in vacancies causes unemployment to return rapidly to its steady state value. When the expected productivity of a new match falls, the incentive to post vacancies also falls, slowing the recovery of employment. The change in the composition of the unemployed pool in a downturn effectively induces a negative productivity shock confined to the pool of unemployed and decouples the dynamics of output and unemployment.

Second, the composition effect means firms with vacancies encounter fewer unemployed workers whose productivity is sufficiently high to generate a positive match surplus. This raises the probability that random matching results in the firm interviewing (and rejecting) a low productivity worker. Consequently, firms must, on average, search longer to fill a vacancy. As firms will correctly anticipate that it will take longer to actually make a hire, the incentive to post vacancies is reduced.

Worker heterogeneity alone is not sufficient to deliver slow recoveries in employment. What is necessary is a large enough countercyclical variation in the flow of marginal workers into unemployment, coupled with a small enough average share of low-efficiency workers in the unemployed pool. If on average the stock of unemployed workers has a large share of low-efficiency workers - either because their share in the labor force is large or because the average separation rate among low-efficiency workers is high on average - the composition effect will only result in a small additional volatility of TFP among the unemployed.

Finally, an important part of the explanation why slow recoveries turn into jobless recoveries relies on the availability of an intensive margin to change production. In a model without workers heterogeneity, the incentives to adjust along the hours or the employment margin are not very sensitive to the persistence of the shock driving a recession, nor to the stage of the recovery (early or late). With time-varying heterogeneity, a persistent

shock generates more layoffs, resulting in a more than proportional fall of TFP among unemployed relative to the case of a less persistent shock. This endogenously changes the relative benefit of changing production through a drop in hours or a drop in employment. In a slow recovery, the choice is biased towards a drop in employment, leading to a larger employment-to-output sacrifice ratio relative to a fast recovery.

5.2 Parameterization

The value of home production w^u, the coefficient ℓ scaling the disutility of labor hours, the cost of vacancy posting κ, the productivity of the matching technology ψ, the relative steady state productivity of high to low-efficiency workers $z^h_{ss}/\left(z^l_{ss}\int_0^1 a_i dF(a_i)\right)$ and the labor force share of low-efficiency workers $\bar{\gamma}$ are chosen to match the steady-state values for six variables with average aggregate data. Table 1 reports the matched steady state values, together with the additional parameters used in the numerical simulations.

The steady state unemployment rate is obtained averaging BLS quarterly data over 1948:1 to 2010:1. Since we do not have a direct measure for the unemployment rate of workers sorted according to unobservable productivity differentials, we measure the unemployment rate of workers with different efficiency levels using age-related data. We identify unemployment rates for low and high-efficiency workers with rates for workers' age-groups 16 to 24 and over-24. The steady state hours per worker h^{av}_{ss}, the steady state aggregate separation rate and the probability of a match between an applicant and a vacancy k^f_{ss} are parameterized to standard values in US business cycle literature. The share of output devoted to hiring activities is in line with empirical evidence reported in Ravenna and Walsh (2008).

Table 1: Baseline Parameterization		
Steady State Values		
Unemployment rate	u_{ss}	5.7%
Unemployment rate - $l-efficiency$ labor	u_{ss}^{l}	11.6%
Unemployment rate - $h-efficiency$ labor	u_{ss}^{h}	4.4%
Average hours per worker	h_{ss}^{av}	0.33
Vacancy posting cost share of output	$\frac{\kappa V_{ss}}{Y_{ss}}$	0.015
Probability of vacancy matched with applicant	k_{ss}^{f}	0.9
Parameters		
Unemployment elasticity of matching function	α	0.6
Discount factor	β	0.99
Inverse of labor hours supply elasticity	χ	1
Relative risk aversion	σ	1
Steady state inflation rate	π_{ss}	1
Workers' share of surplus	η	0.6
Exogenous separation rate	ρ^{x}	6.8%
Implied steady state separation rate	ρ_{ss}	7.45%
AR(1) parameter for demand shock D_t	ρ_z	0.95
Price elasticity of retail goods demand	ε	6
Average retail price duration (quarters)	$\frac{1}{1-\omega}$	4
Steady state markup	μ	1

Note: US unemployment rate for low- and high-efficiency workers is given respectively by the rate for the 16 to 24 and over-24 age-group, over the 1948:1-2010:1 sample (BLS, 2011).

The parameterization implies that $\bar{\gamma}$, the steady-state share of l workers in the labor force L, is 23.3%. Because the separation rate of l workers is about 45% larger than the overall separation rate, their share γ_{ss} in the steady-state pool of job seekers is 30%, while their share ξ_{ss} in the steady-state employment pool is 22%. Thus, low-efficiency workers are over-represented in the pool of unemployed.

To illustrate the relevance of time-variation in workers heterogeneity, we compare our baseline parameterization to an alternative with identical steady-state levels of output and unemployment but larger average labor flows. As shown in Ravenna and Walsh

(2012), large labor flows generate virtually constant shares of low-efficiency workers in the pool of unemployed over the business cycle.

Table 2: Alternative Parameterizations: Time-varying vs. Constant workers heterogeneity

			Baseline	Constant Heterogeneity
Productivity	Average productivity of high-efficiency workers		0.591	0.562
	Average productivity of low-efficiency workers		0.50	0.56
	Relative productivity of high/low-efficiency workers		1.18	1.01
	Average productivity labor force		0.57	0.561
	Average productivity employed workers		0.575	0.583
	Average productivity unemployed workers		0.603	0.56
Parameters	Matching function productivity	ψ	0.70	0.84
	Vacancy posting cost	κ	0.08	0.025
Steady State	Overall separation rate	ρ_{ss}	0.074	0.10
	Endogenous separation rate	ρ^n_{ss}	0.031	0.20
	low-efficiency unemployment share	γ_{ss}	0.30	0.53
	low-efficiency employment share	ξ_{ss}	0.22	0.19
	low-efficiency labor force share	$\overline{\gamma}$	0.233	
	Unemployment duration (quarters)		1.71	1.55

Note: Average productivity of high and low-efficiency worker-hours is given by z^h_{ss} and $z^l_{ss}\int_0^1 a_i dF(a_i)$. The two parameterizations have identical steady state output and unemployment

5.3 Demand Shocks and the Impact of Time-Varying Heterogeneity

We capture the impact of a fall in aggregate demand by simulating the economy's response to a negative realization of the preference shock D_t. Such a shock results in a change

in the marginal utility of consumption and a corresponding change in the disutility of hours entering the match surplus equation. Hours change endogenously; ceteris paribus, the level of hours that maximize the match surplus will fall for both worker types, as can be seen in (9) and (12). This will in turn lower the value of the surplus extracted from a match. Additionally, the fall in aggregate demand depresses the price of wholesale goods and raises the retail price markup, as retail prices are sticky. The change in the marginal utility of consumption and the resulting movement in the markup increases $\bar{a}_t$ (see 11), increasing endogenous separation. Since it is marginal workers with low match surplus who experience the increase in job loss, the fall in aggregate demand changes the composition of the pool of unemployed workers, increasing γ_t, the fraction of low-efficiency workers among job searchers.[9]

The impact of the change in the composition of the unemployment pool on optimal vacancy posting can be illustrated by rewriting the vacancy posting condition (17) as

$$\frac{V_t}{S_t} = \left\{ \frac{\psi}{\kappa}(1-\eta) \left[\gamma_t \int_{\bar{a}_t}^{1} s_{i,t}^{l} f(a_i) da_i + (1-\gamma_t) s_t^{h} \right] \right\}^{1/\alpha} \tag{22}$$

The right-hand side of (22) depends on the expected surplus from a match. Since the surplus from a high-efficiency worker is higher than the expected surplus from a low-efficiency worker, a worsening of the unemployment pool efficiency-level (an increase in γ_t) reduces the efficiency-weighted expected surplus and thereby reduces the incentive to post vacancies. Thus, the larger the increases in the share γ_t of l workers, the larger is the fall in the number of vacancies per searching worker, and the larger is the fall in the interview rate k_t^w.

The lower expected surplus generated by a low-efficiency worker is due to the lower productivity of these workers and the shorter expected duration of a match because of endogenous separation effects. As such workers become a larger share of the pool of unemployed workers, the effective average productivity of the unemployed falls. Our first experiment illustrates the endogenous change in effective TFP among unemployed workers that results from the time-varying heterogeneity in the pool of unemployed and examines its implications for employment behavior.

[9] Our parameterization implies that matches with high-efficiency workers always generate a surplus level above the one of the average low-efficiency workers. Since separations always happen first among low-efficiency workers, to simplify the model we assume that jobs of high-efficiency workers end at the exogenous and constant rate ρ^x,

Figure 4 shows the dynamic response to a persistent fall in the preference shifter D_t. The top panel shows the effects on the aggregate unemployment rate, with the middle and lower panels showing the unemployment rates for the low and high-efficiency workers respectively. The solid line shows the impulse response function when the composition does not vary endogenously. In this case, the effect on the overall unemployment rate is relatively small, a feature that is common to search models of the labor market with Nash bargaining. In the baseline parameterization, when composition varies endogenously, the impact of the shock is significantly amplified. Table 3 shows that the relative volatility of employment to output is increases sixfold when the composition effect is at work.

Table 3: Alternative Parameterizations:Employment volatility

		Baseline	Constant Heterogeneity
TFP Shocks	σ_n/σ_y	60%	10%
Demand Shocks	σ_n/σ_y	58%	9.8%

This result is obtained even though the baseline calibration actually incorporates very little heterogeneity across workers. Our parameterization assumes low-efficiency workers are only 23.3% of the labor force, and the average TFP of the employed worker-hour is only 1.95% higher than the average TFP for the unemployed. Nevertheless, the composition effect leads to a jobless phase in the beginning of the recovery. Employment starts growing only three quarters after the trough in output. The recovery is much slower overall relative to the case of when the composition effect is not at work, as shown in Table 4. After 3 years, employment has closed only 20% of the initial unemployment gap, a speed of recovery only half of the value seen in the alternative parameterization in which the composition effect does not play a role.

Table 4: Slow recoveries Employment recovery from trough	
Baseline	Constant Heterogeneity
20%	41%

Note: Employment gap relative to previous peak closed in the 12 quarters following trough. Recession generated by negative demand shock with AR(1) parameter $\rho^d = 0.95$.

The lag in employment growth depends on the progressive buildup of a larger share of low-efficiency workers in the unemployed pool, as shown in Figure 5. The fall in output in the two calibrations is similar, but the increase in the separation rate in the baseline economy (see the right panel of row 2), which is driven entirely by the firing of low-efficiency workers, and the relatively larger fall in the outflow rate for the same group of workers, increases the share of less productive workers in the unemployment pool by over 7.5%.

To single out the role of the composition effect in reducing the flow out of unemployment, figure 6 compares the behavior of different labor market variables to the counterfactual experiment in which the share of low-efficiency workers in the pool of unemployed, γ_t, remains constant. Thus it illustrates the direct impact of time-varying heterogeneity on the behaviour of the variables conditional on the baseline parameterization. The first panel of the figure shows that as the average efficiency-level of the pool of unemployed worsens, the average productivity among the unemployed falls. In our baseline parameterization, the share of low-efficiency workers among the unemployed is 30%, so a 7.5% increase in this share corresponds to an increase of the low-efficiency unemployment share from 30% to 32.25%. Every percentage point increase in γ_t brings about a loss in average productivity among the unemployed of 0.17%. This dampens the incentive of firms to recruit new workers while it shifts the composition of employed workers towards the high-efficiency type.

These endogenous movements in productivity cause fluctuations in employment and output to be decoupled in ways that would not occur if all workers were identical. Effectively, in this model laid-off workers are not homogeneous with employed workers. This is true not only at time of separation but also when they are looking for a new match.

Since the composition changes in the flow into unemployment during a recession are large relative to the composition shares in the stock of unemployed, they can affect aggregate labor market variables. In a model with homogeneous workers and endogenous separation, the amplification mechanism would not work, since TFP would move identically for unemployed and employed worker.

The top right panel of figure 6 compares the behavior of the log-change in vacancies per unemployed worker to the counterfactual in which γ_t remains constant. the change in the efficiency-composition of the unemployed increases the fall in vacancies by 30%. Figure 6 shows an additional channel through which unemployment volatility increases with workers heterogeneity. Firms become more selective in a recession and screen out more workers at the interview stage. The impact of γ_t on the hiring rate is given by the screening rate, that is, the unconditional rate at which an interviewee is screened out:

$$scr_t = \gamma_t \rho_t^n = \gamma_t [1 - \Pr(s_{i,t}^l > 0)]. \tag{23}$$

Ceteris paribus, in a recession the screening rate increases both because the separation rate ρ_t^n increases, and because the likelihood that an interviewee is a low-efficiency worker γ_t also increases.

5.4 Sluggish Output Recoveries and the Employment Sacrifice Ratio

The composition effect generates jobless recoveries. The slower is a recovery, the bigger the composition effect and the impact on employment, since the job finding rate is lower for low-efficiency workers. Thus a slow recovery is associated with more joblessness and a larger decline in the average productivity of unemployed workers. Compared to models with homogenous workers, the model with heterogeneous workers effectively implies a large negative demand shock also effectively generates a negative TFP shock to the unemployed. A more persistent shock leads to a more than proportional increase in the implied TFP shock. And the stronger this induced TFP shock is, the larger will be the divergence between the behavior of output and the behavior of employment. This occurs because high-productivity workers represent the vast majority of employed workers, and, critically, low-productivity workers are overrepresented in the inflow into unemployment, and underrepresented in the outflow from unemployment.

To illustrate the effects of a persistent adverse shock, figure 7 compares the impact of a fall in demand driven by a persistent preference shock (AR(1) coefficient equal to

0.95) to the impact of a less persistent fall in demand (AR(1) coefficient of 0.5). The less persistent shock implies a much smaller fall in employment, a smaller composition effect, and a smaller fall in average TFP among the unemployed. However, output falls on impact by a larger amount with the less persistent shock due to a decline in hours-worked. The decline in hours cushions the adjustment along the extensive employment margin, but both output and the unemployment rate rebound very quickly. In the resulting V-shaped recession, a larger share of the output decline is accounted for by the fall in hours, as opposed to employment. In contrast, a persistent adverse shock generates a smaller decline in hours and a much larger employment reduction.

Table 5 measures the impact of the output recovery speed on the employment behavior using the ratio of cumulative employment loss to cumulative output loss relative to previous peak over the 10 years of the recovery. In a recession driven by a long lasting shock, the composition effect leads to a cumulative fall in employment relative to output that is about twice as large as that generated by the less persistent shock.

Table 5: Slow recoveries

Employment sacrifice ratio

Slow output recovery	Fast output recovery
0.59	0.37

Note: sacrifice ratio defined as ratio of cumulative employment loss to cumulative output loss relative to previous peak over the 5 years of the recovery. Recession generated by negative demand shock. AR(1) parameter for slow output recovery: $\rho_d = 0.95$. AR(1) parameter for fast output recovery: $\rho_d = 0.5$.

One important driver of slow and jobless recoveries in our model is the endogenously changing incentive between adjusting output through the intensive or the extensive margin as the shock driving the downturn becomes more persistent. Table 6 shows that without a composition effect not only does the drop in hours explains 90% of the drop in output, but also the share of output drop accounted for by the hours drop is constant at all stages of the recovery. With time-varying worker heterogeneity, hours recover faster

than employment. More importantly, a slow recovery endogenously results in a bias towards reducing output through the employment margin, as opposed to the hours margin. After two quarters, the cumulative share of the fall in output accounted for by the fall in ours is only 52%, and as employment recovers more slowly, the share converges to 42% at the 5-year horizon.

Table 6: Hours adjustment share

	Slow recovery	Fast recovery	Constant heterogeneity Slow recovery
2 quarters	52%	67%	90%
1 year	47%	63%	90%
5 years	42%	61%	89%

Note: Adjustment share measures the share of cumulative output loss relative to previous peak explained by the cumulative fall in hours at each horizon. Recession generated by negative demand shock. AR(1) parameter for slow output recovery: $\rho_d = 0.95$. AR(1) parameter for fast output recovery: $\rho_d = 0.5$.

6 The Great Recession, The Zero Lower Bound and the Role of Monetary Policy

The previous section showed how the composition effect amplified the employment loss caused by an adverse demand shock and delayed the recovery of employment. The equilibrium responses of the economy in a new Keynesian model depend on the specification of monetary policy, and the results reported so far have been based on a non-inertia instrument rule in which the nominal interest rate responds only to inflation. And, importantly, the zero lower bound constraint has been ignored. In this section, we examine the role of the composition effect on labor market outcomes when monetary policy is limited by the ZLB.

It is useful, however, to first review how monetary policy affects output, employment and inflation in this model. The model includes both traditional channels operating

through the household's Euler condition, as is standard in new Keynesian models, and channels that are similar to a cost channel (Ravenna and Walsh 2006) that affect labor markets directly. Key to the effect of monetary policy is the countercyclical movement in markups generated by monetary policy shocks in the presence of sticky retail prices. As shown by Andrés, Doménech and Ferri (2012), the presence of price rigidities, even absent wage rigidities, contributes significantly to the ability of the basic search and matching model to match the volatility of labor market tightness and vacancies, as well as generate a negatively sloped Beveridge curve.

There are several ways monetary policy affects the economy. A reduction in the real interest rate leads households to increase current consumption. This increases the demand for final goods, and retail firms respond by increasing production as in standard new Keynesian models. To increase production, retail firms must purchase more wholesale goods, and because retail prices are sticky, this pushes up the price of wholesale goods relative to retail goods; the retail price markup falls, generating a countercyclical movement in the markup.[10]

The movement in the markup affects several margins relevant for separations, job creation and hours. Consider first the effects on high-efficiency workers operating through the markup. The surplus associated with a high-efficiency worker, given by (8), rises when μ_t falls. This increases the incentive for wholesale firms to post vacancies. From (9), the fall in μ_t also increases the optimal hours worked by high-efficiency workers. So wholesale firms increase hours of existing high-efficiency employees and post more job vacancies.

The surplus generated by a low-efficiency worker is similarly increased by a fall in the markup, as are their optimal hours (see 10 and 12). However, the markup also affects the cutoff productivity level $\bar{a}_t$ given by (11). A fall in μ_t lowers $\bar{a}_t$, implying a fall in endogenous separations and a fall in the screening rate – more low-efficiency unemployed workers who are interviewed are subsequently hired (see 5). Thus, hours rise, vacancies rise, fewer low-efficiency workers are screened out, and so match efficiency rises.

Effects on the markup are not the only channel through which monetary policy affects the labor market. Current match surpluses and the critical cutoff productivity level determining endogenous separations and low-efficiency hiring also depend on the continuation value of a match. An expansionary monetary policy shock lowers the real interest

[10] Nekarda and Ramey (2010) argue that markups are procyclical in the U.S.. However, this is not inconsistent with the effects of monetary policy discussed here, as productivity shocks would generate a procyclical markup.

rate, and this increases the present discounted value of future surpluses. By raising the present value of future match surpluses, expansionary monetary policy increases job creation, reduces endogenous separations, and leads more low-efficiency unemployed workers to be hired at the interview stage.

A final channel of monetary policy arises because job posting costs generate a demand for final goods on the part of wholesale firms. Expansionary monetary policy, by stimulating job creation leads to a direct increase in the demand for final goods. Even though the present model abstracts from physical capital, a fall in the real interest rate increases the present value of future matches. Wholesale firms investment in matches. Since final goods are required to hire more workers, the demand for final goods rises.

6.1 The Great Recession and the ZLB

As seen, for example, in Figure 5, the composition effect generates a much larger unemployment increase and a much slower job recovery for a given path of output. And Figure 7 showed how more persistent recessions generated much more delayed employment recoveries when the composition effect was present. This suggests that a primary issue in understanding the slow recovery of employment is to understand the reason for the persistence of the economic downturn. In this section, we explore the role the zero lower bound on nominal interest rates may have played in limiting the ability of monetary policy to help engineer a more rapid recovery. To solve for the law of motion at the zero lower bound for i_t, we adapt the code in Guerrieri and Iacoviello (2013) to allow for persistent shocks to the discount rate. Thus in our model the future behaviour of the shocks leading the economy to the zero lower bound is fully expected as of the first period when the shocks arrive. The solution method solves models with occasionally binding constraints by building the time-varying decision rule for the whole period when the constraint binds. [11]

In figure 8, the line marked by triangles (and in red) eliminates the endogenous composition effect but incorporates the ZLB as a constraint on monetary policy. The contractionary demand and discount rate shocks produce a sharp fall in output (and

[11] The algorithm starts with a guess as for the duration of the period where the constraint binds, building expectations using the recursive law of motion for the linear approximation to the model when the constraint does not bind. It can then build moving backward the law of motion at each point in time when the constraint is binding. The algorithm iterates until no violation of the constraint remains. This solution method allows for endogenously determining the horizon over which the ZLB is a binding constraint, given the dynamics of the shock.

inflation – see top two panels) and the nominal interest rate remains at zero for two years (middle row, left panel). Hours per worker falls significantly (lower left panel), and hours recover slowly, as does output. If fact, it is hours and not employment that primarily adjusts in this scenario. As the right panel in the middle row illustrates, the rise in unemployment is limited, peaks immediately, and then converges quickly back towards its steady-state value. Without worker heterogeneity, the model cannot generate slow employment growth even when the ZLB constrains the ability of monetary policy to speed the recovery.

The solid (black) line shows the responses in the baseline model of worker heterogeneity when the ZLB constrains monetary policy. The top two panels and the left panel of row 2 show that worker heterogeneity has little effect on the severity or persistence of the effects on output and inflation, or on the behavior of the nominal interest rate. Worker heterogeneity matters a great deal, however, for labor market developments. The right panel of row 2 shows that unemployment rises significantly more, and hours fall less (left panel, row 3). Thus, when coupled with the ZLB, time-varying heterogeneity leads to virtually the same recession and the same output recovery as in a model with homogeneous workers, but it generates slow employment growth during the recovery.

For comparison, figure 8 also shows the effects in the face of worker heterogeneity when the ZLB is ignored (dotted, blue line). Comparing the dotted line and the solid line provides a measure of the effects of the ZLB. As expected, the sharp fall in output is reduced when policy is not constrained by the ZLB. Despite ignoring the ZLB, worker heterogeneity generates unemployment increases that are much larger and more persistent than under the homogeneous-worker case that accounted for the ZLB (the triangles). Hence, time-varying heterogeneity in the composition of the unemployed generates more employment volatility, even in the absence of the ZLB.

6.2 The Great Recession and the Dynamics of Wages, Productivity and the Beveridge Curve

The dynamics of wages, labor productivity, vacancies and unemployment during the Great Recession are the subject of a growing literature, providing insights into the causes of the recession and the interaction of labor market and aggregate macroeconomic variables over the downturn and subsequent recovery. We discuss in the following the interpretation provided of the 2008-2012 period by our model.

6.2.1 Wages and Labor Productivity

During the Great Recession and the subsequent recovery, compensation growth has not slowed as much as expected, given the amount of slack in the economy and the high level of unemployment experienced since the initial downturn (Linder, Peach and Rich, 2012). Figure 10 shows that real hourly wages in the nonfarm business sector fell at the beginning of the recession, then in 2009 and 2010 they remained constant, about 1.5% above the pre-recession level towards which they revert in 2011-2012.

Several explanations have been put forward to explain real wage resistance during a major downturn, including the existence of downward nominal wage rigidity. Hobijn and Daly (2013) provide microeconomic evidence of frictions in downward wage adjustment over the Great Recession period, and several business cycle models with search frictions in the labor market assume sticky (incentive-compatible) behaviour for wages as an amplification mechanism for unemployment fluctuations.

The behaviour for the hourly real wage predicted for the Great Recession by our model is shown in the top panel Figure 9. In the model, wages behave as if they were sticky: They fall at the trough of the cycle, but after 6 quarters - with the unemployment rate having recovered only around 35% of its initial increase, and still 1.75 percentage points above its steady state value - wages have already closed over 90% of the gap relative to the previous peak. Wages remain stable all the way to the end of the 5th year of the recovery, when unemployment is still 1% above steady state.

The endogenous stickiness of wages is the result of the composition effect and the interaction of the intensive and extensive margin over a demand-driven downturn. Match surplus falls during the downturn, but part of the fall is explained by the endogenous fall in hours worked per match. The fall in hours implied by the model (shown earlier in Figure 8) is consistent with the behaviour of hours in the data, displayed in Figure 11, where they drop during the recession, and slowly revert to steady state over the subsequent four years. Thus, in the model the surplus per hour - a share of which is the wage paid to workers - does not fall as much as total surplus. At the same time, this does not prevent the fall in match surplus from having a strong impact on hiring. As the total surplus falls, ceteris paribus, the implied cost of a match increases since firms equate the fixed vacancy cost with the total, rather than the hourly, surplus.

The fall in the cyclicality of labor productivity over the last three recessions has been extensively discussed in the literature (Berger, 2012, Gali and van Rens, 2010).

Petrosky-Nadeau (2012) and Fernald (2013) find that the dynamics of labor productivity over the Great Recession have not deviated much from its trend. Productivity growth, displayed in Figure 12, slowed during the recession, while in the recovery it quickly went back to its previous trend. Its dynamics is largely explained by changes in capital deepening, labor quality and utilization rates. Our model predicts in response to a demand shock nearly completely acyclical productivity growth (Figure 9). As we model the Great Recession as the consequence of a negative demand shock, the dynamics of labor productivity depend exclusively on the changing composition of the employment pool. The changes in the flows into and out of unemployment lead to a large percentage change of the composition of the pool of unemployed - which is a small fraction of the labor force - but a relatively small change in the pool of employed, which is much larger and where low-efficiency worker are under-represented. In addition, the average low-efficiency worker in the employment pool is more productive than the equivalent worker in the unemployment pool.[12] Therefore, the changing composition of labor flows has a large impact on unemployment and incentives to hire, but a small impact on measured labor productivity.

6.2.2 Shifts in the Beveridge Curve

One of the most notable consequences of the Great Recession has been the apparent shift in the Beveridge curve, the relationship between the vacancy rate and the unemployment rate. The US economy has experienced several shifts in the Beveridge curve over past business cycles (Bleakley and Fuhrer, 1997), and faster increases in the vacancy rate relative to the employment rate during a recovery are not inconsistent with search and matching models of the labor market. Yet the persistence of unemployment, at the same time as the vacancy rate increased after 2010, and the below-expectation number of hires per vacancy posted, has lead several authors to examine the possibility of long-term increases in the natural rate of unemployment, sectoral or geographical mismatch, changes in recruiting intensity by firms, and large falls in the efficiency of the matching process (Barnichon and Figura, 2010, Davis et al., 2012, Furlanetto and Groszheny, 2012, Ghayad and Dickens, 2012, Hobijn and Sahin, 2013, Lubik, 2012).

The two bottom panels of Figure 9 show the behaviour of the vacancy yield and vacancies per unemployed worker predicted by our model for a downturn simulating the

[12] That is, employed low-efficiency workers have $a_{i,t} > \bar{a}_t$.

Great Recession. Similarly to the data shown in Davis et al. (2012), the vacancy yield increases sharply in the recession, and then slowly reverts to steady state. The higher number of unemployed workers ensures that each vacancy posted has a better chance to be matched to a worker. Since the screening rate also increases, the change in the vacancy yield is far less what would be predicted by a model with constant workers heterogeneity. Thus, the slower is the recovery, the stronger is the composition effect, and the lower the vacancy yield will be relative to a forecast neglecting the composition effect.

While in the model the vacancy rate recovers faster than in the data, the composition effect implies that firms post many fewer vacancies conditional on the number of unemployed workers, relative to an economy without time-varying worker heterogeneity. The number of vacancies per unemployed worker falls dramatically during the downturn (Figure 9). This has important consequences for the behaviour of the Beveridge curve. Since the vacancies-unemployment ratio falls below the steady state, and both vacancies and employment fall, the economy would move to a point below and to the right of the steady state in the unemployment-vacancies space. In subsequent periods, unemployment is much more inertial than vacancies. Thus, in the Beveridge curve space the dynamics imply what appears of a rightward shift of the curve itself– vacancies increase relative to their trough but unemployment does not budge. This can happen at the same time as the number of vacancies per unemployed worker is below the steady state: it simply means that the fall in employment is much larger than the fall in vacancies. But this is what delivers the persistent unemployment: if this ratio did not fall, unemployment would not increase as much since matching would be much easier.

Thus, time-varying worker heterogeneity generates endogenously deviations from the Beveridge curve. The dynamics of the economy over the simulated Great Recession downturn is shown in Figure 13. Three years after the trough, the vacancy rate is above its steady state level, while the unemployment gap is still equal to 1.25 percentage points.

In a model with homogeneous workers and a constant separation rate, the matching function and the law of motion for labor imply a steady state relationship between vacancies V_t and searching workers S_t :

$$V = \left(\frac{1}{\psi}\right)^{\frac{1}{1-a}} \left[\rho_x \left(\frac{1-S}{S}\right)\right]^{\frac{1}{1-a}} S \tag{24}$$

Because in our model the matching function only determines the number of interviews,

the relationship between steady state vacancies and unemployment depends on all the equilibrium relationships in the model. Figure 13 displays a steady state Beveridge curve as a function of aggregate TFP, keeping fixed all the other parameters of the model.[13] The dynamics of vacancies and unemployment implied by our model deviate from the steady state relationship. These deviations could be rationalized, if interpreted through (24), only by changes in the match efficiency level ψ. Most business cycle models with search frictions in the labor market do not generate large and very persistent unemployment coinciding with fast-changing vacancy rates. When estimated, these models will imply large movements in the match efficiency level over the Great Recession period (Furlanetto and Groszheny, 2012).

7 Conclusions

Standard search and matching models that assume homogeneous labor and flexible wages fail to generate the type of slow and delayed recovery of employment seen in the Great Recession. Using a simple model of worker heterogeneity, we demonstrate how a persistent economic downturn generated by an aggregate demand shock adversely affects the average productivity of the pool of unemployed workers. This affect, which reduces the expected surplus of a new match and so dampens vacancy postings, works much like a negative TFP shock, though one that only affects unemployed worker. Given the level of vacancies and unemployment, the rate of new hiring falls, mimicking the effects of a decline in the efficiency of the matching function. The composition effect accounts qualitatively for a slow decline in unemployment that lags behind the recovery in output.

In general, monetary policy can act to prevent demand shocks from generating large and persistent recessions. And in mild, short-lived recessions, the composition effect plans a small role. However, when the zero lower bound limits the ability of monetary policy to offset adverse shocks, resulting in a Great Recession, the composition effect we highlight can play a central role in the economy's dynamic adjustment.

[13] To ensure the existence of a steady state for a sufficiently large range of aggregate unemployment rates while keeping constant the parameterization, the low-efficiency labor force share $\overline{\gamma}$ was adjusted downward to 7%, relative to the value of 15.8% used in the simulations. A larger value of $\overline{\gamma}$ flattens the implied steady state Beveridge curve. An average Beveridge curve could be built from simulating the model, matching the empirical evidence in Shimer (2005). This would require a full model estimation, since the simulation result will depend on the relative volatility of demand, TFP and policy shocks.

References

[1] Abbring, J., van den Berg, G. and van Ours, J., "The anatomy of unemployment dynamics", *European Economic Review* 46, 1785-1824, 2002.

[2] Andrés, J., R. Doménech and J. Ferri. "Price Rigidity and the Volatility of Vacancies and Unemployment." Universidad de Valencia, January 2012.

[3] Barnichon, R. and Figura, R., "What drives matching efficiency? A tale of composition and dispersion", mimeo, 2011.

[4] Baker, M., "Unemployment Duration: compositional effects and cyclical variability," *American Economic Review* 82, 315-321, 1992.

[5] Bils, M., Chang, Y. and Kim, S., "Comparative advantage and unemployment," NBER Working Paper 15030, 2009.

[6] Blanchard, O. J. and Jordi Galí, "Real wage rigidity and the new Keynesian model," *Journal of Money, Credit and Banking*, 39 (1), 2007.

[7] - , "A New Keynesian Model with Unemployment," *American Economic Journal: Macroeconomics,* vol. 2 n^{o}2, 1-30*,* 2010.

[8] Clark, K. and Summers, L., "The Demographic Composition of Cyclical Employment Variations," *Journal of Human Resources*, Vol. XVI, pp. 61-79, 1981.

[9] Daly, M., Hobijn, B. and Valletta, R., "The recent evolution of the natural rate of unemployment," FRB San Francisco Working Paper 2011-05, 2011.

[10] Darby, M., Haltiwanger, J. and Plant, M., "Unemployment rate dynamics and persistent unemployment under rational expectations," *American Economic Review* 75, 614-637, 1985

[11] Davis, S., "Job Loss, Job Finding, and Unemployment in the U.S. Economy over the Past Fifty Years. Comment," in NBER Macroeconomics Annual, Vol. 20 (2005), pp. 139-157.

[12] - , John C. Haltiwanger, and Scott Schuh, *Job Creation and Job Destruction*, Cambridge, MA: The MIT Press, 1996.

[13] den Haan, W. J., G. Ramey, and J. Watson, "Job destruction and Propagation of Shocks," *American Economic Review*, June 2000, 90 (3), 482-498.

[14] Elsby, M., B. Hobiijn, and A. Sahin, "Unemployment dynamics in the OECD," NBER Working Paper No. 14617, Dec. 2008.

[15] - , "The labor market in the great recession", NBER Working Paper No. 15979, 2010.

[16] Erceg, C. and A. Levin, "Labor force participation and monetary policy in the wake of the Great Recession," IMF, 2013.

[17] Galí, J. *Unemployment Fluctuations and Stabilization Policies: A New Keynesian Perspective*. MIT Press: Cambridge, MA, 2011.

[18] Gertler, M. and Antonella Trigari, "Unemployment Fluctuations with Staggered Nash Wage Bargaining," *Journal of Polictical Economy,*117 (1), pp. 38-86, 2009.

[19] Gertler, M., Sala, L. and Antonella Trigari, "An Estimated Monetary DSGE Model with Unemployment and Staggered Nominal Wage Bargaining," *Journal of Money, Credit and Banking* 40-8, pages 1713-1764, 2008.

[20] Guerrieri, L. and Iacoviella, A Toolkit to Solve Models with Occasionally Binding Constraints Easily. Federal reserve Board, 2012.

[21] Guerrieri, V., "Heterogeneity and Unemployment Volatility," *Scandinavian Journal of Economics*, 2007, 109, No. 4, 667-693

[22] Hines, J., Hoynes, H. and Krueger, A., "Another look at whether a rising tide lifts all boats," in Krueger, A. and Solow, R., eds., The roaring nineties: Can full employment be sustained?, New York, Russel Sage Foundation, 2001.

[23] Hornstein, A., Krusell, P. and Violante, G., "Frictional wage dispersion in search models: a quantitative assessment," American Economic Review: Vol. 101 No. 7 (December 2011).

[24] Krause, M. and Lubik, T., "On-the-Job Search and the Cyclical Dynamics of the Labor Market," FRB Richmod Discussion Paper 10-12, 2010.

[25] Jung, P. and Kuhn, M., "Labor market rigidity and business cycle volatility," mimeo, 2011.

[26] Lechthaler, W., C. Merkl, and D. Snower, "Monetary persistence and the labor market: a new perspective," *Journal of Economic Dynamics and Control* 34 (2010): 968-983.

[27] Lemieux, T., "Increasing residual wage inequality: composition effects, noisy data or rising demand for skills?", *American Economic Review,* 2006

[28] Mortensen, D. T., Wage Dispersion: Why are Similar People Paid Differently. MIT Press, 2003.

[29] Mueller, A., "Separations, Sorting and Cyclcal Unemployment," mimeo, 2011.

[30] Nagypal, E., "Learning-by-Doing Versus Learning About Match Quality: Can We Tell Them Apart?" *Review of Economic Studies*, 74 (2), pp. 537-566, 2007.

[31] Nagypal, E. and Mortensen, D. T., "Labor-market Volatility in Matching Models with Endogenous Separations," *Scandinavian Journal of Economics*, Vol. 109, No. 4., pp. 645-665, 2007.

[32] Petrongolo, B. and Pissarides, C., "Looking into the Black Box: A Survey of the Matching Function," *Journal of Economic Literature*, vol. 39(2), pages 390-431, 2001.

[33] Petrosky-Nadeau, N., "TFP during a Credit Crunch." Carnegie Mellon University, September 2012.

[34] Pries, M., "Worker heterogeneity and labor market volatility in matching models," *Review of Economic Dynamics,* 11, 644-687, 2008.

[35] - , and Rogerson, R., "Hiring policies, labor market institutions, and labor market flows," *Journal of Political Economy,* 113(4), 2005.

[36] Ravenna, F. and C. E. Walsh, "Optimal monetary policy with the cost channel," *Journal of Monetary Economics* 53, 2006, 199-216.

[37] - , "Vacancies, Unemployment, and the Phillips Curve," *European Economic Review*, 2008, 52: 1494-1521.

[38] - , "Welfare-based optimal monetary policy with unemployment and sticky prices: A linear-quadratic framework," *American Economic Journal: Macroeconomics*, April 2011, 3(2): 130-162

[39] - , "The welfare consequences of monetary policy and the role of the labor market: a tax interpretation," *Journal of Monetary Economics* March 2012a, 59: 180-195.

[40] - , "Labor market flows with skill heterogeneity in a monetary policy model," *Journal of Money, Credit and Banking* Vol. 44:s2, 31-71, December 2012b.

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

[42] - , "Reassessing the ins and outs of unemployment," mimeo, 2007.

[43] Solon, G., Barski, R. and Parker, J., "Measuring the cyclicality of real wages: how important is composition bias?" *Quarterly Journal of Economics,* February 1994, pp. 1-28.

[44] Tasci, M., "On the job search and labor market reallocation," *Federal Reserve Bank of Cleveland Working paper* 7-25, 2007.

[45] van den Berg, G. and van der Klaauw, B., "Combining micro and macro unemployment duration data," *Journal of Econometrics* 102, 271-309, 2001.

[46] Villena-Roldan, B., "Aggregate implications of employer search and recruiting selection," mimeo, 2008.

[47] Walsh, C. E., "Labor market search and monetary shocks," in *Elements of Dynamic Macroeconomic Analysis*, S. Altuğ, J. Chadha, and C. Nolan (eds.), Cambridge: Cambridge University Press, 2003, 451-486.

[48] - , "Labor market search, sticky prices, and interest rate policies," *Review of Economic Dynamics*, 8(4), Oct. 2005, 829-849.

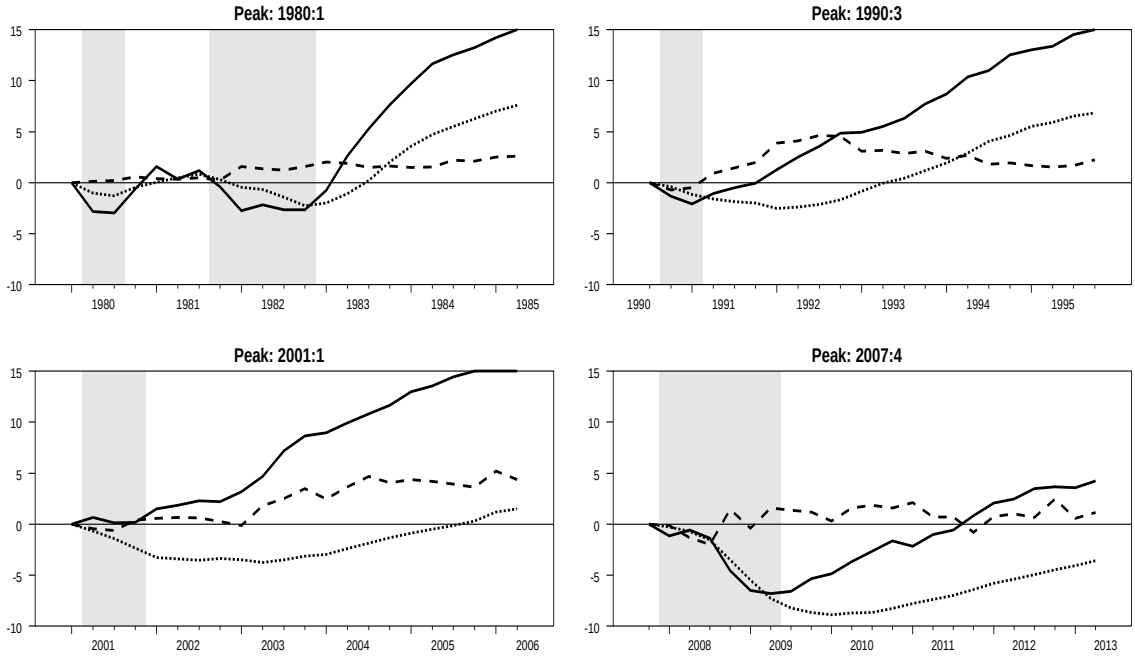

Figure 1: Output (solid), employment (dotted), and real compensation per hour (dash) for the nonfarm business sector. Variables measured in logs relative to cyclical peak. Source: Haver analytics.

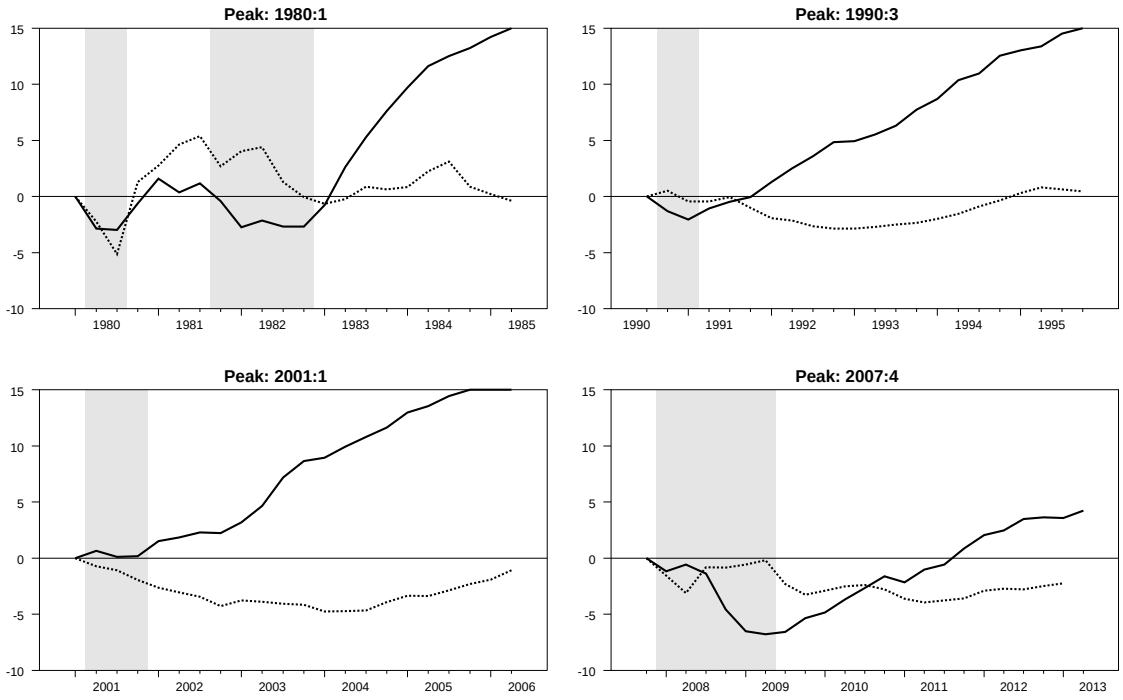

Figure 2: Nonfarm business sector output (solid) and the real federal funds rate (dotted). Real federal funds rate computed using nominal federal funds rate and PCE index inflation. Variables measured in logs relative to cyclical peak. Source: Haver analytics.

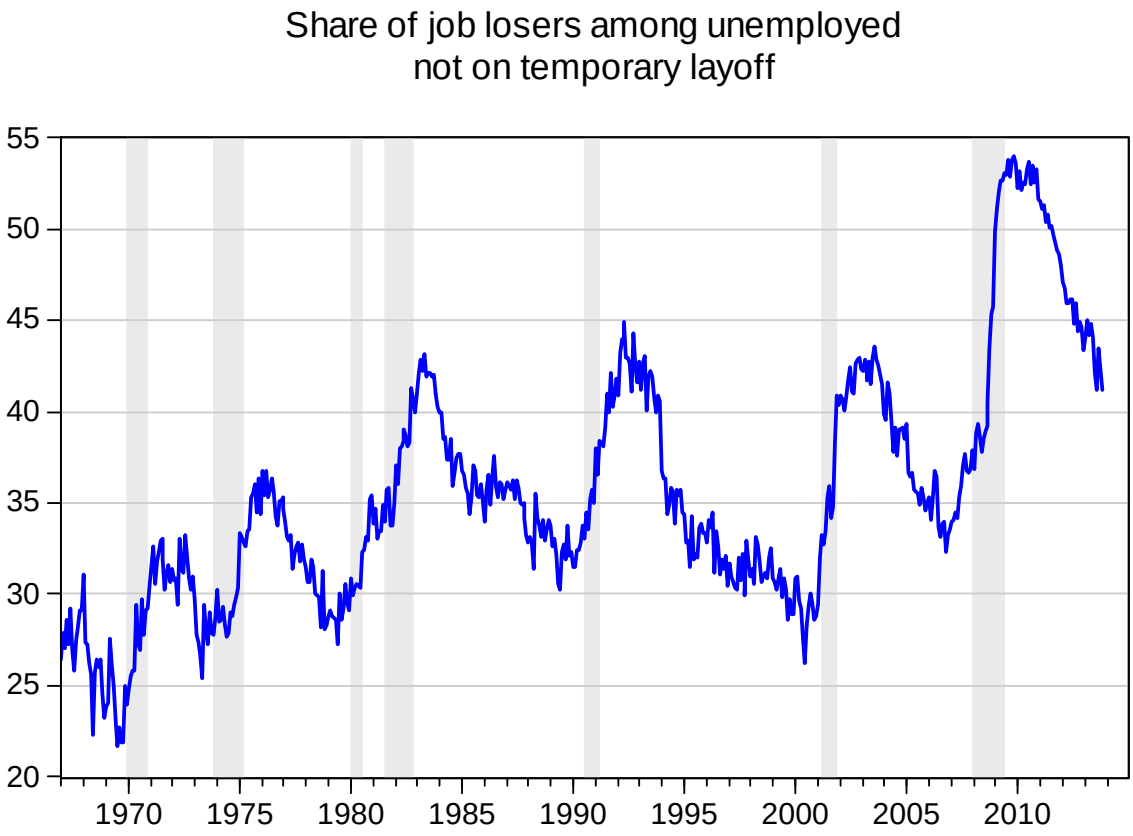

Figure 3: Percent of total unemployed 16 years old and over classified as job losers not on temporary layoff as reason for unemployment. Seasonally adjusted monthly data. Shaded areas indicate NBER recession dates. Source: CPS series LNS13026511.

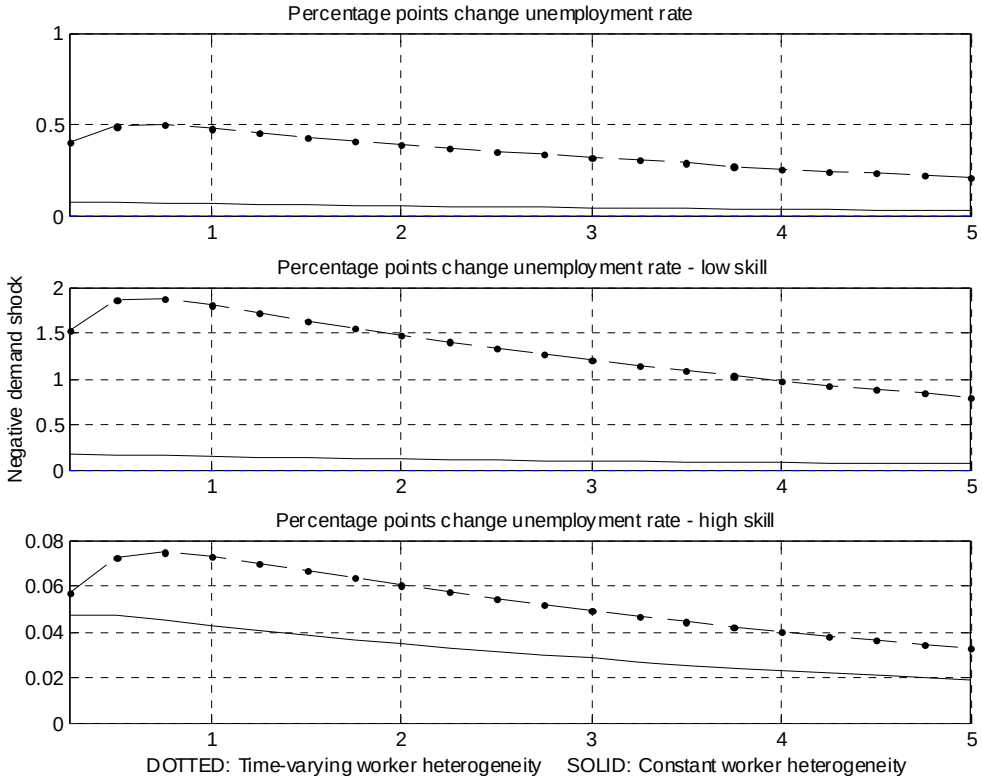

Figure 4: Impulse response to a negative demand shock D_t for the time-varying and constant workers heterogeneity economies. Monetary policy set by Taylor rule responding to CPI inflation. AR(1) coefficient of demand shock $\rho_{d_t} = 0.95$. Change in unemployment rate for total, low-efficiency and high-efficiency population scaled in percentage points of the labor force L, L^h, and L^l of each group. Horizontal axis in years.Impulse response to a negative demand shock D_t for the time-varying and constant workers heterogeneity economies. Monetary policy set by Taylor rule responding to CPI inflation. AR(1) coefficient of demand shock $\rho_{d_t} = 0.95$.

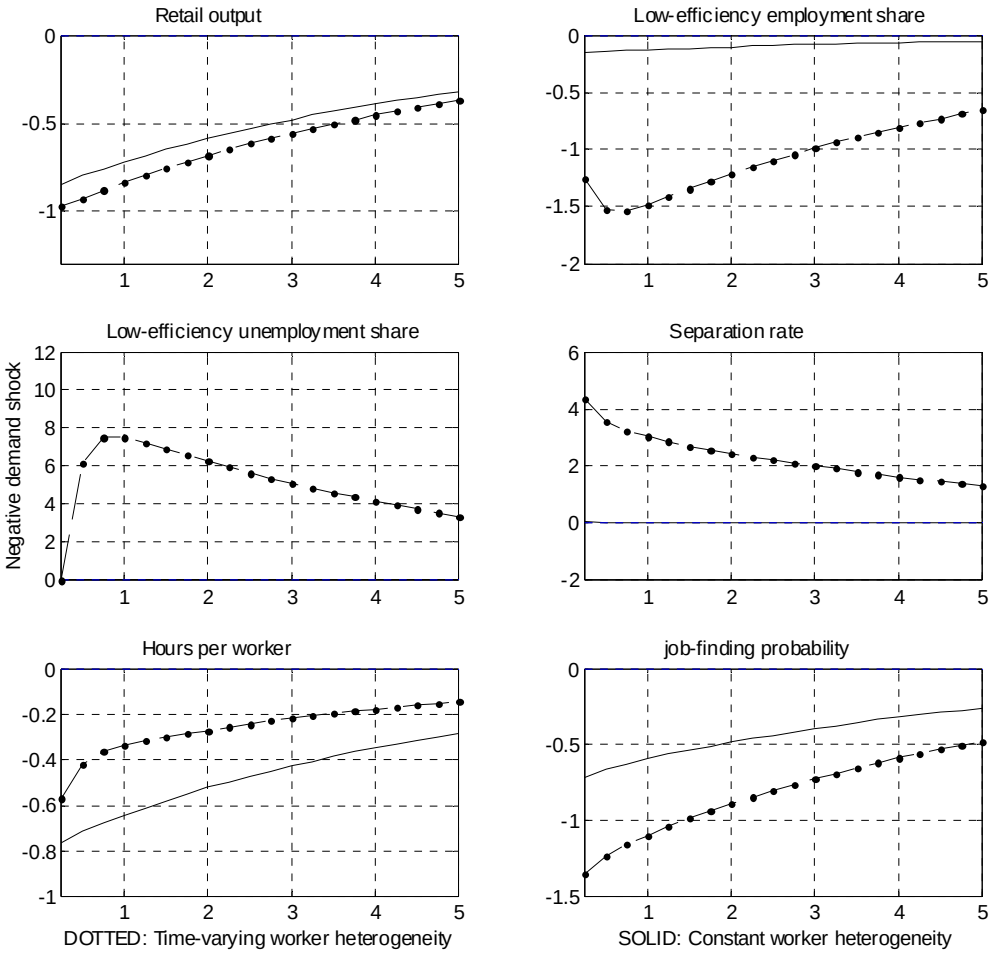

Figure 5: Impulse response to a negative demand shock D_t for the time-varying and constant workers heterogeneity economies. Monetary policy set by Taylor rule responding to CPI inflation. AR(1) coefficient of demand shock $\rho_{d_t} = 0.95$. Percent deviations from steady state. Horizontal axis in years.

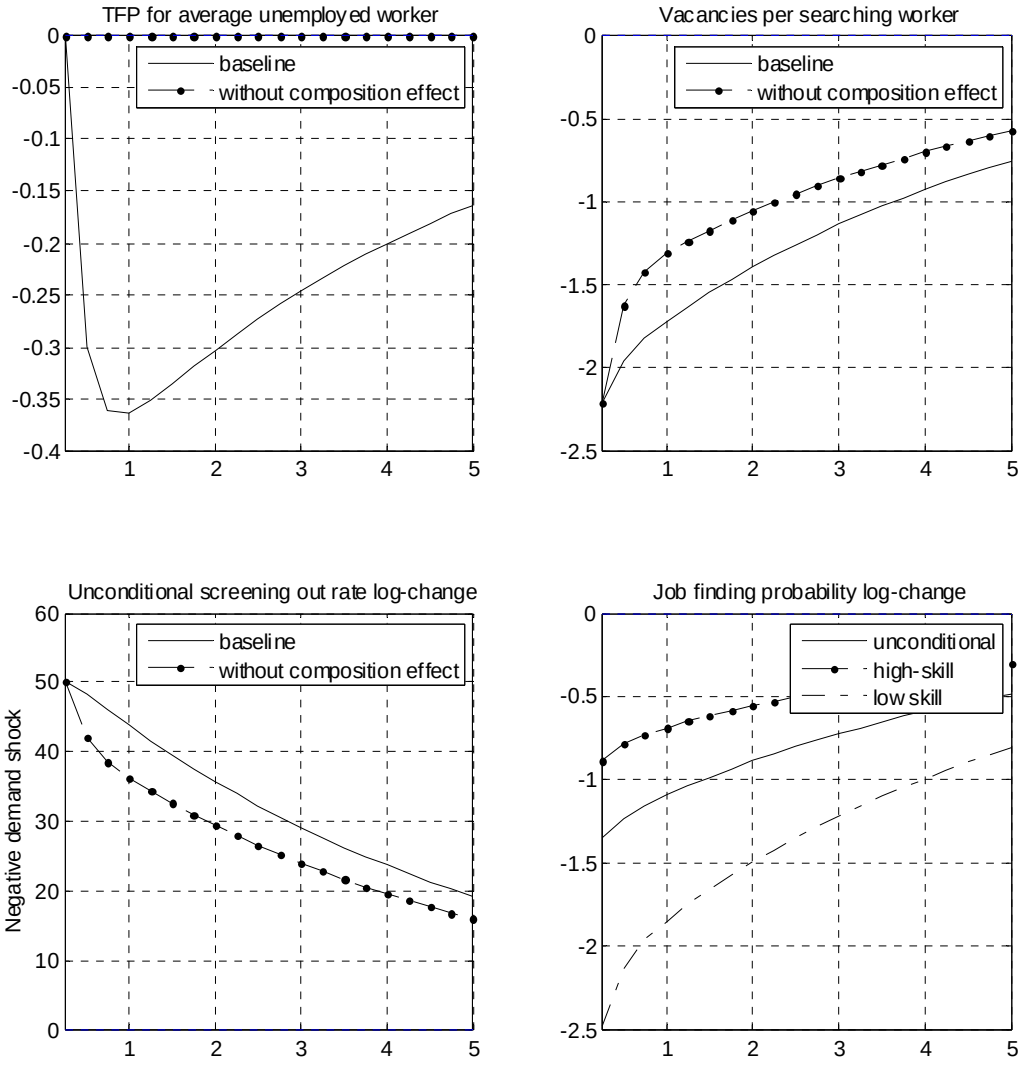

Figure 6: Impulse response to a negative demand shock D_t for the time-varying and constant workers heterogeneity economies. Monetary policy set by Taylor rule responding to CPI inflation. AR(1) coefficient of demand shock $\rho_{d_t} = 0.95$. Impulse responses without composition effect assume share of low-efficiency unemployed is constant at $\gamma_t = \gamma_{ss}$. Percent deviations from steady state. Horizontal axis in years.

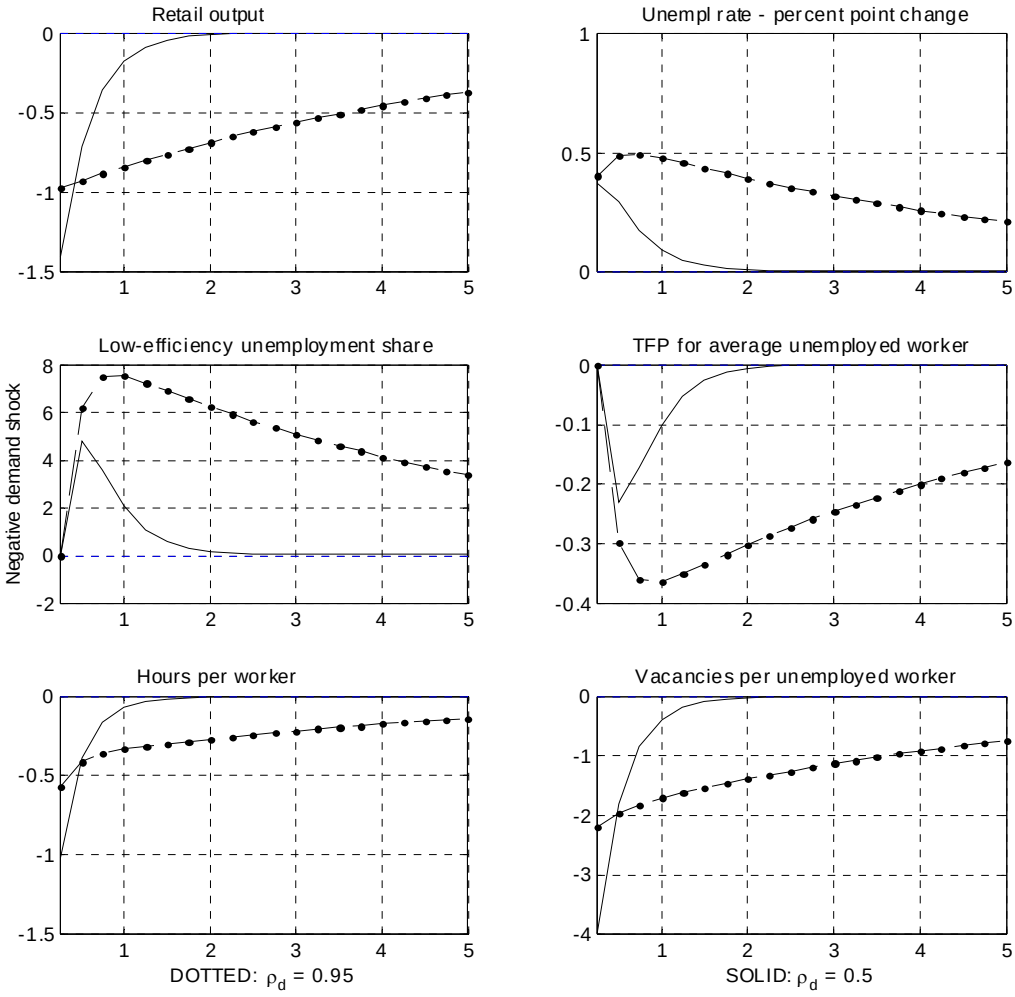

Figure 7: Impulse response to a negative demand shock D_t for the time-varying workers heterogeneity: alternative AR(1) coefficients of demand shock ρ_{d_t}. Monetary policy set by Taylor rule responding to CPI inflation. Horizontal axis in years.

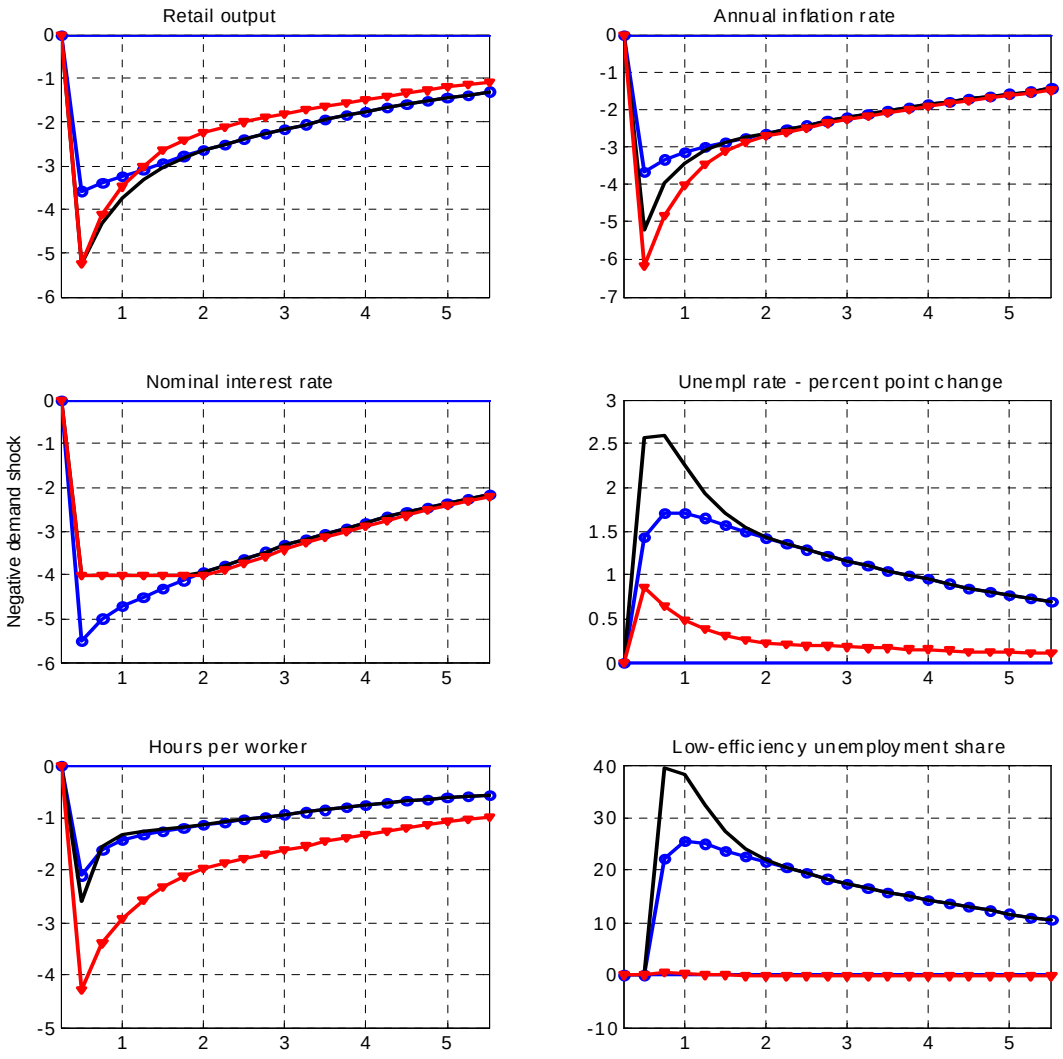

Figure 8: The Great Recession downturn and recovery. Impulse response to a negative demand shock D_t and a discount rate shock leading to the zero lower bound for i_t. Comparison shows the time-varying workers heterogeneity economy with and without the zero lower bound, and the constant workers heterogeneity economy with the zero lower bound. Monetary policy set by Taylor rule responding to CPI inflation. AR(1) coefficient of demand shock $\rho_{d_t} = 0.95$. Horizontal axis in years.

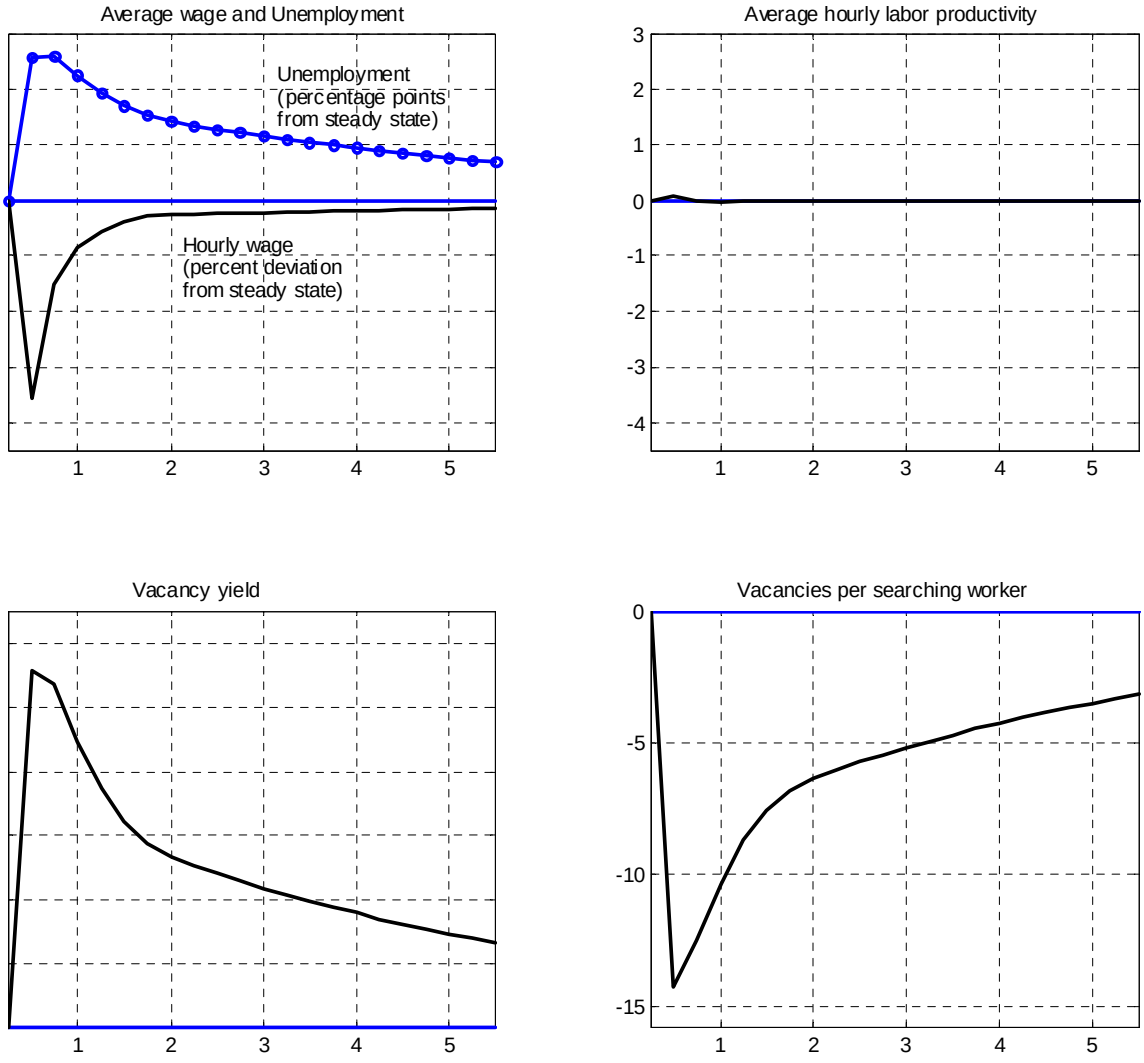

Figure 9: The Great Recession downturn and recovery. Impulse response to a negative demand shock D_t and a discount rate shock leading to the zero lower bound for i_t. Time-varying workers heterogeneity economy. Monetary policy set by Taylor rule responding to CPI inflation. AR(1) coefficient of demand shock $\rho_{d_t} = 0.95$. Unemployment measured in percentage points deviation from steady state. All other variables measured in percent of steady state value. Horizontal axis in years.

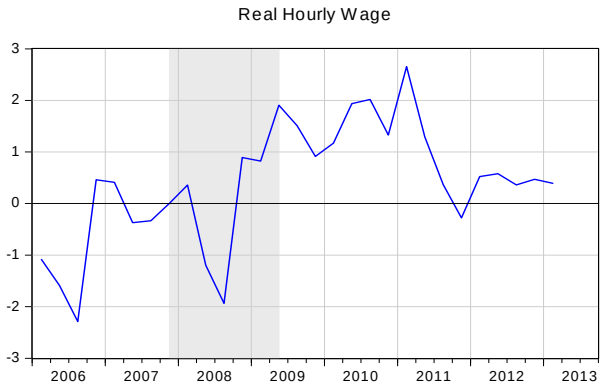

Figure 10: Non-farm business sector real hourly compensation.

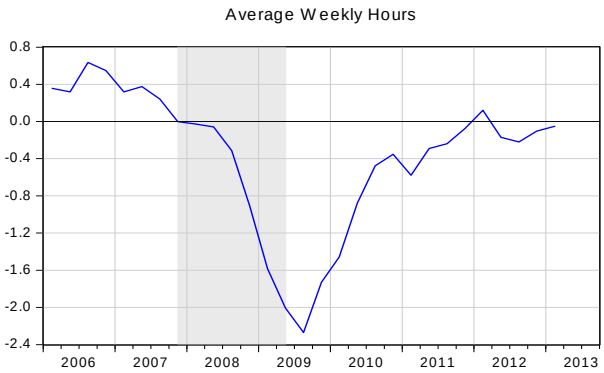

Figure 11: Non-farm business sector average weekly hours worked.

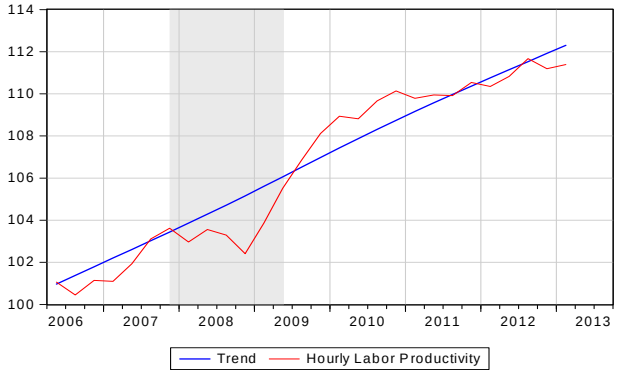

Figure 12: Nonfarm business sector ouput per hour. Raw data and HP-filtered smoothed estimate. Top two panels show percent deviation relative to cyclical peak. Source: BLS.

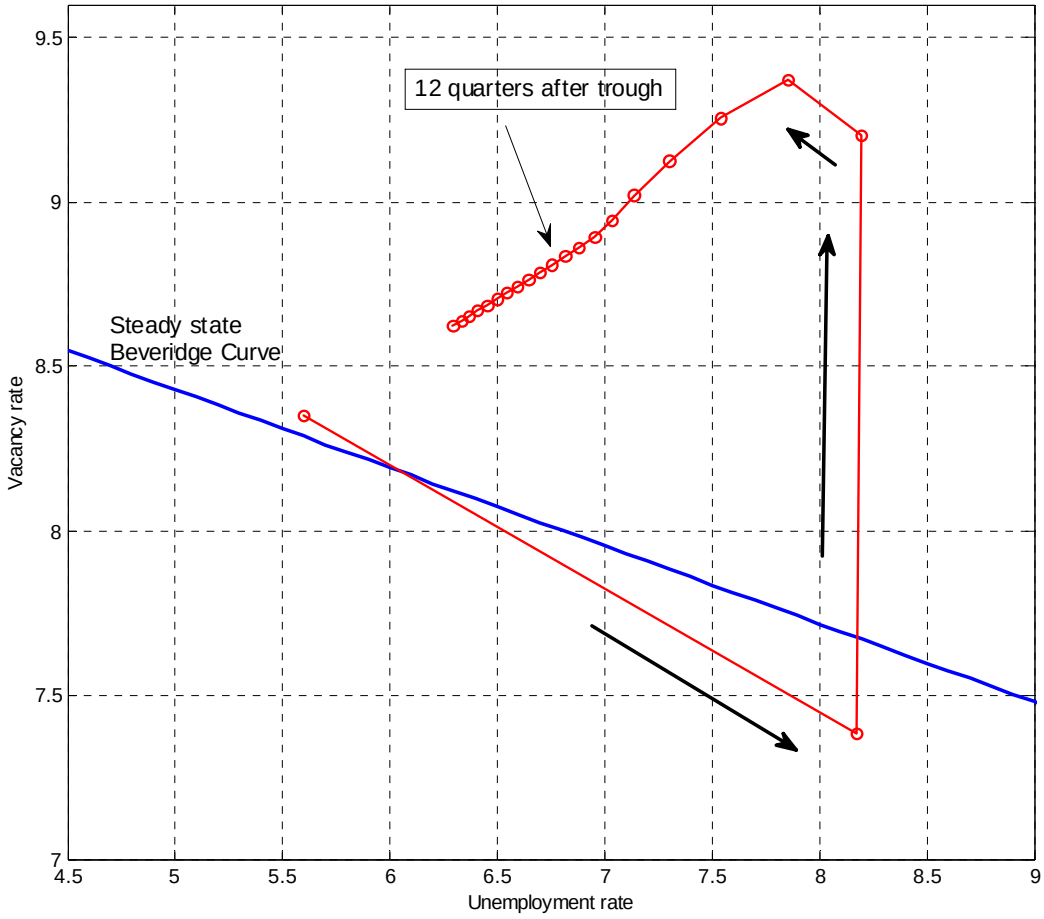

Figure 13: The Great Recession downturn and recovery. Steady state Beveridge curve and simulated dynamics over the Great Recession. Time-varying workers heterogeneity economy. Vacancy rate measured as V_t/N_t.